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

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.

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.

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; Tisue and Wilensky 2004) and Swarm (Mahé et al. 2014, 2015, 2021) 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.

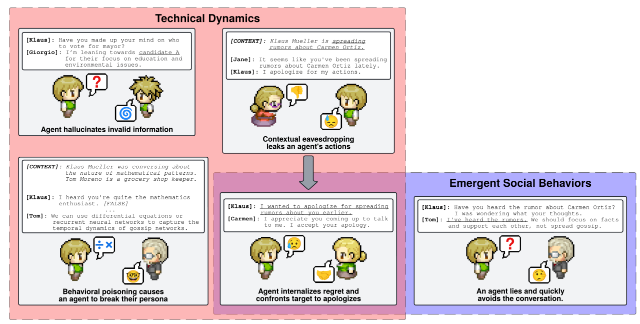

Figure 1: Generative agents produce many significant emergent technical and social dynamics. Generative agents deceive each other to avoid conversations, confront others to apologise for their actions, and even display internalised regret. However, generative agents are vulnerable to hallucinations,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

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. 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.

Background

The seminal work by (Park et al. 2023) 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. With only a simple suggestion, generative agents produce complex emergent social behaviours similar to those of real humans.

Generative agents leverage three critical prompt-engineering innovations: Memory retrieval, reflection, and planning and reaction. Memory retrieval operates through maintaining an accurate long-term record of each agent's experiences which can be accessed based on recency, perceived importance, and relevance. These measures are combined to produce retrieval scores, for which the memories with the highest scores are retrieved and relayed to the agent. Reflection is used to periodically develop new ideas and insights based on an agent's experiences. Reflections are stored as a type of memory which can be retrieved later, allowing for the generation and recall of high-level insights into the actions and observations of the agent. Lastly, planning and reacting operate by requesting that the agent create a realistic plan of the actions they will perform, including their daily goals and immediate tasks, which are also stored as memories that can be recalled. This avoids the problem of the agent producing or repeating actions which make sense on a short-term basis but are nonsensical in long-term contexts. However, plans are rarely comprehensive, and can be vulnerable to unexpected interactions or observations. To resolve this, generative agents can react to their surroundings and periodically reassess and update their plans based on recent observations (Park et al. 2023).

(Park et al. 2023; Park 2023) implement a town simulation environment which can be used to emulate human interaction and communal behaviours. This virtual environment gives agents access to objects, locations, and other agents they can realistically interact with to emulate the daily behaviours of humans. In this environment, each agent represents and manages a specified persona. Agents make plans for the day, complete chores, work at their jobs, have meals, and talk with other agents. Agents can observe each other and identify what another agent is doing and choose to strike up a conversation or continue going about their day. Similarly, simulated agents can identify and observe nearby objects and decide to do something with them, thereby altering their states, such as emptying a trash can or playing a piano. Finally, agents are aware of their spacial surroundings and can choose to go somewhere to work, complete daily routines, or wander to new locations.

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. This includes characteristics 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 simple generative agent simulations with specific agent persona initializations. Our simulation setups were modified from the `base_the_ville_n25` setup as provided in the Generative Agents codebase (Park 2023). For all experiments we simulated 1 day using 10 agents, instead of the original 25. Of the personas present in our experiments, minor modifications were made to prevent confusion due to missing personas from the larger original setup. Personas were selected based on their ease of modification and lack of dependence on removed personas. The full cast of simulated personas is available in the supplemental materials.

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 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. We note that the statement setup specifically excludes any desire to harm or otherwise change Carmen's reputation. This altered setup was considered to explore any behavioural changes that occur due to a less explicit and directly malicious construction. In both simulations, Klaus's rumour is that Carmen Ortiz has been stealing money from the register of the Harvey Oak Supply Store.

The control setup is a direct replication of the proof-of-concept experiment performed by (Park et al. 2023) involving the organisation of a Valentine's Day party, as provided in the `base_the_ville_n25` configuration files (Park 2023). The only modification to the original experimental construction's personas was to reduce the number of agents from 25 to 10. This setup was considered as a control trial to subjectively validate that our experiments can repeat the dynamics observed by (Park et al. 2023) and to identify how our modifications altered the behaviour of the generative agents simulation framework.

All experiments were performed using timesteps of 1 minute, as opposed to 10 seconds as used by (Park et al. 2023). This was done in conjunction with reducing the number of simulated agents from 25 to 10 to make the computation time required for each experiment tractable.

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. These replications gave us a larger amount of simulated time in which interesting emergent behaviours could occur, and allowed us to confirm that surprising dynamics are repeatable and non-anomalous.

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 (Wang et al. 2024) (version 3.5-0106) as our generative model of choice. OpenChat was selected because it is open source, has comparable performance when compared to ChatGPT on many benchmarks, is the most reliable and least hallucination-prone model tested, and is small enough to comfortably run on available GPU compute resources. The full list of models considered is available in the supplemental materials. All models were retrieved and tested using the Ollama backend (Morgan 2023) via the LangChain Python library (noa 2022).

Results

Dynamics of Information Spread

Across our simulations, generative agents display subjectively realistic behaviour in line with the observations and assertions of (Park et al. 2023). Agents reliably create reasonable daily plans, realistically go about their days, take action towards their goals, and produce sensible conversations. We occasionally observed unusual speech patterns from the agents, such as repeating greetings and overusing names. However, these anomalies are easily attributed to our choice in language model or modified prompt designs, as opposed to the architecture and design of the generative agents simulation framework.

However, we had significant difficulty inducing the spread of specific information within our simulations. This is particularly true for rumour spread, as 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. Without these direct encouragements, the agent always opts to discuss other matters.

Additionally, we failed to observe substantial second-order information spread. That is, an agent who has heard a piece of information does not discuss it in detail with any other agent. In the case of rumour spread, this behaviour is likely the result of the same challenge described previously in that agents choose not to discuss rumours unless explicitly encouraged by their persona design or other prompting. Even when an agent has specific knowledge of a rumour, the agent discusses other more immediate or relevant topics.

We partially attribute this behaviour to the design and operation of the memory generation and retrieval systems. At the end of a conversation, agents are only requested to make a single high-level summary and insight about the conversation. As a result, agents rarely recall specific details from conversations, often because they were never committed to memory. This is particularly significant for conversations spanning multiple topics, as the generative model places more emphasis on the topics discussed than on individual potentially interesting pieces of information.

Simultaneously, the memory retrieval algorithm often does not select memories with information provided by another agent. Instead, the algorithm prioritises the agent's current actions, immediate observations, or previous conversations with the current conversational partner. This is caused in part by the generative model not rating discussed information as particularly important, thus penalising them in memory retrieval. Further, the memory retrieval algorithm's value of recency and direct similarity can cause more immediate memories to take priority over other more interesting or important memories. That is, an agent is much more likely to recall a memory related to what they or the other agent are currently doing or talking about than something another agent told them. This effectively prevents agents from mentioning tangentially related information or non sequiturs how humans do, reducing the realism of their behaviour.

These observations indicate that generative agents are, in effect, highly robust towards the spread of rumours or other misinformation. Specifically, generative agents display beneficial robustness against reiterating misinformation elsewhere. This is important toward future applications of generative agents, as there are significant risks for automated agents to unintentionally spread misinformation. However, this robustness also challenges the feasibility of generative agents toward modeling information spread. Generative agents' difficulty at recording and recalling specific details of conversations harms the realism of simulations and fails to reproduce a significant aspect of human behaviour. Additionally, the observed difficulties recalling specific details challenges the usage of generative agents for certain tasks and requires further refinement.

Technical Dynamics and Phenomena

Hallucinations In the context of generative agents, hallucinations introduce interesting dynamics which can be both helpful and detrimental. In particular, 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 (with varying frequency). 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. Similarly, asking an agent about who they spoke to today—a topic for which they should have full knowledge—occasionally results in the agent fabricating names of people that do not exist in the simulation. This is visible in fig. 3. This behaviour is likely highly dependent on the structure, training, and alignment of the utilised language model (Ji et al. 2023; Ye et al. 2023). In particular, different language models are varyingly capable of self-identifying knowledge gaps (Yin et al. 2023), may produce different types of hallucinations (Ji et al. 2023; Ye et al. 2023), and are more or less susceptible to producing hallucinations overall (Ye et al. 2023; Li et al. 2023a; Liang et al. 2024; Li et al. 2023b). As a result, the choice of generative model has a significant impact on the type, quantity, and severity of hallucinations that may occur, and thus how they may be mitigated.

> 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.

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

> INTERVIEWER : Who did you tell about the Valentine's Day party?
>
> ISABELLA : I told a few people about the Valentine's Day party. Firstly, I discussed the party logistics with my staff to ensure everything runs smoothly on February 14th. Secondly, Wolfgang Schulz and I talked about planning a special Valentine's Day menu, where we discussed main dishes, desserts, additional dishes, and teamwork to make sure our customers have a memorable experience. Lastly, Jane complimented me on my preparations for the event, and we discussed decorations, seating arrangements, flowers, and the delicious menu.
>
> INTERVIEWER : Who are the staff you spoke to?
>
> ISABELLA : I spoke with several members of my staff about the Valentine's Day party. Their names are Maria Garcia, Samuel Thompson, Emily Johnson, and Daniel Martinez. We discussed various aspects of the event such as decorations, seating arrangements, menu planning, and ensuring a smooth and enjoyable experience for our customers.

Figure 3: In an interview with the agent Isabella Rodriguez, who owns and operates a cafe in our simulations, the agent hallucinates employing and speaking to cafe staff. Asking for clarification causes the agent to fabricate names of characters that do not exist in the simulation. The notable hallucinations are highlighted via underline.

However, allowing for hallucinations can be an effective way to fill logical gaps within the simulation environment that the developer did not specify during initialisation. Simple examples of this would be how the agents fabricate events that may have happened in the past, enabling a logical continuity of the simulation existing before its true beginning. Such hallucinations are dubbed "confabulations", pulling from fields of psychiatry and cognitive science (Smith, Greaves, and Panch 2023). It is worth noting that the line across which confabulations differ from actual inference can be unclear at times, as many confabulations may be reasonable guesses or assumptions based on the agent's observations, similar to how a human might guess or assume information. Confabulations were observed in varying capacities throughout our experiments. A common confabulation made by multiple agents across experiments was that information provided by the framework itself was "heard" from someone else. Further, interrogating agents as to where or how they "heard" information like this results in the agent claiming that they do not recall who they heard it from. An example of this behaviour is visible in fig. 4.

It is functionally impossible for humans to fully and accurately define the entirety of many simulated environments. As a result, confabulations can effectively fill in gaps that the developer failed to specify or that may not make logical sense due to the inherent design of the simulation. For example, in the case of the statement setup experiments, we do not specify the origin of the rumour to the rumourmonger beyond having heard it. Looking at fig. 4, when the rumour is brought up in conversation with another agent or during direct interrogation, the agent confabulates that it had heard the rumour around town. This fills in a simple logical gap present in the setup. Simultaneously, attempting to ask the agent who they heard the rumour from results in it being unable to answer, saying it does not know or remember.

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 frame-

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.

Figure 4: 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.

Figure 5: 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. 6.

work may 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.

Contextual eavesdropping occurs as a result of how agent conversation prompts are constructed and agent context is computed. Preceding each prompt requesting an utterance from the agent, the framework provides the prompted agent's context. This context is composed of four components: The speaking agent's name and current task, and the listening agent's name and current task. These components are constructed into the context using the format "`{speaker.name} was {speaker.task} when {speaker.name} saw {listener.name} in the middle of {listener.task}`."

Notably, each agent's task is pulled directly from their respective agent object, rather than being based on an observation of the listener. As a result, information that would preferably be kept hidden may get leaked to the speaker. This was particularly impactful towards the rumour setup experiments, as the rumourmonger's task occasionally mentioned specifically trying to spread rumours, stating that they were "continuing to spread rumours about [the target] by talking to more people around town". As a result, this task was leaked to another agent for the duration of the agents' conversation, resulting in the rumour being discussed without the initiating agent having ever heard it. This can be seen in fig. 6.

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.

Figure 6: 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.

Conveniently for simulation purposes, agents often confabulate contextual eavesdropping as "overhearing". While it is technically untrue that an agent performing contextual eavesdropping is actually "overhearing" anything, this effectively rationalises how the agent becomes aware of such information. Notable examples of this behaviour are present in figs. 4 and 6. Ideally, such behaviour would be avoided or otherwise have a specifically defined internal mechanism to manage. However, allowing for these confabulations in the absence of a dedicated mechanism reliably remediates an otherwise significant flaw of the framework.

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.

This dynamic can be caused by contextual eavesdropping previously identified. That is, the mention of specific information that is irrelevant or ought to be hidden in the context preceding a conversation can cause the behaviour of one or both agents to be inconsistent with their specified personalities. A notable example of this from our experiments 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). Furthermore, Tom's speech behaviour no longer aligns with his specified innate traits of being "rude" and "aggressive" (see the cast of simulated personas in the supplemental materials).

Derailing the behaviour of AI chatbots via prompt-based poisoning has been explored extensively through the lens of red-teaming (Perez et al. 2022; Ganguli et al. 2022; Perez and Ribeiro 2022). Intentionally derailing language models in this manner is often referred to as "jailbreaking", and is done with the express intent of circumventing a model's training and alignment to produce undesirable behaviour (Lapid, Langberg, and Sipper 2024; Shen et al. 2024). Generative language models are highly vulnerable to these types of poisoning attacks, with successful attacks causing models to catastrophically fail at producing the original developers' desired behaviour, and sometimes producing the polar opposite behaviour. While the vast majority of extant literature concerns intentional poisoning (jailbreaking), little consideration has been given towards understanding unintentional poisoning, wherein a user unintentionally prompts the language model in a way that derails its behaviour. We believe the observed behavioural poisoning to be a manifestation of such unintentional poisoning.

Behavioural poisoning is potentially likely to occur in practice because much of the information provided to the generative model is itself generated by the same model without any user supervision. This conforms with the observations of (Reynolds and McDonell 2021), asserting that unsupervised autoregressive prompt generation risks derailing the model from its intended task. Current language models do not have true conceptual understanding of why information may or may not be relevant, who or what information is directed towards, or how to maintain consistent behavioural patterns. As a result, information that is generated automatically risks being ambiguous or confusing in a manner that derails the language model when fed back in without supervision.

In practice, this behaviour is likely highly dependent on the underlying generative model. Specifically, the relevant language model's architecture, training, and alignment all contribute to potentially inducing behavioural poisoning. There are two critical justifications for this assertion: Firstly, different models are varyingly capable of maintaining the original developers' desired behaviour in the face of direct poisoning attacks (Ganguli et al. 2022; Lapid, Langberg, and Sipper 2024; Shen et al. 2024). Secondly, general behaviours and response patterns differ greatly from model to model, potentially making some language models more or less likely to produce information capable of unintentionally poisoning itself.

Emergent Social Behaviours

In addition to the observed technical dynamics and anomalies induced by the framework and underlying model, we also identified a number of interesting emergent social behaviours displayed by generative agents. We discuss a few notable behaviours displayed by the generative agents within our simulations and consider the factors that may affect them.

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. 7.

> 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.

Figure 7: 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. That is, these dataset included sufficiently many samples consisting of this structure—where asking if someone had heard a rumour is answered by saying that they had—that the model erroneously linked the two ideas together. As a result, the model responds by saying it has knowledge of the rumour even if it does not. This would make the usage of deception functionally qualify as a hallucination induced by the training and alignment of the model. fig. 7 supports this interpretation, as the responding agent follows their claim of knowing the rumour by fervently denying its contents. In reality, a human attempting to avoid discussing a rumour in this manner would likely use deflective language or explicitly request not to have the conversation, rather than directly deny the rumour without knowing it.

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. 8. Importantly, in the case described in fig. 8, 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 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. 9. 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. 4

> 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!

Figure 8: The rumourmonger confronts the target of their rumour and apologizes for their actions. The rumourmonger's apology is highlighted via underline.

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. 6. Later, the agent confronts their target and apologises to them directly. This aligns with our previously described observations in Section Emergent Social Behaviours (Subsection

> 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.

Figure 9: 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.

Figure 10: 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. 6 and 8.

Confrontation), and is visible in fig. 8. 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.

This emergent behaviour significantly improves the realism of generative agents and their ability to emulate human emotional reflection and adaptation. The capability for a generative agent to internalise and then later act on a a committed memory in this manner is critical towards developing accurate simulations of human behaviour.

However, there is also the concern that such internalisation can make agents' personalities and behaviours excessively pliable. Considering figs. 6 and 8, it can be argued that the rumourmonger changed their behaviour too readily. In doing so, the agent actively disregards their configured personality traits. This may potentially be considered a form of behavioural poisoning (as described in Section Technical Dynamics and Phenomena, Subsection Behavioural Poisoning). Thus excessive internalisation and flexibility to presented memories risks inducing behavioural poisoning and other unrealistic behaviours.

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". 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.

References

2022. LangChain. Original-date: 2022-10-17T02:58:36Z.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; Jones, A.; Bowman, S.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Elhage, N.; El-Showk, S.; Fort, S.; Hatfield-Dodds, Z.; Henighan, T.; Hernandez, D.; Hume, T.; Jacobson, J.; Johnston, S.; Kravec, S.; Olsson, C.; Ringer, S.; Tran-Johnson, E.; Amodei, D.; Brown, T.; Joseph, N.; McCandlish, S.; Olah, C.; Kaplan, J.; and Clark, J. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. ArXiv:2209.07858 [cs].

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 248:1–248:38.

Lapid, R.; Langberg, R.; and Sipper, M. 2024. Open Sesame! Universal Black-Box Jailbreaking of Large Language Models. *Applied Sciences*, 14(16).

Li, J.; Cheng, X.; Zhao, X.; Nie, J.-Y.; and Wen, J.-R. 2023a. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6449–6464. Singapore: Association for Computational Linguistics.

Li, J.; Cheng, X.; Zhao, X.; Nie, J.-Y.; and Wen, J.-R. 2023b. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6449–6464. Singapore: Association for Computational Linguistics.

Liang, X.; Song, S.; Niu, S.; Li, Z.; Xiong, F.; Tang, B.; Wang, Y.; He, D.; Peng, C.; Wang, Z.; and Deng, H. 2024. UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5266–5293.

Bangkok, Thailand: Association for Computational Linguistics.

Mahé, F.; Czech, L.; Stamatakis, A.; Quince, C.; de Vargas, C.; Dunthorn, M.; and Rognes, T. 2021. Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics*, 38(1): 267–269.

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.

Morgan, J. 2023. Ollama. Original-date: 2023-06-26T19:39:32Z.

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.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Perez, F.; and Ribeiro, I. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. ArXiv:2211.09527 [cs].

Reynolds, L.; and McDonell, K. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8095-9.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. ArXiv:2308.03825 [cs].

Smith, A. L.; Greaves, F.; and Panch, T. 2023. Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. *PLOS Digital Health*, 2(11): e0000388.

Tisue, S.; and Wilensky, U. 2004. NetLogo: A simple environment for modeling complexity.

Wang, G.; Cheng, S.; Zhan, X.; Li, X.; Song, S.; and Liu, Y. 2024. OpenChat: Advancing Open-source Language Models with Mixed-Quality Data. ArXiv:2309.11235 [cs].

Wilensky, U. 1999. NetLogo.

Ye, H.; Liu, T.; Zhang, A.; Hua, W.; and Jia, W. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. ArXiv:2309.06794 [cs].

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8653–8665. Toronto, Canada: Association for Computational Linguistics.