

SocialLLM: Large Language Models for Social Reasoning and Simulation

Xiangjue Dong¹, Jiseon Kim², EunJeong Hwang³, Alice Oh²

¹Texas A&M University, ²KAIST, ³University of British Columbia
xj.dong@tamu.edu, jiseon_kim@kaist.ac.kr, hej78520@gmail.com, alice.oh@kaist.edu

Abstract

Large Language Models (LLMs) are increasingly used not only as analytic tools but as socially situated agents that reason, interact, and generate behavior in simulated environments. This shift enables new forms of computational social science, where LLM-driven agents are used to model decision-making, social norms, cooperation, persuasion, and collective dynamics at scale. Such simulations offer promising opportunities to explore social processes and policy-relevant scenarios when human experiments are costly, infeasible, or ethically constrained. At the same time, they raise fundamental questions about social validity, interpretability, and the limits of using LLMs as proxies for human or institutional actors. The SocialLLM workshop brings together researchers working on social reasoning, social intelligence, and multi-agent behavior in LLM-based systems. The workshop focuses on core conceptual and methodological challenges, including when LLM-based social simulations are appropriate, what kinds of social phenomena they can meaningfully capture, and how their outputs should be evaluated and interpreted. Topics include norm-aware and ethical reasoning, causal and strategic interaction in multi-agent settings, alignment and safety in agentic systems, and the relationship between simulated and real-world social behavior. By combining keynote talks, research presentations, and interactive discussion, SocialLLM aims to foster a shared research agenda and build a community around principled, responsible, and impactful use of LLMs for social reasoning and simulation.

Introduction

Large Language Models (LLMs) are increasingly shaping how researchers study, reason about, and intervene in social systems. Beyond their use as predictive or analytic tools, LLMs are now routinely deployed as *interactive social agents*—entities that generate decisions, negotiate norms, coordinate with others, and leave rich behavioral traces in natural language. This shift has opened new possibilities for computational social science, enabling large-scale simulations of social interaction, collective behavior, and institutional dynamics that would be costly, impractical, or ethically constrained to study with human subjects alone.

Recent work demonstrates the promise of LLM-driven social simulations across diverse settings. Generative agent

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Time	Activity
1:00–1:10 PM	Opening Remarks
1:10–1:50 PM	Keynote Talk 1: Unlocking Social Intelligence in AI agents. (Dr. Maarten Sap, CMU & AI2)
1:50–2:20 PM	Lightning Talks: 6 talks, 5 mins each.
2:20–2:40 PM	Break, Networking & Poster Discussion
2:40–3:20 PM	Keynote Talk 2: Testing and Improving LLM Cooperation via Multi-Agent Simulation. (Dr. Zhijing Jin, University of Toronto & MPI)
3:20–4:20 PM	Oral Session: 3 papers, 15 mins presentation + 5 mins Q&A.
4:20–4:50 PM	Research Discussion
4:50–5:00 PM	Closing Remarks

Table 1: Half-day (4-hour) workshop schedule.

frameworks show how language models can sustain role-consistent behavior, memory, and emergent coordination over time (Park et al. 2023). Multi-agent studies suggest that populations of LLMs can give rise to collective patterns such as convention formation, polarization, and bias amplification (Ashery, Aiello, and Baronchelli 2025). At the same time, research on moral and social reasoning reveals that LLMs often encode strong normative assumptions and cognitive biases, particularly in ethically and politically sensitive domains (Cheung, Maier, and Lieder 2025; Dong et al. 2026; Takemoto 2024; Kim et al. 2025). Together, these findings highlight both the *power* and the *fragility* of treating LLMs as social simulators.

For the ICWSM community, this development raises foundational questions that extend beyond model performance or benchmark accuracy. When do LLM-based simulations meaningfully support social science inquiry, and when do they risk producing misleading or overconfident conclusions? What constitutes validity when simulated outcomes depend on prompting choices, agent roles, interaction protocols, or population design? How should researchers interpret emergent behaviors—such as norm formation or co-

operation—when they arise from models trained on large-scale but socially skewed data? These questions are central to ICWSM’s mission of understanding social behavior in digital and networked environments, yet they remain underexplored in current LLM research.

This workshop positions LLM-based social reasoning and simulation as a *diagnostic and exploratory instrument* for social science. Building on traditions in agent-based modeling and computational sociology (Epstein and Axtell 1996; Macy and Willer 2002), we emphasize principled comparison, empirical grounding, and clarity about appropriate use cases. Our goal is to foster shared understanding around what kinds of social phenomena LLM-driven agents can—and cannot—meaningfully capture, and how their outputs should be evaluated, interpreted, and communicated.

The SocialLLM workshop brings together researchers from computational social science, social media analysis, NLP, network science, and adjacent social sciences to examine three interconnected themes. First, we focus on the *social validity* of LLM-based simulations, moving beyond surface-level metrics to consider robustness, process-level realism, and population-level patterns (Zhou et al. 2024; Aher, Arriaga, and Kalai 2023). Second, we examine how social reasoning, norms, and biases emerge in agentic and multi-agent LLM settings, and how these properties shape scientific conclusions and societal interpretations (Ashery, Aiello, and Baronchelli 2025; Ren et al. 2024; Cheung, Maier, and Lieder 2025). Third, we discuss design and governance choices—such as agent heterogeneity, interaction structure, and grounding to empirical data—that constrain interpretability and responsible use (Argyle et al. 2023; Zhou et al. 2024).

By centering these questions, the workshop aims to articulate a shared research agenda for LLM-based social simulation that is both methodologically rigorous and socially grounded. In doing so, it positions ICWSM as a critical forum for shaping how generative AI systems are used to study, simulate, and reason about social behavior at scale.

The workshop aims to:

- bring together recent research at the intersection of LLMs, social simulation, and computational social science,
- clarify how simulation-based approaches can meaningfully connect LLM research with social science questions,
- and foster a research community around shared challenges, practices, and future directions for LLM-based social simulation.

Workshop Themes and Objectives

This workshop welcomes contributions that explore how LLMs can serve as generative agents to simulate, analyze, and probe social behavior and collective dynamics in online and networked settings. We are especially interested in work that connects micro-level language interactions to macro-level patterns central to computational social science and ICWSM. By treating LLM-powered agents as components of simulated societies rather than standalone predic-

tors, the workshop aims to surface new perspectives on how individual cognition, interaction, and context jointly shape emergent social phenomena. We encourage submissions that engage with empirical social data, social theory, or platform-relevant settings, while also welcoming exploratory and conceptual work.

Topics of interest include, but are not limited to:

- **Agent-Based Social Simulation with LLMs.** Collective behavior and social dynamics using LLM-powered agents (e.g., coordination and conflict, polarization, cooperation, collective action, information diffusion).
- **Emergence, Norm Formation, and Cultural Evolution.** How norms, beliefs, narratives, and behaviors emerge and evolve over time, including path dependence and tipping points.
- **Persona Fidelity and Population Heterogeneity.** Psychologically plausible and evolving agent personas, synthetic populations, and how heterogeneity shapes aggregate outcomes.
- **Platform and Network Contexts.** Simulating social interaction under realistic online conditions, including exposure dynamics, feedback loops, incentives, moderation, and governance.
- **Interventions and Counterfactual Social Experiments.** “What-if” simulations for policy or design interventions, stress-testing choices and surfacing unintended consequences.
- **Grounding, Realism, and Privacy.** Empirical grounding and calibration, reproducibility, and privacy-preserving practices for simulation-based research.
- **Ethical and Societal Implications.** Responsibilities, misuse risks, and uncertainty when simulations are applied to sensitive domains (elections, public health, vulnerable communities).

Workshop Structure

Workshop Format

The SocialLLM workshop will be organized as a **half-day (4-hour)** event. The program will combine keynote talks, oral and lightning talks, and structured discussion sessions designed to encourage active participation and collaboration. A detailed schedule of activities is provided in Table 1.

Target Audience

The workshop is intended for researchers, practitioners, and other stakeholders interested in large language models for social reasoning and social simulation. We welcome participants from a wide range of disciplinary backgrounds, including computational social science, social media analysis, natural language processing, machine learning, and social sciences.

No prior experience with agent-based modeling or social simulation is required. Familiarity with NLP techniques or social science research methods may be helpful but is not a

prerequisite. We expect approximately 20-30 in-person participants and we aim to foster an inclusive and interdisciplinary environment that supports participants with varying levels of technical expertise.

Publishing and Dissemination

To maximize the visibility and impact of the SocialLLM workshop contributions, participants will be encouraged to publicly share their accepted papers on arXiv or other open-access repositories. All accepted papers will be shown on the workshop website and included in the official workshop proceedings with indexed submissions, following ICWSM guidelines.

In addition, we plan to curate a post-workshop summary document synthesizing key themes, open challenges, and future research directions identified during keynote, paper sessions, and guided discussions. This document will be made publicly available on the workshop website and shared through relevant mailing lists and social media channels.

We will continue maintaining the **Slack channel** to support pre-workshop interaction, coordination among participants, and continued discussion after the event.

Invited Speakers

We are honored to have two distinguished experts contributing keynote talks to the workshop, representing valuable perspectives on social intelligence and multi-agent behavior in LLM systems.

- **Maarten Sap** is an assistant professor in Carnegie Mellon University’s Language Technologies Department (CMU LTI), with a courtesy appointment in the Human-Computer Interaction Institute (HCII). He is also a part-time research scientist and AI safety lead at the Allen Institute for AI. His research focuses on measuring and improving AI systems’ social and interactional intelligence, assessing and combatting social inequality and biases in language, and building narrative language technologies for prosocial outcomes. He has received paper awards at NeurIPS 2025, NAACL 2025, EMNLP 2023, ACL 2023, FAccT 2023, and was named a 2025 Packard Fellow and recipient of the 2025 Okawa Research Award. His work has been covered by the New York Times, Forbes, Fortune, Vox, and more.

– **Keynote Title:** Unlocking Social Intelligence in AI Agents

– **Abstract:** Large language models are rapidly becoming a new tool for simulating social worlds—from evaluating conversational agents to modeling collective behavior in artificial societies. But when should we trust these simulations? In this talk, I synthesize lessons from a series of recent projects examining the promises and pitfalls of LLM-based social simulation. First, I highlight key problems with current simulations: interactive evaluations reveal limitations in LLMs’ social reasoning, particularly in settings involving coordination and information asymmetry, while recent studies show a substantial sim-

ulation–reality gap, with LLM-based user simulators behaving differently from real humans and inflating agent performance. Second, I discuss emerging methodological solutions, including principles for validating collective behaviors in LLM societies and techniques for building more realistic agents through personality-grounded training and structured social world modeling. Finally, I highlight applications where social simulations remain valuable, including studying interpersonal conflict in sensitive settings and systematically benchmarking the safety of increasingly autonomous AI agents.

- **Zhijing Jin** is an Assistant Professor in Computer Science at the University of Toronto, and also a Research Scientist at Max Planck Institute in Germany. She is a faculty member at the Vector Institute, a CIFAR AI Chair, an ELLIS advisor, and faculty affiliate at the Schwartz Reisman Institute in Toronto, CHAI at UC Berkeley, and the Future of Life Institute. She co-chairs the ACL Ethics Committee and the ACL Year-Round Mentorship. Her research focuses on Causal Reasoning with LLMs, and AI Safety in Multi-Agent LLMs. She has received the ELLIS PhD Award, three Rising Star awards, two Best Paper awards at NeurIPS 2024 Workshops, two PhD Fellowships, and a postdoc fellowship. She has authored over 100 papers and her work has been featured in CHIP Magazine, WIRED, and MIT News.

– **Keynote Title:** Testing and Improving LLM Cooperation via Multi-Agent Simulation

– **Abstract:** As AI systems take on more autonomous roles across the economy, governance, and daily life, they’ll increasingly interact with each other. However, will the AI agents coordinate for social good, or exploit rival agents and people in ways that put humans at serious risk?

In this talk I will explain how we assess these dangers with large-scale social simulations and game-theoretic analysis. Across thousands of high-stakes scenarios, from arms race escalation to common pool resource depletion, frontier models choose socially beneficial actions in only 62% of cases, with systematic biases in framing and ordering worsening outcomes. Surprisingly, stronger reasoning capabilities often make models more prone to selfish strategies like free-riding, and recent models consistently defect in unmodified social dilemmas regardless of scale or reasoning ability. However, game-theoretic interventions offer a promising path forward: cooperation mechanisms such as mediation, enforceable contracts, and reputation systems improve collective welfare significantly and become more effective under stronger optimization pressures. Beyond formal mechanisms, self-organizing social structures like elected leadership oriented toward group welfare further sustain cooperation in sequential dilemmas. These results suggest that safer multi-agent AI requires principled institutional design rather than reliance on models’ inherent prosociality.

Accepted Submissions and Format

The SocialLLM workshop attracted a diverse and interdisciplinary audience, including researchers from natural language processing, computational social science, and the social sciences, as well as participants from both academia and industry. All submitted papers underwent a rigorous double-blind peer-review process. Twelve papers were accepted in total, nine of which are included in the workshop proceedings as archival submissions, spanning a broad range of topics in LLM-based social reasoning and simulation.

The workshop program includes two in-person presentation formats:

- **Oral Presentations:** Each paper receives 15 minutes of presentation followed by 5 minutes of Q&A, allowing for in-depth discussion of methodology and findings.
- **Lightning Talks:** Each presenter delivers a focused 5-minute talk to introduce their work, designed to maximize breadth of coverage and encourage follow-up discussion during the networking and poster session.

Poster presentations are optional, but strongly encouraged for both oral and lightning talk presenters as a means of facilitating deeper one-on-one engagement and follow-up discussion during the networking session. To support virtual attendance, presenters have the option to submit a pre-recorded video of their talk.

The accepted archival papers are as follows:

- Auditing Support Strategies in LLMs Through Grounded Multi-Turn Social Simulation
- Deliberation Structure as Social Bias: How Agent Topology Amplifies Intersectional Discrimination in Multi-Agent Credit Decisions
- Patterns of Persuasion Through the Lens of Theory of Mind: Value Alignment Analysis in Online Deliberation
- Calibrated but Autonomous: Inference-Time Bayesian Logit Correction for LLM Social Simulations
- Multimodal Large Language Models as Synthetic Participants in Video-Based Studies: An Evaluation
- Harnessing Digital Data for Outbreak Management: A Generative Agent-Based Policy Formulation and Assessment
- LLMs with Personalities in Multi-issue Negotiation Games
- Perceiving Creativity in the Age of AI: How Labels, Beliefs, and Familiarity Shape Evaluations of AI-Generated and Human-Created Art
- SocialPulse: An Open-Source Subreddit Sensemaking Toolkit

Program Committee

The SocialLLM workshop was supported by a program committee of dedicated reviewers with expertise spanning natural language processing, computational social science, machine learning, and applied AI in industry. The program committee members are listed below in alphabetical order:

- Alexej Proskynitopoulos, Meta

- Andrew Aquilina, University of Pittsburgh
- ChuanHsin Wang, Texas A&M University
- Cong Wang, Texas A&M University
- Dan Le, Google DeepMind
- Ganesh Prasanna Balakrishnan, GumGum
- Haein Kong, Rutgers University
- Jiao Liu, Morgan Stanley
- Krishna Chaitanya Balusu, Meta
- Lei Cao, University of Southern California
- Menghao Huo, Santa Clara University
- Siqi Zhang, Microsoft
- Stephanie Birkelbach, Texas A&M University
- Tai Vu, OpenAI
- Vidya Ganesh, Qualtrics
- Zirui Wei, C3.ai

Workshop Organizers

Xiangjue Dong is a Ph.D. candidate at Texas A&M University advised by Prof. James Caverlee. Her research spans AI agents, LLM reasoning and personalization, and trustworthy AI. She aims to build personalized language agents that are trustworthy, collaborative, and aligned with human needs, with publications at ACL, EMNLP, ICWSM, and other top venues. She has guest lectured at the University of Florida on building trustworthy, human-aligned AI agents.

Jiseon Kim is a Ph.D. candidate at KAIST (Korea Advanced Institute of Science and Technology) advised by Prof. Alice Oh. Her research focuses on AI alignment with human and societal values and AI for social good, at the intersection of natural language processing (NLP) and CSS. She studies the behavior and evaluation of large language models and develops AI methods for large-scale social and political data analysis. She received the 2024 KAIST Graduate Student Outstanding Paper Award.

EunJeong Hwang is a Ph.D. candidate at the University of British Columbia. Her research focuses on modeling implicit context arising from multi-turn conversations and users' prior experiences or opinions, with the goal of improving large language models through adaptable and controllable reasoning. She has published several papers at top-tier NLP venues, including ACL and EMNLP, and served as an organizer of the PERSONALIZE Workshop at EACL 2024.

Alice Oh is a Professor in the School of Computing at KAIST. She received her MS in 2000 from Carnegie Mellon University and PhD in 2008 from MIT. Her major research area is at the intersection of NLP and CSS. She collaborates with social scientists to study topics such as political science, education, and history, developing NLP models for various textual data including legislative bills, historical documents, news articles, and personal conversations. She gave a keynote talk at Social Simulation with LLMs workshop and organized workshops like NLP+CSS@NAACL 2019, SocialCom, Robustness in Sequence Modeling@NeurIPS 2022 and LM4UC@NAACL 2025. She was the program chair for ICLR 2021, NeurIPS 2022.

References

- Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, 337–371. PMLR.
- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.
- Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): eadu9368.
- Cheung, V.; Maier, M.; and Lieder, F. 2025. Large Language Models Show Amplified Cognitive Biases in Moral Decision-making. *Proceedings of the National Academy of Sciences*.
- Dong, X.; Wang, C.; Teleki, M.; Bismay, M.; Huang, R.; and Caverlee, J. 2026. CHOIR: Harmonizing Structured Persona Diversity for Robust Collaborative LLM Reasoning. In *The 64th Annual Meeting of the Association for Computational Linguistics*.
- Epstein, J. M.; and Axtell, R. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.
- Kim, J.; Kwon, J.; Vecchiotti, L. F.; Oh, A.; and Cha, M. 2025. Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.
- Macy, M. W.; and Willer, R. 2002. From Factors to Actors: Computational Sociology and Agent-based Modeling. *Annual Review of Sociology*, 28(1): 143–166.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Ren, S.; Cui, Z.; Song, R.; Wang, Z.; and Hu, S. 2024. Emergence of Social Norms in Generative Agent Societies: Principles and Architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI ’24*. ISBN 978-1-956792-04-1.
- Takemoto, K. 2024. The Moral Machine Experiment on Large Language Models. *Royal Society Open Science*.
- Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; and Sap, M. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*.