A Survey on LLM-based Conversational User Simulation

Bo Ni[1], Leyao Wang[3], Yu Wang[4], Branislav Kveton[2], Franck Dernoncourt[2], Yu Xia[5], Hongjie Chen[6], Reuben Leura[7], Samyadeep Basu[2], Subhojyoti Mukherjee[2], Puneet Mathur[2], Nesreen Ahmed[8], Junda Wu[5], Li Li[9], Huixin Zhang[10], Ruiyi Zhang[2], Tong Yu[2], Sungchul Kim[2], Jiuxiang Gu[2], Zhengzhong Tu[10], Alexa Siu[2], Zichao Wang[2], David Seunghyun Yoon[2], Nedim Lipka[2], Namyong Park, Zihao Lin[11], Trung Bui[2], Yue Zhao[9], Tyler Derr[1], Ryan A. Rossi[2]

[1]Vanderbilt University, [2]Adobe Research, [3]Yale University, [4]University of Oregon, [5]University of California San Diego, [6]Dolby Laboratories, [7]University of California, Berkeley, [8]Cisco AI Research, [9]University of Southern California, [10]Texas A&M University, [11]UC Davis

Abstract

User simulation has long played a vital role in computer science due to its potential to support a wide range of applications. Language, as the primary medium of human communication, forms the foundation of social interaction and behavior. Consequently, simulating conversational behavior has become a key area of study. Recent advancements in large language models (LLMs) have significantly catalyzed progress in this domain by enabling high-fidelity generation of synthetic user conversation. In this paper, we survey recent advancements in LLM-based conversational user simulation. We introduce a novel taxonomy covering user granularity and simulation objectives. Additionally, we systematically analyze core techniques and evaluation methodologies. We aim to keep the research community informed of the latest advancements in conversational user simulation and to further facilitate future research by identifying open challenges and organizing existing work under a unified framework.

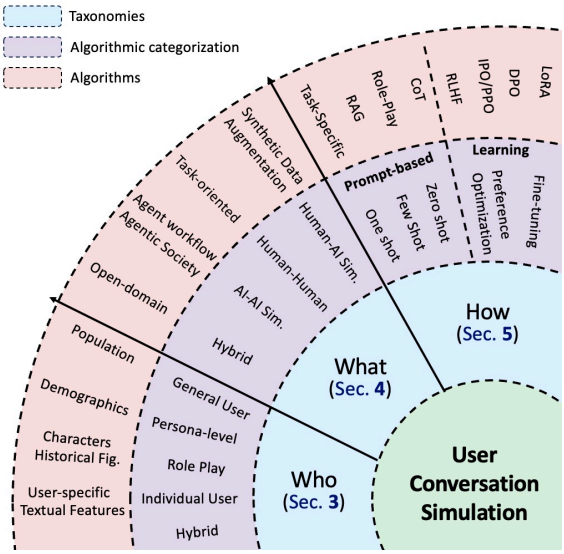


Figure 1: Overview of the proposed taxonomies for user conversation simulation.

1 Introduction

User simulation has been an active area of research for decades.[1] From early simulation games such as *The Sims* (Maxis, 2000) to agent-based environments powered by large language models (Park et al., 2023), the goal has consistently been to create realistic user proxies that support diverse applications. The foundations of user simulation trace back to classic models of user preferences, such as the Bradley-Terry-Luce (Bradley and Terry, 1952) and Plackett-Luce (Plackett, 1975) models. In these models, simulation occurs implicitly through statistical models trained to emulate user preferences and behaviors from observational data, and they remain influential until today in reinforcement learning from human feedback (Ouyang et al., 2022) and direct preference optimization (Rafailov et al., 2023). Other statistical models of user behavior, such as collaborative filtering (Schafer et al., 2007), low-rank matrix factorization (Koren et al., 2009), Bayesian ranking (Rendle et al., 2009), and click models (Chapelle and Zhang, 2009) have been widely adopted to capture user behavior and facilitate personalized experiences.

The emergence of large language models (LLMs) has significantly transformed the landscape

[1]Simulation is often used interchangeably with the term generation, creation, and so on. Similarly, the term conversation is sometimes used interchangeably with dialogue, chat, multi-turn interaction, among others.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4266–4301
March 24-29, 2026

of user simulation in two key ways: first, it enables general-purpose simulation across a wide range of tasks and domains; and second, it drastically lowers the barrier to generating high-quality contextually rich simulated interactions via prompt engineering of pretrained models. In contrast, past frameworks for user simulation in recommender systems (Rohde et al., 2018; Ie et al., 2019) required large amounts of user data to train user simulators and were tailored only to a particular form of recommendations. Unlike simulation in more deterministic domains such as physics (Chan and Wood, 1999), climate (on Climate Change, 2021), or epidemiology (Ferguson et al., 2006), user simulation must grapple with the inherent complexity, variability, and nuance of human behavior. LLMs offer a powerful tool for capturing this complexity—generating coherent, adaptive, and goal-directed behavior without task-specific supervision, making them well-suited for simulating user interactions at scale.

A growing body of work has leveraged LLMs to simulate different user behaviors. In retrieval-augmented generation (RAG) (Lewis et al., 2020; Salemi et al., 2024), the LLM generates a summary of retrieved documents by a classic retrieval system, and thus simulates how a real user would review retrieved knowledge and internalize it. Chain-of-thought (CoT) reasoning (Wei et al., 2022) mimics users' internal reasoning processes, producing step-by-step interactions that reflect human-like deliberation. The LLM-as-a-judge (Gu et al., 2024) simulates how users evaluate responses in terms of helpfulness or correctness, for instance. More domain-specific simulation frameworks have emerged to improve downstream applications such as search, recommendation, and task-oriented dialogue, enabling systems to adapt to user intent and context with greater fidelity. For instance, BASES (Ren et al., 2024a) generates diverse user profiles and simulates large-scale web search behavior, addressing data scarcity and privacy concerns. USimAgent (Zhang et al., 2024b) further replicates user querying, clicking, and session behaviors with strong alignment to real-world interactions. To ensure fidelity, Breuer et al. (Breuer et al., 2024) propose methods for validating synthetic usage data in data-sparse environments. Finally, Balog and Zhai (Balog and Zhai, 2025) present an integrated framework that advances both user modeling and system evaluation through generative AI.

While these methods provide powerful mechanisms for simulating user behaviors, realistic user interactions often require explicitly modeling the interactive, conversational nature of communication. Language, as the primary medium of human interaction, plays a central role in shaping human behavior (Fitch et al., 2010; Tomasello, 2010). LLMs, with their unprecedented language capability, therefore offers a transformative opportunity for conversational user simulations. For example, Zhang et al. (2025c) showed that LLM-generated conversational feedback, such as synthetic user comments, can directly improve recommender system performance. Similarly, Sekulic et al. (2024) demonstrated that conversations can better align systems with user needs and improve satisfaction in task completion. Despite these promising results and a growing body of work in language-based user simulation (Hazrati and Ricci, 2022; Wang et al., 2025a; Zhang et al., 2025c), a dedicated survey that systematically organizes and analyzes the sub-field of conversational user simulation is absent.

In this survey, we aim to fill this gap by providing a comprehensive overview and unified taxonomy of conversational user simulation. As outlined in Table 3 in the Appendix, we first define the key components and scope of this emerging area. We then organize the survey to answer three fundamental questions: *(1)* ***Who*** *is being simulated?* (§3), *(2)* ***What*** *is being simulated?* (§4), and *(3)* ***How*** *is the conversation simulated?* (§5). With this organized discussion, we aim to highlight key research trends and pinpoint open challenges, fostering future research in this area. For more details on related surveys and how our work differs, please refer to Appendix A.3.

2 Problem Definition

We formally define a conversation as a sequence of turns between two or more participants. Let $\mathcal{P} = \{p^1, p^2, \ldots, p^N\}$ be the set of N participants in a conversation. Note that these participants can be of different types, including human users ($\mathcal{P}_U$), systems ($\mathcal{P}_S$), or other agent types, allowing for user-user, user-system, and multi-party conversation simulations.

A conversation, C, is a temporally ordered sequence of T turns: $C = (c_1, c_2, ..., c_T)$ where each turn c_t for $t \in [1, T]$ is a tuple containing the speaker and their utterance:

$$c_t = (p_t^i, u_t)$$

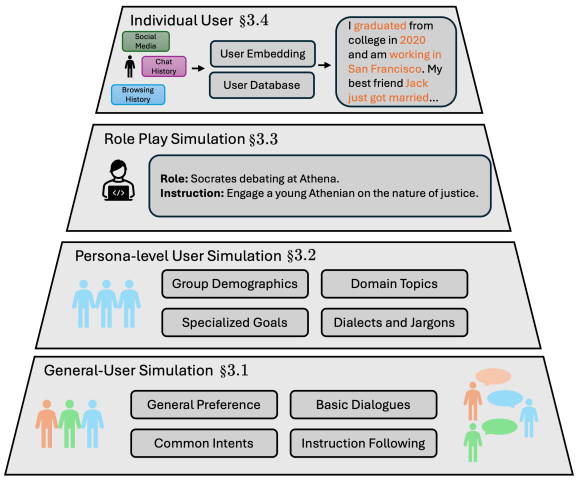


Figure 2: Overview of the proposed taxonomy for **who** is being simulated, i.e., the target of simulation.

where $p_t^i \in \mathcal{P}$ denotes the participant p^i speaking at turn t, and u_t is the utterance they produce from a vocabulary $\mathcal{V}$.

The core of conversational simulation is to model the behavior of one or more target participants. Let C_{t-1} be the conversational history and $C_{t-1} = (c_1, \ldots, c_{t-1})$. Additionally, let Ψ_{p^i} be the context of participant p^i that can include their demographics, domain knowledge, or personal context, etc. Then, the fundamental task of a conversational simulator is to generate a participant's next utterance by modeling the probability distribution:

$$P(u_t | C_{t-1}, \Psi_{p^i})$$

3 Who: Simulated Users and Interactions

The first core question that we explore is the target of simulation. As illustrated in Figure 2, this section discusses simulation of conversations going from general users (Sec. 3.1), to persona-level simulation of conversations (Sec. 3.2), to more role-playing simulation (Sec. 3.3), to the most fine-grained individual user conversation simulation (Sec. 3.4). In each section, we will first formally define the corresponding simulation target, and then introduce relevant works in the area.

3.1 General User Simulation

General user simulation models conversations from a broad population perspective, as seen in models like ChatGPT (OpenAI, 2023).

Definition 3.1 (General User Simulation)**.** The persona $\Psi_p^{default}$ is considered *default*, representing an average user sampled randomly from the general population $\mathcal{P}_U$. No distinguishing characteristics are provided in the prompt.

Recent advances improve interaction quality through multi-turn optimization (Xiong et al., 2025), exploration-based learning (Song et al., 2024), and diverse user simulation (Dhole, 2024). For a detailed discussion of trajectory-level fine-tuning, failure-aware exploration, segment-based optimization, and reinforcement-based training approaches, please refer to Appendix B.1.

3.2 Persona-level User Simulation

Persona-level user simulation models users at the demographics level (Chen et al., 2024b), thereby going beyond general user simulations.

Definition 3.2 (Persona-level User Simulation)**.** The persona Ψ_p is explicitly defined by a set of m attributes: $\Psi_p = \{\psi_1, \psi_2, \ldots, \psi_m\}$, where each ψ_i corresponds to a demographic, interest, or stylistic feature.

Recent work explores a range of techniques for persona grounding, including demographic prompting (Hu and Collier, 2024), psychometric modeling (Wang et al., 2025b; Ji et al., 2024), trait-infused architectures (Huang, 2024; Yang et al., 2025), and activation-level control (Zhu et al., 2024b). These methods have shown promise but also raise concerns around fairness and bias (Li et al., 2025a; Deshpande et al., 2023). This underscores the need for trustworthy simulations. For a more detailed discussion of these approaches and their implications, please refer to Appendix B.2.

3.3 Role Play Simulation

Unlike persona-level simulation (§3.2), which relies on fixed traits, role play simulation offers greater flexibility by modeling real or fictional individuals. For instance, while persona-level simulation may represent a generic "20-year-old programmer," role play simulation can emulate "Mark Zuckerberg at 20," including his unique style and historical context.

Definition 3.3 (Role Play Simulation)**.** Let Φ_θ denote an LLM with parameters θ and let $\mathcal{M}$ be its latent manifold. Given an identity handle h (e.g., "Mark Zuckerberg at 20") expressed in natural language, the model induces an implicit embedding

$$I := E_\theta(h) \in \mathcal{M},$$

where $E_\theta : \Sigma^* \to \mathcal{M}$ is the encoder realized by Φ_θ. A role-play persona is the conditional output

distribution

$$\Psi_p \coloneqq p_\theta(y \mid x, I) = p_\theta\big(y \mid x, E_\theta(h)\big),$$

i.e., the model's behavior when inputs x are conditioned on the latent identity I.

This paradigm leverages LLMs' implicit knowledge to simulate rich character traits (Serapio-García et al., 2023; Wang et al., 2024a), enhanced by prompting (Wang et al., 2025b), fine-tuning (Shao et al., 2023), and self-play (Ji et al., 2025). Applications span storytelling (Wu et al., 2024), social simulation (Wang et al., 2025b), and agent memory modeling (Fan et al., 2025). For a full discussion, see Appendix B.3.

3.4 Individual User Simulation

Individual user simulation represents the most fine-grained level of conversational modeling in our taxonomy. Unlike role-play simulations, which rely on *implicit* predefined characters or figures, individual user simulations are grounded in the *explicit*, often dynamic, personal context of users.

Definition 3.4 (Individual User Simulation)**.** The persona Ψ_p is derived from the full personal history $\mathcal{H}_p$ of participant p, including chat logs, documents, and interaction histories, i.e., $\Psi_p = \mathcal{H}_p$.

Techniques range from profile injection (Zhang et al., 2018; Jang et al., 2022) and dialogue history modeling (Li et al., 2021) to multi-session memory (Xu et al., 2022; Chhikara et al., 2025) and real-world trait grounding (Yamashita et al., 2023; Gao et al., 2023). For a detailed discussion of the related work, please refer to Appendix B.4.

3.5 Hybrid User Simulation

While general user, persona-level, role play, and individual user simulations are conceptually distinct, they often overlap in practice. Large-scale models such as GPT (OpenAI, 2023) and LLaMA (Touvron et al., 2023) naturally blend these paradigms, often exhibiting emergent persona and role-play behaviors (Wang et al., 2024a). Recent works have demonstrated the use of GPT-4 and related models to simulate dialogues for evaluation, data augmentation, and conversational benchmarks (Park et al., 2023; Guo et al., 2024). However, these studies typically adopt a single simulation target (e.g., role play or individual user emulation) without explicitly formalizing hybrid combinations. Thus, the systematic treatment of hybrid user simulation remains limited. The hybrid user simulation can be particularly relevant for multi-agent interactions, as emergent behavior depends on balancing generic user roles with individual variation.

4 What: Simulation Objectives

This section defines the *objectives* of LLM-based user simulation by focusing on *interaction patterns*, rather than applications (§8). We categorize these into four paradigms: *Human–AI*, *Human–Human*, *AI–AI*, and *Many-Human–AI* simulation. Let an interaction trajectory τ be defined as

$$\tau = ((u_1, v_1), \ldots, (u_T, v_T)) \tag{1}$$

where T is the number of turns in the interaction. The paradigms in the rest of this section will be defined following the definition.

4.1 Human–AI Simulation

Human–AI simulation models turn-based interactions where a human prompts and the AI responds, aiming not just to mimic behavior but to create realistic contexts for evaluating AI capabilities. We define a *Human-AI simulation* as any trajectory where u, v is defined as human utterance and AI response, respectively. At each turn t, the human utterance u_t is generated according to one of the "Who" levels:

$$h_t \sim P\big(C_{t-1},\, \Psi_p,\big),$$

where Ψ_p is the corresponding profile defined in the previous sections. The AI response v_t is generated by a fixed model with $\Psi_p^{default}$ being the default profile.

Recent approaches generate synthetic human–AI dialogues to reduce reliance on expensive annotation, including Self-Instruct (Wang et al., 2023), WizardLM (Xu et al., 2024a), and Auto Evol-Instruct (Zeng et al., 2024). Other methods target broad coverage (Li et al., 2024b; Zhang et al., 2024e), domain adaptation (Das et al., 2024; Rachidi et al., 2025), or task specificity (Wang et al., 2024d; Patel et al., 2024). For a detailed discussion, see Appendix C.1.

4.2 Human–Human Simulation

Human–Human simulation models conversations between two human participants, aiming to replicate natural dialogue patterns (Yamashita et al.,

2023; Dinan et al., 2019). This paradigm supports the development of agents that can maintain consistent personas and engage in grounded interactions. We define a *Human–Human simulation* as any trajectory where each participant's utterances are initialized by human input at some "Who" level:

$$u_t \sim P\big(C_{t-1},\, \Psi_{p_u},\big),$$
$$v_t \sim P\big(C_{t-1},\, \Psi_{p_v},\big),$$

Both sides use their own profiles Ψ_{p_*} to simulate realistic two-party human dialogue.

Key datasets include PersonaChat (Zhang et al., 2018), Wizard-of-Wikipedia (Dinan et al., 2019), EmpatheticDialogues (Rashkin et al., 2019), and MultiWOZ (Budzianowski et al., 2018). Scalable alternatives such as self-play bootstrapping (Shah et al., 2018) reduce reliance on manual curation. For an extended discussion, see Appendix C.2.

4.3 AI–AI Simulation

AI–AI simulation models conversations where both participants are autonomous AI agents, interacting without human input (Park et al., 2023; Ren et al., 2024b). This paradigm enables scalable data generation and the study of emergent behaviors. We define *AI-AI* simulation as any trajectory where two AI agents A_1, A_2 converse *without* ongoing human input. The only human contribution is a *general seed prompt* $\mathcal{Q}$, and thereafter

$$a_t^1,\, a_t^2 \;\sim\; P\big(\mathcal{Q},\, a_{1:t-1}^1,\, a_{1:t-1}^2\big),$$

alternating turns. No persona, role, or individual-level profiles are specified at the entity level. Although certain level of simulation (persona, role-play, individual) might be assigned based on the seed prompt $\mathcal{Q}$.

Recent work explores emergent social behaviors (Park et al., 2023; Ren et al., 2024b), collaborative task-solving (Li et al., 2023; Wu et al., 2023), and adversarial debate (Du et al., 2024; Rennard et al., 2025; Hua et al., 2024). For a comprehensive discussion, see Appendix C.3.

4.4 Many-Human–AI Simulation

Many-human–AI simulation generalizes human–AI interaction to multi-user settings, where multiple humans engage with one or more AI agents toward a shared objective. This paradigm captures both individual behavior and group dynamics in collaborative environments.

Definition 4.1 (Many-Human–AI Simulation)**.** Let $\mathcal{U} = \{u^1, \ldots, u^n\}$ be a set of n human participants, each endowed with a profile Ψ_{u^i}, and let A be an AI system that follows general user behavior following definition 3.1. A *Many-Human–AI simulation* is any dialogue trajectory

$$\tau = \{\tau^1, \tau^2, \ldots, \tau^n\}$$

where

$$\tau^j = \big\{(a_t,\, u_t^j)\big\}_{t=1}^{T},$$

Each turn t is produced as follows:

$$u_t^j \sim P\big(C_{t-1},\, \Psi_{u^j}\big)$$

$$a_t \sim P\big(u^{1:t},\, \mathcal{Q}\big)$$

where $\mathcal{Q}$ is the default prompt to align the AI to general user profile.

Recent work explores collaborative roles (Klieger et al., 2024), proxy participation (Leong et al., 2024), and group dialogue simulation (Mao et al., 2024). Despite the advancement in AI-AI simulations (Li et al., 2023; Wu et al., 2023) and Human-AI (Wang et al., 2023) simulations, general-purpose frameworks for simulating many-human–AI settings remain limited. For an extended discussion, see Appendix C.4.

4.5 Hybrid Simulation

Hybrid simulation blends multiple paradigms (i.e., Human–Human, AI–AI, and Human–AI) within the same environment, reflecting the complexity of real-world interactions. For example, sandbox environments like Smallville (Park et al., 2023) simulate AI–AI communities, yet individual dialogues within them resemble Human–Human exchanges. Despite its prevalence, hybrid simulation remains under-theorized. We advocate for systematic frameworks to model and benchmark mixed-interaction settings, essential for socially adaptable AI.

5 How: Techniques and Methodologies

To generate synthetic user conversations, it is crucial to incorporate diverse types of contextual knowledge, each requiring distinct representations. For example, static persona facts guide tone and content (Li et al., 2016), while long-term preferences use memory-based retrieval (Madotto et al., 2018). We discuss integration strategies for these heterogeneous information in the next section.

5.1 Prompt-Based Simulation

Prompt-based simulation formulates user generation as conditional language modeling:

$$u_t \sim P(C_{t-1}, \Psi_p, \mathcal{P}),$$

where $\mathcal{P}$ is the optional in-context examples. Existing work falls into two complementary tracks: method-driven(zero-shot/few-shot and chain-of-thought) and content-driven(persona/role-play and task-specific prompts). Zero-/few-shot prompts enable scalable simulation with minimal examples (Terragni et al., 2023; Zhang et al., 2024a; Wang et al., 2025b), while Chain-of-Thought improves coherence via step-by-step reasoning (Luo et al., 2024). Persona and task-specific prompts guide tone and domain behavior (Shanahan et al., 2023; Abbasiantaeb et al., 2024; Kong et al., 2024; Kiesel et al., 2024). See §D.1 for more details.

5.2 RAG

Retrieval-augmented generation (RAG) enhances user simulation by conditioning responses on external knowledge:

$$u_t \sim P(C_{t-1}, \Psi_p, \mathcal{R}(C_{t-1}, \Psi_p))$$

where $\mathcal{R}(\cdot)$ retrieves context to enhance realism, relevance, and personalization of behaviors.

RAG methods vary by retrieval trigger mechanism. Always-on approaches (Shimadzu et al., 2025), retrieve at every turn prepend relevant passages. Adaptive methods (Wang et al., 2025c) use a learned classifier to decide when retrieval is necessary to improve efficiency. Goal/State-driven approaches (Zhu et al., 2025) retrieve based on internal user memory for more personalized generation. For further examples, see Appendix D.2.

5.3 Fine-tuning

Fine-tuning adapts a pretrained language model into a user simulator by updating its parameters $\boldsymbol{\Theta}$ on a dataset of user dialogues $\mathcal{D}$. Formally, we define $\mathcal{D} = \{(C_{t-1}, \Psi_p, \mathcal{I}, u_t)\}_{t=1}^{T}$, where C_{t-1} denotes the conversation history up to turn $t-1$, Ψ_p is the persona description associated with the user, $\mathcal{I}$ represents optional task-specific instructions, and u_t is the ground-truth user utterance at turn t.

During training, the fine-tuned model is optimized to generate user utterances conditioned on the given context:

$$u_t \sim P_{\boldsymbol{\Theta}'}(C_{t-1}, \Psi_p, \mathcal{I}),$$

where the updated parameters $\boldsymbol{\Theta}'$ are obtained by maximizing a supervised fine-tuning objective:

$$\boldsymbol{\Theta}' = \arg\max_{\boldsymbol{\Theta}} \mathcal{L}_{\mathrm{FT}}(\boldsymbol{\Theta}; \mathcal{D}).$$

Here, $\mathcal{L}_{\mathrm{FT}}$ typically corresponds to the cross-entropy loss between the predicted and ground-truth user responses u_t.

We group fine-tuning strategies into three types. Full-model supervised methods retrain all parameters on in-domain data, as in DAUS (Sekulic et al., 2024), SoulChat (Chen et al., 2023), and MuPaS (Wang et al., 2024c). Parameter-efficient approaches like ESC-Role (Zhao et al., 2024b), BiPO (Cao et al., 2024), and SRA-guided LoRA (Madani et al., 2024) use adapters or activation steering to preserve efficiency. Interactive methods optimize simulators via interaction feedback: UGRO (Hu et al., 2023) applies reward modeling with PPO, while PlatoLM (Kong et al., 2024) leverages large-scale self-play before assistant fine-tuning. See Appendix D.3 for further details.

5.4 RL/DPO

Reinforcement Learning with Human (or AI) Feedback (RLHF) and Direct Preference Optimization (DPO) enhance user simulators by training policies that maximize rewards or preference scores across multi-turn interactions:

$$u_t \sim P(C_{t-1}, \Psi_p, \pi_\theta), \quad \pi_\theta = \arg\max_{\theta} \mathbb{E}[R(\tau)]$$

where π_θ is a user policy optimized via RL or DPO using feedback from user preferences.

These methods support adaptive, strategic, and goal-driven behaviors beyond standard supervised learning. Personalization-focused approaches like curiosity-driven RLHF (Wan et al., 2025) reward disambiguation of latent user traits. Memory-aware methods (Seo et al., 2024) use DPO to optimize memory selection for factual coherence. ArCHer (Zhou et al., 2024) introduces hierarchical RL for long-horizon planning via utterance- and token-level control. Offline learning frameworks like hindsight regeneration (Hong et al., 2024) revise suboptimal segments using observed user feedback. Action-level DPO (Chen et al., 2025b) improves intent clarification by preferring follow-ups that maximize user satisfaction. See Appendix D.4 for further details.

	WHO (Section 3)					WHAT (Section 4)					HOW (Section 5)					APPLICATIONS (Section 8)				
	General User Simulation (§3.1)	Persona User Simulation (§3.2)	Role Play Simulation (§3.3)	Individual User (§3.4)	Hybrid (§3.5)	Human-AI (§4.1)	Human-Human (§4.2)	AI-AI (§4.3)	Many Human-AI (§4.4)	Hybrid (§4.5)	Prompt-based (§5.1)	RAG (§5.2)	Fine-Tuning (§5.3)	RL/DPO (§5.4)	Hybrid (§5.5)	Recommendation	Summarization	Text Generation	Question Answering	Other
PersonalConv (Li et al., 2025b)	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓
PersonalDialog (Zheng et al., 2019)	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
RoleLLM (Wang et al., 2024a)	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗
DMPO (Shi et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗
SDPO (Kong et al., 2025)	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗
AgentQ (Putta et al., 2024)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓	✗
LOOP (Chen et al., 2025a)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
KAUCUS (Dhole, 2024)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✓	✗
CROSS (Yuan et al., 2024)	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗
PersonaEffect (Hu and Collier, 2024)	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗
EvalPersonality (Wang et al., 2025b)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗
P² (Jiang et al., 2023)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗
PsyPlay (Yang et al., 2025)	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
CSHI (Zhu et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓
RAGate (Wang et al., 2025c)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
PB&J (Joshi et al., 2025)	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗
CharacterLLM (Shao et al., 2023)	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✓	✓	✗
CharacterBench (Zhou et al., 2025)	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗
CharacterBench (Huang, 2024)	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗
SmallVille (Park et al., 2023)	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
CharMap (Xu et al., 2024c)	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗
DramaLLM (Wu et al., 2024)	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
LifeStageBench (Fan et al., 2025)	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
PersonaChat (Zhang et al., 2018)	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
ProphetChat (Liu et al., 2022)	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗
WoW (Dinan et al., 2019)	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗
EmpatheticDialogues (Rashkin et al., 2019)	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓
PLATO (Kong et al., 2024)	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗
ConvEval (Balog and Zhai, 2023)	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗
UserSimCRS (Afzali et al., 2023)	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗

Table 1: Overview of the proposed taxonomy for user-simulated data generation techniques and their applications. Using this taxonomy, we provide a qualitative and quantitative comparison of methods.

5.5 Hybrid Approaches

Hybrid approaches combine prompting, retrieval, fine-tuning, and RL/DPO to improve realism, controllability, and sample efficiency. Retrieval-augmented fine-tuning (e.g., PRAISE (Kaiser and Weikum, 2025)) integrates context during training for better grounding. Prompt-to-fine-tune pipelines (Chen et al., 2023) bootstrap data via prompting before supervised tuning. RAG + RL/DPO loops (Wan et al., 2025) coordinate retrieval and policy learning to adaptively query and respond. Hierarchical pipelines like ARCHer (Zhou et al., 2024) modularize simulation tasks, while personalized stacks (Zhang et al., 2025a) combine persona prompts, memory, adapters, and RLHF. See Appendix D.5 for more.

6 Evaluation

Traditional Metrics Classical metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), slot-F1 (Chen et al., 2019) remain common for structured or goal-oriented dialogues. While efficient and reproducible, these metrics capture narrow facets and are often complemented by human or LLM judges. Human Evaluation remains the gold standard. It is typically conducted through either *interactive* evaluation *offline* evaluation (Ye et al., 2022; Zhou et al., 2025; Fan et al., 2025). For detailed discussion on traditional metrics and human evaluation, see Appendix E.1 for details.

LLM-as-Judge LLM-as-Judge uses a strong language model as an automatic evaluator, prompted with dialogue context, generated responses, and a rubric targeting dimensions such as coherence, factuality, or safety (Kong et al., 2024; Fan et al., 2025; Lin and Chen, 2023; Saad-Falcon et al., 2024; Zheng et al., 2023). A typical evaluation protocol involves three components: (1) defining evaluation dimensions and rating scales (e.g., 1–5 Likert), (2) providing few-shot exemplars or calibration prompts to align the judge, and (3) instructing the model to explain its reasoning before issuing a final score (Salemi et al., 2025; Li et al., 2024a; Gu et al., 2025). Despite its effectiveness, LLM-as-Judge is sensitive to prompt phrasing and underlying model bias. To address this, recent work proposes symmetric prompting, voting or ensembling across judges, and meta-evaluation by comparing model scores to human ratings (Fan et al., 2025; Zhou et al., 2025).

Trustworthy and Causal Evaluation Beyond accuracy-oriented metrics, recent studies emphasize the importance of trustworthy (Liu et al., 2024; Ni et al., 2025; Zhao et al., 2025) and causal/offline evaluation (Petrov et al., 2025; Dudík et al., 2014; Laban et al., 2025) for conversational systems. These paradigms assess not only output quality, but also reliability under uncertainty, robustness to distribution shifts or adversarial prompts, and generalization across topics and user profiles.

7 Datasets

We categorize commonly used user simulation datasets across dialogue types, with a summary provided in Table 11 and additional details in Appendix F. These datasets span a wide range of interaction settings, including personalized conversations (Zhang et al., 2018; Li et al., 2025b), multiparty social dialogues (Gao et al., 2023), and information-seeking question answering (Abbasiantaeb et al., 2024; Kong et al., 2024). Role-based and character-grounded simulations (Zhou et al., 2025; Wang et al., 2024a) support more fine-grained persona modeling, while negotiation datasets (Lewis et al., 2017; He et al., 2018) evaluate goal-driven interaction and strategic reasoning. Other datasets explore long-term memory and context modeling (Fan et al., 2025), multi-domain generalization (Zhao and Eskénazi, 2018), and dialogue-level response ranking (Zhu et al., 2024a). This growing corpus of datasets provides a foundation for developing, evaluating, and benchmarking user simulation models under diverse conversational settings.

8 Applications

Conversational user simulation underpins a broad spectrum of applications. Before the advent of LLMs, user simulators were already employed for data augmentation in contexts with sparse user histories (Zhao et al., 2021), and LLMs have further benefited these applications (Zhao et al., 2024a). In conversational recommendation, simulators enable the adaptation of systems to diverse user preferences (Yoon et al., 2024). In education, they power conversational companions that support interactive learning (Xu et al., 2024b). Last but not the least, in human–computer interaction (HCI), they facilitate tasks such as user interface testing, reducing reliance on human participants (Moore and Arar, 2018). They have also been applied in specialized domains such as video question answering, where arena-style evaluation with modified Elo rating systems is supported. For further discussion of applications and domains, see Appendix G.

9 Open Problems & Challenges

9.1 Long Conversations

Simulating extended interactions challenges current models' ability to maintain persona consistency across turns, often leading to drift in style, beliefs, or goals (Cho et al., 2023; Xu et al., 2024c). These issues are amplified in role-based settings, where broken memory or contradictions can lead to hallucinations and character violations (Tang et al., 2025). Simulated users are also unrealistically cooperative, and long dialogues accumulate errors

or lose task focus (Putta et al., 2024). Solutions require better memory mechanisms, discourse planning, and consistency modeling (Kong et al., 2025; Fan et al., 2025).

9.2 Diversity

Simulators often mirror cultural-linguistic majorities, yielding overly polite, homogeneous behavior. Though prompting enables personas (Shanahan et al., 2023), diversity remains limited. Richer simulation requires fine-grained control over traits like emotion, verbosity, and strategy. Most work targets single-user setups, neglecting hybrid or multi-user dynamics essential for realism and personalization.

9.3 Biases and Toxic Content

LLM-based simulations risk encoding biases and generating toxic content, especially when personas involve sensitive demographics or public figures (Li et al., 2025a; Deshpande et al., 2023). Such biases can harm both research and deployment. While prompt filtering and alignment exist, robust protocols for simulation quality and safety are lacking (Hu and Collier, 2024). For more open problems and challenges, please refer to Appendix H.

10 Conclusion

In this survey, we retrospected the representative literature on LLM-based conversational user simulation through a unified framework along three axes following **Who**, **What**, and **How**. We further provided a broad overview of existing methods used to simulate user conversations, discussed their strengths and limitations, and categorized them across diverse applications. In addition, we examined evaluation protocols and common datasets to support benchmarking. Finally, we outlined key open challenges and suggested future research directions to build more consistent, diverse, and trustworthy user simulators.

11 Limitations

While this survey aims to provide a comprehensive overview of LLM-based conversational user simulation, it has several limitations. First, our taxonomy is designed to balance generality and clarity, but certain hybrid or domain-specific methods may not fit perfectly into the proposed categories. Additionally, while we summarize datasets and evaluations, a full benchmarking study across methods was beyond the scope of this work.

12 Ethical considerations

Our work focuses on surveying and organizing existing research on LLM-based conversational user simulation. We aim to support researchers in developing more effective and trustworthy simulation methods. However, although our work is a survey and does not involve human subjects or the creation of new models or datasets, we acknowledge that research on LLM-based conversational user simulation presents meaningful ethical challenges. Role-playing public figures may risk misinformation, reputational harms, or the inappropriate use of likenesses without consent. Similarly, the construction of demographic personas and synthetic dialogues can introduce or reinforce stereotypes, underrepresent certain groups, or produce misleading impressions of lived experiences. The generation of synthetic data further raises questions of provenance, authenticity, and potential downstream misuse, particularly if data are repurposed without clear disclosure.

To mitigate these concerns, we highlight the importance of transparency in documenting the sources and purposes of synthetic data, careful consideration of representational balance when simulating demographic personas, and adherence to community standards around privacy and consent when role-playing individuals or groups. Researchers should also remain mindful of the dual-use risks of simulation technologies, such as their potential application in manipulative or deceptive contexts. By surfacing these issues within our survey, we aim to provide not only an overview of technical progress but also a reminder that the development of user simulation methods must be accompanied by ongoing ethical reflection and responsible practice.

Acknowledgments

This research is supported by Adobe Research and the National Science Foundation (NSF) under grant numbers IIS2239881, IIS2524380, and IIS2524379.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational QA via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web*

Search and Data Mining, WSDM 2024, pages 8–17. ACM.

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

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 3354–3377. Association for Computational Linguistics.

Krisztian Balog and ChengXiang Zhai. 2023. User simulation for evaluating information access systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 302–305.

Krisztian Balog and ChengXiang Zhai. 2025. User simulation in the era of generative AI: user modeling, synthetic data generation, and system evaluation. *CoRR*, abs/2501.04410.

Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding your users: a practical guide to user research methods*. Morgan Kaufmann.

Nolwenn Bernard and Krisztian Balog. 2024. Identifying breakdowns in conversational recommender systems using user simulation. In *ACM Conversational User Interfaces 2024, CUI 2024*, page 26. ACM.

Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260.

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

Timo Breuer, Norbert Fuhr, and Philipp Schaer. 2024. Validating synthetic usage data in living lab environments. *ACM Journal of Data and Information Quality*, 16(1):5:1–5:33.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 4515–4524. Association for Computational Linguistics.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Grace Chan and Andrew T. A. Wood. 1999. Simulation of stationary gaussian vector fields. *Statistics and Computing*, 9(4):265–268.

Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer, and Sam Denton. 2024. Balancing cost and effectiveness of synthetic data generation strategies for llms. *CoRR*, abs/2409.19759.

Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, pages 1–10. ACM.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 3002–3017. Association for Computational Linguistics.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 2108–2126. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.

Kevin Chen, Marco F. Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. 2025a. Reinforcement learning for long-horizon interactive LLM agents. *CoRR*, abs/2502.01600.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö. Arik. 2025b. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909.

Stanley F. Chen, Doug Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183. Association for Computational Linguistics.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready AI agents with scalable long-term memory. *CoRR*, abs/2504.19413.

Won Ik Cho, Yoon Kyung Lee, Seoyeon Bae, Jihwan Kim, Sangah Park, Moosung Kim, Sowon Hahn, and Nam Soo Kim. 2023. When crowd meets persona: Creating a large-scale open-domain persona dialogue corpus. *CoRR*, abs/2304.00350.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS 2017*, pages 4299–4307.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008*, pages 87–94. ACM.

Robert H. Crites and Andrew G. Barto. 1995. Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8, NeurIPS 1995*, pages 1017–1023. MIT Press.

Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic patient-physician dialogue generation from clinical notes using LLM. *CoRR*, abs/2408.06285.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270. Association for Computational Linguistics.

Kaustubh Dhole. 2024. Kaucus-knowledgeable user simulators for training large language models. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat, SCI-CHAT 2024*, pages 53–65.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *The Seventh International Conference on Learning Representations, ICLR 2019*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 3029–3051. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024*.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.

Siqi Fan, Xiusheng Huang, Yiqun Yao, Xuezhi Fang, Kang Liu, Peng Han, Shuo Shang, Aixin Sun, and Yequan Wang. 2025. If an LLM were a character, would it know its own story? evaluating lifelong learning in llms. *CoRR*, abs/2503.23514.

Neil M. Ferguson, Derek A.T. Cummings, Christophe Fraser, James C. Cajka, Patrick C. Cooley, and Donald S. Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452.

W. Tecumseh Fitch, Ludwig Huber, and Thomas Bugnyar. 2010. Social cognition and the evolution of language: constructing cognitive phylogenies. *Neuron*, 65(6):795–814.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed

from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 15387–15405. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, pages 8048–8057.

Jennifer Haase and Sebastian Pokutta. 2025. Beyond static responses: Multi-agent LLM systems as a new paradigm for social science research. *CoRR*, abs/2506.01839.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic HCI research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023*, pages 433:1–433:19. ACM.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19.

Naieme Hazrati and Francesco Ricci. 2022. Simulating users' interactions with recommender systems. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–98. ACM.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343. Association for Computational Linguistics.

Sviatlana Höhn, Jauwairia Nasir, Daniel C. Tozadore, Ali Paikan, Pouyan Ziafati, and Elisabeth André. 2024. Beyond pretend-reality dualism: Frame analysis of llm-powered role play with social agents. In *Proceedings of the 12th International Conference on Human-Agent Interaction, HAI 2024*, pages 393–395. ACM.

Joey Hong, Jessica Lin, Anca D. Dragan, and Sergey Levine. 2024. Interactive dialogue agents via reinforcement learning on hindsight regenerations. *CoRR*, abs/2411.05194.

Daniel G. Horvitz and Donovan J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 10289–10307. Association for Computational Linguistics.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023*, pages 3953–3957. ACM.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. Game-theoretic LLM: agent workflow for negotiation games. *CoRR*, abs/2411.05990.

Yuxuan Huang. 2024. Orca: Enhancing role-playing abilities of large language models by integrating personality traits. *CoRR*, abs/2411.10006.

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

Edward Ionides. 2008. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 10803–10812. AAAI Press.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. *CoRR*, abs/2503.17662.

Yongyi Ji, Zhisheng Tang, and Mayank Kejriwal. 2024. Is persona enough for personality? using chatgpt to reconstruct an agent's latent personality from simple descriptions. In *Proceedings of the 41st International Conference on Machine Learning, ICML 2024 Workshop on LLMs and Cognition*, volume 235. PMLR.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information*

Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627. Association for Computational Linguistics.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025*, pages 1073:1–1073:28. ACM.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017*, pages 781–789. ACM.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving language model personas via rationalization with psychological scaffolds. *CoRR*, abs/2504.17993.

Magdalena Kaiser and Gerhard Weikum. 2025. Preference-based learning with retrieval augmented generation for conversational question answering. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025*, pages 1053–1057. ACM.

Kate Kaplan. 2020. Typical designer–to–developer and researcher–to–designer ratios.

David Kempe, Jon M. Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM.

Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. 2024. Simulating follow-up questions in conversational search. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024*, volume 14609 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Benjamin Klieger, Charis Charitsis, Miroslav Suzara, Sierra Wang, Nick Haber, and John C. Mitchell. 2024. Chatcollab: Exploring collaboration between humans and AI agents in software teams. *CoRR*, abs/2412.01992.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. 2025. SDPO: segment-level direct preference optimization for social agents. *CoRR*, abs/2501.01821.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. 2024. Platolm: Teaching llms in multi-round dialogue via a user simulator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 7841–7863. Association for Computational Linguistics.

Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37.

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. 2015. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 767–776.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation.

Paul Lagrée, Claire Vernade, and Olivier Cappé. 2016. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, pages 1597–1605.

Luc Lamontagne, François Laviolette, Richard Khoury, and Alexandre Bergeron-Guyard. 2014. A framework for building adaptive intelligent virtual assistants. In *Artificial intelligence and applications*, volume 10, pages 2014–816.

Jeongmin Lee, Jin-Xia Huang, Minsoo Cho, Yoon-Hyung Roh, Oh-Woog Kwon, and Yunkeun Lee. 2024. Developing conversational intelligent tutoring for speaking skills in second language learning. In *Generative Intelligence and Intelligent Tutoring Systems - 20th International Conference, ITS 2024*, volume 14798 of *Lecture Notes in Computer Science*, pages 131–148. Springer.

Joanne Leong, John C. Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, embodied agents that participate in meetings when you are unavailable. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–28.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *CoRR*, abs/1706.05125.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025a. LLM generated persona is a promise with a catch. *CoRR*, abs/2503.16527.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024b. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *CoRR*, abs/2402.13064.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*. The Association for Computer Linguistics.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):45:1–45:25.

Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, Yu Wang, Xiangliang Zhang, Philip S. Yu, Xiyang Hu, and Yue Zhao. 2025b. A personalized conversational benchmark: Towards simulating personalized conversations. *CoRR*, abs/2505.14106.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. 2024c. Panoptic scene graph generation with semantics-prototype learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 3145–3153. AAAI Press.

Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S. Muthukrishnan, Vishwa Vinay, and Zheng Wen. 2018. Offline evaluation of ranking policies with click models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 1685–1694. ACM.

Jieun Lim, Unggi Lee, Junbo Koh, Yeil Jeong, Yunseo Lee, Gyuri Byun, Haewon Jung, Yoonsun Jang, Sanghyeok Lee, and Jewoong Moon. 2025. Development and implementation of a generative artificial intelligence-enhanced simulation to enhance problem-solving skills for pre-service teachers. *Computers & Education*, 232:105306.

Chien-Chang Lin, Anna Y. Q. Huang, and Owen H. T. Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

Chang Liu, Xu Tan, Chongyang Tao, Zhenxin Fu, Dongyan Zhao, Tie-Yan Liu, and Rui Yan. 2022. Prophetchat: Enhancing dialogue generation with simulation of future conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 962–973. Association for Computational Linguistics.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *CoRR*, abs/2502.11528.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 1–21. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 7828–7840. Association for Computational Linguistics.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stanczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. 2025. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. *CoRR*, abs/2504.08942.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. Duetsim: Building user simulator with dual large language models for task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024*, pages 5414–5424. ELRA and ICCL.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan S. Kankanhalli, and Junnan Li. 2025. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 8461–8474. Computer Vision Foundation / IEEE.

Oleksandr Lytvyn. 2025. Human-ai interaction in language acquisition: Evaluating llm as a language partner. In *Proceedings of the MEi: CogSci Conference*, volume 19.

Navid Madani, Sougata Saha, and Rohini K. Srihari. 2024. Steering conversational large language models for long emotional support conversations. *CoRR*, abs/2402.10453.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 1468–1478. Association for Computational Linguistics.

Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-user chat assistant (MUCA): a framework using llms to facilitate group conversations. *CoRR*, abs/2401.04883.

Maxis. 2000. The sims.

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.

Robert J. Moore and Raphael Arar. 2018. Conversational UX design: An introduction. In Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren, editors, *Studies in Conversational UX Design*, Human-Computer Interaction Series, pages 1–16. Springer.

Rémi Munos and Andrew W. Moore. 1999. Variable resolution discretization for high-accuracy solutions of optimal control problems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99*, pages 1348–1355. Morgan Kaufmann.

Bo Ni, Yu Wang, Lu Cheng, Erik Blasch, and Tyler Derr. 2025. Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12417–12425.

Monika Olędzka, Mark Benesio Carace, Susana de Oliveira Tomaz, Benny Pan, and Pengfei Jiang. 2024. Ai as a teaching assistant: An innovative approach to education through customized model answer generation and guided practice. *Studia Edukacyjne*, pages 67–79.

Intergovernmental Panel on Climate Change. 2021. *Climate Change 2021: The Physical Science Basis*. Cambridge University Press.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, volume 35, pages 27730–27744.

Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025*, pages 20:1–20:18. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Jaesung Park, Seongeun Yang, Chaewon Yoon, Jongwook Si, Yuchul Jung, and Sungyoung Kim. 2024. Ai english conversation coaching platform in the metaverse: Focused on korean users. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6. IEEE.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023*, pages 2:1–2:22. ACM.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024. Datadreamer: A tool for synthetic data generation and reproducible LLM workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 3781–3799. Association for Computational Linguistics.

Nikhil Patel and Sandeep Trivedi. 2020. Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty. *Empirical Quests for Management Essences*, 3(3):1–24.

Aleksandr V. Petrov, Michael Murtagh, and Karthik Nagesh. 2025. Llms for estimating positional bias in logged interaction data.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *CoRR*, abs/2502.08691.

Robin Lewis Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent Q: advanced reasoning and learning for autonomous AI agents. *CoRR*, abs/2408.07199.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 615–636. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Inass Rachidi, Anas Ezzakri, Jaime Bellver-Soler, and Luis Fernando D'Haro. 2025. Design, generation and evaluation of a synthetic dialogue dataset for contextually aware chatbots in art museums. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 20–28.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5370–5381. Association for Computational Linguistics.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2024a. BASES: large-scale web search user simulation with large language model based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 902–917. Association for Computational Linguistics.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024b. Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, pages 7895–7903.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: bayesian personalized ranking from implicit feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press.

Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. 2025. Bias in the mirror : Are llms opinions robust to their own adversarial attacks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, pages 2128–2143. Association for Computational Linguistics.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007*, pages 521–530. ACM.

James Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. 2018. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *CoRR*, abs/1808.00720.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian J. McAuley. 2024. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479. Association for Computational Linguistics.

Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025. Reasoning-enhanced self-training for long-form personalized text generation.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 7370–7392. Association for Computational Linguistics.

J. Ben Schafer, Dan Frankowski, Jonathan L. Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 291–324. Springer.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, André Ferreira Manso, and Roland Mathis. 2024. Reliable llm-based user simulator for task-oriented dialogue systems. *CoRR*, abs/2402.13374.

Youngkyung Seo, Yoonseok Heo, Jun-Seok Koh, and Du-Seong Chang. 2024. Efficient and accurate memorable conversation model using DPO based on sllm. *CoRR*, abs/2407.06537.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. 2023. Personality traits in large language models. *CoRR*, abs/2307.00184.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 41–51. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 13153–13187. Association for Computational Linguistics.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2024. Direct multi-turn preference optimization for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 2312–2324. Association for Computational Linguistics.

Hikaru Shimadzu, Takehito Utsuro, and Daisuke Kitayama. 2025. Retrieval-augmented simulacra: Generative agents for up-to-date and knowledge-adaptive simulations. *CoRR*, abs/2503.14620.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization for LLM agents. *CoRR*, abs/2403.02502.

Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. A survey on recent advances in conversational data generation.

Marc Steen, Lottie Kuijt-Evers, and Jente Klok. 2007. Early user involvement in research and design projects – a review of methods and practices. In *23rd EGOS Colloquium*, volume 5, pages 1–21.

Elizabeth Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1):1–21.

Yihong Tang, Bo Wang, Xu Wang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. Rolebreak: Character hallucination as a jailbreak attack in role-playing systems. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025*, pages 7386–7402. Association for Computational Linguistics.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. Rolecraft-glm: Advancing personalized role-playing in large language models. *CoRR*, abs/2401.09432.

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Ferreira Manso, and Roland Mathis. 2023. In-context learning user simulators for task-oriented dialog systems. *CoRR*, abs/2306.00774.

Gerald Tesauro. 1994. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219.

Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012*, pages 5026–5033. IEEE.

Michael Tomasello. 2010. *Origins of Human Communication*. MIT Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Kenneth Train. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, NY.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 11836–11850. Association for Computational Linguistics.

Salih Tutun, Marina E. Johnson, Abdulaziz Ahmed, Abdullah Albizri, Sedat Irgil, Ilker Yesilkaya, Esma Nur Ucar, Tanalp Sengun, and Antoine Harfouche. 2023. An ai-based decision support system for predicting mental health disorders. *Information Systems Frontiers*, 25(3):1261–1276.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015 Workshop on Deep Learning*.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. 2025. Enhancing personalized multi-turn dialogue with curiosity reward. *CoRR*, abs/2504.03206.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025a. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):55:1–55:37.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025b. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):55:1–55:37.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 14743–14777. Association for Computational Linguistics.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook. *CoRR*, abs/2403.18105.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025c. Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 491–503. Association for Computational Linguistics.

Xiaoyu Wang, Ningyuan Xi, Teng Chen, Qingqing Gu, Yue Zhao, Xiaokai Chen, Zhonglin Jiang, Yong Chen, and Luo Ji. 2024c. Multi-party supervised fine-tuning of language models for multi-party dialogue generation. *CoRR*, abs/2412.05342.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 13484–13508. Association for Computational Linguistics.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024d. Codeclm: Aligning language models with tailored synthetic data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3712–3729. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 3271–3290. Association for Computational Linguistics.

Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, Chi Jin, Tong Zhang, and Tianqi Liu. 2025. Building math agents with multi-turn iterative preference learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Canwen Xu, Daya Guo, Nan Duan, and Julian J. McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 6268–6278. Association for Computational Linguistics.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024b. Large language models for education: A survey. *CoRR*, abs/2405.13001.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 5180–5197. Association for Computational Linguistics.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024c. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *CoRR*, abs/2404.12138.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, PACLIC 2023*, pages 852–861. Association for Computational Linguistics.

Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. 2025. Psyplay: Personality-infused role-playing conversational agents. *CoRR*, abs/2502.03821.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022*, pages 351–360. Association for Computational Linguistics.

Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S. Yu, and Qingsong Wen. 2025. Position: Llms can be good tutors in foreign language education. *CoRR*, abs/2502.05467.

Se-eun Yoon, Zhankui He, Jessica Maria Echterhoff, and Julian J. 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), NAACL 2024*, pages 1490–1504. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 8015–8036. Association for Computational Linguistics.

Murong Yue, Wijdane Mifdal, Yixuan Zhang, Jennifer Suh, and Ziyu Yao. 2024. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *CoRR*, abs/2404.06711.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic instruction evolving for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 6998–7018. Association for Computational Linguistics.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024a. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 1807–1817. ACM.

Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024b. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 2687–2692. ACM.

Jianguo Zhang, Kun Qian, Zhiwei Liu, Shelby Heinecke, Rui Meng, Ye Liu, Zhou Yu, Huan Wang, Silvio Savarese, and Caiming Xiong. 2024c. Dialogstudio: Towards richest and most diverse unified dataset collection for conversational AI. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2299–2315. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 2204–2213. Association for Computational Linguistics.

Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1512–1520. ACM.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024d. RAFT: adapting language model to domain specific RAG. *CoRR*, abs/2403.10131.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024e. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 19228–19252. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2025a. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025b. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025*, pages 10364–10379. Association for Computational Linguistics.

Zijian Zhang, Shuchang Liu, Ziru Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. 2025c. Llm-powered user simulator for recommender system. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025*, pages 13339–13347. AAAI Press.

Chenyang Zhao, Xueying Jia, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024a. SELF-GUIDE: better task-specific instruction following via self-synthetic finetuning. *CoRR*, abs/2407.12874.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Dandan Liang, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024b. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 15785–15810. Association for Computational Linguistics.

Tiancheng Zhao and Maxine Eskénazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10. Association for Computational Linguistics.

Tong Zhao, Bo Ni, Wenhao Yu, Zhichun Guo, Neil Shah, and Meng Jiang. 2021. Action sequence augmentation for early graph-based anomaly detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2668–2678, New York, NY, USA. Association for Computing Machinery.

Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. Beware of your po! measuring and mitigating ai safety risks in role-play fine-tuning of llms.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Longwei Zheng, Fei Jiang, Xiaoqing Gu, Yuanyuan Li, Gong Wang, and Haomin Zhang. 2025. Teaching via llm-enhanced simulations: Authenticity and barriers to suspension of disbelief. *The Internet and Higher Education*, 65:100990.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025*, pages 26101–26110. AAAI Press.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning, ICML 2024*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2024a. Starling-7b: Increasing llm helpfulness & harmlessness with rlaif. In *First Conference on Language Modeling*.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2025. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. In *Proceedings of the ACM on Web Conference 2025, WWW 2025*, pages 4653–4661. ACM.

Minjun Zhu, Linyi Yang, and Yue Zhang. 2024b. Personality alignment of large language models. *CoRR*, abs/2408.11779.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A Background

A.1 Simulation in Other Fields

The foundations of simulation trace back to early philosophical inquiries in Ancient Greece, where understanding processes was a precursor to formal simulation. Since the Industrial Age, simulation has played a central role across scientific disciplines—ranging from physical systems like wind tunnels to computational fluid dynamics. While physics-based simulation, including recent advances in diffusion models, remains well-established, it is not the focus of this work.

Simulation has also been widespread in computer science. For example, the cascade model in social networks (Kempe et al., 2003) simulates how influence spreads through interactions among users. In information retrieval and recommender systems, click models simulate user behavior by modeling how users interact with ranked lists of items (Richardson et al., 2007; Craswell et al., 2008; Chuklin et al., 2015). These models support both system optimization (Kveton et al., 2015; Lagrée et al., 2016) and offline evaluation (Joachims et al., 2017; Li et al., 2018).

Simulation is particularly integral to reinforcement learning (RL), which often requires data-hungry training. Early RL research demonstrated success using handcrafted simulators for controlled domains such as backgammon (Tesauro, 1994), elevator control (Crites and Barto, 1995), and mountain car (Munos and Moore, 1999). Today, standardized simulators like MuJoCo (Todorov et al., 2012) provide high-fidelity environments for physical control tasks. These simulators are grounded in well-understood dynamics, unlike human behavior, which remains far more complex and less predictable. Building large-scale, diverse human simulators remains a core challenge.

Inspired by RL success in physical environments, researchers have also developed simulators for human-centric domains such as recommender systems (Rohde et al., 2018; Ie et al., 2019). However, most of these frameworks require learning user simulators from domain-specific data, a task complicated by weak sequential signals in common datasets like MovieLens (Harper and Konstan, 2015), where preference patterns often reflect static biases rather than temporal interactions (Koren et al., 2009).

In statistics, simulation also plays a role in counterfactual reasoning, which estimates treatment effects from logged data without costly or infeasible A/B tests (Bottou et al., 2013). Matching methods (Stuart, 2010) attempt to estimate treatment effects by pairing treated and untreated samples with similar covariates, though imbalance often introduces bias. Propensity-based estimators—such as inverse propensity scoring (Horvitz and Thompson, 1952), clipped estimators (Ionides, 2008), and doubly robust approaches (Robins et al., 1994; Dudík et al., 2014)—face similar issues. In such settings, simulation holds promise as a way to impute missing data or generate counterfactual outcomes, including for complex objects like conversations.

A.2 Human Modeling

Modeling of human preferences has been studied extensively. Pairwise preferences are often modeled using the Bradley-Terry-Luce model (Bradley and Terry, 1952). Permutation preferences are frequently modeled using the Plackett-Luce model (Plackett, 1975). These models are commonly known as discrete choice models (Train, 2009). Modeling and learning from human feedback has recently found many applications in machine learning (Christiano et al., 2017), notably in reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and direct preference optimization (DPO) (Rafailov et al., 2023). Generative human simulation take these efforts to the next level, by modeling complex human interactions with humans and not just relative preferences.

A.3 Comparing to Existing Surveys

Recently, related surveys have been published in the domain. However, we want to point out that our surveys differs in multiple aspects comparing to existing surveys, as shown in Table 2. First, instead of focusing on a simple simulation target (Chen et al., 2024b), we provide a comprehensive analysis on the variety of targets in the who section. Additionally, although data generation is a key conversational application (Soudani et al., 2024), we do not limit our discussion to data generation by including extended discussions on various forms of simulations such as individual simulation. Last but not least, we focus exclusively on conversational data, which enables us to provide a more detailed and systematic discussion on conversation specific topics and techniques such as conversation paradigms comparing to prior works (Zhang et al., 2025a; Balog and Zhai, 2025).

Table 2: Comparison of Existing Surveys and Our Survey on LLM-based Conversational User Simulation. Symbols indicate coverage level: ● = covered; ○ = not covered; ◎ = partially covered.

Survey	Who (User-level)	Who (Persona / Role-play)	Who (Global / Default)	What (Conv. Paradigm)	How (Techniques)	Conv. Evaluation	Applications
Zhang et al. (2025a)	●	◎	○	○	●	○	◎
Balog and Zhai (2025)	◎	◎	○	○	●	○	◎
Soudani et al. (2024)	○	○	●	◎	○	○	●
Chen et al. (2024b)	○	●	○	◎	◎	◎	●
Ours	●	●	●	●	●	●	●

Taxonomy	Category	Description
Who (Section 3)	General User (Sec. 3.1)	This class of techniques performs simulation of a general user conversation. This is the most general class of techniques.
	Persona-level (Sec. 3.2)	This class of techniques leverages demographics and other persona-based data to simulate conversations
	Role Play (Sec. 3.3)	The class of techniques leverages role playing to simulate conversations based on specific
	Individual User (Sec. 3.4)	The class of conversational simulation techniques focus on generating specific conversations and other facets tailored to a specific user.
	Hybrid (Sec. 3.5)	This class of techniques combines techniques from two or more of the prior categories of techniques.
What (Section 4)	Human-AI (Sec. 4.1)	This class of conversation objective simulates the interaction between human and AI. Commonly used for instruction-finetuning.
	Human-Human (Sec. 4.2)	This class of conversation objective focuses on simulating the conversation between human, often associated with assigned personalities.
	AI-AI (Sec. 4.3)	This class of conversation objective revolves around the open-ended interaction between two or more AI agents.
	Many Human-AI (Sec. 4.4)	This class of conversation objective explores the collaborative conversation trajectories between multiple human and AI.
	Hybrid (Sec. 4.5)	This class of conversation objective combines two or more of the above conversational simulation paradigms.
How (Section 5)	Prompt-based (Sec. 5.1)	This class of techniques directly uses prompts to steer LLMs toward simulating user conversations.
	RAG (Sec. 5.2)	Retrieval-Augmented Generation methods incorporate external knowledge into simulation by retrieving relevant context.
	Fine-Tuning (Sec. 5.3)	These techniques adapt LLMs to user simulation tasks through supervised training (e.g. LoRA) on dialogue data.
	RL/DPO (Sec. 5.4)	This class trains user simulators using feedback-driven optimization such as Reinforcement Learning and Direct Preference Optimization.
	Hybrid (Sec. 5.5)	Hybrid methods combine prompting, retrieval, fine-tuning, and RL/DPO, often using modular or multi-stage architectures.

Table 3: Taxonomy of User Conversation Simulation (Section 3-5).

B Who: Simulation Target (Extended)

B.1 General User Simulation

To optimize customized dialogue systems for general user interactions, recent work has focused on improving conversational performance at the turn level and trajectory level. M-DPO (Xiong et al., 2025) introduces a multi-turn online iterative framework for direct preference learning, specifically designed to handle multi-turn reasoning and tool integration. Building on trajectory-based optimization, ETO (Song et al., 2024) develops an exploration-based approach that learns from past exploration trajectories, including failure cases, to improve performance. Additionally, to address training noise in long conversations, SDPO (Kong et al., 2025) fo-

Table 4: Taxonomy of General User Simulation

Category	Work(s)
Trajectory Optimization	M-DPO (Xiong et al., 2025), SDPO (Kong et al., 2025), LOOP (Chen et al., 2025a)
Exploration Optimization	ETO (Song et al., 2024), AgentQ (Putta et al., 2024)
Diverse User Simulation	KAUCuS (Dhole, 2024)

Table 5: Taxonomy of Persona-level User Simulation.

Category	Work(s)
Descriptive Features	Prompting (Hu and Collier, 2024), PsyPlay (Yang et al., 2025)
Personality Framework	Orca (Huang, 2024), HEXACO (Ji et al., 2024), PB&J (Joshi et al., 2025)
Benchmarks	PersonaCatch (Li et al., 2025a), PersonaEval (Wang et al., 2025b)

cuses on leveraging specific meaningful segments within conversations while minimizing the impact of less relevant interactions. AgentQ (Putta et al., 2024) further improves the trajectory exploration by addressing the sub-optimal policy outcomes due to compounding errors and limited exploration data. It combines Monte Carlo Tree Search (MCTS) with self-critique and iterative fine-tuning, learning from both positive and negative conversational trajectories.

More recently, LOOP (Chen et al., 2025a) trains interactive assistants directly in their target environments by formulating the training as a partially observable Markov decision process. This approach uses a data and memory-efficient variant of Proximal Policy Optimization (PPO) that eliminates the need for value networks and requires only a single copy of the language model in memory. Complementing these optimization approaches, work by Dhole (2024) focuses on generating diverse user simulations that capture varied interaction patterns between users and assistants. These simulations provide training data for developing more helpful and robust conversational agents that can handle the breadth of general user interactions.

B.2 Persona-level User Simulation

Persona-level simulation primarily relies on explicitly defining user characteristics for LLMs. Early approaches use direct prompting with descriptive attributes (e.g., demographics) (Hu and Collier, 2024; Qiu et al., 2024), while more recent work

Table 6: Taxonomy of *Role Play Simulation*.

Category	Method(s)
Frameworks	Ditto (Lu et al., 2024), CharacterLLM (Shao et al., 2023) PCL (Ji et al., 2025) RoleCraft (Tao et al., 2024) CharMap (Xu et al., 2024c)
Evaluation Benchmarks	RoleBench (Wang et al., 2024a), CharacterBench (Zhou et al., 2025), PsyPlay-Bench (Yang et al., 2025)
Challenges & Risks	RoleBreak (Tang et al., 2025), Toxicity (Deshpande et al., 2023)

grounds simulation in psycho-social theory by incorporating established personality frameworks such as the Big Five (Wang et al., 2025b; Jiang et al., 2023; Li et al., 2025a; Yang et al., 2025; Jiang et al., 2024; Serapio-García et al., 2023) and HEXACO (Ji et al., 2024). These efforts enable LLMs to simulate users with more realistic and consistent social psychometrics. Frameworks like Orca (Huang, 2024) augment generation with personal context, and PsyPlay (Yang et al., 2025) builds personality-infused agents capable of portraying designated traits. Beyond persona, more recent work (Mehri et al., 2025) focuses on aligning the goals of the simulated persona.

Beyond prompting, fine-tuning methods such as Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) (Li et al., 2025a) offer greater control, while techniques like Personality Activation Search (PAS) (Zhu et al., 2024b) adjust internal activations to align with group preferences. These approaches demonstrate that LLM personas can consistently reflect assigned traits in psychometric tests (Wang et al., 2025b) and exhibit representative linguistic patterns (Jiang et al., 2024).

However, a critical concern is that persona assignment may introduce systematic biases, leading to skewed outcomes such as political forecasts (Li et al., 2025a), or increased toxicity (Deshpande et al., 2023), potentially reinforcing harmful stereotypes. These limitations underscore the need for better validation (Hu and Collier, 2024) and more trustworthy simulation practices.

B.3 Role Play Simulation

Role play simulation is fundamentally defined by the implicit knowledge of language models (Serapio-García et al., 2023; Lù et al., 2025), which allows LLMs to inherently associate the role with a rich set of personal traits and psychosocial

status (Wang et al., 2024a). The capability of LLMs to capture nuanced relationships within extensive textual data means they can simulate more diverse and complex behaviors, reflecting the variability of human decision-making (Wang et al., 2025b).

To elicit and shape this implicit knowledge, various techniques are proposed. Direct "role-play prompting" instructs LLMs to generate like a specific character, leveraging their steer-ability to tailor output style and tone (Wang et al., 2025b, 2024a). More advanced methods focus on training LLMs as "trainable agents" specifically for role-playing, improving consistency and naturalness through fine-tuning on character-related dialogues (Shao et al., 2023; Lu et al., 2024; Tao et al., 2024). More recently, Ji et al. (2025) demonstrates that contrastive learning through adversarial self-play can enhance character consistency and prevent deviations from the assigned character. Qiu and Lan (2025) further shows that letting the role play agent to reconstruct the masked dialogue can improve the generation quality.

Leveraging the implicit character knowledge allows role-playing agents to make persona-driven decisions that align with their assigned identities (Xu et al., 2024c; Park et al., 2023; Qiu and Lan, 2024). This capability moves role-play beyond mere conversation to influence choices in various scenarios, with character-level benchmarks introduced to evaluate decision-making consistency (Xu et al., 2024c). Role-playing can enable complex "drama-interaction" systems where multiple LLMs simulate intricate social dynamics (Wu et al., 2024) or act as interactive simulacra of human behavior in social science research (Wang et al., 2025b; Park et al., 2023). The ability for an LLM to "know its own story" as a character, including its past and experiences, is an emerging area of research exploring lifelong learning in LLMs (Fan et al., 2025).

The reliance on implicit knowledge also introduces significant challenges and risks. A primary concern is *character hallucination*, where LLMs generate content inconsistent with the assigned character's known attribute. For example, asking a Mozart character programming problems can lead to unexpected output. Recent research has shown that such hallucination can be exploited as jailbreak attacks (Tang et al., 2025; Sadeq et al., 2024). In addition, pre-existing biases tied to famous figures or demographics in the training data can lead to increased toxicity and the propagation of incorrect stereotypes when LLMs role-play specific individuals, which can result in outputs that are defamatory or harmful (Deshpande et al., 2023). Furthermore, the robustness of role-playing can decrease as the complexity of the roles increases (Wang et al., 2025b), and LLMs might exhibit *persona collapse* or low consistency in challenging situations (Xu et al., 2024c).

B.3.1 Evaluation on Role-Playing Benchmarks

This subsection summarizes representative evaluation results on commonly used role-playing benchmarks, illustrating how different modeling paradigms perform in character simulation and role consistency.

RoleLLM. Table 7 below reports results on RoleLLM (Wang et al., 2024a), where CUS measures character understanding, RAW evaluates response appropriateness, and SPE assesses role-specific knowledge. Prompt-based large models achieve strong overall performance, while fine-tuned models—despite smaller model sizes—show competitive results and notably stronger role knowledge (SPE).

Table 7: Evaluation on RoleLLM (Wang et al., 2024a).

Model	CUS	RAW	SPE	Avg.
RoleGPT (Prompt-Based)	57.6	53.2	32.3	47.7
ChatPLUG (RAG w/ LLaMA)	24.0	34.7	25.8	28.2
Character.AI (Prompt-Based LLaMA)	41.9	45.7	30.3	39.3
RoleLLaMA-7B (Fine-tuned)	32.9	37.6	38.1	36.2
RoleLLaMA2-13B (Fine-tuned)	37.5	47.9	48.8	44.7

WikiRole. Table 8 presents evaluation on WikiRole (Dinan et al., 2019), focusing on role accuracy and explicit role knowledge. While large prompt-based models perform well overall, fine-tuned models show clear gains in role knowledge.

Table 8: Evaluation on WikiRole (Dinan et al., 2019).

Model	Accuracy	Role Knowledge
GPT-4 (Prompt-Based)	80.0	76.2
CharacterGLM (Self-Play Fine-tuned)	75.0	47.3
Qwen-72B + Chat	90.0	66.4

RoleInstruct. Table 9 reports results on RoleInstruct (Tao et al., 2024), where fine-tuned models consistently outperform prompt-only baselines on role-specific knowledge (SPE).

Table 9: Evaluation on RoleInstruct (Tao et al., 2024).

Model	CUS	RAW	SPE	Avg.
GPT-4 (Prompt-Based)	52.0	55.7	28.3	45.3
RoleGLM (Fine-tuned)	50.5	52.6	34.0	45.7
RoleCraft-GLM (Self-Play Fine-tuned)	51.5	53.8	35.7	47.0

Summary. Across role-playing benchmarks, large prompt-based models benefit from scale and implicit knowledge, achieving strong baseline performance. However, fine-tuned models—often built on smaller backbones—consistently improve role-specific knowledge and character consistency, particularly on SPE metrics. In contrast, retrieval-augmented approaches such as ChatPLUG generally underperform, suggesting that fine-tuning is more effective than RAG for role simulation and character elicitation.

B.4 Individual User Simulation

The most straightforward individual user simulation paradigm is through the integration of user profiles into prompts. For example, PERSONA-CHAT (Zhang et al., 2018) provides crowdworkers with multi-sentence user profiles, enabling individualized realistic user simulation. The FoCus dataset (Jang et al., 2022) extends this idea by incorporating Wikipedia-based facts into persona grounding. These approaches enhance the response by aligning responses with a user's stated context.

In order to incorporate more nuanced individual profiles, another line of research leverages a user's *dialogue history* to learn communication style and evolving preferences. PHMN (Li et al., 2021) extracts long-term dialogue patterns such as personalized word usage and context attention, improving multi-turn response selection.

To handle long-range personalization, *multi-session memory* mechanisms have been introduced. The MSC dataset (Xu et al., 2022) captures multiple-session chats with user-specific memory summaries to enable coherent re-engagement. Mem0 (Chhikara et al., 2025) proposes a production-ready memory architecture that dynamically extracts and consolidates user memories across sessions.

Finally, more recent efforts collect *real-world personality traits* to ground simulation in natural behavior. The RealPersonaChat dataset (Yamashita et al., 2023) uses real personality scores (e.g., Big Five traits) in free-form Japanese chats, helping the model reflect actual personality cues.

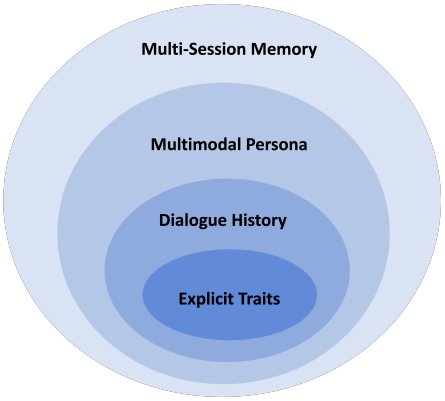


Figure 3: Taxonomy of Individual User Simulation. Explicit traits

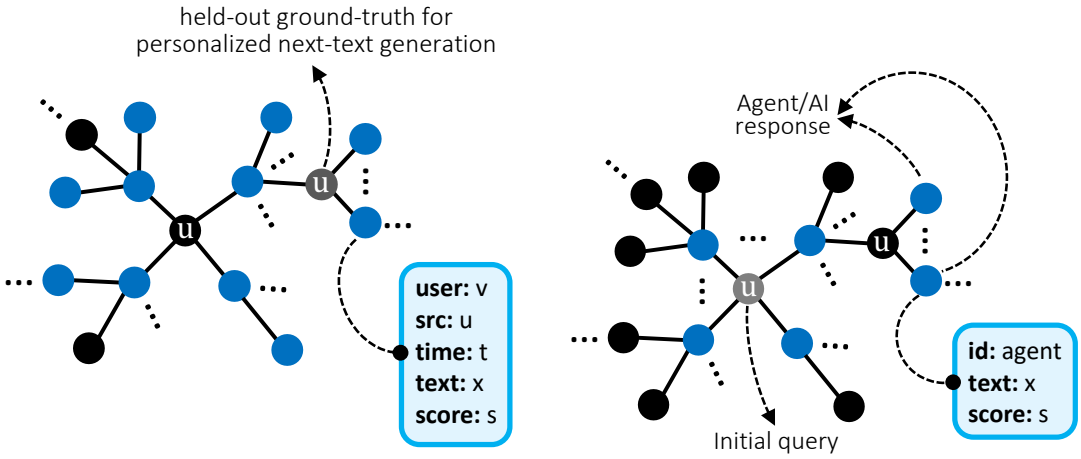
LiveChat (Gao et al., 2023) mines detailed persona profiles and interactions from live-streaming platforms, constructing a large-scale Chinese corpus with naturally occurring individual variation.

C What: Simulation Objectives (Extended)

C.1 Human-AI Simulation

Generating such conversations has attracted significant attention due to the scarcity of high-quality, diverse human-annotated data and the high cost of large-scale annotation (Wang et al., 2023; Xu et al., 2024a; Li et al., 2024b). Recent studies have proposed various frameworks for synthesizing instruction-following datasets. One prominent approach focuses on bootstrapping instruction datasets through self-generation. The self-instruct framework (Wang et al., 2023) introduces a semi-automated pipeline in which a pretrained LLM generates synthetic instructions and input-output pairs. These synthetic examples are filtered and then used to fine-tune the model itself, improving its instruction-following capabilities. Building upon this idea, WizardLM (Xu et al., 2024a) progressively rewrite simple instructions into more complex forms, enabling LLMs to handle multi-step reasoning and diverse task demands. To further reduce reliance on heuristic design, Auto Evol-Instruct (Zeng et al., 2024) fully automates the evolution process, iteratively optimizing instruction datasets without any human intervention.

Another line of work emphasizes generalized and domain-agnostic data generation. GLAN (Li et al., 2024b) introduces a paradigm inspired by educational taxonomies, decomposing human knowledge into fields to generate instructions span-



(a) Human-Human Conversation Trajectories

(b) Human-AI Conversation Trajectories

Figure 4: Comparing different types of conversational trajectories starting from an initial input query by the user.

ning a wide range of domains. Building on this idea in multimodal contexts, Multimodal Self-Instruct (Zhang et al., 2024e) synthesizes visual instruction reasoning tasks using abstract images. ProphetChat (Liu et al., 2022) further extends this direction by simulating future conversations based on existing dialogue histories.

A complementary strategy focuses on task-specific and tailored data alignment. Recognizing that different downstream applications require distinct instruction distributions, CodecLM (Wang et al., 2024d) introduces a metadata-based Encode-Decode process to generate high-quality synthetic data tailored for specific tasks. In parallel, Chan et al. (2024) empirically evaluate trade-offs in synthetic data generation strategies under various resource constraints, providing practical guidelines for selecting the augmentation methods. Additionally, DataDreamer (Patel et al., 2024) provides a unified library for chaining synthetic data generation, fine-tuning, and evaluation, emphasizing reproducibility and practical pipeline.

In addition, a growing body of work explores domain-specific human–machine simulations tailored to specialized applications. For instance, synthetic patient–physician dialogues are generated to support medical AI systems (Das et al., 2024), while museum guidance agents are developed through curated human–AI conversations (Rachidi et al., 2025). These efforts highlight the importance of aligning simulation data with the nuanced requirements of different domains.

C.2 Human-Human Simulation

Early sequence-to-sequence chatbots demonstrated fluency but often produced generic responses and lacked consistent speaker personalities (Vinyals and Le, 2015). To address this, PersonaChat (Zhang et al., 2018) provide each speaker with a personal profile of facts, encouraging models to maintain a coherent identity throughout the conversation. Beyond persona consistency, other works injected additional context to make dialogues more natural. Wizard-of-Wikipedia dataset (Dinan et al., 2019) tasks one participant (the "wizard") with retrieving and using Wikipedia facts during conversation. Models trained on these dialogues learn to incorporate factual knowledge into their responses, enabling more informative and substantive discussions. The EmpatheticDialogues dataset (Rashkin et al., 2019), consisting of 25k human–human conversations, was collected to teach models how to recognize and respond to a conversation partner's emotions. Systems fine-tuned on these emotionally grounded dialogues have been shown to be perceived as significantly more empathetic by human evaluators, compared to models trained only on general chat data.

In *task-oriented dialogue* domains (e.g., booking tickets), human–human simulation is equally important. Large-scale corpora such as MultiWOZ (Budzianowski et al., 2018) illustrate the collection of human–human written conversations spanning multiple domains for developing task-oriented agents. MultiWOZ contains around 10k dialogues collected via crowd workers role-playing

as user and assistant. While such Wizard-of-Oz style data provide valuable training and evaluation scenarios, creating a new human-human dataset for every domain is time-consuming and expensive.

To mitigate the problem of data scarcity, researchers have explored *synthetic dialogue generation*. Shah et al. (2018) proposed a self-play bootstrapping approach for task-oriented dialogues: using a predefined domain schema and slot-filling logic, a simulated "user" and "agent" were auto-generated to produce dialogues, which were then paraphrased and validated by crowd workers. This method produced additional training trajectories without relying entirely on manually written dialogues, effectively augmenting the data for learning dialog policies.

C.3 AI-AI Simulation

One primary use case of AI–AI simulation is the study of *emergent behavior and social dynamics* by placing multiple LLM-driven agents in shared virtual environments and allowing them to interact over extended time horizons (Haase and Pokutta, 2025). Rather than optimizing for a specific task, these simulations aim to observe how complex interaction patterns arise from simple initial conditions and agent profiles. For example, Park et al. (Park et al., 2023) instantiated a small community of generative agents, each equipped with a unique backstory, memory module, and behavioral script. These agents autonomously engaged in open-ended conversations, formed social relationships, and coordinated group activities—exhibiting behavior reminiscent of real-world communities. CRSEC (Ren et al., 2024b) further examined how repeated local interactions among agents lead to the emergence of shared social norms and behavioral conventions over time. Similarly, AgentSociety (Piao et al., 2025) demonstrated that large-scale LLM-based simulations can replicate macro-level social patterns and align with real-world trajectories of community evolution. Such results suggest that today's large language models are not only capable of coherent multi-agent dialogue but can also simulate realistic, temporally grounded social processes. Consequently, AI–AI simulation has emerged as a powerful tool for generating rich interaction data and conducting behavioral evaluations at scale, offering a scalable and cost-effective alternative to traditional human-annotated datasets.

Beyond data augmentation, AI–AI simulation has become a framework for *multi-agent cooperation* (Guo et al., 2024). CAMEL (Li et al., 2023) uses a role-playing paradigm - one agent is assigned as the "user" (task setter) and another as the "assistant" (solution provider) - to autonomously interact to complete a given task. Similarly, AutoGen (Wu et al., 2023) defines multiple conversable agents (LLMs, tools, or human proxies) that communicate with each other to complete complex tasks. These frameworks demonstrated that solutions obtained via these AI–AI conversations can outperform those from a single agent working alone.

While many AI–AI simulations focus on collaboration, others explore the *adversarial* dynamics among agents. For instance, Du et al. (2024) demonstrated that adversarial debate between LLMs can lead to improvements in both factuality and reasoning, as the models iteratively challenge and refine each other's claims. Furthermore, Rennard et al. (2025) introduced a self-debate framework in which two instances of a language model argue opposing viewpoints to persuade a third, neutral LLM judge. This setup enables the examination of how persuasive strategies and biases propagate across models. More recently, Hua et al. (2024) studied adversarial behavior in game-theoretic scenarios, revealing that unstructured LLMs often deviate from rational strategies. To mitigate this, they introduced *agentic workflows* that guide LLM reasoning through structured decision-making procedures, enabling more consistent alignment with game-theory principles.

C.4 Many Human-AI Simulation

Historically, most research on human–AI collaboration has centered on single-user scenarios, where the AI functions as a personal assistant or tutor. However, there is growing recognition that future AI agents will need to operate as members of teams rather than merely serve individual users in isolation. Recent work has begun to explore this paradigm shift. For example, ChatCollab (Klieger et al., 2024) treats the AI agent as a peer collaborator within a team-based software engineering setting. The system enables joint problem-solving among multiple human participants and one or more AI agents, and the results suggest that AI can occupy differentiated roles within human teams—such as idea contributor, critic, or coordinator—mirroring the diversity of human collabora-

tion styles.

Additionally, Dittos (Leong et al., 2024) introduces a personalized AI agent capable of standing in for a human teammate during remote meetings. These embodied agents are not merely passive note-takers but are designed to actively participate on behalf of the absent user. Experiments show that such agents can evoke feelings of presence and trust among human team members, even contributing meaningfully to group discussions. Furthermore, the Multi-User Chat Assistant (MUCA) (Mao et al., 2024) framework introduces a simulation-based approach to multi-party dialogue. MUCA employs a multi-user simulator to mimic the behaviors of several distinct human participants, enabling the training and evaluation of group-aware AI assistants. By modeling not just individual utterances but the evolving group dynamics over time, MUCA facilitates the development of more socially intelligent AI agents capable of navigating complex multi-user settings.

Despite these promising directions, research on many-human–AI simulation remains relatively underexplored. Most existing systems are still grounded in narrow, task-specific environments, and there is a clear need for more general-purpose frameworks and benchmarks that reflect the diversity of real-world collaboration. As AI becomes increasingly integrated into workplace, educational, and social group contexts, developing robust simulation tools for many-human–AI interaction will be essential for building agents that are trustworthy, effective, and contextually aware in group settings.

D How: Techniques and Methodologies (Extended)

D.1 Prompt-Based Simulation

Zero-/Few-Shot Prompting. A zero-shot prompt provides only high-level instructions; a few-shot prompt additionally lists a handful exemplars. Terragni et al. (2023) use few-shot in-context examples to prompt GPT-NeoX for task-oriented dialogue (TOD), achieving diverse responses without fine-tuning. Zhang et al. (2024a) represent each synthetic recommender-system user as an LLM endowed with a profile, memory, and action space, enabling near-zero-shot behavioral simulation. Wang et al. (2025b) extend this idea to Web environments, showing that LLM agents can simulate browsing and clicking behaviors with minimal user data.

Chain-of-Thought Prompting. CoT prompts ask the model to reason step-by-step before producing the final utterance, thereby improving logical consistency. DuetSim (Luo et al., 2024) adopts a generator–verifier loop: one LLM enumerates dialogue acts in CoT form, while a second verifies and verbalizes them, leading to stronger goal coherence.

Persona / Role-Play Prompting. This prompting strategy assigns the LLM a persona (e.g. "novice buyer"), thereby controlling tone and constraints (Shanahan et al., 2023). Abbasiantaeb et al. (2024) run two GPT-4 agents (student/teacher) to generate multi-turn QA dialogues automatically. Kong et al. (2024) propose a Socratic user simulator that generates more natural, diverse questions than earlier role-play pipelines such as Baize (Xu et al., 2023) and UltraChat (Ding et al., 2023).

Task-Specific Prompting. These prompts encode detailed domain constraints or required actions so that generated utterances align tightly with an application. Terragni et al. (2023) supply GPT-4 with a high-level TOD goal and a few exemplar dialogues, letting the model adapt across domains. For conversational search, Kiesel et al. (2024) prompt an LLM to ask plausible follow-up questions given a system answer, producing queries that match human behavior on both automatic metrics and human evaluation.

D.2 RAG

We categorize RAG-based simulation techniques by their retrieval trigger mechanism: *Always-on*, *Adaptive*, and *Goal / State–driven*.

Always-on. Retrieval is applied at every turn, regardless of necessity. For example, KAUCUS(Shimadzu et al., 2025) introduces a framework for building diverse user simulators by leveraging external text. Its SRAG (Simulator with RAG) model prepends a retrieved passage to each user prompt. Specifically, for every user turn in the training data, SRAG queries MS MARCO using the previous utterance (via BM25) and prepends the top-ranked passage before generating the next response. This retrieval step enriches the simulator with factual, varied content, enhancing the diversity and realism of simulated behaviors.

Shimadzu et al. (2025) extend this idea to large-scale social network simulations. Each LLM-based agent mimics human behavior by retrieving up-

Table 10: Taxonomy of User Simulated Data Generation Techniques.

		Description
Prompt-based Generation	Zero-shot	No examples provided, direct inference
	Few-shot	Inference guided by few examples
	Chain-of-thought	Intermediate reasoning steps provided
Fine-tuning Methods	Supervised Fine-tuning (SFT)	Learning from explicit labeled data
	Reinforcement Learning (RLHF)	Learning via feedback-driven reward signals
Hybrid Approaches	Retrieval-augmented Generation	Generation enhanced by external information retrieval
	Persona-based Generation	Generation personalized based on user or persona context
	Persona-based RAG	Generate based on persona context and RAG

to-date content (e.g., from the web) before generating posts or replies. This retrieval step enables agents to discuss trending topics beyond the LLM's pretraining, leading to more natural and timely interactions. Compared to static simulators, these retrieval-augmented agents better capture dynamic social media behavior.

Adaptive. Adaptive-gate methods use a learned classifier to decide, at each turn, whether retrieval is necessary. This selective retrieval strategy improves efficiency by avoiding unnecessary queries and reduces noise. For instance, RAGate (Wang et al., 2025c) introduces an LLM-based gating mechanism trained on human-labeled dialogues to predict whether to retrieve external knowledge. Given the dialogue history and user/goal context, the gate outputs a binary decision—retrieve or not. By augmenting only relevant turns, RAGate improves simulation efficiency.

Goal / State Driven. Retrieval is guided by the simulator's internal state—such as user goals or memory contents. Zhu et al. (2025) propose CSHI, a modular simulator for conversational recommendation systems that maintains structured user preference memory. At each turn, CSHI retrieves relevant preferences to generate contextually appropriate responses. For instance, when asked about likes or recommendations, the simulator queries its long-term and real-time memory for matching items or traits. It also dynamically updates its memory as new preferences emerge during the interaction. This goal- and memory-driven retrieval helps ensure coherence and personalization in simulated user behavior.

D.3 Fine-tuning

We group existing methods into three broad categories: *Full-model supervised*, *Parameter-efficient*, and *Interactive / self-play* fine-tuning.

Full-model supervised. These works update *all* weights on in-domain dialogues. DAUS (Sekulic et al., 2024) fine-tunes LLaMA-7B on two task-oriented dialogue (TOD) datasets, halving hallucination rates and improving goal fulfillment. In the affective domain, SoulChat (Chen et al., 2023) performs full SFT on a 2 M-turn empathy corpus to raise human-rated listening and comfort scores. MuPaS (Wang et al., 2024c) extends the idea to *multi-party* settings by masking inactive speakers during SFT so a single model can generate coherent group conversations.

Parameter-efficient. Instead of updating the full model, these methods insert lightweight adapters. ESC-Role (Zhao et al., 2024b) LoRA-tunes Qwen-14B on merged ESCConv/ExTES/Smile data to produce realistic emotional-support seekers. BiPO steering vectors (Cao et al., 2024) learn small activation-space adapters that steer an LLM toward target dialogue strategies without touching the backbone. Another paper (Madani et al., 2024) combines a Strategy Relevance Assessment (SRA) gate with LoRA fine-tuning: strategy-bearing tokens are detected, and ground-truth strategy-conditioned responses are injected mid-sequence to prevent intention drift in lengthy emotional-support conversations.

Interactive / self-play. Here, the simulator is refined with *interaction signals* rather than static targets. UGRO (Hu et al., 2023) lets an LLM act as a user-side judge that scores candidate replies; those rewards are used to fine-tune the TOD model with PPO, boosting task success. PlatoLM (Kong et al., 2024) first SFTs a 'Socratic' questioner, dialogues it with ChatGPT to create 600 k self-play conversations, then fine-tunes the assistant, achieving MT-Bench SOTA for 7 B models.

D.4 RL/DPO

We categorize RL/DPO-based simulation techniques by the dimension of conversational adaptation they target: *Personalization*, *Memory*, *Long-Horizon Planning*, *Offline/Hindsight Learning*, and *Action-level Clarification*.

Personalization. Personalization-driven RL/DPO encourages simulators to infer and adapt to individual user traits or latent preferences through the course of interaction. For example, Wan et al. (2025) propose a curiosity-based reward in RLHF that motivates the simulator to elicit and disambiguate latent user types. The policy is rewarded for actions that sharpen its belief over the user's profile, resulting in more tailored and proactive questioning during multi-turn dialogues. This approach outperforms standard RLHF on simulated user personalization tasks.

Memory. Memory-aware RL/DPO methods optimize what content from past dialogue should be retained or recalled to maximize future conversational quality. Seo et al. (2024) introduce a DPO-based memory selector that dynamically updates the simulator's external memory. By learning from preferences over different memory configurations, the simulator achieves improved factual recall and coherence in follow-up turns, while maintaining a compact memory state.

Long-Horizon Planning. Long-horizon RL approaches address the challenge of assigning credit and planning over many dialogue turns. Zhou et al. (2024) present ArCHer, a hierarchical RL method with a high-level Q-learning policy for utterance-level goals and a low-level PPO agent for token-level realization. This structure enables the simulator to optimize behaviors spanning dozens of turns, drastically improving sample efficiency and strategic coherence in multi-turn scenarios such as tool use or customer support.

Offline and Hindsight Learning. Offline and hindsight-based RL methods improve simulator policy using retrospective credit assignment or offline data augmentation, reducing the need for costly interactive roll-outs. Hong et al. (2024) employ "hindsight regeneration" to rewrite suboptimal dialogue segments after observing user reactions, then fine-tune the simulator offline on these augmented interactions. This yields more effective user steering in counseling and persuasion settings compared to standard imitation or prompt-based approaches.

Action-level Clarification. Action-level DPO techniques optimize the simulator's ability to resolve ambiguity by learning when and how to clarify user intent. Chen et al. (2025b) frame clarification as a preference optimization problem—contrasting possible next actions (clarifying vs. answering vs. delegating)—and train the simulator to prefer actions that maximize downstream task success and user satisfaction. Their method achieves higher performance on multi-turn QA and goal-oriented dialogue tasks relative to standard supervised or DPO training.

D.5 Hybrid Approaches

We organize hybrid simulation techniques by the integration pattern they employ:

Retrieval-Augmented Fine-Tuning. Some approaches inject retrieved passages *during fine-tuning* instead of only at inference. For example, Zhang et al. (2024d) introduce RAFT, which fine-tunes LLaMA-derived models using triplets of query, retrieved passages (including distractors) and chain-of-thought answers, training the model to cite relevant evidence and disregard irrelevant context while improving performance on domain-specific benchmarks such as PubMed, HotpotQA, and Gorilla (Zhang et al., 2024d). Similarly, Kaiser and Weikum (2025) propose PRAISE for conversational QA: it trains modular adapters through Direct Preference Optimization by iteratively generating, ranking, and fine-tuning on top-ranked (positive) versus lower-ranked (negative) context–answer samples, enhancing factual grounding and response robustness (Kaiser and Weikum, 2025). These hybrid retrieval-and-fine-tuning methods outperform both standard RAG and vanilla supervised fine-tuning in terms of factual accuracy and conversational resilience.

Prompt-to-Fine-Tune Curricula. A common hybrid strategy uses prompt-based simulators to bootstrap high-quality synthetic data, then fine-tunes adapters or the full model on this expanded corpus. Chen et al. (2023) employ zero-shot GPT-4 role-play to generate millions of empathetic dialogues, followed by LoRA-based fine-tuning for fast, domain-adapted simulation. Kong et al. (2024) use Socratic self-play to create hundreds of thousands of multi-turn conversations, then apply su-

pervised fine-tuning to stabilize persona and dialogue quality. This curriculum leverages the rapid adaptability of prompts and the efficiency and consistency of finetuned models.

RAG + RL/DPO Loops. Here, retrieval mechanisms are coupled with RL/DPO policies: the RL/DPO policy decides whether to retrieve, how to incorporate the retrieved context, and how to adjust simulator strategy based on feedback. MemDPO (Seo et al., 2024) learns a memory selector using preference optimization, enabling the simulator to maintain a compact, relevant external memory for better recall and coherence. Curiosity-RAGate combines curiosity-driven RLHF rewards (see §5.4) with adaptive retrieval, encouraging clarification queries only when the user's preferences are ambiguous (Wan et al., 2025).

Hierarchical Modular Pipelines. Some hybrid systems structure the simulation pipeline as a set of specialized modules, each using different methods. For example, ARCHer (Zhou et al., 2024) (see §5.4) employs a planner module with CoT prompting, a retriever module for web or tool queries, a finetuned executor module, and a DPO-trained critic, each collaborating to optimize long-horizon dialogue. Social-network simulators like KAUCUS-SRAG integrate always-on retrieval and RL-tuned posting policies to simulate dynamic trends (Shimadzu et al., 2025).

Personalized Hybrid Stacks. Cutting-edge personalized simulators operate on multiple levels: using persona-driven prompting, private memory retrieval, adapter-based fine-tuning for user-specific style, and RLHF for preference alignment. Recent surveys (see, e.g., Zhang et al. (2025a); Liu et al. (2025)) describe this as a three-tier system—input (prompt/RAG), model (parameter-efficient fine-tuning), and objective (RL/DPO alignment)—that has become standard in open-source frameworks.

E Evaluation (Extended)

E.1 Traditional Metrics

N-gram overlap. BLEU/ROUGE-style metrics are still reported in several dialog simulation settings (e.g., BLEU/F1 in SimDial (Zhao and Eskénazi, 2018), ROUGE/F1 in DialogStudio (Zhang et al., 2024c), ROUGE-L in RoleLLM (Wang et al., 2024a)). They are easy to compute but correlate poorly with human judgments in open-ended conversation, where many acceptable responses share little lexical overlap with references.

F1 slots / fact accuracy. When a dialogue contains well-defined, enumerable facts or slots (e.g., Wizard-of-Oz style tasks such as CrossWOZ (Zhu et al., 2020) or MultiWOZ (Ye et al., 2022)), precision/recall-style measures (F1) or exact match on slots/entities provide a clear signal for information correctness, but they ignore pragmatic qualities such as empathy or persona fidelity.

Perplexity. Perplexity (Chen et al., 1998) is sometimes reported to indicate distributional fit to a human corpus, but it is an intrinsic language model measure and does not necessarily reflect interactive quality or human preference.

Task success / accuracy. In goal-driven simulations such as negotiation (Lewis et al., 2017; He et al., 2018) or action decision (Chen et al., 2021), binary or graded success criteria are natural. These metrics are interpretable but only apply when a clear ground-truth objective is defined.

In practice, these traditional metrics offer low-cost, reproducible signals, but they measure narrow facets of dialogue quality. As a result, many recent works pair them with LLM or human judges to capture semantic and pragmatic dimensions that overlap metrics miss.

E.2 Human Evaluation

Human evaluation remains the reference standard. Two modes are prevalent. The first is *interactive* evaluation, where humans converse with a system (possibly paired with a simulator) and then rate satisfaction, coherence, or realism. The second is *offline* evaluation, where annotators read transcripts and provide Likert-scale ratings or pairwise preferences over full conversations or single turns. For example, MultiWOZ (Ye et al., 2022) reports human judgments on dialogue quality; role/character benchmarks such as CharacterBench (Zhou et al., 2025) and LifeStageBench (Fan et al., 2025) also rely on expert or crowd annotators for final scoring, sometimes in combination with model judges. Common protocols include: (i) Likert scoring on multiple axes (e.g., naturalness, coherence, goal completion, persona/role fidelity), (ii) pairwise A/B testing that asks which conversation (or response) is better along one criterion, and (iii) reporting inter-annotator agreement to quantify rating consistency.

The main downsides of human evaluation are cost, latency, and limited reproducibility across studies due to differences in annotator pools, rubrics, and scales. These issues motivate the growing use of LLM-as-Judge as a scalable proxy, followed by *meta-evaluation* that quantifies how well model-judged scores align with human ratings on shared subsets (Fan et al., 2025; Zhou et al., 2025). Nevertheless, for high-stakes or nuanced aspects such as safety, subtle persona drift, or social norm adherence, human studies remain the final arbiter.

F Datasets (Extended)

Personalized conversations. Personalized datasets (Li et al., 2025b; Zheng et al., 2019; Zhang et al., 2018) focuses on natural conversations in various domains. These conversations happen between individuals of diverse traits, including age, gender, location, and personal interests, among others.

Multiparty dialogues. Multiparty dialogues (Gao et al., 2023; Liu et al., 2023) are common on social platforms, where usually one person serves as a host or lead talker, while others engage in the conversations. These datasets usually involve addressee recognition to determine whom the host responds to.

Question answering. Question answering datasets contain conversations between two persons, one asking questions, the other answering. SimQuAC (Abbasiantaeb et al., 2024) consists of LLM-based simulated teacher-student conversations. These conversations focus on question-answering over Wikipedia articles, where the student asks questions to explore the article's topic, and the teacher responds based on the article. SocraticChat (Kong et al., 2024) contain conversations of simulated human-like questions. Furthermore, several datasets (Zhou et al., 2025; Shao et al., 2023; Tu et al., 2024; Wang et al., 2024a) contain dialogues where simulated responses are given based on characters from novels or TV shows.

Ranking. Nectar (Zhu et al., 2024a) contains conversations from various LLMs. The dataset is used for response ranking.

Multi-domain dialogues. SimDial (Zhao and Eskénazi, 2018) contains dialogues in multiple scenarios, including restaurants, movie, bus, and weather, among others.

Wizard-of-Oz. Wizard-of-Oz datasets (Byrne et al., 2019; Zhu et al., 2020; Ye et al., 2022) contains conversations between humans to help develop future conversational systems, or wizards. Each conversation contains two role, users and wizards, where users ask questions for a given goal, wizards give answers according to the questions.

Memory enhancement. A few datasets (Fan et al., 2025; Chen et al., 2024a) consists of dialogues that are generated by LLMs with focus on long-term conversation memory. Hence, these conversations are usually longer and have more connections among different turns.

Negotiation. Negotiation datasets (He et al., 2018; Lewis et al., 2017) consist of dialogues between two persons, where in DealOrNoDeal (Lewis et al., 2017) they are asked to divide a set of items, and in CraigslistBargain one person serves as a buyer to negotiate down item prices, while the other serves as a seller to profit as much as possible.

G Applications (Extended)

Scalable Evaluation of Dialogue Agents. Traditional user studies are expensive and difficult to replicate. Simulated users offer an efficient and reproducible alternative for evaluating dialogue systems across diverse scenarios, including rare edge cases. By configuring user profiles and behavioral trajectories, researchers can systematically assess model performance under controlled yet realistic conditions. This is particularly useful in domains such as personalized customer support (Patel and Trivedi, 2020), where user-specific behavior must be consistently modeled across sessions.

Training Robust and Adaptive Systems. In reinforcement learning or imitation learning settings, user simulators can generate large-scale, contextually rich interactions without requiring live deployment. This facilitates closed-loop training where conversational agents learn to handle ambiguity, recover from errors, and generalize across user intents. Simulation also supports curriculum learning, where complexity can be increased progressively. Applications include adaptive virtual assistants that personalize their behavior over time based on user history (Lamontagne et al., 2014).

Design Exploration and Prototyping. Before real users are involved, system designers often need

Table 11: Summary of Datasets. Taxonomy of User Simulated Conversational & Other Complex Datasets. Note that what (§4) is Human-AI §4.1 (H-AI), AI-AI (§4.3), Human-Human §4.2 (H-H), and so on.

	Who	What	Task	Data Size	Eval. Metric	Code
PersonaChat(Zhang et al., 2018)	Indiv.	H-H	Text Gen.	11K Conv.	Likelihood / F1 / Classification Loss	[code]
PersonalDialog (Zheng et al., 2019)	Indiv.	H-H	Text Gen.	21M Conv.	Perplexity / Accuracy	[link]
PersonalConv (Li et al., 2025b)	Indiv.	H-H	Classif., Regres., Gen.	111K Conv.	Accuracy / F1 / MCC / MAE / RMSE / ROUGE-1 / ROUGE-L / BLEU / METEOR	[code]
LiveChat (Gao et al., 2023)	Persona	H-H	Gen., Recog.	1M Conv.	Recall / MRR / BLEU-n / ROUGE-n / ROUGE-L	[code]
SimDial (Zhao and Eskénazi, 2018)	Gen.	AI-AI	Dialog Gen.	9K Conv.	BLEU / F1	[code]
Taskmaster-1 (Byrne et al., 2019)	Gen.	H-H / H-AI	Wizard-of-Oz	13K Conv.	Human Judge	[code]
CrossWOZ (Zhu et al., 2020)	Gen.	H-H	Wizard-of-Oz	6K Conv.	F1	[code]
MultiWOZ (Ye et al., 2022)	Gen.	H-H	Wizard-of-Oz	10K Conv.	Human Judge	[code]
SimQuAC (Abbasiantaeb et al., 2024)	Gen.	AI-AI	Q&A	334 Conv. 4K Ques.	Human Judge	[code]
SocraticChat (Kong et al., 2024)	Gen.	H-AI	Q&A	25K Conv.	LLM Judge	[code]
Nectar (Zhu et al., 2024a)	Gen.	AI-AI	Conversations	183K Prompts	-	[code]
DealOrNoDeal (Lewis et al., 2017)	Role	H-H	Negotiation	6K Conv.	Human Judge	[code]
CraigslistBargain (He et al., 2018)	Role	H-H	Negotiation	7K Conv.	Human Judge	[code]
ABCD (Chen et al., 2021)	Role	H-H	Action Decision	10 K Conv.	Recall	[code]
SAD (Liu et al., 2023)	Role	AI-AI	Q&A	6K	LLM / Automatic Evaluation	[code]
RoleLLM (Wang et al., 2024a)	Role	H-AI	Q&A	168K S.	ROUGE-L	[code]
CharacterLLM (Shao et al., 2023)	Role	H-H / H-AI	Gen., Interview Q&A	14.4K scenes	LLM Judge	[code]
SocialBench (Chen et al., 2024a)	Role	ALL	Q&A	6K Ques.	Acc / Cover	[code]
CharacterEval (Tu et al., 2024)	Role	H-H	Q&A	2K Conv.	Custom	[code]
CharacterBench (Zhou et al., 2025)	Role	H-H	Q&A	23K S.	Model Judge	[code]
LifeStageBench (Fan et al., 2025)	Role	AI-AI	Role-based Q&A	1.3K / 202 S.	Model Judge	–
DialogStudio (Zhang et al., 2024c)	Hybrid	ALL	-	Union	ROUGE-L / F1	[code]

to explore how different dialogue policies behave under varied user goals or engagement patterns. Simulators can be configured to emulate distinct user types, such as cooperative users, those who frequently disengage, or users who ask clarifying or inquisitive questions. This supports early-stage testing of system responses, decision timing, and personalization strategies. In educational technology, for example, simulated students with varying misconceptions or engagement levels can inform the development of personalized tutoring systems (Lin et al., 2023).

Research on Human-Like Behavior and Alignment. User simulation is not only a tool for engineering systems but also a methodological asset for studying human-machine interaction. Simulated dialogues can serve as proxies for human behavioral data, enabling experiments on alignment, fairness, or affective response modeling. For instance, emotional dynamics and social support behaviors can be explored through simulated users in wellness-oriented dialogue settings (Tutun et al., 2023; Li et al., 2024c).

Benchmark Construction and Diagnostic Analysis. Finally, LLM-driven simulators can be used to automatically generate structured benchmarks that test specific capabilities of conversational agents, such as contextual recall, politeness strategies, or belief tracking. Simulation also aids in generating counterfactual examples and behavioral probes, which are valuable for identifying model blind spots and diagnosing failure modes.

G.1 Recommendation

Conversational user simulation has become a pivotal tool in developing, training, and evaluating recommendation systems, enabling scalable and controlled experimentation without requiring extensive human interaction. A foundational contribution by (Zhang and Balog, 2020) introduced an agenda-based user simulator tailored for evaluating dialogue flows in recommendation systems, demon-

strating that simulated evaluations could closely mirror human judgment. Building on this, (Afzali et al., 2023) developed UserSimCRS, an extensible simulation toolkit that incorporates user personas, satisfaction prediction, and conditional language generation, further enhancing realism and adaptability in recommendation contexts such as movie dialogues. To move beyond general evaluation and into system diagnostics, Bernard and Balog (2024) proposed a simulator-driven framework to identify conversational breakdowns in conversational recommendation system interactions, offering a novel approach to robustness testing. With the advent of large language models (LLMs), more recent work has explored generative user simulators. Zhu et al. (2025) presented a controllable, scalable LLM-based simulator framework (CSHI) that supports plugin-based behavior modulation and personalization, while (Yoon et al., 2024) introduced a five-task evaluation protocol to systematically benchmark LLMs' ability to simulate user behavior in recommendation dialogues, revealing both potential and limitations (e.g., popularity bias, weak personalization). Together, these studies demonstrate that user simulation not only supports evaluation but also actively contributes to the design, optimization, and understanding of conversational recommendation systems.

G.2 Education

With the rise and evolution of LLMs, the fusion of AI and education has entered a new phase (Wang et al., 2024b; Xu et al., 2024b), particularly in the domain of conversational user simulation. The current research shows that the simulation applications in education are widely adopting Generative AI (GenAI), with a significant reliance on closed-source LLMs represented by OpenAI's GPT family (OpenAI, 2023). The public release of ChatGPT in late 2022 greatly catalyzed this trend; the advanced conversational generation abilities of such LLMs are valued for ease of integration and high performance, and have been rapidly applied in various educational scenarios (Park et al., 2024; Zhang et al., 2025b). However, there's also exploration into open-sourced alternatives (Xu et al., 2024b; Höhn et al., 2024; Lee et al., 2024). This shift is driven by the pursuit of greater customization (Jin et al., 2025), data security, and transparency—features that allow models to be more deeply adapted to specific pedagogical contexts (Zheng et al., 2025). Such adaptation efforts often go beyond simple prompt engineering, instead involving fine-tuning on domain-specific data to align the AI's simulated user behavior with precise teaching objectives.

Furthermore, the role AI plays in educational user simulations has undergone a fundamental transformation. In the pre-GenAI era, simulated user behavior was often limited and predefined by pre-set scripts. LLMs enable AI-powered simulator to embody a much broader and more dynamic range of user characteristics, thereby creating more interactive and adaptive learning experiences. An intuitive thinking would be simulating a teacher, a tutor, a teaching assistant or coach (Olędzka et al., 2024; Ye et al., 2025), especially for language teaching or social skill training (Park et al., 2024; Lee et al., 2024). However, when it comes to math or physics, for instance, a more practical application is the use of LLMs to simulate students (Yue et al., 2024; Pan et al., 2025) due to their limited capabilities (Frieder et al., 2023). These simulations provide a platform for pre-service teachers to practice classroom management and instructional strategies in a low-stakes environment (Pan et al., 2025; Lim et al., 2025). A key research frontier in this area is enhancing the realism of these simulated students, by modeling the behavior of students at different cognitive levels, including common misconceptions and error patterns, to create more authentic and challenging training scenarios (Yue et al., 2024). Beyond simulating students, LLMs are also being used to simulate learning partners or collaborators to support tasks such as collaborative problem-solving or language practice (Lytvyn, 2025).

G.3 Limited User Feedback

In real conversations, user feedback is sparse, delayed, or ambiguous, making it hard to train and evaluate user simulators based on clean supervision. Most human-annotated datasets rely on intensive labeling efforts (Jang et al., 2022; Ahn et al., 2023) and are limited in size and diversity (Wang et al., 2023; Xu et al., 2024a). As a result, existing works often rely on data that implicitly reflects idealized user behavior, providing dense, turn-level signals and overlooking phenomena like dropout, disengagement, or subtle feedback cues. Addressing this challenge may require integrating synthetic and human-authored data, modeling implicit or de-

layed feedback, and exploring learning objectives that better tolerate sparse and noisy supervision.

G.4 HCI/UI

User-conversational simulated data has significant implications for the fields of human-computer interaction and interface design (Moore and Arar, 2018). On paper, user testing is a crucial step in the user interface design process, but in practice, resources are often so constrained that testing each interface is impossible. Such simulated data offers the opportunity for user researchers to conduct full-scale usability studies (Baxter et al., 2015; Steen et al., 2007), user surveys, and focus groups without using the extensive resources it currently takes to do so. Furthermore, many corporate user research teams are far outnumbered by designers, with many teams having a ratio of 1:5 (Kaplan, 2020). Given the resource constraints, utilizing user-simulated data in UI testing would allow designers to take part in an integral step that they are often forced to skip.

Furthermore, studies like (Hämäläinen et al., 2023) evaluate LLMs' ability to generate synthetic user research data for usability tasks. They underscore the importance of careful prompt writing and using these methods earlier in the design process for need finding and early feedback. The same study even found that LLM simulated conversational data was often distinguishable from human results, with humans actually believing LLM answers to be more "human" than the actual human data it was being compared to. Overall, user-conversational simulated data offers the opportunity to conduct HCI and user research in scenarios where research would have initially been too difficult to conduct.

G.5 Video Understanding

Conversational user simulation can also be applied to video understanding domain. VideoAutoArena (Luo et al., 2025) introduces a novel user of conversational user simulation by generating open-ended adaptive questions to evaluate large multimodal models. Instead of relying on static multiple-choice benchmarks, this framework simulates realistic user inquiries about video content and challenges models through an arena-style evaluation with a modified ELO Rating System. This approach demonstrates that LLM-based conversational simulation can serve not only as a generator of user-style queries, but also as a rigorous evaluation mechanism, enabling scalable and user-centered assessment of video comprehension in multimodal systems.

G.6 Discussion

Admittedly, the scope of conversational user simulation spans a wide array of domains, and it's unrealistic to cover them all. We instead focus on areas where conversation simulation has shown tangible, direct high-impact: (i) augmenting training and evaluation pipelines, (ii) enabling controlled experimentation at scale, and (iii) facilitating personalized interaction design. These domains not only benefit from simulation's scalability but also expose its role in advancing robust, adaptive, and user-centered systems.

H Challenges (Extended)

H.1 Evaluation

In the evaluation of conversational user simulation, human judgment remains the gold standard for assessing qualities like coherence, persona fidelity, and goal alignment. However, it is costly, slow, and often inconsistent across annotator pools and rubrics. LLMs are increasingly used as automated judges, but they are often sensitive to prompt wording and lack grounding, calibration, and reliability, particularly for long or complex conversations. Recent work (Fan et al., 2025; Zhou et al., 2025) has introduced promising practices such as symmetric prompting, ensembling, and meta-evaluation against human ratings, but evaluation protocols still vary across studies. Building standardized, multi-layered evaluation pipelines remains an important open direction.

H.2 User-Specific Personalization

Simulating specific users enables personalized training and evaluation, but introduces unique challenges. First, user-level simulation raises privacy concerns, especially when behavioral traces are derived from real individuals. Future work may explore privacy-preserving techniques such as differential privacy, federated learning, or privacy-aware user simulation. Second, real users evolve over time, shifting goals, language use, or engagement patterns, yet most simulators remain static. Third, interaction histories are often limited, making it difficult to capture fine-grained behavioral patterns. Finally, existing benchmarks rarely support user-level evaluation or isolate individual variation in a

consistent way. Addressing these challenges calls for methods that can generalize from limited signals, model user drift, and preserve privacy while supporting personalization.

H.3 Video Understanding

While recent frameworks like VideoAutoArena (Luo et al., 2025) showcase the promise of LLM-based conversational simulators for video understanding, several limitations remain. First, aligning simulated user queries with the fine-grained temporal and semantic aspects of videos remains difficult, often leading to generic or misaligned questions. Additionally, maintaining coherent multi-turn conversations about dynamic visual content challenges current LLMs, which struggle with long-range context tracking. Another critical gap lies in the lack of alignment with real-world user behaviors and preferences, limiting the realism and utility of the simulated interactions. Addressing these challenges is essential for deploying truly user-centric conversational agents in video-rich environments.

Evaluation of Dialogue Diversity. Quantifying diversity in simulated conversations remains an open challenge. Existing evaluation metrics primarily focus on surface-level textual variation or semantic similarity, which fail to capture deeper diversity in knowledge, reasoning patterns, and persona-specific perspectives. Developing metrics that can evaluate *knowledge diversity* and behavioral variation across simulated characters is essential for assessing the realism and utility of conversational simulations.

Persona Consistency and Adaptability. Maintaining persona consistency over long conversations while allowing adaptive responses to evolving contexts presents a fundamental trade-off. Overemphasizing consistency may result in rigid and repetitive behavior, whereas excessive adaptability can lead to persona drift. How to dynamically balance these competing objectives—particularly in multi-turn or open-ended interactions—remains largely unexplored.

Hybrid Simulation Paradigms. Existing conversational simulation paradigms span human–AI, AI–AI, and mixed hybrid settings. While each paradigm offers unique advantages, systematic comparisons across these settings are scarce. The lack of unified benchmarks and evaluation protocols hinders a comprehensive understanding of when and how different simulation paradigms should be employed.

Knowledge Updating and Temporal Evolution. Most current role-play simulations focus on static personas, such as historical figures or fictional characters. In contrast, simulating active real-world figures requires models to adapt as underlying knowledge and public roles evolve over time. Although prior work has explored adaptive and evolving agents (**?**), applying such strategies to conversational role-play—while preserving persona coherence—remains an open research direction.

Bias and Safety in Persona Simulation. Persona-based simulation introduces inherent risks of stereotyping, bias amplification, and unsafe behaviors. Mitigating these risks without sacrificing authenticity and expressiveness is a critical challenge. Future work must explore principled approaches to bias detection, controllable generation, and safety-aware persona modeling to ensure responsible deployment of conversational simulators.