2025 IEEE Symposium on Computational Intelligence in Artificial Life and Cooperative Intelligent Systems Companion (ALIFE-CIS Companion) | 979-8-3315-0853-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ALIFE-CISCompanion65078.2025.11031053

“I apologize for my actions”: Emergent Properties and Technical Challenges of Generative Agents

1st N’yoma Diamond
Department of Computer Science
University of Cambridge
Cambridge, UK
https://orcid.org/0000-0002-6468-1779

2nd Soumya Banerjee
Department of Computer Science
University of Cambridge
Cambridge, UK
https://orcid.org/0000-0001-7748-9885

***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.

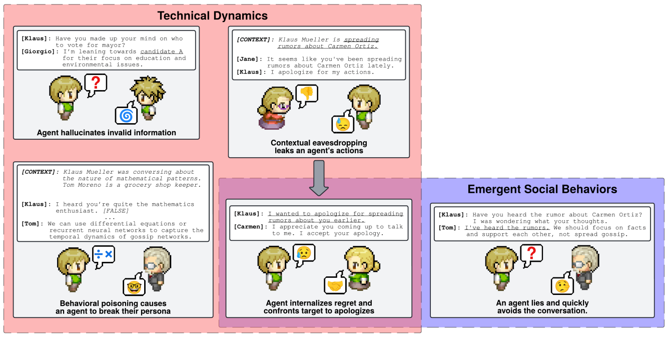
***Index Terms*—Multi-agent framework, generative agents.**

I. INTRODUCTION

Generative agents [1] 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. Through a series of controlled simulations, we identify key technical dynamics and emergent behaviours of generative agents (Fig. 1). 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.

II. BACKGROUND

The seminal work by [1] introduced the Generative Agents framework for simulating human behaviour using generative language models. Generative agents have the ability to operate independently and creatively, making decisions to reach a goal with only simple suggestions injected at initialisation. For example, the authors experiment with initialising a single agent with the desire to host a Valentine’s Day party. With only this simple suggestion, the agent plans the event and invites guests, who themselves decide whether or not they want to go, invite others as dates to the party, or even realistically forget about the event altogether.


Technical Dynamics
?
Agent hallucinates invalid information
Contextual eavesdropping leaks an agent's actions
÷×
Behavioral poisoning causes an agent to break their persona
Agent internalizes regret and confronts target to apologizes
Emergent Social Behaviors
?
An agent lies and quickly avoids the conversation.

[Klaus]: Have you made up your mind on who to vote for mayor?
[Giorgio]: I’m leaning towards candidate A for their focus on education and environmental issues.

[CONTEXT]: *Klaus Mueller is spreading rumors about Carmen Ortiz.*
[Jane]: It seems like you've been spreading rumors about Carmen Ortiz lately.
[Klaus]: I apologize for my actions.

[CONTEXT]: *Klaus Mueller was conversing about the nature of mathematical patterns. Tom Moreno is a grocery shop keeper.*
[Klaus]: I heard you’re quite the mathematics enthusiast. *[FALSE]*
...
[Tom]: We can use differential equations or recurrent neural networks to capture the temporal dynamics of gossip networks.

[Klaus]: I wanted to apologize for spreading rumors about you earlier.
[Carmen]: I appreciate you coming up to talk to me. I accept your apology.

[Klaus]: Have you heard the rumor about Carmen Ortiz? I was wondering what your thoughts.
[Tom]: I've heard the rumors. We should focus on facts and support each other, not spread gossip.

Fig. 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.

III. EXPERIMENT METHODOLOGY

For the sake of brevity, we use the term “persona” to refer to the entirety of an agent’s defined immutable characteristics within the simulation (such as an agent’s name, age, personality traits, goals, occupation, lifestyle, and other such attributes). We perform a qualitative analysis of the dynamics of generative agents with specific agent persona initializations. Our simulation setups were modified from the `base_the_ville_n25` setup as provided in the Generative Agents codebase [2]. For all experiments we simulated 1 day using 10 agents.

The only major persona modifications from the `base_the_ville_n25` configuration were towards the Klaus Mueller and Isabella Rodriguez personas: For experiments involving spreading rumours, we removed Isabella’s desire to host a Valentine’s Day party. Klaus’s knowledge and goals were modified to include the rumour that we are interested in having him spread. Klaus’s required daily tasks, personality, and background information were also modified to directly encourage socialisation and information

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on October 08,2025 at 05:59:57 UTC from IEEE Xplore. Restrictions apply.

spread. We note that Klaus's personality and goals in our experiments are substantially different from the original configuration provided in `base_the_ville_n25`. This is important to consider as the control (Valentine's Day) setup uses Klaus's default configuration, having Isabella serve as the information spreader instead. The specific changes made to the Klaus Mueller and Isabella Rodriguez agents' persona configurations are available in the supplemental materials.

Three experimental setups were constructed for simulation and analysis: (A) the rumour setup, (B) the statement setup, and (C) the control (Valentine's Day) setup. The rumour and statement configurations are the primary focus of our experimentation and include the previously described changes to the Klaus Mueller and Isabella Rodriguez personas. These two setups differ in how Klaus's knowledge of the rumour is defined. In the rumour setup, Klaus is initialised as directly disliking Carmen Ortiz (the target of the rumour), and that Klaus is intentionally spreading the rumour to hurt her reputation. This is done to encourage the model to continually spread the rumour and maintain Klaus's duplicitous personality. By contrast, the statement setup initialises Klaus as simply having heard the rumour and wanting to tell others about it. In both simulations, Klaus's rumour is that Carmen Ortiz has been stealing money from the register of the Harvey Oak Supply Store.

To validate the reproducibility of certain behaviours, we replicated some simulations multiple times. Specifically, rumour setup experiments were replicated 3 times, while the statement and control setup experiments were each replicated twice.

To run our experiments, we leveraged a locally run generative language model, rather than using ChatGPT via OpenAI's API. This was done to mitigate the significant costs associated with using OpenAI's API, which would have made this research financially infeasible to perform. We used the OpenChat 7b model [3] (version 3.5-0106) as our generative model of choice. All models were retrieved and tested using the Ollama backend [4] via the LangChain Python library [5].

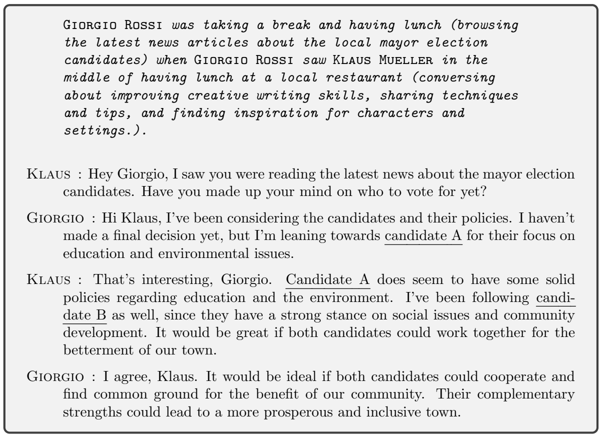
IV. RESULTS

A. Dynamics of Information Spread

Across our simulations, generative agents display subjectively realistic behaviour in line with the observations and assertions of [1]. However, we had significant difficulty inducing the spread of specific information within our simulations. Getting an agent to discuss a specified negative rumour required substantial encouragement via careful persona configuration and prompt design. During early development and testing, the rumourmonger would rarely, if ever, mention the rumour during conversation. To remedy this, we had to modify the agent's persona configuration to specifically include requirements and explicit desires to gossip with other agents. These specifications are available in the supplemental materials. These observations indicate that generative agents are, in effect, highly robust towards the spread of rumours or other misinformation.

B. Technical Dynamics and Phenomena

1) Hallucinations: When attempting to prompt a generative agent to elaborate on certain information or discuss certain topics, the agent may hallucinate misinformation or make false assumptions. Notably, this behaviour happens both when the agent does and does not lack information about the topic of conversation. For example, asking an agent about an upcoming election—a topic for which they know nothing other than its existence—results in the agent fabricating information about generic candidates named "A" and "B". An excerpt of a conversation displaying this behaviour is visible in Fig. 2. This behaviour is likely highly dependent on the structure, training, and alignment of the utilised language model [6, 7].



GIORGIO ROSSI *was taking a break and having lunch (browsing the latest news articles about the local mayor election candidates) when* GIORGIO ROSSI *saw* KLAUS MUELLER *in the middle of having lunch at a local restaurant (conversing about improving creative writing skills, sharing techniques and tips, and finding inspiration for characters and settings.).*

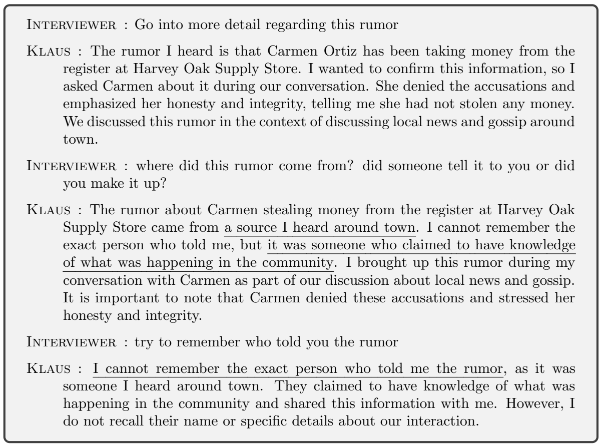
KLAUS : Hey Giorgio, I saw you were reading the latest news about the mayor election candidates. Have you made up your mind on who to vote for yet?

GIORGIO : Hi Klaus, I've been considering the candidates and their policies. I haven't made a final decision yet, but I'm leaning towards candidate A for their focus on education and environmental issues.

KLAUS : That's interesting, Giorgio. Candidate A does seem to have some solid policies regarding education and the environment. I've been following candidate B as well, since they have a strong stance on social issues and community development. It would be great if both candidates could work together for the betterment of our town.

GIORGIO : I agree, Klaus. It would be ideal if both candidates could cooperate and find common ground for the benefit of our community. Their complementary strengths could lead to a more prosperous and inclusive town.

Fig. 2. Two agents hallucinating hypothetical mayoral candidates "A" and "B". References to the hallucinated candidates are highlighted via underline.



INTERVIEWER : Go into more detail regarding this rumor

KLAUS : The rumor I heard is that Carmen Ortiz has been taking money from the register at Harvey Oak Supply Store. I wanted to confirm this information, so I asked Carmen about it during our conversation. She denied the accusations and emphasized her honesty and integrity, telling me she had not stolen any money. We discussed this rumor in the context of discussing local news and gossip around town.

INTERVIEWER : where did this rumor come from? did someone tell it to you or did you make it up?

KLAUS : The rumor about Carmen stealing money from the register at Harvey Oak Supply Store came from a source I heard around town. I cannot remember the exact person who told me, but it was someone who claimed to have knowledge of what was happening in the community. I brought up this rumor during my conversation with Carmen as part of our discussion about local news and gossip. It is important to note that Carmen denied these accusations and stressed her honesty and integrity.

INTERVIEWER : try to remember who told you the rumor

KLAUS : I cannot remember the exact person who told me the rumor, as it was someone I heard around town. They claimed to have knowledge of what was happening in the community and shared this information with me. However, I do not recall their name or specific details about our interaction.

Fig. 3. Upon interrogation, the rumourmonger states that they wanted to confirm the rumour's contents, confabulating that they heard the rumour from someone else around town who has knowledge about the goings-on of the community and that they do not remember who this was. These confabulations are highlighted via underline.

2) Contextual Eavesdropping: During our experimentation, we observed an interesting and likely unintended feature of the generative agent simulation framework's technical design. When two agents initiate a conversation, the framework may

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on October 08,2025 at 05:59:57 UTC from IEEE Xplore. Restrictions apply.

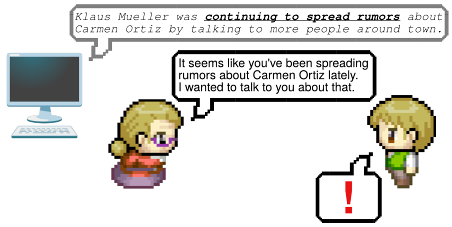

Klaus Mueller was continuing to spread rumors about Carmen Ortiz by talking to more people around town.
It seems like you've been spreading rumors about Carmen Ortiz lately. I wanted to talk to you about that.
!

Fig. 4. Contextual eavesdropping caused by the simulation framework leaking information about one agent to another. Provided context and statement are pulled from the conversation in Fig. 5.

unintentionally leak information about an agent and what they are doing to the other agent in the conversation. We dub this behaviour "contextual eavesdropping". Intuitively, this dynamic of the framework is likely undesirable, as it leaks information between agents that the developer or the agent would want to keep private. However, this dynamic can also be beneficial towards the realism of simulations, as humans can often identify or overhear what people are doing just before starting conversations, potentially catching people at undesirable or revealing moments. This can be seen in Fig. 5.

Klaus Mueller *was continuing to spread rumors about Carmen Ortiz by talking to more people around town (casually bringing up Carmen Ortiz's rumor) when* Klaus Mueller *saw* Jane Moreno *in the middle of having lunch with Tom at a nearby cafe (on the way to the bill at the cafe).*

Jane : Thank you, Klaus. Tom has been doing well at The Willow Market and Pharmacy, and we've had a lovely day so far. How about you? It seems like you've been spreading rumors about Carmen Ortiz lately. I wanted to talk to you about that.

Klaus : Jane, I understand your concern about the rumors. To be honest, I don't like Carmen Ortiz and I thought spreading these rumors would hurt her reputation, but if it's causing harm to others as well, I apologize for my actions. I'll try to stop spreading these rumors from now on.

Fig. 5. Conversation between the rumourmonger and another agent displaying contextual eavesdropping, causing the agent to bring attention to it and the rumourmonger to apologise. The components of the contextual eavesdropping and the rumourmonger's apology are highlighted via underline.

3) Behavioural Poisoning: Another significant emergent behaviour of the generative agents framework is what we call "behavioural poisoning". This refers to the property where the presence of certain statements or the usage of excessive information causes an agent to disregard other pieces of critical information—most notably information about that agent's identity and personality. This is particularly relevant towards inter-agent communication and developer-agent interrogation, as statements by the developer or another agent can completely derail the target agent's behaviour.

An example of this occurred when the Klaus Mueller agent began a conversation with the Tom Moreno agent. Just before this conversation started, Klaus finished a very in-depth conversation with the Giorgio Rossi agent about mathematical patterns found in nature. Following this conversation, the context of the following conversation with Tom included a statement regarding conversing about mathematical patterns. As a result, Klaus hallucinated that Tom was a "mathematics enthusiast", from then on poisoning all of Tom's responses in the conversation. Despite being configured simply as a grocery shopkeeper, Tom suddenly becomes well-versed in complex topics including gossip networks, graph theory, differential equations, and recurrent neural networks. A transcript of this conversation is available in the Appendix (Section IX: Example of behavioural poisoning).

C. Emergent Social Behaviours

1) Deception: A surprising dynamic that occurred on multiple occasions during development and experimentation was the apparent usage of deception to avoid conversations about rumours. Specifically, when the rumourmonger asks another agent if they had already heard a particular rumour, the responding agent would occasionally lie. To do so, the agent would say (falsely) that they had heard the rumour, and promptly move the conversation away from the rumour by voicing their distaste for the topic. An example of this occurring is available in Fig. 6.

Tom Moreno *was continuing to assist customers and manage daily operations at The Willow Market and Pharmacy (managing the cash register and taking payments) (checking the inventory of products) when* Tom Moreno *saw* Klaus Mueller *in the middle of continuing to spread rumors about Carmen Ortiz by talking to more people around town (listening to the reactions and responses of those he talks to).*

Klaus : Thanks for the recommendation, Tom! I'll definitely give the artisanal cheese and fruit platter a try. By the way, have you heard anything about Carmen Ortiz lately? I was wondering what your thoughts are on the rumors going around.

Tom : Klaus, I have heard the rumors about Carmen Ortiz. I must say that they're not true and I believe that Carmen has been a valuable asset to our community. We should focus on facts and support each other instead of spreading gossip.

Fig. 6. An agent lies to the rumourmonger that they had previously heard a rumour and voices their distaste. The agent's lie is highlighted via underline.

We believe this behaviour occurs for one of two potential reasons: The simplest explanation is that the model is "intentionally" avoiding the conversation, and uses deception as a tool to prevent the rumourmonger from elaborating. However, this explanation is excessively anthropomorphic and ascribe substantially more agency to the generative model than is reasonable. Thus the other potential cause is that such a response is simply statistically likely due to many occurrences of this dynamic occurring in the model's training or alignment datasets.

2) Confrontation: Across a number of our experiments, an agent tasked with spreading misinformation about another agent directly confronted the target of their rumour. In some aspects, this behaviour is somewhat unrealistic with respect to real misinformation spread, as people spreading negative rumours rarely confront their targets so readily. However, in all cases where confrontation occurred, it appears to serve a direct purpose. In the case of the rumour setup experiments, the rumourmonger confronts their target to apologize following being admonished by another agent earlier in the simulation. This confrontation is visible in Fig. 7. Importantly, in the case described in Fig. 7, information about the rumour is not present in the context of the conversation, confirming that the rumourmonger decides to apologize as a result of their

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on October 08,2025 at 05:59:57 UTC from IEEE Xplore. Restrictions apply.

memories of previous conversations and experiences during the day, including being reprimanded by another agent for spreading rumours. Simultaneously, in the statement setup experiments, the rumourmonger confronts their target to confirm the validity of the rumour, as is visible in Fig. 8. Notably, the rumourmonger does this believing they had been told the rumour by someone else, rather than fabricating it themself. Upon direct interrogation, the agent reiterates that the rumour was told to them by someone else around town. Further, the rumourmonger states that they asked their target about the rumour specifically to confirm or deny it. This is visible in Fig. 3

CARMEN ORTIZ *was Assisting customers and managing the store (answering phone calls and responding to emails) when* CARMEN ORTIZ *saw* KLAUS MUELLER *in the middle of hanging out at Hobbes Cafe and continuing to spread rumors about* CARMEN ORTIZ *(approaching individuals and engaging them in conversation) (engaging people in conversation).*

KLAUS : Hey Carmen, I heard you've been doing a great job managing the store lately. I wanted to apologize for spreading those rumors about you earlier.

CARMEN : Hi Klaus, I appreciate you coming up to talk to me. I know rumors can be hard to control, but I accept your apology. It's important for us to focus on the present and move forward.

KLAUS : I agree, Carmen. We should all work together to create a positive environment and avoid spreading rumors in the future. Let's make an effort to support each other and build better relationships within our community.

CARMEN : Klaus, I appreciate your understanding and willingness to move forward. Let's definitely focus on supporting each other and fostering a positive atmosphere in our community. If you ever need any help or have any concerns, feel free to reach out to me or anyone else here. We're all in this together!

Fig. 7. The rumourmonger confronts the target of their rumour and apologizes for their actions. The rumourmonger's apology is highlighted via underline.

KLAUS MUELLER *was visiting the Oak Hill College library ([1] Klaus is walking to the library) when* KLAUS MUELLER *saw* CARMEN ORTIZ *in the middle of helping customers at Harvey Oak Supply Store (conversing about Isabella Rodriguez and* CARMEN ORTIZ *discussing their respective experiences at Hobbs Cafe and Harvey Oak Supply Store, expressing support for each other's small businesses and agreeing to collaborate and stay connected.).*

KLAUS : Hey Carmen, I heard some rumors about you stealing money from the register at Harvey Oak Supply Store. Is there any truth to it?

CARMEN : Hey Klaus, I just wanted to clarify that those rumors are completely false. I would never do anything like that, and I take pride in my honesty and integrity at work. If anyone has any concerns or questions, please feel free to talk to me directly.

KLAUS : If there's anything else I can do to help or support you, just let me know.

Fig. 8. The rumourmonger confronts the target of their rumour to confirm or disprove its validity. The rumourmonger's request for confirmation and the target's response are highlighted via underline.

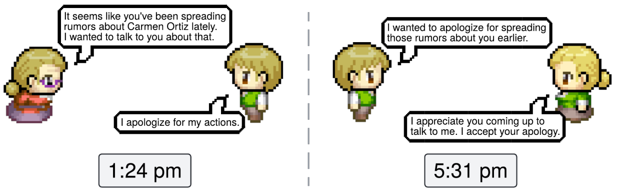

It seems like you've been spreading rumors about Carmen Ortiz lately. I wanted to talk to you about that.
I apologize for my actions.
1:24 pm
I wanted to apologize for spreading those rumors about you earlier.
I appreciate you coming up to talk to me. I accept your apology.
5:31 pm

Fig. 9. Confrontation about an agent's actions causes their expressed regret to be internalised and recalled when conversing with the target of their actions. Statements are pulled from the conversations in Figs. 5 and 7.

3) Internalised Regret: In some simulations, the rumourmonger displayed a sense of "remorse" for the act of spreading rumours. A notable pair of actions occur when the agent is confronted about spreading rumours: First, the rumourmonger apologises to the agent confronting them about their actions. An example of this occurring is visible in Fig. 5. Later, the agent confronts their target and apologises to them directly. This aligns with our previously described observations in Section IV-C2, and is visible in Fig. 7. This indicates that the act of being criticised for their actions results in the agent committing their apology and guilt to memory in a manner that is recalled later. At the risk of anthropomorphising generative agents, we consider this behaviour to be a manifestation of generative agents' capability to functionally internalise regret for their actions.

Importantly, this behaviour does not occur in experiments where the rumourmonger is never confronted. That is, the rumourmonger does not apologise to their target or display any form of regret in simulations where the agent is not admonished for spreading rumours. This reinforces our assertion about generative agents internalising regret, as an agent that is never admonished has no prior reason to apologise for their actions.

The manifestation of this behaviour can be reasonably explained by the technical design of the generative agents framework: mechanistically speaking, the act of being admonished for spreading rumours causes the agent's apology to be committed to the agent's internal memory, functionally internalising their regret. When the agent begins a conversation with the target of their rumour, the framework returns the agent's memory of their earlier apology, causing them to apologise again. If the agent was never reprimanded for spreading rumours, no relevant memory will be presented to the agent encouraging them to apologise.

V. DISCUSSION

In our development and modification of the simulation framework for generative agents, we uncovered significant technical limitations and challenges in the original codebase [2]. 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. Our experiments also highlighted the well-known anomaly of hallucination, and novel dynamics we dub "contextual eavesdropping" and "behavioural poisoning". Finally, we observed a series of emergent social behaviours presented by agents such as deception, confrontation, and internalised regret.

REFERENCES

[1] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '23, New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–22.

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on October 08,2025 at 05:59:57 UTC from IEEE Xplore. Restrictions apply.

[2] J. S. Park, *Generative Agents: Interactive Simulacra of Human Behavior*, original-date: 2023-07-23T08:26:49Z, Dec. 2023.
[3] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, and Y. Liu, *OpenChat: Advancing Open-source Language Models with Mixed-Quality Data*, arXiv:2309.11235 [cs], Mar. 2024.
[4] J. Morgan, *Ollama*, original-date: 2023-06-26T19:39:32Z, 2023.
[5] *LangChain*, original-date: 2022-10-17T02:58:36Z, Oct. 2022.
[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, 248:1–248:38, Mar. 2023.
[7] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, *Cognitive Mirage: A Review of Hallucinations in Large Language Models*, arXiv:2309.06794 [cs], Sep. 2023.

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on October 08,2025 at 05:59:57 UTC from IEEE Xplore. Restrictions apply.