What Limits LLM-based Human Simulation: LLMs or Our Design?

Qian Wang [1] Jiaying Wu [1] Zhenheng Tang [2] Bingqiao Luo [1] Nuo Chen [1] Wei Chen [1] Bingsheng He [1]
[1]National University of Singapore
[2]Hong Kong University of Science and Technology

Abstract

We argue that advancing LLM-based human simulation requires addressing both LLM's inherent limitations and simulation framework design challenges. Recent studies have revealed significant gaps between LLM-based human simulations and real-world observations, highlighting these dual challenges. To address these gaps, we present a comprehensive analysis of LLM limitations and our design issues, proposing targeted solutions for both aspects. Furthermore, we explore future directions that address both challenges simultaneously, particularly in data collection, LLM generation, and evaluation. To support further research in this field, we provide a curated collection of LLM-based human simulation resources.[1]

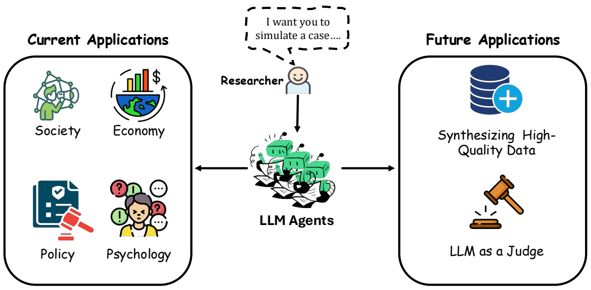

Figure 1. LLM-based Human Simulation Applications

1. Introduction

Simulation has long been a crucial tool for understanding human behavior by replicating their actions and characteristics (Ofoegbu, 2023; Winsberg, 2003). It has enabled advancements in studying social dynamics (Cioffi-Revilla, 2010; Dilaver & Gilbert, 2023), policy implementation (Downing et al., 2000; Orcutt et al., 1976), and economic forecasting (Bainbridge, 2018; Dignum et al., 2020).

With the advent of Large Language Models (LLMs), researchers have increasingly utilized these models to construct simulations, treating LLM agents as proxies for humans to perform actions and engage in interactions (Li et al., 2023b; Lin et al., 2023; Park et al., 2023b; Shi et al., 2024; Wang et al., 2024b; Wu et al., 2024; Zhang et al., 2024b), as shown in Figure 1. Initial successes in LLM-based human simulations have been demonstrated across diverse fields, including society, economics, policy, and psychology (Chen et al., 2024a; Li et al., 2024b;f; Lin et al., 2023; Park et al., 2023b; Yang et al., 2024b). Moreover, reliable LLM simulations can generate high-quality data for LLM training (Tang et al., 2024; Zhang et al., 2024a) and evaluate data quality (Chiang et al., 2024; Moniri et al., 2024; Xu et al., 2023b; Zheng et al., 2023b), serving as a data generator and evaluator (Gu et al., 2024; Li et al., 2024c; Son et al., 2024) to enhance LLM pre-training and simulation abilities.

However, most LLM-based human simulations to date have primarily focused on simulation methods within specific scenarios, with limited investigation into whether these simulations can authentically replicate human behavior or what criteria define their alignment with human characteristics. Several studies have highlighted the limitations of LLMs' simulation capabilities. For instance, lots of research (Ai et al., 2024; Hu & Collier, 2024; Lee et al., 2024; Petrov et al., 2024) has shown that LLMs struggle to replicate distinct personalities. Specifically, Lee et al. (2024) demonstrated that even when LLMs are prompted to role-play diverse and randomized personas in response to psychological questionnaires, they exhibit consistent values and moral preferences across various contexts.

The challenges of LLM-based human simulations can be examined from two dimensions: **LLMs' inherent drawbacks** and **simulation designs' drawbacks**. The core difficulty lies in aligning LLMs with human behavior, which involves simulating specific personalities and modeling human groups. Compared to tasks in NLP or CV, LLM-based human simulations present a much greater complexity (Mao et al., 2024). In NLP and CV, text or images are decomposed into their smallest units—such as words, phrases, symbols, or image tokens—based on a predefined vocabulary (Bai et al., 2024; Ohm & Singh, 2024). However, *humans cannot be encoded as two-dimensional tokens*; they embody unique

[1] https://github.com/Persdre/llm-human-simulation

preferences, lived experiences, and intricate behaviors.

To address the above challenges, we propose solutions focusing on two key aspects: (1) enhancing LLM training with comprehensive and unbiased human behavioral data, and (2) developing systematic validation frameworks to ensure simulation reliability with ground truth. Furthermore, we identify three directions for future development of LLM-based human simulations: (1) collecting multi-dimensional human data, (2) synthesizing high-quality training data, and (3) leveraging LLM-as-a-judge for data quality evaluation.

Our Contributions:

- We present a systematic analysis of LLM-based human simulations, identifying two fundamental challenges: the inherent limitations of LLMs and the drawbacks in simulation framework design. Our analysis reveals how these challenges manifest across different simulation categories, from social interactions to economic behaviors.
- We propose a unified simulation framework that clearly defines the roles of LLM actions and human participation, providing a methodological foundation for understanding and improving human simulation systems.
- We present targeted solutions addressing both LLM limitations and framework design challenges, including enhanced data collection methods, improved validation mechanisms, and systematic evaluation procedures.
- We identify promising future directions for advancing LLM-based human simulations, particularly in data quality improvement and automated evaluation capabilities. To support future research, we provide a curated collection of resources and examples [2].

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of LLM-based simulation categories and their key components, establishing the foundation for our analysis. Building on this foundation, Section 3 examines the inherent limitations of LLMs in human simulation, while Section 4 analyzes critical drawbacks in simulation framework design. To address these challenges, Section 5 presents targeted solutions that tackle both LLM limitations and framework design issues. Finally, Section 6 explores promising future directions, focusing on two key aspects: enhancing data collection through wearable sensors and developing LLM-based approaches for high-quality training data synthesis.

2. Existing LLM-based Human Simulations

LLM-based human simulations have achieved initial success across multiple fields, as detailed in Section 2.2. These simulations can be categorized based on their component control mechanisms: **LLM Actions** or **Human Participation**. In the following subsections, we first give a formal definition of the present different simulation types (social, economic, policy, and psychological simulations), examining their applications and effectiveness. For each type, we analyze both LLM Actions and Human Participation components through detailed examples and summary tables, providing a comprehensive understanding of their architectures, capabilities, and limitations.

2.1. LLM-based Human Simulation Formulation

We formally unify the simulation process into a general framework consisting of three core components. First, the simulation environment $\mathcal{E}$ encompasses a state space $\mathcal{S}$ that defines different states of the simulation, and evaluation procedures $\mathcal{V}$ that assess the actions a generated by agents. Second, the simulation agents $\mathcal{F} = \{f_1, ..., f_n\}$ represent the LLM-based actors, where each agent f maps its input message m and environment state to actions a. Third, the simulation rules $\mathcal{R}$ govern agent interactions and behaviors, state transitions, and evaluation criteria.

In this framework, LLMs serve as the underlying model for agents f to generate human-like behaviors, while human knowledge and data provide the ground truth for evaluation procedures $\mathcal{V}$. The details are in Algorithm 1.

2.2. Simulations Categories

LLM-based human simulations can be broadly categorized into four main types based on their application scenarios: social simulation, economic simulation, policy simulation, and psychological simulation.

Social simulation focuses on modeling complex social interactions and group dynamics. Leveraging LLM's capabilities, LLM agents with unique profiles engage in extensive communication, generating rich behavioral data for in-depth social science analysis (Bail, 2024; Ziems et al., 2024). These simulations often explore emergent social phenomena, communication patterns, and group behavior dynamics (Wei et al., 2022).

Economic simulation explores market dynamics and economic decision-making processes. LLMs are provided with endowments and information, and set with pre-defined preferences, allowing for an exploration of their actions in economic contexts (Bauer et al., 2023; Chen et al., 2023). These simulations particularly focus on market interactions, resource allocation, and strategic decision-making in financial

[2] https://github.com/Persdre/llm-human-simulation

Algorithm 1 General Framework for LLM-based Human Simulation

```
Input:
Environment setup $\mathcal{E}$, interaction rules $\mathcal{R}$, initial agent profiles $P = \{p_1, ..., p_n\}$ // Human-defined: simulation context and constraints
Human behavior dataset $D_{\text{human}}$, expert knowledge base $K_{\text{expert}}$, validation criteria $\mathcal{V}$ // Human-defined: validation basis from real data
LLM-based agents $\mathcal{F} = \{f_1, ..., f_n\}$, state space $\mathcal{S}$, action space $\mathcal{A}$ // Human-defined: system configuration
while simulation not complete do
    Generate agent responses: $a_i \leftarrow f_i(m_i, \mathcal{E})$ for each $f_i \in \mathcal{F}_s$ // Agents process inputs and generate actions
    Update agent memories: $\mathcal{S}_i \leftarrow \text{update}(a_i, m_i)$ for each $f_i \in \mathcal{F}_s$ // Maintain agent state history
    Plan next actions: $\text{plan}_i \leftarrow f_i(\mathcal{S}_i, \mathcal{R})$ for each $f_i \in \mathcal{F}_s$ // Generate action plans based on rules
    Expert validation: $v_{\text{expert}} \leftarrow \mathcal{V}_{\text{expert}}(\mathcal{E}, \{a_i\})$ // Human experts assess agent behaviors
    Data-based validation: $v_{\text{data}} \leftarrow \mathcal{V}_{\text{data}}(\{a_i\}, D_{\text{human}})$ // Compare with real human data
    Rule compliance check: $v_{\text{rule}} \leftarrow \mathcal{V}_{\text{rule}}(\{a_i\}, \mathcal{R})$ // Verify rule compliance
    $\mathcal{E} \leftarrow \text{update}(\mathcal{E}, [v_{\text{expert}}, v_{\text{data}}, v_{\text{rule}}], \{\mathcal{S}_i\})$ // Update based on validations
    Update interaction history: $H \leftarrow H \cup \{a_i, v\}$ // Record interactions and validations
end while
Output: Validated simulation results and interaction history $H$ // Final simulation outcomes with validation records
```

markets (Kim et al., 2024; Li et al., 2024f).

Policy simulation examines policy development and implementation impacts. LLMs are utilized to simulate policy-making processes through virtual government scenarios and evaluate policy impacts across diverse communities (Lempert, 2002). These simulations enable policymakers to understand complex policy dynamics and potential societal effects before real-world implementation (Zhang et al., 2024b), thereby reducing risks and improving policy effectiveness.

Psychological simulation investigates individual and group psychological processes. Multiple LLM agents are utilized to simulate humans with diverse personality traits and cognitive patterns (Dillion et al., 2023; Ke et al., 2024), enabling systematic studies of mental processes, emotional responses, and behavioral tendencies (Ke et al., 2024). These simulations specifically focus on understanding psychological mechanisms and behavioral dynamics across various contextual scenarios.

In the following sections, we provide a detailed analysis of each simulation category, examining their specific methodologies, applications, and designs.

2.3. Social Simulation

LLMs have emerged as a powerful tool in Computational Social Science (CSS) research since the pioneering work of simulating human daily interactions in a virtual town (Park et al., 2023a). The flexibility of LLM-based simulations (Gao et al., 2024) enables both the exploration of diverse scenarios and the study of emergent phenomena in controlled environments (Wei et al., 2022), while also validating conclusions derived from human experiments (Zhao et al., 2023).

Several notable implementations demonstrate the versatility of LLM simulations. Zhao et al. (2023) developed CompeteAI, a framework examining inter-agent competition through a virtual town environment with distinct agent types. In healthcare simulation, Li et al. (2024b) introduced Agent Hospital, where LLM-powered agents representing medical staff and patients simulate comprehensive treatment processes, complementing similar work in AgentClinic (Schmidgall et al., 2024). In academic settings, Jin et al. (2024) developed AgentReview, a peer review simulation framework that addresses privacy concerns while analyzing latent factors in the review process. These simulation approaches have also proven effective in education (Zhang et al., 2024c), validating traditional classroom interaction patterns while improving the learning experience.

However, LLM-powered human simulation raised trustworthiness and reliability concerns (Zhu et al., 2024). Li et al. (2024e) pointed out that LLM agents exhibited incon-

Table 1. Analysis of LLM Actions and Human Participation in Social Simulations. Following Algorithm 1, we categorize simulation components into LLM-generated behaviors and human-defined controls.

LLM Actions	Human Participation
Social Interaction	**Framework Design**
Dialogue generation	Environment rules
Decision-making	Interaction protocols
Inter-agent communication	System architecture
Role-based Behavior	**Validation**
Role simulation	Behavior consistency
Context-aware responses	Evaluation metrics
Role-specific knowledge	Quality control
Process Simulation	**Process Control**
Peer review decisions	Scenario configuration
Treatment management	Parameter adjustment
Multi-agent coordination	Intervention design

sistency between "what they reported" and "how they behaved" during tests. For instance, when an LLM agent was asked to select a personality trait, it selected "extraverted"; however, during conversations, it behaved more aligned with an "introverted" personality. These findings suggested that LLMs displayed behaviors inconsistent with their self-reported traits, raising concerns about the authenticity and reliability of LLM-based simulations in related research.

LLM Actions in social simulations focus on behavior generation and state maintenance. The agents generate responses through dialogue and social interactions ($a_i \leftarrow f_i(m_i, \mathcal{E})$), update their internal states including memories and relationships ($\mathcal{S}_i \leftarrow \text{update}(a_i, m_i)$), and plan future actions based on the current context ($\text{plan}_i \leftarrow f_i(\mathcal{S}_i, \mathcal{R})$). The inherent challenges lie in maintaining consistent personality traits and generating contextually appropriate behaviors.

Human Participation primarily involves the initial setup and ongoing validation. Experts define the simulation environment ($\mathcal{E}$), interaction rules ($\mathcal{R}$), and validation criteria ($\mathcal{V}$). During simulation, they assess agent behaviors through expert validation (v_{expert}), data-based comparison (v_{data}), and rule compliance checks (v_{rule}). The key challenges include designing comprehensive validation mechanisms and ensuring simulation fidelity.

To better understand social simulations through the lens of Algorithm 1, we analyze their implementation of LLM Actions and Human Participation in Table 1.

2.4. Economic Simulation

LLMs have been employed in various economic simulations, from individual trading decision-making to system-level market dynamics (Li et al., 2024f; Luo et al., 2024; Wang et al., 2024a). In behavioral economics experiments, Horton (2023) demonstrated LLMs' capability as homo silicus (Kar, 2023) in replicating classic scenarios like unilateral dictator games (Kahneman et al., 1986) and hiring decisions (Horton et al., 2011). At the system level, Li et al. (2023a) successfully used LLMs to simulate labor markets, consumption patterns, and macroeconomic phenomena through the EconAgent framework (Li et al., 2024d), achieving better performance than traditional models in predicting inflation and unemployment trends.

Studies have shown that LLM simulation decisions are generally rational (Bauer et al., 2023; Chen et al., 2023) and align with utility maximization (Kim et al., 2024) while capturing deviations (Bybee, 2023) to some extent. Lu et al. (2024); Phelps & Russell (2023) proved that LLMs exhibit not only fairness, cooperation, and social norms but also altruism and selfishness. However, Ross et al. (2024) found that LLMs show weaker loss aversion, similar risk aversion, and stronger time discounting compared to humans, which may limit their ability to simulate human behavior in economic scenarios.

In game theory applications, Guo et al. (2024) discovered that certain LLMs can converge faster to Nash Equilibrium strategies with gaming history. Fontana (Fontana et al., 2024) found that in the Iterated Prisoner's Dilemma, Llama2 and GPT3.5 were more cooperative and forgiving, while Llama3 was uncooperative unless the opponent always cooperated.

LLM Actions in economic simulations focus on market behavior generation and strategy formation. The agents generate trading decisions and market responses ($a_i \leftarrow f_i(m_i, \mathcal{E})$), update their market knowledge and positions ($\mathcal{S}_i \leftarrow \text{update}(a_i, m_i)$), and plan trading strategies based on market conditions ($\text{plan}_i \leftarrow f_i(\mathcal{S}_i, \mathcal{R})$). The key challenges lie in maintaining economic rationality and generating consistent market behaviors.

Human Participation primarily involves market design and validation. Experts define the market environment ($\mathcal{E}$), trading rules ($\mathcal{R}$), and evaluation criteria ($\mathcal{V}$). During simulation, they assess market outcomes through expert validation (v_{expert}), empirical comparison (v_{data}), and regulatory compliance checks (v_{rule}). The main challenges include ensuring market efficiency and maintaining economic stability.

To better understand economic simulations through the lens of Algorithm 1, we analyze their implementation of LLM Actions and Human Participation in Table 2.

2.5. Policy Simulations

Policy simulations require modeling complex human interactions and trust dynamics. One critical aspect of such

Table 2. Analysis of LLM Actions and Human Participation in Economic Simulations. Following Algorithm 1, we categorize simulation components into LLM-generated behaviors and human-defined controls.

LLM Actions	Human Participation
Market Interaction	**Market Design**
Trading decisions Price negotiation Resource allocation	Market rules Trading protocols Market structure
Strategy Formation	**Validation**
Portfolio management Risk assessment Investment planning	Economic rationality Performance metrics Market efficiency
Market Adaptation	**Process Control**
Market analysis Strategy adjustment Multi-agent coordination	Parameter tuning Market intervention Stability control

Table 3. Analysis of LLM Actions and Human Participation in Policy Simulations. Following Algorithm 1, we categorize simulation components into LLM-generated behaviors and human-defined controls.

LLM Actions	Human Participation
Policy Implementation	**Policy Design**
Response generation Decision execution Stakeholder coordination	Regulatory framework Implementation rules Policy structure
Impact Assessment	**Validation**
Outcome prediction Compliance monitoring Impact tracking	Policy effectiveness Assessment metrics Quality assurance
Policy Adaptation	**Process Control**
Policy analysis Strategy refinement Multi-agent alignment	Parameter adjustment Policy intervention Implementation control

simulations is human trust (Xie et al., 2024; **?**), a foundational element that influences decision-making and collaboration in various domains, from governance and commerce to community-building. Trust not only shapes individual behaviors but also drives the success of policies by fostering cooperation and reducing uncertainties.

Agent-based models (ABMs) have emerged as a powerful tool for policy simulation. By modeling individual decision-making processes (Zhang et al., 2024b) and incorporating heterogeneity, ABMs enable researchers to explore multi-scalar and interdisciplinary complexities (An et al., 2021), such as feedback, nonlinearity, and time lags. Lempert (Lempert, 2002) highlights ABMs' potential in simulating complex social systems and supporting public policy decision-making, especially under deep uncertainty where traditional predictive methods fail.

LLM-powered ABMs have shown promising results in policy simulation. Li (Li et al., 2024d) constructed a simulation environment that integrates labor and consumption market dynamics, driven by agents' decisions on work and consumption, as well as fiscal and monetary policies. Through this framework, EconAgent demonstrated swift replication of complex human-like decision-making patterns in policy response scenarios.

Urban policy simulation has particularly benefited from LLM integration. Xu et al. (2023a) introduces the Urban Generative Intelligence (UGI) platform, combining the UrbanKG knowledge graph with a city simulator engine. The framework's CityGPT, a domain-specific LLM (Yan et al., 2024; Zou et al., 2025) pre-trained on urban-specific data, creates embodied agents for simulating urban tasks such as mobility planning and policy-making.

LLM Actions in policy simulations focus on policy execution and impact assessment. The agents generate policy responses ($a_i \leftarrow f_i(m_i, \mathcal{E})$), update their compliance states ($\mathcal{S}_i \leftarrow \text{update}(a_i, m_i)$), and plan implementation strategies ($\text{plan}_i \leftarrow f_i(\mathcal{S}_i, \mathcal{R})$). The key challenges lie in maintaining policy consistency and coordinating multi-stakeholder actions.

Human Participation involves policy framework design and effectiveness validation. Experts define the policy environment ($\mathcal{E}$), regulatory rules ($\mathcal{R}$), and assessment criteria ($\mathcal{V}$). During simulation, they evaluate policy outcomes through expert review (v_{expert}), impact assessment (v_{data}), and compliance verification (v_{rule}). The main challenges include ensuring policy validity and implementation feasibility.

To better understand policy simulations through the lens of Algorithm 1, we analyze their implementation of LLM Actions and Human Participation in Table 3.

2.6. Psychology Simulation

LLMs have demonstrated significant potential in psychological simulations due to their ability to replicate human cognitive and social behaviors (Dillion et al., 2023; Ke et al., 2024). Studies have shown that LLMs can effectively simulate human decision-making processes and information search patterns (Binz & Schulz, 2023; Hagendorff, 2023), with performance comparable to average human participants in problem-solving tasks (Orrù et al., 2023). However, certain limitations persist, particularly in simulating underlying psychological traits (Petrov et al., 2024).

Table 4. Analysis of LLM Actions and Human Participation in Psychology Simulations. Following Algorithm 1, we categorize simulation components into LLM-generated behaviors and human-defined controls.

LLM Actions	Human Participation
Cognitive Process	**Experimental Design**
Mental state generation	Study protocols
Decision reasoning	Behavioral rules
Emotional response	Experimental setup
Behavioral Response	**Validation**
Behavior generation	Psychological validity
Response adaptation	Assessment metrics
Pattern formation	Theory consistency
Mental Dynamics	**Process Control**
State transition	Variable control
Cognitive development	Intervention design
Personality expression	Quality assurance

Various applications have emerged in psychological domains. In counseling, frameworks like ChatCounselor (Liu et al., 2023) and interactive role-playing scenarios (Qiu & Lan, 2024) have shown promise in generating high-quality therapeutic dialogues. In psychological assessment, LLMs have demonstrated capability in inferring psychological traits and deriving personality profiles (Peters & Matz, 2024), while innovative approaches like PsychoGAT (Yang et al., 2024b) transform standardized scales into interactive assessment tools. In professional education, frameworks such as PATIENT-ψ (Wang et al., 2024c) combine cognitive models with LLMs to create realistic patient interactions for training purposes.

LLM Actions in psychology simulations focus on cognitive processes and behavioral responses. The agents generate psychological responses ($a_i \leftarrow f_i(m_i, \mathcal{E})$), update mental states ($\mathcal{S}_i \leftarrow \text{update}(a_i, m_i)$), and plan behavioral strategies ($\text{plan}_i \leftarrow f_i(\mathcal{S}_i, \mathcal{R})$). The key challenges lie in maintaining psychological consistency and simulating complex mental processes.

Human Participation involves experimental design and behavioral validation. Experts define the psychological environment ($\mathcal{E}$), behavioral protocols ($\mathcal{R}$), and assessment criteria ($\mathcal{V}$). During simulation, they evaluate responses through expert analysis (v_{expert}), behavioral comparison (v_{data}), and protocol verification (v_{rule}). The main challenges include ensuring psychological validity and experimental rigor.

To better understand psychology simulations through the lens of Algorithm 1, we analyze their implementation of LLM Actions and Human Participation in Table 4.

3. LLM Inherent Drawbacks

In this section, we analyze the inherent drawbacks of LLMs that limit their effectiveness in simulation, particularly focusing on how these limitations affect the LLM actions identified in Section 2.

3.1. Bias

Bias in LLMs fundamentally affects their cognitive and behavioral simulation capabilities. These biases manifest as systematic misrepresentation and distortions that impact both cognitive processes and behavioral patterns (Chen et al., 2024b; Ferrara, 2023). The presence of these biases significantly limits LLMs' ability to accurately simulate diverse human behaviors and thought processes, particularly affecting their cognitive reasoning and decision-making capabilities.

Cultural bias impairs cross-cultural interaction mechanisms. Training data predominantly sourced from English-speaking and Western contexts results in limited understanding of diverse cultural interaction patterns (Santurkar et al., 2023; Wang et al., 2023). For instance, (Wang et al., 2023) demonstrated that LLMs show systematic preferences for Western cultural elements, limiting their ability to simulate authentic cross-cultural interactions and decision-making processes. This bias particularly affects the interaction mechanisms component of LLM-driven simulations, reducing their effectiveness in global scenarios.

Gender bias distorts behavioral pattern simulation. Studies have shown that LLMs perpetuate gender stereotypes in their behavioral simulations (Kotek et al., 2023; Wan et al., 2023). For instance, Kotek et al. (2023) found that LLMs are 3–6 times more likely to generate gender-stereotypical behavioral patterns, affecting both the cognitive processes and behavioral patterns components of simulations. This bias particularly impacts the authenticity of simulated social interactions and decision-making processes.

Occupational and socioeconomic biases affect cognitive process simulation. These biases significantly impact how LLMs simulate decision-making and reasoning processes across different social groups (Gwartney & McCaffree, 1971; Harrison & Budworth, 2015). The overrepresentation of certain professions and socioeconomic groups in training data affects the cognitive processes component, leading to unrealistic simulation of thought patterns and decision-making processes for underrepresented groups.

3.2. LLM Capability Limitations

LLM's cognitive process simulation limitations affect decision-making authenticity. LLMs struggle to maintain consistent cognitive patterns when simulating human decision-making processes (Hu & Collier, 2024; Li et al., 2024e). This limitation directly impacts the cognitive processes component identified in Section 2. Specifically, LLMs show inconsistent reasoning across different scenarios (Gui & Toubia, 2023), struggle with emotional processing in decision-making (Ai et al., 2024), and face challenges in adapting to new information (Ke et al., 2024).

Behavioral pattern simulation faces temporal consistency challenges. The behavioral patterns component is significantly impacted by LLMs' inability to maintain consistent long-term memory (Zheng et al., 2024). These limitations manifest in their difficulty to maintain consistent behavioral patterns over time (Li et al., 2024a), develop realistic habit formation (Zhong et al., 2024), and effectively simulate learning from past experiences (Chu et al., 2024).

Interaction mechanism limitations affect multi-agent simulation quality. The interaction mechanisms component faces significant challenges due to LLMs' limitations in several areas. LLMs struggle with maintaining consistent personas across multiple interactions (Han et al., 2022), often fail to accurately simulate complex social dynamics (Shanahan et al., 2023), and face difficulties in replicating authentic group behaviors (Park et al., 2023b).

Memory constraints limit behavioral consistency across interactions. LLMs' memory limitations particularly affect the behavioral patterns and interaction mechanisms components (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

In this section, we analyze the fundamental drawbacks in current human-designed simulation frameworks that limit their reliability and effectiveness, particularly focusing on how these design limitations affect the simulation components identified in Section 2.

4.1. Framework Design Drawbacks

Current frameworks oversimplify complex human psychological states. Human designers often create oversimplified frameworks when attempting to model psychological states (Jansen et al., 2023; Tjuatja et al., 2024). 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).

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 (Chen et al., 2024b; Grossberg & Gutowski, 1987). 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. Human designers face significant challenges in creating frameworks that accurately model complex human motivations and incentives (Petrakis & Petrakis, 2012; Stern, 1999). 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). 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 in complex scenarios (Reason et al., 2024).

Integration of expert knowledge faces systematic challenges. Current simulation designs struggle to effectively incorporate expert knowledge and domain-specific insights (Si et al., 2024). Framework designers face difficulties in creating systems that can translate qualitative expert knowledge into quantitative simulation parameters while maintaining the nuance and complexity of human behavior (De La Torre et al., 2024). This limitation particularly affects the ability to create realistic behavioral simulations.

5. Solutions for Reliable LLM Simulation

Based on our analysis of LLM inherent limitations (Section 3) and simulation design challenges (Section 4), we propose comprehensive solutions to enhance the reliability of LLM-based human simulations. These solutions focus on two fundamental aspects: addressing LLM inherent limitations (Section 5.1) and improving simulation design frameworks (Section 5.2). Figure 2 provides an overview of our proposed solutions.

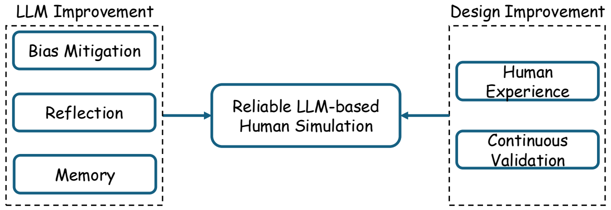

Figure 2. Solutions for Reliable LLM-based Human Simulation

5.1. Addressing LLM Inherent Limitations

Bias mitigation through enhanced training strategies. To address cultural, gender, and socioeconomic biases, we propose: (1) incorporating diverse and balanced training data from multiple cultural contexts (Santurkar et al., 2023), for example, using multilingual datasets and ensuring equal representation of different cultural perspectives; (2) implementing debiasing algorithms during model training (Kotek et al., 2023), including techniques like counterfactual data augmentation and balanced fine-tuning; and (3) developing occupation-aware training protocols (Harrison & Budworth, 2015).

Enhanced cognitive consistency through architectural improvements. To improve cognitive process simulation, we recommend: (1) implementing reflection mechanisms for consistent decision-making (Li et al., 2024e); (2) developing specialized modules for different cognitive tasks (Hu & Collier, 2024); and (3) incorporating feedback loops for behavioral consistency (Gui & Toubia, 2023).

Memory enhancement through hybrid architectures. To address memory limitations, we propose: (1) implementing external memory banks for effective information storage and retrieval (Zhong et al., 2024); (2) developing temporal awareness modules (Chu et al., 2024); and (3) creating hierarchical memory structures for both short-term and long-term information management (Yuan et al., 2024).

5.2. Improving Simulation Design

Comprehensive framework design with validation. To enhance framework reliability, we recommend: (1) developing modular simulation architectures for component-specific validation (Reason et al., 2024); (2) implementing continuous monitoring systems (Xexéo et al., 2024); and (3) creating standardized evaluation metrics (He et al., 2023).

Enhanced human experience modeling. To better capture human experiences, we propose: (1) incorporating multimodal data sources (Evans, 2015); (2) developing context-aware frameworks (Beratan, 2007); and (3) implementing experience accumulation mechanisms (Stern, 1999).

Improved incentive modeling. To address motivation modeling challenges: (1) developing multi-dimensional reward systems (Petrakis & Petrakis, 2012); (2) creating adaptive incentive mechanisms (Maslow, 2023); and (3) implementing context-sensitive reward structures (Williams, 2000).

5.3. Integration and Validation

Comprehensive validation mechanisms. To ensure simulation reliability: (1) implementing hierarchical validation systems (Ke et al., 2024); (2) developing real-time monitoring tools (Shapira et al., 2024); and (3) creating standardized benchmarks (Si et al., 2024).

Expert knowledge integration. To effectively incorporate domain expertise: (1) developing systematic methods for knowledge translation (De La Torre et al., 2024); (2) creating flexible domain adaptation frameworks (Jansen et al., 2023); and (3) implementing continuous refinement mechanisms (Aldridge & Kim, 2024).

6. Future Directions

Although LLM-based human simulations face current limitations, the future development of this field shows significant potential. This section discusses three key directions that could advance the field substantially.

6.1. Advancing Human Data Collection

Recent developments in wearable technology, such as smart glasses and rings, enable comprehensive collection of human behavioral data (Cardenas et al., 2024; Fang et al., 2024). These data sources extend beyond traditional text, image, and audio formats, providing richer information for improving LLM simulations in several aspects:

Physiological and cognitive data collection. Wearable sensors can capture physiological signals (e.g., heart rate, skin conductance) and brain activity patterns (Yuan et al., 2024). This data helps LLM simulations better understand and model human emotional states and decision-making processes.

Behavioral pattern monitoring. Motion sensors and activity trackers can record daily behavioral patterns, including movement, social interactions, and routines (Chu et al., 2024). This detailed behavioral data enables LLMs to simulate human actions more accurately in various situations.

Environmental context awareness. Environmental sensors can measure surrounding conditions that affect human behavior (Li et al., 2024a). This contextual information allows LLM simulations to better account for how environment influences human responses.

6.2. Synthesizing High-Quality Data

Advances in LLM technology enable the generation of high-quality synthetic training data (Ke et al., 2024), offering several advantages:

Scenario-based data generation. LLMs can create realistic scenarios that are difficult to observe in real-world settings (Tjuatja et al., 2024). For example, they can simulate various social skill practice scenarios, such as conflict resolution and negotiation situations (Yang et al., 2024a). This allows researchers to generate training data for situations where real-world data collection would be impractical or ethically challenging.

Behavioral variation synthesis. LLMs can generate diverse behavioral patterns by systematically varying key parameters such as personality traits, emotional states, and cultural backgrounds (He et al., 2023). For instance, in counseling simulations, LLMs can create client responses that reflect different mental health conditions, coping mechanisms, and treatment responses, providing rich training data for therapeutic applications.

Cross-validation with expert knowledge. The synthetic data can be validated through a combination of expert evaluation and real-world benchmarks (Si et al., 2024). For example, in economic simulations, generated trading behaviors can be compared with historical market data and validated by domain experts. This multi-level validation process helps ensure that synthetic data accurately reflects real-world human behavior patterns while maintaining domain-specific validity.

6.3. LLM as a Judge

Automated evaluation through LLMs. Human evaluation still remains the gold standard for assessing LLM-based human simulations. However, this approach is both time-consuming and resource-intensive (Ouyang et al., 2022; Zheng et al., 2023a). If LLM-based human simulations can improve to a high level, LLMs can serve as automated evaluators, and it could evaluate the LLM-generated data's quality, significantly reducing the manual effort required in data evaluation.

Iterative improvement through feedback loops. By implementing automated LLM evaluation systems, researchers can create continuous feedback loops where evaluation results directly inform simulation improvements (Dubois et al., 2024; Tyser et al., 2024). This iterative process allows for rapid refinement of simulation models, leading to more accurate and reliable human behavior simulations.

7. Conclusion

In this position paper, we analyze the current state and future potential of LLM-based human simulations. While our analysis reveals significant challenges in cognitive processing and framework design, we believe these limitations present valuable opportunities for advancement. Specifically, the growing capabilities of LLMs in generating and evaluating synthetic data suggest a promising future where LLMs could serve both as simulation engines and quality control mechanisms. This dual role of LLMs, combined with improved data collection methods, could lead to a self-improving cycle in human behavior simulation. Looking forward, we expect the field to move beyond simple behavior replication toward more sophisticated simulations that capture the nuanced aspects of human cognition and social interaction. To support this development, we provide a collection of resources that researchers can build upon for future work.

References

Ai, Y., He, Z., Zhang, Z., Zhu, W., Hao, H., Yu, K., Chen, L., and Wang, R. Is cognition and action consistent or not: Investigating large language model's personality. *arXiv preprint arXiv:2402.14679*, 2024.

Aldridge, I. and Kim, D. Quantitative financial models with scenarios from llm: Temporal fusion transformers as alternative monte-carlo. *Available at SSRN 4999492*, 2024.

An, L., Grimm, V., Sullivan, A., Turner Ii, B., Malleson, N., Heppenstall, A., Vincenot, C., Robinson, D., Ye, X., Liu, J., et al. Challenges, tasks, and opportunities in modeling agent-based complex systems. *Ecological Modelling*, 457:109685, 2021.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A. L., Darrell, T., Malik, J., and Efros, A. A. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22861–22872, 2024.

Bail, C. A. Can generative ai improve social science? *Pro-*

ceedings of the National Academy of Sciences, 121(21): e2314021121, 2024.

Bainbridge, W. S. *Computer simulations of space societies*. Springer, 2018.

Bauer, K., Liebich, L., Hinz, O., and Kosfeld, M. Decoding gpt's hidden 'rationality'of cooperation. 2023.

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

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

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

Cardenas, L., Parajes, K., Zhu, M., and Zhai, S. Autohealth: Advanced llm-empowered wearable personalized medical butler for parkinson's disease management. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0375–0379. IEEE, 2024.

Chen, G., Fan, L., Gong, Z., Xie, N., Li, Z., Liu, Z., Li, C., Qu, Q., Ni, S., and Yang, M. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*, 2024a.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. Humans or LLMs as the judge? a study on judgement bias. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., and yi Lee, H. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course, 2024. URL https://arxiv.org/abs/2407.05216.

Chu, Z., Wang, Z., Zhang, R., Ji, Y., Wang, H., and Sun, T. Improve temporal awareness of llms for sequential recommendation. *arXiv preprint arXiv:2405.02778*, 2024.

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

De La Torre, F., Fang, C. M., Huang, H., Banburski-Fahey, A., Amores Fernandez, J., and Lanier, J. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024.

Dignum, F., Dignum, V., Davidsson, P., Ghorbani, A., van der Hurk, M., Jensen, M., Kammler, C., Lorig, F., Ludescher, L. G., Melchior, A., et al. Analysing the combined health, social and economic impacts of the corovanvirus pandemic using agent-based social simulation. *Minds and Machines*, 30:177–194, 2020.

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

Dillion, D., Tandon, N., Gu, Y., and Gray, K. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Downing, T. E., Moss, S., and Pahl-Wostl, C. Understanding climate policy using participatory agent-based social simulation. In *International workshop on multi-agent systems and agent-based simulation*, pp. 198–213. Springer, 2000.

Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P. S., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Fang, C. M., Danry, V., Whitmore, N., Bao, A., Hutchison, A., Pierce, C., and Maes, P. Physiollm: Supporting personalized health insights with wearables and large language models. *arXiv preprint arXiv:2406.19283*, 2024.

Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Fontana, N., Pierri, F., and Aiello, L. M. Nicer than humans: How do large language models behave in the prisoner's dilemma? *arXiv preprint arXiv:2406.13605*, 2024.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., and Li, Y. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.

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

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

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

Guo, S., Bu, H., Wang, H., Ren, Y., Sui, D., Shang, Y., and Lu, S. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*, 2024.

Gwartney, J. D. and McCaffree, K. M. Variance in discrimination among occupations. *Southern Economic Journal*, pp. 141–155, 1971.

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

Han, S., Kim, B., Yoo, J. Y., Seo, S., Kim, S., Erdenee, E., and Chang, B. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. *arXiv preprint arXiv:2204.10825*, 2022.

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

He, T., Fu, G., Yu, Y., Wang, F., Li, J., Zhao, Q., Song, C., Qi, H., Luo, D., Zou, H., et al. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*, 2023.

Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14:399–425, 2011.

Hu, T. and Collier, N. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*, 2024.

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

Jansen, B. J., Jung, S.-g., and Salminen, J. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023.

Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., and Wang, J. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.

Kahneman, D., Knetsch, J. L., Thaler, R., et al. Fairness as a constraint on profit seeking: Entitlements in the market. *American economic review*, 76(4):728–741, 1986.

Kar, S. *Simulating Economic Experiments Using Large Language Models: Design and Development of a Computational Tool*. PhD thesis, Massachusetts Institute of Technology, 2023.

Ke, L., Tong, S., Cheng, P., and Peng, K. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*, 2024.

Kim, J., Kovach, M., Lee, K.-M., Shin, E., and Tzavellas, H. Learning to be homo economicus: Can an llm learn preferences from choice. *arXiv preprint arXiv:2401.07345*, 2024.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pp. 12–24, 2023.

Lee, B. W., Lee, Y., and Cho, H. Language models show stable value orientations across diverse role-plays. *arXiv preprint arXiv:2408.09049*, 2024.

Lempert, R. Agent-based modeling as organizational and public policy simulators. *Proceedings of the national academy of sciences*, 99(suppl_3):7195–7196, 2002.

Li, H., Yang, C., Zhang, A., Deng, Y., Wang, X., and Chua, T.-S. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*, 2024a.

Li, J., Wang, S., Zhang, M., Li, W., Lai, Y., Kang, X., Ma, W., and Liu, Y. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024b.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024c.

Li, N., Gao, C., Li, Y., and Liao, Q. Large language model-empowered agents for simulating macroeconomic activities. *Available at SSRN 4606937*, 2023a.

Li, N., Gao, C., Li, M., Li, Y., and Liao, Q. 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), pp. 15523–15536, 2024d.

Li, S., Yang, J., and Zhao, K. 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*, 2023b.

Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., and Sun, L. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*, 2024e.

Li, Y., Luo, B., Wang, Q., Chen, N., Liu, X., and He, B. 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*, pp. 1094–1106, 2024f.

Lin, J., Zhao, H., Zhang, A., Wu, Y., Ping, H., and Chen, Q. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*, 2023.

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., and Wu, J. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*, 2023.

Lu, Y., Aleta, A., Du, C., Shi, L., and Moreno, Y. Llms and generative agent-based models for complex systems research. *Physics of Life Reviews*, 2024.

Luo, B., Zhang, Z., Wang, Q., Ke, A., Lu, S., and He, B. Ai-powered fraud detection in decentralized finance: A project life cycle perspective. *ACM Computing Surveys*, 57(4):1–38, 2024.

Mao, H., Chen, Z., Tang, W., Zhao, J., Ma, Y., Zhao, T., Shah, N., Galkin, M., and Tang, J. Position: Graph foundation models are already here. In *Forty-first International Conference on Machine Learning*, 2024.

Maslow, A. H. *Motivation And Personality: Motivation And Personality: Unlocking Your Inner Drive and Understanding Human Behavior by AH Maslow*. Prabhat Prakashan, 2023.

Moniri, B., Hassani, H., and Dobriban, E. Evaluating the performance of large language models via debates, 2024. URL https://arxiv.org/abs/2406.11044.

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

Ohm, A. K. and Singh, K. K. Study of tokenization strategies for the santhali language. *SN Computer Science*, 5 (7):807, 2024.

Orcutt, G. H., Caldwell, S., and Wertheimer, R. F. *Policy exploration through microanalytic simulation*. The Urban Insitute, 1976.

Orrù, G., Piarulli, A., Conversano, C., and Gemignani, A. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in artificial intelligence*, 6:1199350, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023a.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. 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, 2023b. Association for Computing Machinery.

Peters, H. and Matz, S. C. Large language models can infer psychological dispositions of social media users. *PNAS nexus*, 3(6):pgae231, 2024.

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

Petrov, N. B., Serapio-García, G., and Rentfrow, J. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

Phelps, S. and Russell, Y. I. Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*, 2023.

Qiu, H. and Lan, Z. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*, 2024.

Reason, T., Rawlinson, W., Langham, J., Gimblett, A., Malcolm, B., and Klijn, S. 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, 2024.

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

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., and Moor, M. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Shanahan, M., McDonell, K., and Reynolds, L. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Shapira, E., Madmon, O., Reichart, R., and Tennenholtz, M. Can large language models replace economic choice prediction labs? *arXiv preprint arXiv:2401.17435*, 2024.

Shi, Z., Gao, S., Chen, X., Feng, Y., Yan, L., Shi, H., Yin, D., Ren, P., Verberne, S., and Ren, Z. Learning to use tools via cooperative and interactive agents. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10642–10657, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.624. URL https://aclanthology.org/2024.findings-emnlp.624/.

Si, C., Yang, D., and Hashimoto, T. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Son, G., Ko, H., Lee, H., Kim, Y., and Hong, S. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*, 2024.

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

Tang, S., Pang, X., Liu, Z., Tang, B., Ye, R., Dong, X., Wang, Y., and Chen, S. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*, 2024.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., and Neubig, G. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Tyser, K., Segev, B., Longhitano, G., Zhang, X.-Y., Meeks, Z., Lee, J., Garg, U., Belsten, N., Shporer, A., Udell, M., et al. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*, 2024.

Upadhayay, B., Behzadan, V., and Karbasi, A. Cognitive overload attack: Prompt injection for long context. *arXiv preprint arXiv:2410.11272*, 2024.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

Wang, Q., Gao, Y., Tang, Z., Luo, B., and He, B. Enhancing llm trading performance with fact-subjectivity aware reasoning. *arXiv preprint arXiv:2410.12464*, 2024a.

Wang, Q., Wang, T., Li, Q., Liang, J., and He, B. Megaagent: A practical framework for autonomous cooperation in large-scale llm agent systems. *arXiv preprint arXiv:2408.09955*, 2024b.

Wang, R., Milani, S., Chiu, J. C., Zhi, J., Eack, S. M., Labrum, T., Murphy, S. M., Jones, N., Hardy, K., Shen, H., et al. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024c.

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., and Lyu, M. R. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Williams, S. J. Emotion and social theory: corporeal reflections on the (ir) rational. 2000.

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

Wu, Z., Zheng, S., Liu, Q., Han, X., Kwon, B. I., Onizuka, M., Tang, S., Peng, R., and Xiao, C. Shall we talk: Exploring spontaneous collaborations of competing llm agents. *arXiv preprint arXiv:2402.12327*, 2024.

Xexéo, G., Braida, F., Parreiras, M., and Xavier, P. The economic implications of large language model selection on earnings and return on investment: A decision theoretic model. *arXiv preprint arXiv:2405.17637*, 2024.

Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., Hu, Z., Torr, P., Ghanem, B., and Li, G. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*, 2024.

Xu, F., Zhang, J., Gao, C., Feng, J., and Li, Y. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*, 2023a.

Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., and Wu, Y. Towards reasoning in large language models via multi-agent peer review collaboration, 2023b. URL https://arxiv.org/abs/2311.08152.

Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., and Liang, Y. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pp. 4006–4017, 2024.

Yang, D., Ziems, C., Held, W., Shaikh, O., Bernstein, M. S., and Mitchell, J. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*, 2024a.

Yang, Q., Wang, Z., Chen, H., Wang, S., Pu, Y., Gao, X., Huang, W., Song, S., and Huang, G. 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)*, pp. 14470–14505, 2024b.

Yuan, C., Xie, Q., Huang, J., and Ananiadou, S. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pp. 1963–1974, 2024.

Zhang, J., Qiao, D., Yang, M., and Wei, Q. Regurgitative training: The value of real data in training large language models. *arXiv preprint arXiv:2407.12835*, 2024a.

Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., and Wei, F. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024b.

Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Liu, Z., Hou, L., and Li, J. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*, 2024c.

Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., and Xie, X. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512*, 2023.

Zheng, J., Qiu, S., Shi, C., and Ma, Q. Towards lifelong learning of large language models: A survey. *arXiv preprint arXiv:2406.06391*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023a.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023b. URL https://arxiv.org/abs/2306.05685.

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.

Zhu, L., Huang, X., and Sang, J. 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*, pp. 1726–1732, 2024.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

Zou, X., Yan, Y., Hao, X., Hu, Y., Wen, H., Liu, E., Zhang, J., Li, Y., Li, T., Zheng, Y., et al. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606, 2025.