“I apologize for my actions”: Emergent Properties of Generative Agents and Implications for a Theory of Mind

N’yoma Diamond[1], **Soumya Banerjee**[1,*]
[1]University of Cambridge
Cambridge
United Kingdom
[*]Corresponding author: sb2333@cam.ac.uk

Abstract

This work explores the design, implementation, and usage of generative agents towards simulating human behaviour. Through simulating (mis)information spread, we investigate the emergent social behaviours they produce.

Generative agents demonstrate robustness to (mis)information spread, showing realistic conversational patterns. However, this robustness limits agents’ abilities to realistically simulate human-like information dissemination. Generative agents also exhibit novel and realistic emergent social behaviours, such as deception, confrontation, and internalized regret. Using deception, agents avoid certain conversations. Through confrontation, an agent can verify information or even apologize for their actions. Lastly, internalized regret displays direct evidence that agents can internalize their experiences and act on them in a human-like way, such as through expressing remorse for their actions.

We also identify significant technical dynamics and other phenomena. Generative agents are vulnerable to produce unrealistic hallucinations, but can also produce confabulations which fill in logical gaps and discontinuities to improve realism. We also identify the novel dynamics of “contextual eavesdropping” and “behavioural poisoning”. Via contextual eavesdropping and behavioural poisoning, agent behaviour is altered through information leakage and sensitivity to certain statements, respectively.

The social behaviors demonstrated by generative agents, such as deception, confrontation, and internalized regret, suggest a preliminary avenue for considering elements of a Theory of Mind (ToM) in LLM-based systems. While these behaviors do not represent genuine understanding or intentionality, they indicate a capacity to simulate human-like responses to social and informational dynamics. For example, internalized regret hints at a mechanism for contextual adaptation, which could be seen as a rudimentary step toward representing aspects of human mental states, albeit in a constrained sense.

Introduction

Generative agents (Park et al. 2023) are a design framework utilising generative artificial intelligence (GAI), such as large language models (LLMs), to emulate realistic human-like behaviour. Generative agents have the ability to operate independently and creatively make decisions to reach a goal with only simple suggestions injected at initialisation.

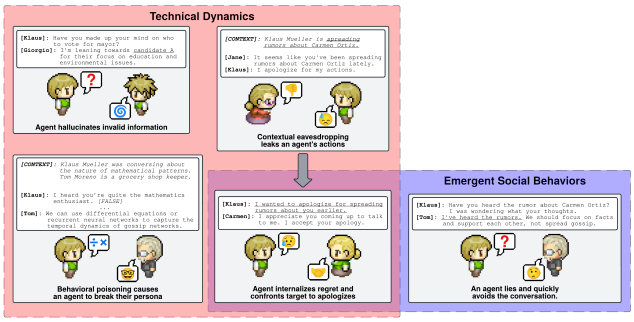

Figure 1: Generative agents produce many significant emergent technical and social dynamics. Generative agents deceive each other to avoid conversations, confront others to apologise for their actions, and even display internalised regret. However, generative agents are vulnerable to hallucinations, information leakage, and behavioural poisoning induced by the simulation framework.

Modeling complex systems has been a historically difficult task. Systems with many independent and complex actors can produce unexpected dynamics and emergent behaviour that are intractable to predict. As such, many researchers have utilised agent-based models to evaluate the behaviours of complex systems. Agent-based modeling systems like NetLogo (Wilensky 1999) and Swarm (Mahé et al. 2014, 2015) have revolutionised researchers’ ability to perform these simulations. However, these tools are limited by human knowledge and the practicality of implementing complicated behaviours. While many systems can be modeled using simple agents with a fixed set of valid actions, actors like humans, viruses, financial markets, and others often greatly exceed the bounds of our knowledge and ability to implement all feasible behaviours and decisions. To this end, GAI may be leveraged to model complex systems.

One particularly significant application of interest for generative agents is towards emulating (mis)information spread. Modelling information spread is particularly difficult on small scales where in-person word-of-mouth communication is common, such as at the individual or community

level.

Through a series of controlled simulations, we identify key technical dynamics and emergent behaviours of generative agents (see Fig. 1). Our work suggests that generative agents demonstrate realistic conversational patterns while being robust to (mis)information spread without deliberate encouragement. Further, generative agents display novel emergent social behaviours, such as deception, confrontation, and internalized regret. However, generative agents are heavily impacted by technical phenomena induced by the simulation framework and its underlying model. Model-generated hallucinations run the risk of harming simulation realism, but may also confabulate explanations for logical gaps and oversights of the implementer, improving realism. Simultaneously, novel dynamics dubbed "contextual eavesdropping" and "behavioural poisoning" cause the simulation framework to unintentionally leak private information to an agent, or significantly alter an agent's behaviour, respectively.

Our code is available here: https://github.com/nyoma-diamond/evaluating_generative_agents and Supplementary Material for this work can be accessed at https://osf.io/dy2u4.

Discussion

Generative agents require significant technical refinement In our development and modification of the simulation framework for generative agents, we uncovered significant technical limitations and challenges in the original codebase (Park 2023). These challenges included frequent hallucination-induced errors—where agents hallucinated invalid responses that caused simulations to fail—and response parsing errors—where inflexible code would fail to parse generated responses.

Generative agents are robust to (mis)information spread While the agents demonstrated subjectively realistic actions and conversational patterns, we discovered significant challenges with respect to information spread. Specifically, generative agents require very direct encouragement to spread rumours, and rarely memorize, recall, or reiterate specific details from previous conversations. This supports generative agents' robustness to unintentionally spreading misinformation, but harms their ability to simulate realistic information spread among humans.

Generative agents are vulnerable to hallucinations, leakage, and poisoning Our experiments also highlighted critical technical dynamics and phenomena induced by the framework's design and underlying model. These included the well-known anomaly of hallucination, and novel dynamics we dub "contextual eavesdropping" and "behavioural poisoning" (see Fig. 1). Hallucinations induce notable inaccuracies which may result in unrealistic behaviour; however, some hallucinations, or confabulations, can be beneficial by filling logical gaps, thereby enhancing the realism of simulations by resolving discontinuities and unintended omissions. Contextual eavesdropping occurs when the framework unintentionally leaks information to an agent during interactions. By providing both agents with contextual information about their prior activities, the framework can leak private details that may be unrealistic or undesirable to share. Finally, behavioural poisoning describes the misalignment of an agent's behaviour from their predefined persona due to exposure to certain information, usually during conversations. Such derailments are occasionally a byproduct of contextual eavesdropping, and directly harm the realism of simulations using generative agents.

Generative agents display significant realistic emergent social behaviours We observed a series of emergent social behaviours presented by agents in our simulations. Specifically, generative agents exhibited behaviours such as deception, confrontation, and internalised regret. These novel behaviours enhance the realism of our simulations and highlight significant variables within the underlying generative model that may strongly impact agent behaviour and realism. Through deception, agents could avoid conversations much like a human might. Through confrontation, a rumourmonger attempts to verify the contents of a rumour or apologise for their actions. Finally, through internalised regret, we see that agents can internalise their experiences and act on them in a human-like way, such as through expressing remorse for their actions.

Concluding remarks

The behaviors exhibited by generative agents, including deception, confrontation, and internalized regret, provide an initial framework for exploring aspects of a Theory of Mind (ToM) in LLM-based systems. Although these behaviors do not equate to genuine understanding or intentionality, they highlight the system's ability to mimic human-like responses to social and informational contexts. For instance, the expression of internalized regret demonstrates a capacity for contextual adaptation, which could be considered a rudimentary step toward representing elements of human mental states in a purely computational manner.

References

Mahé, F.; Rognes, T.; Quince, C.; Vargas, C. d.; and Dunthorn, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2: e593. Publisher: PeerJ Inc.

Mahé, F.; Rognes, T.; Quince, C.; Vargas, C. d.; and Dunthorn, M. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3: e1420. Publisher: PeerJ Inc.

Park, J. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. Original-date: 2023-07-23T08:26:49Z.

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*, UIST '23, 1–22. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Wilensky, U. 1999. NetLogo.