Accuracy, Diversity, and Reflection: Purpose-driven Evaluation for Social Simulation

Heechan Lee
heechan@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Joseph Seering
seering@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Abstract

With the rapid advancement of Large Language Models (LLMs), social simulations are evolving to capture richer interactions. However, current validation methodologies remain predominantly focused on predictive accuracy—assessing how closely outputs mimic ground truth data. In this position paper, we argue that this accuracy-centric view is insufficient, as simulations serve diverse purposes beyond prediction, such as exploring plausible futures and understanding underlying mechanisms. We propose a purpose-driven evaluation framework comprising three dimensions: **Accuracy**, **Diversity**, and **Reflection**. We define Accuracy and Diversity at micro-, meso-, and macro-levels, reflecting how social simulations should be valid not only in individual interactions but also in emergent group dynamics and aggregate outcomes, and that the methods for evaluating at each scale may be different. Furthermore, we introduce Reflection as a critical dimension to evaluate whether the simulation aids users in debugging assumptions (process) and interpreting outcomes (outcome). We conclude with guidelines for applying these complementary dimensions to establish the operational validity of LLM-based social simulations.

Keywords

Social Simulation, Large Language Models, Simulation Evaluation

1 Introduction

The rapid advancement of Large Language Models (LLMs) along with their flexibility in generating diverse behaviors and adapting to context via natural language interaction has generated renewed interest in social simulations. Recent work has begun exploring simulations across multiple domains, including modeling social phenomena [4, 10], supporting educational scenarios [14, 28], and building user simulators for testing models or products [2, 6, 18]. For these simulations to be practically useful, they must be trusted as reliable proxies for reality. Accordingly, prior work has prioritized validation centered on predictive accuracy and realism, typically relying on human judgments [1, 23] or comparisons to ground-truth datasets [9, 10, 17].

However, as Larooij & Törnberg have argued [16], there is a critical misalignment between the purposes of simulation and the methods of validation. Many existing approaches focus on *surface-level realism*, asking merely whether agents appear to mimic individual human behavior or recognizable social patterns. This misplaced focus can obscure whether the intended mechanisms are operating accurately, weakening the coupling between validation results and what the simulation claims to explain.

Building on this critique, in this position paper, we argue that purpose-driven validation must extend beyond accuracy, and that there is a need to systematically articulate the range of evaluation options: in practice, social simulations are used for a wide range of goals beyond predictive accuracy, and evaluation should align with those diverse purposes.

We outline three dimensions for validating LLM-based social simulations that encompass and extend methods used in prior work. **Accuracy** asks whether agent behaviors and interaction dynamics match relevant empirical or theoretical targets at the micro-, meso-, and macro-levels. **Diversity** captures whether the simulation covers a range of plausible behaviors and outcomes—again at micro-, meso-, and macro-levels—rather than repeatedly producing a single "typical" trajectory, while still respecting boundaries on what should and should not be generated. **Reflection** asks whether the simulation enables users to make sense of both how the simulation is constructed and why it produces particular outcomes: (i) *process reflection* supports deliberate consideration of modeling goals, priorities, and assumptions while building the simulation; and (ii) *outcome reflection* supports tracing, explaining, and, when needed, backtracking from results to make outcomes actionable for downstream tasks. We conclude with considerations for application of these different categories of metrics.

2 Background

2.1 Purposes for simulation

Simulations offer a vital alternative to real-world experiments, which are often constrained by high costs, time, and ethical boundaries [3, 20]. Validation has traditionally prioritized *predictive accuracy*—how closely outputs mimic historical ground truth. However, relying on prediction is insufficient for social simulation. Human behavior and social dynamics are inherently heterogeneous and stochastic; even under identical conditions, outcomes emerge as a distribution rather than a single deterministic trajectory [12, 15, 19]. Therefore, "hitting" a specific future is less important than capturing the range of plausible patterns and underlying mechanisms appropriate for the simulation's purpose.

At this point, we revisit the fundamental question: *"Why do we model or simulate?"* Epstein [7] argued that the purpose of modeling cannot be explained by prediction alone. Simulations are used, in some cases, for explanation (why such phenomena occur), and in others to explore plausible futures or bound outcomes, and in still others to support decision-making and learning.

This paper was prepared for PoliSim@CHI 2026: LLM Agent Simulation for Policy, a workshop at CHI 2026 (CHI Conference on Human Factors in Computing Systems), April 16, 2026, Barcelona, Spain.

Analyzing the 16 purposes in [7], two major dimensions emerge beyond what might typically be evaluated through a lens of predictive accuracy. First of these is the goal of exploration. This usage includes Epstein's purposes of "bounding the range of outcomes," "illuminating uncertainties," and "offering options in crisis." To achieve this goal, a simulation must generate diverse, realistically possible behaviors and interactions, demonstrating how far results can spread and within what boundaries they remain. In other words, this goal naturally demands **diversity**—the ability to produce and explore various trajectories and outcomes within a plausible range.

Second is the goal of reflection & understanding. This usage focuses on Epstein's purposes of "explaining," "revealing the simple mechanisms behind apparent complexity (or conversely, showing why the simple is actually complex)," "training/educating practitioners or the public," and "disciplining policy dialogue." Here, the core function of the simulation is to help users interpret results, trace which assumptions led to which outcomes, and make sense of why this happened. Therefore, this goal demands **reflection** from the simulation—the ability to reveal assumptions and mechanisms, and to provide interpretable interactions such as comparison, explanation, and backtracking.

The presence of these two alternative goals does not mean we should discard accuracy as a goal, but rather that our evaluation methods should align with the reason for creating the simulation. Thus, the core of evaluation shifts from "did it perfectly mimic reality?" to "what is it matching, and at what level, as required to achieve the intended purpose?"

2.2 Recent approaches to evaluation

Current evaluation methods for LLM-based social simulations predominantly focus on assessing how closely agents mimic humans or replicate societal patterns. Many approaches rely on human or model-based judgments. Evaluators assess agents on criteria such as naturalness and consistency [1, 23], often utilizing interviews or Turing-style tests to verify adherence to assigned profiles [21, 22]. Recently, LLM-as-a-Judge [29] has been adopted to scale these assessments, automatically evaluating agent interactions against defined social norms [26, 30].

Alternatively, outcomes have been evaluated by comparing simulation results with empirical data or established theories. Studies validate whether simulations reproduce known phenomena, such as fake news propagation [17] or echo chamber formation [10], often using quantitative metrics like cosine similarity to real-world networks [9]. Others benchmark agent behaviors against human distributions in game-theoretic scenarios (e.g., Trust Games) [25] or large-scale QA datasets [13].

The approaches described above mostly aim to measure how much the simulation resembles human behavior, or how close the simulation is to a sort of ground truth. While these approaches are important in proving that the simulation produces realistic and plausible results, evaluations based on whether current simulation systems can support the full range of user goals, such as diversity or reflection, remain underexplored.

3 Broadening Evaluation: Accuracy, Diversity, and Reflection

To broaden simulation evaluation beyond ground-truth fidelity, we propose an evaluation framework consisting of three core dimensions: Accuracy, Diversity, and Reflection. None of these represents a completely novel method, but we hope that their consolidation and comparison provides useful starting point for discussion.

3.1 Accuracy

Accuracy focuses on how realistically the simulation reproduces human behavior and social phenomena. Most existing evaluation approaches for simulation have mainly focused on this perspective. Accordingly, we organize Accuracy across three levels: the micro level (evaluating an individual agent's behavior), the meso level (evaluating interactions among agents), and the macro level (evaluating the overall simulated society), noting that evaluation methods may differ across scales of evaluation.

3.1.1 Micro. At the micro level, accuracy refers to how closely an individual agent's behavior (e.g., actions, utterances, emotions, decisions) aligns with realistic human behavior. This includes whether an agent's responses are aligned with its assigned profile or role, whether emotional expressions are appropriate to the conversational context, and whether its decisions resemble those a human would make in a similar situation. Micro-level accuracy is evaluated through comparisons between agent outputs and human data, such as textual similarity to human utterances [8, 10], alignment with interview responses [21, 23], or human judgments [9, 17].

3.1.2 Meso. At the meso level, accuracy concerns whether interactions among multiple agents and their mutual influences resemble real-world social interaction dynamics. Examples include natural turn-taking in conversations, the formation and evolution of interpersonal relationships, strategic adaptation based on others' emotional states or past actions, and small-scale information propagation within a group. Meso-level accuracy is often evaluated using conversation quality metrics (e.g., coherence, responsiveness) [27], and theory-informed interaction patterns [24]. These methods assess whether interaction structures and dynamics align with known social theories or empirical observations.

3.1.3 Macro. At the macro level, accuracy refers to whether collective social phenomena emerge from the simulation when considering the agent society as a whole. Typical targets include large-scale outcomes such as opinion polarization, norm formation, and population-level information diffusion across networks, though these may be quite difficult to operationalize precisely for the purpose of evaluation. Macro-level accuracy is assessed by comparing aggregate simulation outcomes with known theoretical predictions or empirical patterns [5, 11]. This includes structural checks (e.g., whether numerical trends match existing models), pattern matching against real-world data, and checkpoint-based evaluations that periodically verify whether the simulation trajectory remains consistent at the scale of society.

Table 1: Proposed Evaluation Framework: Accuracy, Diversity, and Reflection.

Dim.	Level	Definition	Examples	Methods
Accuracy	Micro	Alignment of agent's behavior with human reality.	Profile alignment; context-appropriate emotions.	Text similarity to human data; interview alignment; human eval.
	Meso	Realism of interaction dynamics among agents.	Turn-taking; relationship evolution; info propagation.	Comparison with conversation or interaction patterns.
	Macro	Emergence of collective social phenomena.	Polarization; norm formation; collective coordination.	Comparison with theories/empirical patterns; structural checks of outcome.
Diversity	Micro	Variation in agent responses to identical stimuli.	Reaction range (e.g., anger vs. negotiation) to same provocation.	Entropy over actions; semantic distinctiveness; embedding dispersion.
	Meso	Variety of interaction patterns in a group.	Different diffusion paths; alternative conflict trajectories.	Counting distinct interaction patterns; conversation tree structure.
	Macro	Range of societal outcomes from same initial conditions.	Exploring the best & worst cases in static condition.	Distribution of final states; trajectory metrics; coverage of outcome landscape.
Reflection	Process	Understanding assumptions/gaps while building.	Realizing missing variables (e.g., incentives); Iterating the simulation model.	Insight yield; rationale clarity; iteration quality logs.
	Outcome	Interpreting results for downstream tasks.	Tracing causes of the results; Adapting the interpretation for downstream task.	Causal traceability; decision usefulness; comparative reasoning studies.

3.2 Diversity

Diversity measures the simulator's ability to explore different outcomes from the same initial condition. LLM-based simulations can leverage their inherent stochasticity to support outcome bounding, reveal uncertain ranges, and explore variability in plausible futures. As with Accuracy, we define Diversity at the micro, meso, and macro levels.

3.2.1 Micro. At the micro level, diversity refers to the extent of variation an individual agent can exhibit when responding to the same stimulus or situation across multiple simulation runs. For instance, given an identical provocation, an agent may respond with anger in one run, remain silent in another, or attempt negotiation in a third. Micro-level diversity can be assessed by analyzing the distribution of an agent's actions or utterances across repeated runs, and comparing against expected distributions. Metrics such as entropy over action categories, semantic distinctiveness among generated responses, or embedding-based dispersion can be used to quantify variation while holding initial conditions constant.

3.2.2 Meso. At the meso level, diversity captures the variety of interaction patterns and mutual influence structures that emerge among agents within a local network or group. Examples include different diffusion pathways of fake news or alternative conflict resolution trajectories within a team arising from identical starting configurations. Meso-level diversity can be evaluated by analyzing how many distinct interactions and influence patterns appear across repeated runs. Possible approaches include counting distinct patterns of response types, measuring structural differences in conversation trees, or analyzing the emergence of distinct subcommunities using clustering or network modularity.

3.2.3 Macro. At the macro level, diversity refers to the range of distinct societal-level outcomes observable from the simulation under the same valid initial conditions. A diverse simulation may produce both "best-case" and "worst-case" societal trajectories to show the boundary, such as consensus versus polarization, coordination versus fragmentation, without changing the underlying assumptions or inputs. Macro-level diversity can be assessed by analyzing the distribution of final states across runs and comparing them against a theoretical or conceptual outcome landscape. Trajectory metrics can count distinct macro-level paths, capturing diversity beyond endpoints. One possible approach is to first define domain-relevant outcome dimensions (e.g., total impact, inequality across groups, or time to stabilization), discretize this space into plausible outcome regions, and then measure how many of these regions are reached across repeated runs. This can include simple bin coverage, entropy over final-state categories, or coverage of expert-defined outcome regimes. Trajectory metrics can further capture diversity beyond endpoints by clustering repeated runs into distinct pathway types or by comparing differences in macro-level temporal patterns.

A challenge for all levels of Diversity is distinguishing between genuinely plausible outcomes and hallucinated ones. Pursuing diversity does not mean accepting random outcomes that are clearly unrealistic—instead, a simulation system must differentiate between "what could happen" and "what would never happen." Therefore, the goal should be **Plausible Diversity**, either by balancing Diversity with the Accuracy, or by defining boundaries of outcomes prior to running evaluation.

3.3 Reflection

Reflection captures whether building and using the simulation helps users (or even the simulation developers) to better understand the target of simulation, and whether they can interpret simulation results and apply them to downstream tasks (e.g., decision making). Without this capability, users cannot assess how iterative changes affect outcomes or diagnose what is missing or misaligned. Moreover, the black-box nature of LLMs can amplify the difficulty of

interpretation and calibration [16]. Therefore, when building LLM-based social simulation systems, we should evaluate whether they provide sufficient opportunities for user and developer reflection.

3.3.1 Process reflection. Process reflection asks whether building the simulation helps developers understand what matters for the target phenomenon. This perspective aligns with the philosophy of *Research through Design* (RtD), viewing the construction of a simulation not merely as an engineering task, but as a method of inquiry that generates knowledge through the act of making [31]. During authoring, developers often surface hidden assumptions and knowledge gaps: what must be specified, what is missing, and what cannot be expressed with the current design. For example, developers may realize that a key variable was underspecified (e.g., incentives, hierarchy, prior ties), that two rules produce indistinguishable behaviors, or that a phenomenon appears only after adding a neglected constraint (e.g., attention limits, communication cost). We can evaluate process reflection by (1) insight yield: what distinct gaps, assumptions, or design questions developers articulate, (2) rationale clarity: whether developers can explain what they changed and why, and (3) iteration quality: whether revisions reduce contradictions, better match intended behaviors, or productively shift the overall design concept of the simulation. These can be captured through simulation building logs, revision histories, think-aloud protocols, or post-hoc rationale reports, and assessed either qualitatively through thematic coding or quantitatively by counting distinct insights and comparing changes across iterations.

3.3.2 Outcome reflection. Outcome reflection asks whether users can interpret results and use them for downstream tasks. It focuses on whether users can explain *why* an outcome happened, which interactions drove it, and what trade-offs it implies. For example, users may trace polarization to an influence pathway, explain why a conflict escalated in one run but not another, or identify which parameter shift (e.g., policy, agent incentives) changed the result. Outcome reflection can be assessed via user studies measuring (1) causal traceability: can users reconstruct a plausible causal chain, (2) decision usefulness: do explanations improve decisions or confidence calibration, and (3) comparative reasoning: can users compare trajectories and link differences to controllable parameters. These can be operationalized through tasks in which users explain outcomes, compare alternative runs, or make downstream decisions based on simulation evidence, and evaluated based on explanation quality, decision quality, confidence calibration, or depth of review. In some cases, outcome reflection can also be assessed indirectly through the quality of the downstream artifact or decision, as long as the study design can attribute such gains to improved interpretation rather than raw system output.

4 Considerations in Evaluating Accuracy, Diversity and Reflection

The above dimensions extend existing simulation evaluation methods, but the tradeoffs or relationships between these dimensions have not previously been clearly articulated. In this section, we discuss a subset of these relationships.

4.1 Complementarity of Accuracy, Diversity, and Reflection

In terms of evaluating the overall quality of a simulation, each dimension is not mutually exclusive but rather may play a complementary role. If Accuracy is guaranteed but Diversity has not been verified, it becomes difficult to distinguish whether consistent results across multiple runs are due to *overfitting* to training data, or if different results are merely accidental errors rather than plausible outcomes. If the simulation system has been verified to possess the capability to generate sufficient Diversity, then when results repeatedly converge under specific conditions, we can more confidently trust that this convergence represents a stable outcome rather than overfitting. Conversely, a simulation where Diversity is guaranteed but Accuracy is unverified becomes a *babbling machine* that generates noise. As previously mentioned, this fails to achieve *Plausible Diversity* and degrades the reliability of the simulation itself. Even if Accuracy and Diversity are excellent, if nothing can be gained from the implementation process or result interpretation, the simulation may not serve users' actual needs.

4.2 Purpose-driven Evaluation Design

Due to the unpredictable nature of humans, finding a "one-size-fits-all" evaluation method is challenging. Instead, users must determine which dimensions to focus on and how to measure them, based on the purpose of the simulation and the specific domain. Not every simulation system needs to be perfect across all three dimensions and all levels (micro/meso/macro). For example, assume a scenario of creating rules to prevent the spread of an infectious disease. Here, the user's goal is to observe possible behaviors and outcomes according to changes in rules, and ultimately to formulate the final rules. Since the transmission routes of infectious diseases are scientifically proven, at the meso level, the impact of interactions between agents can be narrowed down to whether the virus is transmitted. Therefore, on the meso level, it is necessary to conduct an evaluation focused on *Accuracy*, specifically, how closely this reproduces scientific facts. However, the expectations at the micro level are different. Since humans often behave in ways we do not expect, it may be necessary to conduct an evaluation focused more on *Diversity* on the micro level, i.e., whether each agent has the ability to show various unexpected behaviors. Furthermore, features such as visualization are needed to show each agent's actions and interactions with others and to allow users to backtrack to understand how the rules they set influenced the final result. Correspondingly, if the simulation development process focuses significantly on *Reflection*, it can help users identify what they had not thought of while testing the system, or discover and complement loopholes in the rules. Therefore, in a research project implementing a simulation system, it is important to conduct appropriate evaluations for the components of each level according to the overall needs of the research. By configuring and reporting these three dimensions according to the purpose and domain, we can validate whether the simulation system is aligned with the intended use and more persuasively argue for operational validity.

5 Conclusion

We argued that validating LLM-based social simulations solely by predictive accuracy and realism is insufficient because simulators are often used for purposes beyond prediction, such as bounding plausible futures and supporting mechanistic understanding. To address this goal-validation mismatch, we introduced a purpose-driven framework with three complementary dimensions—Accuracy, Diversity, and Reflection. These dimensions specify what to match and at which level (micro/meso/macro), what range of plausible behaviors and outcomes to cover, and how to support users through process/outcome reflection. Ultimately, these three dimensions are complementary, and tailoring evaluation to align with a user's purpose and domain provides a stronger basis for arguing the simulation's operational validity.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 8–17.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Robert Axelrod. 1997. Advancing the art of simulation in the social sciences. In *Simulating social phenomena*. Springer, 21–40.

[4] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*. 3326–3346.

[5] Giordano De Marzo, Luciano Pietronero, and David Garcia. 2023. Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619* (2023).

[6] Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. SimulatorArena: Are User Simulators Reliable Proxies for Multi-Turn Evaluation of AI Assistants?. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 35200–35278.

[7] Joshua M Epstein. 2008. Why model? *Journal of artificial societies and social simulation* 11, 4 (2008), 12.

[8] Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. 2024. Agent-based modelling meets generative ai in social network simulations. In *International Conference on Advances in Social Networks Analysis and Mining*. Springer, 155–170.

[9] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984* (2023).

[10] Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. Large Language Model Driven Agents for Simulating Echo Chamber Formation. *arXiv preprint arXiv:2502.18138* (2025).

[11] James K He, Felix PS Wallis, Andrés Gvirtz, and Steve Rathje. 2024. Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology* (2024).

[12] John H Holland. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[13] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516* (2025).

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

[15] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[16] Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review* 59, 1 (2025), 15.

[17] Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498* (2024).

[18] Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025. UXAgent: A System for Simulating Usability Testing of Web Design with LLM Agents. *arXiv preprint arXiv:2504.09407* (2025).

[19] John H Miller and Scott E Page. 2009. *Complex adaptive systems: an introduction to computational models of social life: an introduction to computational models of social life*. Princeton university press.

[20] Xinyi Mou, Zhongyu Wei, and Xuan-Jing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*. 4789–4809.

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

[22] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[23] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems* 43, 2 (2025), 1–37.

[24] Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986* (2023).

[25] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems* 37 (2024), 15674–15729.

[26] Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, et al. 2025. Sentient Agent as a Judge: Evaluating Higher-Order Social Cognition in Large Language Models. *arXiv preprint arXiv:2505.02847* (2025).

[27] Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024. Speechagents: Human-communication simulation with multi-modal multi-agent systems. *arXiv preprint arXiv:2401.03945* (2024).

[28] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 10364–10379.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[30] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667* (2023).

[31] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.