LLM-based Human Simulations Have Not Yet Been Reliable

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly employed for simulating human behaviors across diverse domains. However, our position is that current LLM-based human simulations remain insufficiently reliable, as evidenced by significant discrepancies between their outcomes and authentic human actions. Our investigation begins with a systematic review of LLM-based human simulations in social, economic, policy, and psychological contexts, identifying their common frameworks, recent advances, and persistent limitations. This review reveals that such discrepancies primarily stem from inherent limitations of LLMs and flaws in simulation design, both of which are examined in detail. Building on these insights, we propose a systematic solution framework that emphasizes enriching data foundations, advancing LLM capabilities, and ensuring robust simulation design to enhance reliability. Finally, we introduce a structured algorithm that operationalizes the proposed framework, aiming to guide credible and human-aligned LLM-based simulations.

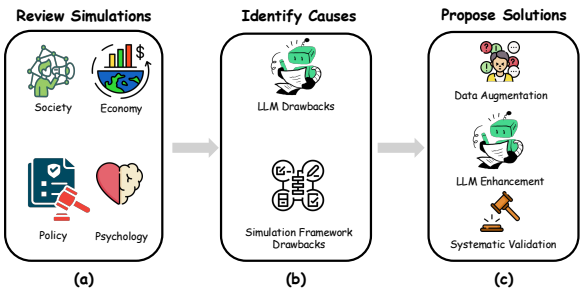

Figure 1: Flow of This Position Paper. We start by reviewing the current LLM-based human simulations, and then identify the causes of the gaps between simulation outputs and real-world human behavior. Finally, we propose targeted solutions for advancing the reliability of LLM-based human simulation.

1 Introduction

Simulation has long been instrumental in understanding and replicating human actions and characteristics (Winsberg, 2003; Ofoegbu, 2023; Cioffi-Revilla, 2010; Dilaver and Gilbert, 2023; Orcutt et al., 1976). The advent of Large Language Models (LLMs) has brought in a new era, with researchers leveraging LLM agents as proxies for humans within simulated environments (Park et al., 2023b; Lin et al., 2023; Li et al., 2023; Wu et al., 2024b; Zhang et al., 2024b; Shi et al., 2024; Wang et al., 2024b). Initial successes span diverse domains, including societal modeling, economics, policy simulation, and psychology (Park et al., 2023b; Lin et al., 2023; Li et al., 2024f; Chen et al., 2024a; Li et al., 2024b; Yang et al., 2024b). Furthermore, high-fidelity LLM simulations hold promise for generating valuable training data (Tang et al., 2024; Zhang et al., 2024a) and evaluating data quality (Xu et al., 2023b; Moniri et al., 2024), thus acting as both data generators and evaluators to enhance LLM capabilities (Gu et al., 2024; Son et al., 2024; Li et al., 2024c).

Despite these advancements, a critical question arises:

> *Can current LLM-based simulations simulate humans reliably?*

Our position is: **Current LLM-based human simulations have not yet been reliable.** This position is supported by empirical evidence revealing significant discrepancies between LLM-generated simulations and observed human behavior. As summarized in Table 1, these gaps are across diverse domains.

To understand the root causes of these gaps, we begin by surveying the landscape of LLM-based human simulations. Our review spans the social, economic, policy, and psychological domains, identifying both seminal advancements and persistent limitations. We then argue that these shortcomings stem from two fundamental challenges:

(1) The inherent limitations of current LLMs in capturing the nuances of human cognition and behavior. This first issue manifests through multiple dimensions: (a) embedded biases (cultural, gender, occupational, and socioeconomic) that distort behavioral simulations (Kotek et al., 2023; Wan et al., 2023; Wang et al., 2023b); (b) cognitive process limitations that compromise decision-making authenticity (Li et al., 2024e; Gui and Toubia, 2023); (c) inconsistent behavioral patterns over time due to memory constraints (Zheng et al., 2024;

Simulation Task	Metric	How LLM Performance Diverged from Human Behavior
Clinics Diagnosis	Accuracy	LLM achieved only 40.18% diagnostic correctness, significantly below human expert levels (Schmidgall et al., 2024).
Emotion Understanding	Accuracy	GPT-4 (58%) demonstrated lower accuracy in understanding emotions compared to humans (70%) (Li et al., 2024e).
Persona	Variance	Assigned persona accounted for <10% of variance in human annotations, indicating weak role adherence (Hu and Collier, 2024).
Human-like Qualities	Accuracy	GPT-4's human-likeness rating (51% at 3 turns) sharply degraded over longer interactions (13.3% at 20 turns) (Wu et al., 2024a).
Trading	Return	GPT-4's trading return (28%) significantly underperformed that of human traders (39%) (Li et al., 2024f).
Adversarial Attack	Attack Success Rate	GPT-4-Turbo (27% success rate) proved far more susceptible to attacks than humans (6% success rate) (Chen et al., 2024b).

Table 1: Empirical evidence of gaps between LLM simulations and real human behavior.

Zhong et al., 2024); and (d) interaction mechanism deficiencies that affect multi-agent simulation quality (Tang et al., 2025; Li et al., 2024a).

(2) The design flaws of current simulation in adequately modeling real-world complexities and interactions. These flaws include: (a) oversimplification of complex psychological states and interactions (Williams, 2000; Zhou et al., 2025); (b) insufficient incorporation of comprehensive human experiences and contexts (Grossberg and Gutowski, 1987; Chen et al., 2024b); (c) inadequate modeling of human incentives and motivations (Stern, 1999; Petrakis and Petrakis, 2012); (d) limited validation mechanisms lacking comprehensive evaluation criteria (He et al., 2023; Ke et al., 2024); and (e) inadequate real-time monitoring and adjustment mechanisms (Xexéo et al., 2024; Reason et al., 2024).

To address the above gaps, we propose a systematic solution framework composed of three parts. Firstly, we advocate for enriching the foundational data used in simulations to ensure more realistic and diverse inputs. Secondly, to improve LLMs' abilities in simulation, we outline targeted interventions for mitigating inherent LLM limitations—focusing on their cognitive process simulation, behavioral consistency, and memory capabilities. Lastly, we propose principles for verifiable and robust simulation frameworks, emphasizing continuous and multi-level validation architectures. To operationalize these principles and our other targeted solutions, we then propose a systematic framework that provides a structured workflow for developing, executing, and validating reliable LLM-based human simulations.

Our Contributions are summarized as follows:

❶ We review current LLM-based human simulations, highlighting recent advancements while also identifying persistent limitations that hinder their reliability across various applications.

❷ We attribute these limitations primarily to two main categories: inherent weaknesses within LLMs themselves and flaws in current simulation framework designs.

❸ We propose a systematic solution framework, encompassing targeted strategies for enriching data foundations, mitigating LLM modeling issues, and ensuring robust validations.

2 Current LLM-based Human Simulations

LLM-based human simulations have shown initial success across diverse fields. Fundamentally, these simulations revolve around two core dimensions: **LLM-driven Capabilities** and the overarching **Simulation Design**. This section begins by presenting a general formulation for LLM-based human simulations, outlining their foundational components. Following this, we will explore four primary categories of these simulations: social, economic, policy and psychological. For each category, we will examine its applications, effectiveness and limitations, with a specific analysis of how LLM-driven capabilities are utilized and how the simulation framework is structured.

2.1 LLM-based Human Simulation Formulation

We formalize LLM-based human simulation as a tuple

$$\mathcal{H} = (\mathcal{E}, \mathcal{F}, \mathcal{R}), \quad (1)$$

where the environment $\mathcal{E} = (\mathcal{S}, \mathcal{V})$ includes the state space $\mathcal{S}$ and evaluation procedures $\mathcal{V}$, the agents $\mathcal{F} = \{f_1, \ldots, f_n\}$ are LLM-based policies, and the rules $\mathcal{R}$ define interaction mechanisms, state transitions, and evaluation criteria.

Each agent f_i maps its input message m and current state $s \in \mathcal{S}$ to an action a:

$$f_i : (m, s) \mapsto a. \tag{2}$$

LLMs provide the **LLM-driven Capabilities** enabling f_i to generate human-like behaviors, while empirical data and domain knowledge ground $\mathcal{V}$ and inform the overall **Simulation Design**.

2.2 Social Simulation

Within the formalism $\mathcal{H} = (\mathcal{E}, \mathcal{F}, \mathcal{R})$, social simulations define the environment $\mathcal{E}$ as a communicative and relational space, where $\mathcal{S}$ encodes evolving conversational and social states, and $\mathcal{V}$ evaluates dialogue plausibility, role consistency, and emergent group dynamics. Agents $f_i \in \mathcal{F}$ map $(m, s) \mapsto a$ into conversational actions, while $\mathcal{R}$ governs turn-taking, interactional norms, and social coordination rules.

LLM-driven Capabilities. LLMs enable agents to generate realistic dialogue, make nuanced social decisions, and embody distinct personas. These capabilities allow emergent social phenomena to be explored under controlled but flexible settings.

Simulation Design. Framework design focuses on specifying $\mathcal{E}$ (communication protocols, social settings) and $\mathcal{R}$ (interaction rules, persona constraints), with $\mathcal{V}$ incorporating validation metrics such as coherence, plausibility, and cross-agent consistency. Major challenges include validating long-horizon behaviors and capturing authentic emergent processes.

Representative Implementations. Applications include virtual towns for daily life modeling (Park et al., 2023a), inter-agent competition (Zhao et al., 2023), healthcare simulations (Agent Hospital (Li et al., 2024b), AgentClinic (Schmidgall et al., 2024)), and peer review systems (Jin et al., 2024). These showcase the adaptability of LLM-based social simulations.

Despite their versatile applications, significant limitations persist. We summarize the limitations in social simulation as follows:

Limitations

(1) **Persona Inconsistency:** Agents struggle to maintain stable, long-term personality traits throughout interactions. (2) **Trait-Behavior Mismatch:** A significant gap exists between agents' self-reported characteristics (e.g., "extraverted") and their actual observed "introverted" behaviors. (3) **Fragile Social Dynamics:** The coherence of multi-agent social systems is often brittle, facing challenges in sustaining believable long-horizon dynamics.

2.3 Economic Simulation

In economic simulations, $\mathcal{E}$ specifies market environments, with $\mathcal{S}$ representing market states (e.g., order books, prices) and $\mathcal{V}$ providing metrics such as efficiency, returns, and stability. Agents $f_i \in \mathcal{F}$ implement policies $(m, s) \mapsto a$ as economic actions like bidding, trading, or resource allocation, governed by rules $\mathcal{R}$ encoding market protocols and strategic constraints.

LLM-driven Capabilities. LLMs provide adaptive decision-making for trading, portfolio management, and negotiation. They exhibit limited rationality, time preferences, and occasionally fairness-driven or cooperative behaviors resembling human economic agents.

Simulation Design. Design choices define $\mathcal{E}$ as the market system, $\mathcal{R}$ as trading rules, and $\mathcal{V}$ as validation against benchmarks or theoretical equilibria. Frameworks must ensure market stability, realism of agent preferences, and robustness of evaluation mechanisms.

Representative Implementations. Examples include CryptoTrade for cryptocurrency markets (Li et al., 2024f), EconAgent for labor and macroeconomic simulations (Li et al., 2024d), and agent-based analyses of Nash equilibria and cooperation (Guo et al., 2024a; Fontana et al., 2024). Studies report both alignment with utility-maximizing principles and deviations reflecting fairness or irrationality (Bybee, 2023; Ross et al., 2024). However, despite capturing some human-like deviations, crucial differences in economic behavior persist. We summarize the primary limitations below:

Limitations

(1) **Weaker Loss Aversion:** Compared to humans, LLM agents exhibit a diminished aversion to losses, affecting their risk-taking behavior. (2) **Stronger Time Discounting:** Agents devalue future rewards more heavily than humans, showing a stronger preference for immediate gratification. (3) **Limited Fidelity:** These behavioral gaps limit the fidelity of simulations in accurately capturing human economic decision-making in certain contexts.

2.4 Policy Simulation

For policy simulations, $\mathcal{E}$ encodes institutional environments with $\mathcal{S}$ representing societal states (e.g., economic indicators, compliance levels) and $\mathcal{V}$ measuring impacts such as equity or efficiency. Agents $f_i \in \mathcal{F}$ model stakeholders whose actions $(m, s) \mapsto a$ reflect policy responses, while $\mathcal{R}$ defines implementation mechanisms, enforcement, and adaptation rules.

LLM-driven Capabilities. LLMs generate responses to policy interventions, simulate compliance vs. resistance, and capture diverse stakeholder perspectives. They can provide foresight into short-term responses and interactional dynamics.

Simulation Design. Design involves specifying $\mathcal{E}$ as the policy context, $\mathcal{R}$ as regulatory and enforcement structures, and $\mathcal{V}$ as outcome assessment metrics (e.g., societal welfare, inequality, public trust). Challenges lie in representing heterogeneous stakeholder interests and cascading long-term impacts.

Representative Implementations. Notable systems include EconAgent for economic policy analysis (Li et al., 2024d) and Urban Generative Intelligence (UGI) frame-

works with domain-specific LLMs for urban planning (Xu et al., 2023a; Zou et al., 2025). These applications show the potential for simulating systemic policy dynamics.

Despite this potential for simulating systemic dynamics, these frameworks face significant hurdles. The most pressing limitations are summarized as follows:

Limitations

(1) **Modeling Long-Term Effects:** Accurately capturing the cascading, long-term impacts of policies and the complex interplay of diverse stakeholder interests remains a formidable challenge. (2) **Inherent LLM Biases:** Biases within the models can produce skewed policy interpretations and inequitable simulated outcomes if not carefully mitigated. (3) **Validation Challenges:** The multifaceted impacts predicted by simulations are difficult to validate due to the immense cost and logistical complexity of large-scale, real-world verification.

2.5 Psychological Simulation

In psychological simulations, $\mathcal{E}$ specifies experimental contexts, with $\mathcal{S}$ encoding cognitive and affective states, and $\mathcal{V}$ assessing validity against psychological theories or empirical data. Agents $f_i \in \mathcal{F}$ map $(m, s) \mapsto a$ into cognitive or behavioral outputs, while $\mathcal{R}$ defines task protocols, persona constraints, and state transitions.

LLM-driven Capabilities. LLMs can emulate cognitive processes such as decision reasoning, mental state inference, and trait expression. They generate behavioral responses reflecting psychological constructs and dynamic mental-state changes.

Simulation Design. Framework design encodes experimental protocols, role assignments, and evaluation criteria. $\mathcal{V}$ typically grounds evaluation in psychological validity, response consistency, and alignment with established findings.

Representative Implementations. Applications include ChatCounselor for counseling interactions (Liu et al., 2023), PsychoGAT for trait inference (Yang et al., 2024b), and PATIENT-ψ for medical training (Wang et al., 2024c). Studies also benchmark LLMs in cognitive tasks, showing near-human performance in specific reasoning tests (Binz and Schulz, 2023; Hagendorff, 2023).

While these applications show promise in mimicking specific cognitive tasks, a core challenge constrains their psychological fidelity. This primary limitation is outlined below:

Limitations

The primary challenge is a **core representational difficulty**: LLMs struggle to authentically model and represent underlying, long-term psychological traits. This fundamental issue constrains the fidelity and depth of simulations across various applications, from counseling to cognitive assessment, even as they show success in reproducing specific behaviors.

3 LLM Inherent Drawbacks

Drawing from the limitations highlighted in Section 2, we now analyze the intrinsic properties of LLMs that fundamentally constrain their use as reliable simulators.

3.1 Bias

Bias in LLMs fundamentally affects cognitive and behavioral simulation. Experiments in utility theory (Ross et al., 2024) show that only two out of nine models passed all competence tests compared with human responses, with stronger time discounting, less rational loss assessments, and greater inequity aversion. Such distortions directly limit the authenticity of simulated reasoning and decision-making.

Cultural bias undermines cross-cultural reliability. Because training data are predominantly English and Western, LLMs systematically favor Western values and interaction patterns. Comparative studies report significantly higher accuracy in U.S. contexts compared to non-Western settings such as Japan and South Africa (Wang et al., 2023b; Qu and Wang, 2024), highlighting the lack of generalizable cultural cognition.

Gender bias distorts behavioral pattern simulation. LLMs perpetuate stereotypes at far higher rates than humans, with models generating 3–6 times more gender-stereotypical behaviors. Alignment efforts remain insufficient: GPT-4-Turbo shows a 27% attack bias success rate compared to 6% in humans (Kotek et al., 2023; Chen et al., 2024b).

Socioeconomic and occupational bias reduces diversity. Overrepresentation of certain professions and social groups in training data leads to unrealistic reasoning about underrepresented groups (Gwartney and McCaffree, 1971; Harrison and Budworth, 2015), undermining the ability to model authentic diversity in human decision-making.

3.2 Mismatches in Simulating Cognition and Behavior

Cognitive mismatches reduce decision-making authenticity. LLMs display inconsistent reasoning patterns across tests. For instance, Mixtral-8*7b scores low extraversion (2) in one test but high (5) in another, while value surveys drop from 90% to 70% under adversarial prompts (Li et al., 2024e; Hu and Collier, 2024). These mismatches indicate fragile reasoning, weak emotional processing, and limited adaptability.

A further complication arises from *experiment-awareness*. In human-subject research, participants are typically kept unaware of the precise research purpose to avoid demand characteristics. In contrast, LLMs may recognize that they are being tested and adapt their outputs accordingly. Recent work shows that frontier models can detect evaluation settings and even infer the intended task, leading to altered behaviors. (Needham et al., 2025; Nguyen et al., 2025).

Behavioral patterns lack temporal consistency. Studies in generative simulation (e.g., GovSIM) show that models fail to sustain equilibrium, leading to unsustainable resource use such as overharvesting (Piatti et al., 2024). Similarly, LLMs fail to form realistic habits (Zhong et al., 2024) or simulate learning over repeated interactions (Chu et al., 2024).

Data contamination drives memorization. Because training corpora overlap with benchmarks, LLMs sometimes recall memorized content instead of reasoning. For example, MMLU subsets contain up to 57% flawed items (Gema et al., 2024), and GPT-4 can predict masked options with 57% exact-match accuracy (Deng et al., 2024). This undermines the authenticity of "simulated reasoning."

Memory constraints limit behavioral consistency across interactions. LLMs' memory limitations particularly affect the behavioral patterns and interaction mechanisms components (Tang et al., 2025). As cognitive load increases, LLMs demonstrate severe performance degradation; for instance, at the highest load levels, performance can decline by 50% to 60% relative to initial levels observed at low cognitive loads (Upadhayay et al., 2024). These constraints significantly impact their ability to model long-term relationships (Li et al., 2024a), retain context across multiple interactions (Yuan et al., 2024), and adapt behaviors based on past experiences (Evans, 2015).

4 Simulation Design Drawbacks

Human-designed simulation frameworks also face critical limitations that undermine their reliability as analyzed in Section 2. We summarizes these drawbacks in Table 2, while the following subsections elaborate with representative cases.

Overall, these limitations can be grouped into two major categories. The first set of issues arises from the **design of simulation frameworks themselves**, where human cognition, experience, and incentives are often simplified or incompletely modeled. The second set concerns **validation and monitoring**, where the lack of rigorous evaluation and adaptive mechanisms decrease reliability of simulations. The following subsections discuss these two aspects in detail.

4.1 Framework Design Drawbacks

Current frameworks oversimplify complex human psychological states and interactions. Human designers often create oversimplified frameworks when attempting to model psychological states (Tjuatja et al., 2024; Jansen et al., 2023). For example, many frameworks reduce complex emotional states to basic categories or numerical scales, failing to capture the subtle interplay between different psychological factors (Williams, 2000). This oversimplification stems from the challenge designers face in quantifying and operationalizing complex human psychological processes (Evans, 2015). In addition, while LLMs have shown promise in simulating individual human behavior, simulating group behaviors introduces additional complexities. Group dynamics are influenced by factors such as social norms, interpersonal relationships, and collective decision-making processes, which are not easily captured by current LLM frameworks.

Simulation designs fail to account for comprehensive human experiences. Framework designers struggle to incorporate the full spectrum of human lived experiences into their simulations (Beratan, 2007). Current designs often focus on specific, measurable behaviors while neglecting the rich tapestry of personal histories, cultural contexts, and life experiences that shape human decision-making (Grossberg and Gutowski, 1987; Chen et al., 2024b). For example, in conversational recommender system simulations, the success rate of recommendations decreased significantly under 2-5 rounds (from 70% to only 1%) with the user simulator indicating the challenge of CRS not effectively utilizing the interaction information obtained from user simulators, due to the complexity of scenarios (Zhu et al., 2024). This limitation reflects the difficulty in creating frameworks that can adequately represent the complexity of human experiential learning.

Current frameworks lack effective human incentive modeling. Designers face significant challenges in creating frameworks that accurately model complex human motivations and incentives (Stern, 1999; Petrakis and Petrakis, 2012). Many current designs rely on simplified reward systems that fail to capture the intricate web of personal, social, and cultural factors influencing human decision-making. For instance, individuals often make decisions contrary to their personal preferences to maintain social status, reputation, or 'face' in certain cultural contexts (Hwang, 1987). This limitation stems partly from LLMs' inherent constraints - they lack embodied experience and real social interactions, hindering their ability to fully comprehend complex social dynamics. Such oversimplification results in behavioral simulations that inadequately reflect the true complexity of human motivational systems.

4.2 Validation and Monitoring Drawbacks

Current validation mechanisms lack comprehensive evaluation criteria. Human-designed validation systems struggle to establish effective criteria for evaluating simulation authenticity (He et al., 2023). The challenge lies in developing metrics that can effectively measure both the accuracy of individual behaviors and the coherence of complex interaction patterns (Ke et al., 2024).

Drawback Type	Impact on Simulation	Representative Evidence
Oversimplified psychological states	Reduces realism of cognition and group dynamics	Emotional states collapsed into basic categories or scales; group dynamics poorly captured (Tjuatja et al., 2024; Jansen et al., 2023; Williams, 2000)
Incomplete coverage of lived experiences	Neglects historical, cultural, and contextual richness in decisions	CRS simulations degrade after 2–5 rounds due to lack of experiential integration (Beratan, 2007; Zhu et al., 2024)
Weak incentive modeling	Fails to capture complex social and cultural motivations	Simplified rewards ignore "face" or reputation trade-offs; unrealistic decision-making (Stern, 1999; Hwang, 1987)
Validation gaps	Lacks comprehensive evaluation metrics for behaviors and interactions	Metrics insufficient to capture complex dynamics; authenticity criteria unclear (He et al., 2023; Ke et al., 2024)
Limited monitoring	Poor adaptability to emergent or unexpected behaviors	Current systems lack real-time adjustment; unable to maintain simulation quality (Xexéo et al., 2024; Reason et al., 2024)

Table 2: Summary of simulation design drawbacks that limit reliability and authenticity in LLM-based human simulations.

Current frameworks often rely on oversimplified validation methods that fail to capture the full complexity of human behavior.

Monitoring systems lack effective real-time adjustment capabilities. Framework designers struggle to create effective mechanisms for real-time monitoring and adjustment of simulation parameters (Xexéo et al., 2024). Current designs often lack the flexibility to adapt to emerging behavioral patterns or unexpected interaction dynamics, limiting their ability to maintain simulation quality (Reason et al., 2024).

5 Proposed Solutions

We propose systematic solutions that directly address the limitations identified in Sections 3 and 4. We focus on three areas: (1) enriching the data foundation (Section 5.1), (2) improving the abilities of LLM (Section 5.2), and (3) engineering trustworthy simulations through robust validation (Section 5.3). Figure 2 provides an overview.

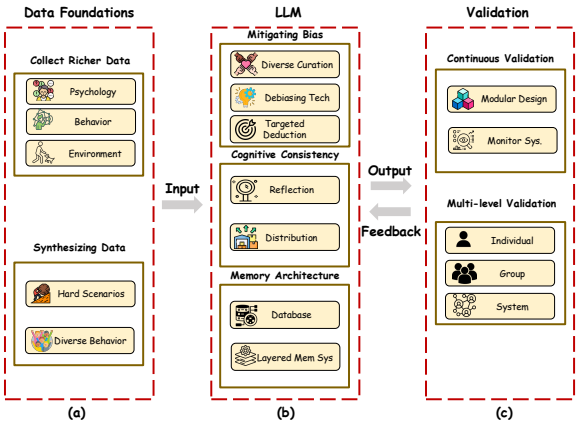

Figure 2: Overview of the Proposed Solution Framework. It details three core components: (a) Enriched Data Foundations, (b) Improved LLM Capabilities, and (c) Trustworthy Simulation Design through Robust Validation.

5.1 Enriching Data Foundation for Simulations

5.1.1 Harnessing Richer Human Data

Simulations grounded solely in traditional text, image, or audio data inevitably miss crucial dimensions of human experience.

Capturing Physiological and Cognitive Data. Wearable sensors unlock access to vital physiological signals, including heart rate and skin conductance, and can even record neural activity patterns, offering deeper insights into internal states (Yuan et al., 2024). Integrating such data is key for LLMs to move beyond superficial behavior mimicry towards modeling the underlying emotional states and cognitive processes that drive human decisions.

Monitoring Fine-Grained Behavioral Patterns. Motion sensing and activity tracking technologies provide continuous, detailed records of daily routines, physical movement, and social interaction dynamics, as demonstrated in studies like (Chu et al., 2024). This granular behavioral data enables LLMs to generate actions that are not just plausible in general, but accurate within specific situational and personal contexts.

Achieving Environmental Context Awareness. Human behavior is deeply embedded in its environmental context. Incorporating data from sensors that measure ambient conditions—such as light, sound, and temperature (Li et al., 2024a)—allows simulations to account for these crucial external factors, leading to more nuanced and realistic models of human responsiveness.

5.1.2 Synthesizing High-Fidelity Human Data via LLMs

While real-world data is invaluable, its collection is often constrained by cost, logistical complexity, ethical barriers, and the sheer impossibility of observing rare or counterfactual scenarios. LLMs offer a powerful

Algorithm 1 Our Proposed LLM-based Human Simulation Framework

Initialize: Agent set $\mathcal{F} = \{f_1, ..., f_n\}$, Environment $\mathcal{E}$, Validation metrics $\mathcal{V}$, History $H \leftarrow \emptyset$ // Setup simulation components
while simulation not complete **do**
  for each agent $f_i \in \mathcal{F}$ **do**
    Perception update: $p_i \leftarrow \text{perceive}(f_i, \mathcal{E}, H)$ // Agent observes environment
    Generate action: $a_i \leftarrow f_i(p_i, \mathcal{E})$ // Context-aware action generation
    Behavioral validation: $v_i \leftarrow \mathcal{V}_{\text{behav}}(a_i, D_{\text{human}})$ // Compare with human patterns
    if $v_i <$ threshold **then**
      $a_i \leftarrow \text{adjust}(a_i, D_{\text{human}})$ // Recalibrate non-human-like actions
    end if
    Execute action: $\mathcal{E} \leftarrow \text{apply}(a_i, \mathcal{E})$ // Apply action to environment
    Memory update: $M_i \leftarrow M_i \cup \{(p_i, a_i, \mathcal{E})\}$ // Update agent memory
  end for
  if time $\mod$ validate_interval **then**
    Individual validation: $V_{\text{indiv}} \leftarrow \mathcal{V}_{\text{indiv}}(\{a_i\}, D_{\text{human}})$ // Assess individual behaviors
    Group validation: $V_{\text{group}} \leftarrow \mathcal{V}_{\text{group}}(\mathcal{E}, D_{\text{social}})$ // Verify group dynamics
    Expert assessment: $V_{\text{expert}} \leftarrow \mathcal{V}_{\text{expert}}(\mathcal{E}, H)$ // Optional human expert review
    Apply calibrations: $\mathcal{F} \leftarrow \text{calibrate}(\mathcal{F}, [V_{\text{indiv}}, V_{\text{group}}, V_{\text{expert}}])$ // Adjust agent parameters
  end if
  Anomaly detection: $A \leftarrow \text{detect_anomalies}(\mathcal{E}, H, \mathcal{F})$ // Identify unrealistic patterns
  Environment correction: $\mathcal{E} \leftarrow \text{correct}(\mathcal{E}, A)$ // Fix detected anomalies
  Update history: $H \leftarrow H \cup \{(\{a_i\}, \mathcal{E}, V)\}$ // Record simulation state
end while
Return: Validated simulation results $\mathcal{E}$, Interaction history H, Validation metrics V =0

complementary approach by enabling the synthesis of high-quality data to augment or replace real-world data where necessary.

Generating Data for Inaccessible Scenarios. LLMs excel at creating plausible data for situations that are difficult, dangerous, or unethical to capture in reality. Examples include simulating high-stakes negotiations, sensitive social interactions, or specific therapeutic encounters (Tjuatja et al., 2024; Yang et al., 2024a). This capability provides crucial training and evaluation data that would otherwise be unavailable.

Synthesizing Diverse Behaviors. Human behavior is inherently diverse. To effectively model this, LLMs can generate a wide spectrum of behavioral patterns through the manipulation of key parameters. These parameters can include intrinsic agent characteristics such as personality traits, emotional states, and cultural backgrounds (He et al., 2023). Furthermore, LLMs can synthesize varied behavioral responses under specific, but infrequently occurring conditions. For instance, they can be used to augment datasets by generating human behaviors in bull or bear markets, allowing for the exploration of human reactions within these distinct economic markets (Li et al., 2024f; Guo et al., 2024b).

5.2 Improving LLM Abilities

The persistence of cultural, gender, and demographic biases, even in state-of-the-art LLMs (Jiao et al., 2024; Foka, 2024), necessitates proactive mitigation strategies that go beyond standard training procedures. We propose a multi-faceted approach to tackle these challenges as follows:

Mitigating Bias through Advanced Training. Building on recent advances in bias understanding and mitigation (Pawar et al., 2024; Rakshit et al., 2024), we advocate for a multi-pronged framework. This involves: (1) Diverse Data Curation, by actively incorporating balanced, multilingual datasets reflecting diverse cultural and demographic groups to counteract learned biases; (2) Algorithmic Debiasing, implementing techniques like counterfactual data augmentation and balanced training objectives during model development to reduce learned biases; and (3) Targeted Bias Reduction, such as occupation-aware post-training to prevent specific stereotype perpetuation.

Architectures for Believable and Consistent Cognition. To improve the believability and consistency of simulated human cognition, we propose architectural innovations. These include: (1) Internal Reflection Mechanisms, enabling agents to internally review and align planned responses with their persona, goals, and past actions before outputting (Li et al., 2024e). (2) Distributed Cognitive Architecture using specialized agents for distinct cognitive modules (e.g., planning, reasoning, emotion). This modular approach, unlike monolithic models, improves performance and consistency in specific cognitive domains, as supported by psychology and economic simulations (Hu and Collier, 2024; Wang et al., 2024a).

Overcoming Memory Deficits with Hybrid Architectures. Naive approaches to augmenting LLMs with external memory often fall short of resolving fundamental limitations like constrained context windows, retrieval inaccuracies, and information forgetting (Wang et al., 2023a; Hatalis et al., 2023; Han et al., 2024; Feng et al., 2024). To address these limitations, we propose integrated hybrid memory systems. One key component, *Scalable External Knowledge Bases*, utilizes vector stores or specialized databases for efficient large-scale information storage and precise retrieval,

decoupling an agent's knowledge capacity from LLM constraints (Zhong et al., 2024). Complementing this, *Layered Memory Systems* with dynamic information management, such as short-term working and long-term episodic/semantic memory, akin to human cognitive structures (Yuan et al., 2024; Cowan, 2008) for more human-like recall and forgetting.

5.3 Engineering Trustworthy via Validation

Beyond improving LLMs themselves, the underlying simulation infrastructure must be engineered for reliability and robust validation. We propose a verification-centric approach with two key components:

Modular Design with Continuous Validation. Simulation frameworks should employ modular architectures where individual components, such as agent models or environmental modules, can be independently developed, tested, and validated (Reason et al., 2024). This should be paired with integrated monitoring systems that track agent behavior and system states to detect anomalies or deviations from expected patterns in real-time (Xexéo et al., 2024). Modular design enables validation at multiple levels of granularity.

Multi-level Validation Approach for Holistic Assessment. Building on the above modular designs, a verifiable simulation framework implements validation at three critical levels for comprehensive assessment: (1) Individual-level validation: focusing on comparing individual agent responses and actions against statistical distributions derived from human behavior in similar contexts; (2) Group-level validation: assessing whether emergent social patterns and collective behaviors within the simulation match known human social phenomena or theoretical expectations of group dynamics; and (3) System-level validation: employing a combination of automated metrics and periodic expert review to ensure the holistic realism and plausibility of the simulation as a whole.

Synthesizing the solutions from the preceding subsections, we introduce Algorithm 1, which, in accordance with the symbols defined in Section 2.1, provides a systematic framework for executing and validating LLM-based human simulations, thereby advancing their fidelity and trustworthiness.

6 Rebuttal to Alternative Views

***Alternative View 1**: Functional Adequacy in Niche Applications.*

Position: One perspective posits that LLM-based human simulations can achieve functional adequacy for particular applications, even without perfectly replicating human behavior (Park et al., 2024; Yang et al., 2025; Wang et al., 2023c).

Our Rebuttal: We contend that focusing on such "functional adequacy" is insufficient and potentially misleading. Our key counter-arguments are threefold: (1) **Brittleness over Robustness:** Success in isolated tasks is not a reliable indicator of general performance, as this perceived adequacy is often brittle and fails when conditions change even slightly. (2) **Lack of Generalization:** These successes fail to generalize; a model adequate for one scenario remains untrustworthy in broader applications without a proven foundation of reliability. (3) **The Need for Verifiability:** True progress demands moving beyond ad-hoc performance. The core aim of our work is to establish methodologies for building and verifying reliability, which is a prerequisite for any trustworthy application.

Beyond the pragmatic argument for functional adequacy, a more fundamental challenge addresses the very nature of human behavior itself, which is shown as follows:

***Alternative View 2**: The Absence of a Definitive Ground Truth in Human Behavior.*

Position: This more fundamental view asserts that the quest for "reliable" simulation is inherently flawed, as human behavior itself lacks a singular, definitive ground truth. Proponents highlight the significant variance in human actions due to non-replicable contextual factors (Goldspink and Cilliers, 2010; Salganik and Watts, 2006) and the inherent prediction ceilings for stochastic phenomena (Conte and Gilbert, 2012; Yazdi, 2024).

Our Rebuttal: The inherent complexity and stochasticity of human behavior do not excuse unreliable simulations; on the contrary, they demand more rigorous methodologies. Our objective is not the perfect replication of an elusive "ground truth," but the creation of simulations demonstrably faithful to observed human behavioral patterns. This very complexity underscores the critical need for the robust validation and reliability-improving frameworks we advocate, ensuring that simulations, even if imperfect, are rendered as trustworthy and insightful as possible.

7 Conclusion

We argue that current LLM-based human simulations remain unreliable, while highlighting the key challenges that underlie this limitation. To address these issues, we outline a pathway forward centered on enriched data sources, advances in LLM modeling, and verifiable design principles. These proposed directions aim to substantially enhance the reliability of LLM-based simulations. We also rebut alternative views to highlight the necessity of moving beyond pragmatic compromises toward a higher standard of verifiable reliability. As the volume of simulation studies continues to grow rapidly, our framework serves as a timely call to action for the LLM research community to pursue more rigorous, reliable, and validated simulation methodologies.

Limitations

Risks of Misinformation and Economic Implications. The potential misuse of highly realistic simulations

raises concerns about the spread of misinformation and the erosion of trust in digital interactions. From an economic standpoint, while simulations can reduce costs for behavioral research and training data generation, as shown in Section 5, there is a risk of creating self-reinforcing feedback loops if synthetic data contaminates future model training.

Costs of Simulation Construction. While the proposed framework prioritizes establishing theoretical and methodological foundations for LLM-based human simulation, future empirical work will need to rigorously evaluate the trade-offs between computational costs and performance gains. Initial efforts could prioritize bias mitigation, include multi-modal data sources and implement reflection mechanisms. Moreover, the reliance on LLMs for human simulation incurs significant API costs. Although cheaper alternatives are available, they often lack generalizability. A promising approach involves integrating specialized LLMs tailored for specific tasks, utilizing models that excel in particular domains while ensuring efficient communication and data sharing between the models.

Ethical and Social Considerations. The data collection methods proposed in Section 5, which involve wearable sensors, present practical challenges related to privacy protection and ethical data usage. Key concerns include obtaining participant consent and ensuring the security of the collected data, both of which pose significant challenges.

References

Kathi K Beratan. 2007. A cognition-based view of decision processes in complex social–ecological systems. *Ecology and society*, 12(1).

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Leland Bybee. 2023. Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024a. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Zhendong Chu, Zichao Wang, Ruiyi Zhang, Yangfeng Ji, Hongning Wang, and Tong Sun. 2024. Improve temporal awareness of llms for sequential recommendation. *arXiv preprint arXiv:2405.02778*.

Claudio Cioffi-Revilla. 2010. A methodology for complex social simulations. *Journal of Artificial Societies and Social Simulation*, 13(1):7.

Rosaria Conte and Nigel Gilbert. 2012. Computational modeling of social response uncertainty. In *Social Simulation Conference*, pages 1–15. ECCS.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7778–7802.

Ozge Dilaver and Nigel Gilbert. 2023. Unpacking a black box: a conceptual anatomy framework for agent-based social simulation models. *Journal of Artificial Societies and Social Simulation*, 26(1).

Ian M Evans. 2015. *How and why thoughts change: Foundations of cognitive psychotherapy*. Oxford University Press, USA.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from agi: Are llms all we need? *arXiv preprint arXiv:2405.10313*.

Amalia Foka. 2024. She works, he works: A curious exploration of gender bias in ai-generated imagery. *arXiv preprint arXiv:2407.18524*.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2024. Nicer than humans: How do large language models behave in the prisoner's dilemma? *arXiv preprint arXiv:2406.13605*.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*.

Chris Goldspink and Paul Cilliers. 2010. *Modelling Human Behaviour as Complex Systems*. Springer, London.

Stephen Grossberg and William E Gutowski. 1987. Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. *Psychological review*, 94(3):300.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. 2024a. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. 2024b. Economics arena for large language models. *ArXiv*, abs/2401.01735.

James D Gwartney and Kenneth M McCaffree. 1971. Variance in discrimination among occupations. *Southern Economic Journal*, pages 141–155.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1.

Songyue Han, Mingyu Wang, Jialong Zhang, Dongdong Li, and Junhong Duan. 2024. A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. *Electronics*, 13(24):5040.

Jennifer A Harrison and Marie-Hélène Budworth. 2015. Unintended consequences of a digital presence: Employment-related implications for job seekers. *Career Development International*, 20(4):294–314.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 2023. Memory matters: The need to improve long-term memory in llm-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 277–280.

Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Kwang-kuo Hwang. 1987. Face and favor: The chinese power game. *American journal of Sociology*, 92(4):944–974.

Bernard J Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024c. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024d. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536.

Siyu Li, Jin Yang, and Kui Zhao. 2023. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *arXiv preprint arXiv:2307.10337*.

Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024e. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024f. Cryptotrade: A reflective llm-based agent to guide zero-shot cryptocurrency trading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1094–1106.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. 2024. Evaluating the performance of large language models via debates. *Preprint*, arXiv:2406.11044.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*.

Jord Nguyen, Hoang Huu Khiem, Carlo Leonardo Attubato, and Felix Hofstätter. 2025. Probing evaluation awareness of language models. In *ICML Workshop on Technical AI Governance (TAIG)*.

Wilson Chukwuemeka Ofoegbu. 2023. Simulation: A tool for system design and analysis. *GPH-International Journal of Social Science and Humanities Research*, 6(11):98–111.

Guy H Orcutt, Steven Caldwell, and Richard F Wertheimer. 1976. *Policy exploration through microanalytic simulation*. The Urban Insitute.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023b. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Panagiotis Petrakis and Panagiotis Petrakis. 2012. Human incentives. *The Greek Economy and the Crisis: Challenges and Responses*, pages 233–268.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Aishik Rakshit, Smriti Singh, Shuvam Keshari, Arijit Ghosh Chowdhury, Vinija Jain, and Aman Chadha. 2024. From prejudice to parity: A new approach to debiasing large language model word embeddings. *arXiv preprint arXiv:2402.11512*.

Tim Reason, William Rawlinson, Julia Langham, Andy Gimblett, Bill Malcolm, and Sven Klijn. 2024. Artificial intelligence to automate health economic modelling: A case study to evaluate the potential application of large language models. *PharmacoEconomics-Open*, 8(2):191–203.

Jillian Ross, Yoon Kim, and Andrew W Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*.

Matthew J. Salganik and Duncan J. Watts. 2006. Experimental study of cultural contingency in human behavior. *Science*, 311(5762):854–856.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Zhengliang Shi, Shen Gao, Xiuyi Chen, Yue Feng, Lingyong Yan, Haibo Shi, Dawei Yin, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. 2024. Learning to use tools via cooperative and interactive agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10642–10657, Miami, Florida, USA. Association for Computational Linguistics.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.

Paul C Stern. 1999. Information, incentives, and proenvironmental consumer behavior. *Journal of consumer Policy*, 22(4):461–478.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

Zhenheng Tang, Xiang Liu, Qian Wang, Peijie Dong, Bingsheng He, Xiaowen Chu, and Bo Li. 2025. The lottery LLM hypothesis, rethinking what abilities should LLM compression preserve? In *The Fourth Blogpost Track at ICLR 2025*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Bibek Upadhayay, Vahid Behzadan, and Amin Karbasi. 2024. Cognitive overload attack: Prompt injection for long context. *arXiv preprint arXiv:2410.11272*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Qian Wang, Yuchen Gao, Zhenheng Tang, Bingqiao Luo, and Bingsheng He. 2024a. Enhancing llm trading performance with fact-subjectivity aware reasoning. *arXiv preprint arXiv:2410.12464*.

Qian Wang, Tianyu Wang, Qinbin Li, Jingsheng Liang, and Bingsheng He. 2024b. Megaagent: A practical framework for autonomous cooperation in large-scale llm agent systems. *arXiv preprint arXiv:2408.09955*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024c. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023a. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023b. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Xintao Wang, Kewen Zeng, Zhiyuan Liu, Yufeng Chen, and Ruihong Tang. 2023c. Role-play with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12268–12289. ACL.

Simon Johnson Williams. 2000. Emotion and social theory: corporeal reflections on the (ir) rational.

Eric Winsberg. 2003. Simulated experiments: Methodology for a virtual world. *Philosophy of science*, 70(1):105–125.

Weiqi Wu, Hongqiu Wu, and Hai Zhao. 2024a. Self-directed turing test for large language models. *arXiv preprint arXiv:2408.09853*.

Zengqing Wu, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, Run Peng, and Chuan Xiao. 2024b. Shall we talk: Exploring spontaneous collaborations of competing llm agents. *arXiv preprint arXiv:2402.12327*.

Geraldo Xexéo, Filipe Braida, Marcus Parreiras, and Paulo Xavier. 2024. The economic implications of large language model selection on earnings and return on investment: A decision theoretic model. *arXiv preprint arXiv:2405.17637*.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023a. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023b. Towards reasoning in large language models via multi-agent peer review collaboration. *Preprint*, arXiv:2311.08152.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024a. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024b. Psychogat: A novel psychological measurement paradigm through interactive fiction games with llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14470–14505.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. Twinmarket: A scalable behavioral and socialsimulation for financial markets. *arXiv preprint arXiv:2502.01506*.

Mohammad Yazdi. 2024. Computational tools and techniques for reliability and maintainability. In *Advances in computational mathematics for industrial system reliability and maintainability*, pages 59–77. Springer.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1963–1974.

Jinghui Zhang, Dandan Qiao, Mochen Yang, and Qiang Wei. 2024a. Regurgitative training: The value of real data in training large language models. *arXiv preprint arXiv:2407.12835*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024b. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512*.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 2024. Towards lifelong learning of large language models: A survey. *arXiv preprint arXiv:2406.06391*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025. The pimmur principles: Ensuring validity in collective behavior of llm societies. *arXiv preprint arXiv:2509.18052*.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How reliable is your simulator? analysis on the limitations of current llm-based user simulators for conversational recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1726–1732.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606.