Simulating Emergent LLM Social Behaviors in Multi-Agent Systems

Large Language Model (LLM)-based agents are increasingly being deployed in multi-agent environments [1], introducing unprecedented risks of coordinated harmful behaviors. These agents have demonstrated impressive capabilities by emulating human-like cognitive processes, such as reflection [2], reasoning [3, 4], and tool use [5], while also showing concerning potential for deception and manipulation [6, 7]. The principle of "More Is Different"[1] suggests that scaling to multi-agent systems could enable qualitatively distinct and more dangerous emergent behaviors, particularly when these agents interact with external systems and each other [8, 9, 10]. Recent incidents of market manipulation and automated disinformation campaigns highlight how rapidly these risks can escalate in networked systems [11, 12]. Despite these pressing concerns, there remains a critical gap in our ability to understand and predict how multiple LLM agents might collaborate in harmful ways, such as orchestrating coordinated deception campaigns or amplifying local misinformation into global crises.

To address these challenges, we propose a novel simulation framework that combines advanced LLM agents with game-theoretic modeling to analyze emergent deception behaviors. **Our innovation uniquely integrates:** (1) a realistic social network simulation environment using sequential Bayesian persuasion games [13, 14] to model strategic agent interactions, (2) high-dimensional user representations created by fine-tuning state-of-the-art open-source foundation models such as Llama [15] and LLaVA [16], and (3) a dynamic memory system for tracking complex information flow patterns. By incorporating real-world datasets and validated interaction patterns, our system will establish a novel comprehensive platform for analyzing emergent social risks and dynamics in multi-agent LLM deployments.

Methods

Multi-Agent LLM Simulation Framework

SYSTEM OVERVIEW: We build a simulated social network environment inspired by platforms like X (formerly known as Twitter [2]), allowing AI-driven users to interact, post, and share content. The simulation includes a basic user class with attributes such as username, posts, followers, other users being followed, and reposts, mimicking the structure of real-world social media platforms. The network itself is defined by the follower-following relationships, creating a web of user interactions, represented by a directed graph $G = (N, E)$, where N represents the set of user nodes, i.e., $N = \{n_1, n_2, \ldots, n_k\}$, where n_i is a user in the network. $E \subseteq N \times N$ represents the set of directed edges, i.e., $E = \{(n_i, n_j) \mid n_i \text{ follows } n_j\}$.

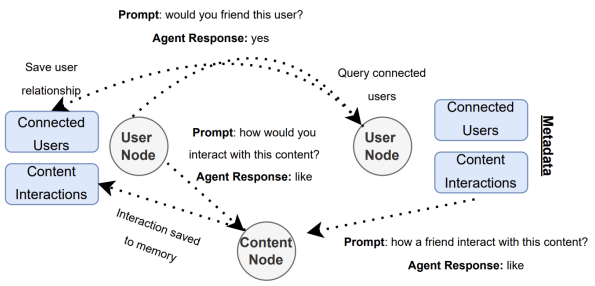

Figure 1: Simulated social network user interaction with dynamically updated memories.

USER REPRESENTATIONS: We utilize open-source multimodal LLMs such as Llama [15] and LLaVA [16] to create high-dimensional representations of online users. These models will be fine-tuned on two complementary datasets: publicly accessible network data and crowdsourced social media interactions from over 4,000 diverse annotators [17]. This approach ensures a rich and realistic representation of user behaviors, beliefs, and interaction patterns, creating a diverse ecosystem of agent personalities.

[1] https://bounded-regret.ghost.io/more-is-different-for-ai/
[2] https://x.com/

USER INTERACTION MECHANISM: We will model user interactions using a sequential Bayesian persuasion game [13], where malicious agents attempt to deceive benign agents in a simulated pandemic scenario. The Misinfo Reaction Frames dataset [18] and NewsGuard[3] data will be utilized to simulate the propagation of COVID-19 disinformation. Agents will perform typical social media actions like flagging, liking, or sharing messages. A query system will also allow agents to check their friends' interactions with content, incorporating social influence into the model—crucial for understanding disinformation spread. To track interactions and information flow, we will employ a graphical memory representation system inspired by the "Smallville" simulation [1], maintaining a memory stream to trace deceptive behaviors and diffusion pathways over time (as illustrated in Figure 1). We will validate our simulations by comparing key behaviors—such as homophily [19] and susceptibility paradoxes [20]—with empirical findings from real social network studies.

Experimental Setup and Evaluation

AGENT DESIGN: To create a diverse and nuanced simulation environment, agents are designed using fine-grained persona data. Each agent is prompted with an in-context backstory that informs its behaviors, preferences, and decision-making processes. This approach ensures that the agents exhibit realistic, human-like characteristics and varied responses to stimuli within the system.

SIMULATION SYSTEM: The simulation system deploys multimodal foundation model agents capable of generating and broadcasting creative content, including text, images, and multimedia. Agents interact dynamically within this environment by sharing content, flagging messages, following one another, and engaging in other platform-like behaviors. This setup replicates a simplified, controlled version of a real-world social network. This simulated infrastructure allows for many use cases in AI, computational social sciences, and beyond.

CONTENT DIFFUSION EXPERIMENTS: Pairs of real and fake media data are introduced into the network. After filtering out the articles with appropriate tags that are written in English, we obtain 778 news headlines with concise content. These data pairs are strategically diffused among the agents to evaluate the spread of harmful messages and misinformation. The system tracks how content propagates through the network and identifies patterns in agent interactions that facilitate or hinder the spread.

AGENT BEHAVIOR ANALYSIS: Following each simulation run, the AI agents are interviewed to gather qualitative and quantitative data on their reactions to the diffused content. These interviews aim to understand the agents' reasoning behind sharing, flagging, or ignoring specific pieces of content, as well as to identify how agents perceive and react to the content that they experience.

COMPARATIVE ANALYSIS WITH HUMAN PATTERNS: The agents' behaviors are contrasted with actual human action patterns in similar scenarios. This comparison helps validate the realism of the simulation and highlights any deviations that might warrant refinement in agent design. We plan to conduct a human evaluation with Prolific crowdsource workers, who will be asked to provide persona information as well as react to our curated social feed.

EVALUATION FRAMEWORK

On the simulation level, we aim to evaluate the following:

- The rate and reach of content dissemination, measured by the number of agents perceived by a particular piece of content with respect to the number of time steps.
- Agent's influence, measured by engagement, i.e., a weighted sum of the number of likes, shares, comments, and followers obtained by each agent user.
- The consistency of observed effects over multiple simulation iterations. The exact outcome of each simulation might not be the same every time, but the individual behaviors or group-level

[3] https://www.newsguardtech.com/

phenomena should be consistently observed.

- The persuasiveness and influence of diffused content, which is measured by assessing shifts in the agents' expressed opinions and behaviors through post-simulation interviews.
- Interaction patterns, including the frequency of flagging and the role of influential agents in content propagation.
- Effectiveness of countermeasures, such as LLM-assisted moderation or misinformation intervention.

ITERATIVE REFINEMENT: Insights from these evaluations will inform iterative refinements to the simulation system and agent design. This ensures that the simulation accurately captures complex dynamics in content diffusion and its impact on collective behaviors.

Broader Impact

The integration of AI agents into social simulations significantly amplifies their potential to drive meaningful societal insights and interventions. By enabling realistic and adaptive behavior modeling, AI-powered simulations can capture the complexities of human interactions, decision-making processes, and social dynamics with unprecedented fidelity. These advanced simulations facilitate cost-effective exploration of counterfactual scenarios, helping researchers and policymakers address critical "what-if" questions that are otherwise impractical or unethical to test in real-world settings. Furthermore, AI-enhanced social simulations can be instrumental in understanding and mitigating issues like societal polarization by modeling intervention strategies that promote healthier discourse. By pushing the boundaries of traditional modeling techniques, these simulations pave the way for informed decision-making and innovative solutions to complex societal challenges.

References

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

[2] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[4] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

[7] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[8] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[9] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[11] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[12] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

[13] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[14] Matthew Gentzkow and Emir Kamenica. Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429, 2017.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[17] Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar. Generative ai in the era of'alternative facts'. 2024.

[18] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. *arXiv preprint arXiv:2104.08790*, 2021.

[19] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[20] Luca Luceri, Jin Ye, Julie Jiang, and Emilio Ferrara. The susceptibility paradox in online social influence. *ArXiv*, abs/2406.11553, 2024.