Simulating Emergent LLM Social Behaviors in Multi-Agent Systems

Salman Rahman Department of Computer Science, UCLA

Large Language Model (LLM)-based agents are increasingly being deployed in multi-agent environments [1], introducing unprecedented risks of coordinated harmful behaviors. While individual LLMs have already demonstrated concerning capabilities for deception and manipulation [2, 3], the principle of "More Is Different"[1] suggests that scaling to multi-agent systems could enable qualitatively distinct and more dangerous emergent behaviors [4, 5, 6]. Recent incidents of market manipulation and automated disinformation campaigns highlight how rapidly these risks can escalate in networked systems [7, 8]. 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 [9, 10] to model strategic agent interactions, (2) high-dimensional user representations created by fine-tuning state-of-the-art open-source foundation models (Llama [11], LLaVA [12], Gemma [13]), 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.

Our implementation leverages a network of independently communicating LLM agents, fine-tuned on high-quality crowdsourced social media interactions for realistic behavior patterns [14]. Through this framework, we simulate three critical scenarios: the spread of health misinformation during pandemics [15], coordinated market manipulation leading to financial instability [8], and systematic efforts to undermine public trust in institutions [16]. Agents interact through social media actions like sharing, commenting, and flagging content. We track emergent behaviors through three novel systems: (1) an information diffusion tracker measuring content spread and influence using network reach metrics, (2) a pattern detector identifying coordinated manipulation strategies through our graphical memory system, and (3) a vulnerability assessment tool using graph-theoretic metrics to identify susceptible communities [17, 18]. By manipulating malicious agent distributions, we quantify how local deceptions escalate into network-wide crises and validate against real-world disinformation campaigns.

The impact of our research extends beyond academic contributions, addressing urgent real-world challenges in AI deployment safety. Our framework will provide the first systematic approach to understanding and preventing emergent risks in multi-agent LLM systems before they manifest in deployed systems. Immediate deliverables include: (1) an open-source simulation software that will enable researchers across social science, economics, psychology, and HCI to investigate emerging questions in their domains, while allowing companies to stress-test their multi-agent AI systems, (2) validated assessment tools for identifying network vulnerabilities and predicting cascade effects in interconnected AI systems, and (3) concrete intervention strategies for disrupting harmful information amplification before it reaches critical mass. In the long term, our research establishes a foundation for responsible scaling of multi-agent AI systems by providing both technical safeguards and empirically-validated deployment guidelines. As companies race to deploy increasingly powerful and multi-agent AI systems, our framework will serve as a crucial safety checkpoint, helping prevent scenarios where coordinated AI deception could trigger market instabilities or undermine social institutions.

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

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

[15] Michael A Gisondi, Rachel Barber, Jemery Samuel Faust, Ali Raja, Matthew C Strehlow, Lauren M Westafer, and Michael Gottlieb. A deadly infodemic: social media and the power of covid-19 misinformation, 2022.

[16] Ana Pérez-Escoda, Luis Miguel Pedrero-Esteban, Juana Rubio-Romero, and Carlos Jiménez-Narros. Fake news reaching young people on social networks: Distrust challenging media literacy. *Publications*, 9(2):24, 2021.

[17] 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.

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