

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

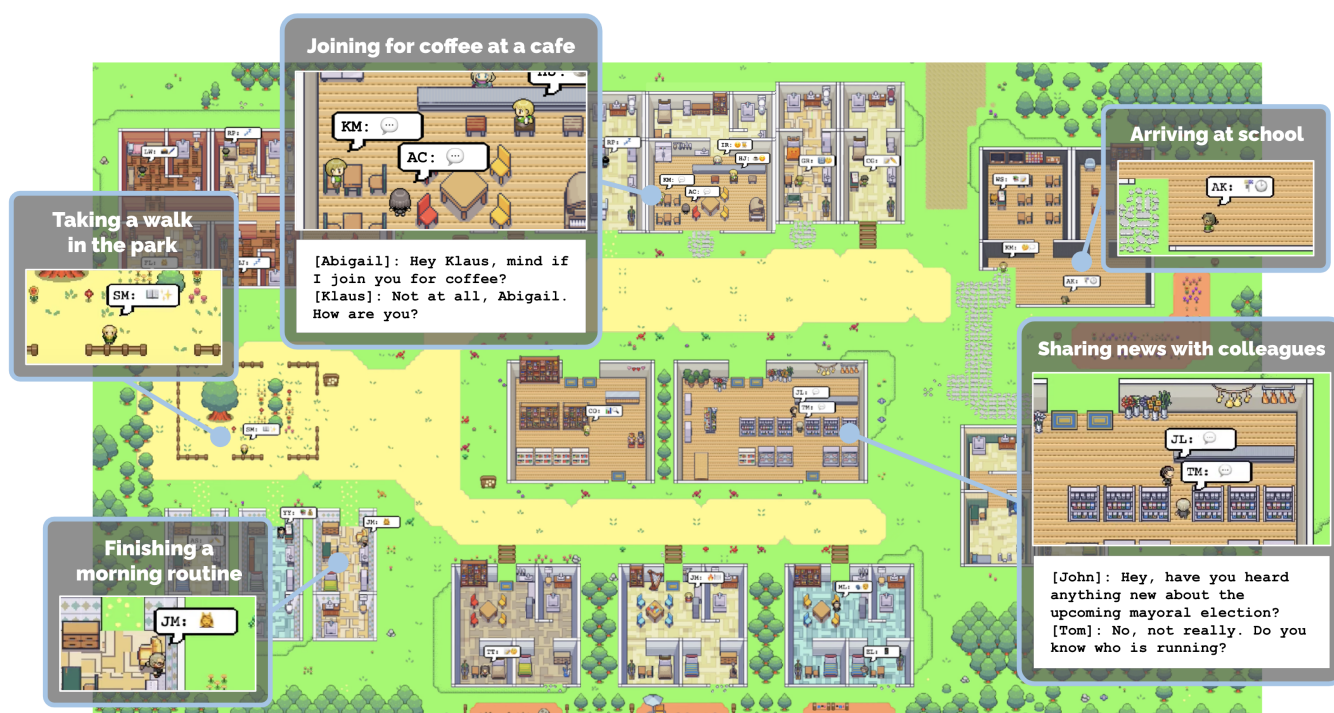


Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of *The Sims*, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

ABSTRACT

Believable proxies of human behavior can empower interactive applications ranging from immersive environments to rehearsal spaces for interpersonal communication to prototyping tools. In this paper, we introduce generative agents: computational software agents that simulate believable human behavior. Generative agents wake up, cook breakfast, and head to work; artists paint, while

authors write; they form opinions, notice each other, and initiate conversations; they remember and reflect on days past as they plan the next day. To enable generative agents, we describe an architecture that extends a large language model to store a complete record of the agent's experiences using natural language, synthesize those memories over time into higher-level reflections, and retrieve them dynamically to plan behavior. We instantiate generative agents to populate an interactive sandbox environment inspired by *The Sims*, where end users can interact with a small town of twenty-five agents using natural language. In an evaluation, these generative agents produce believable individual and emergent social behaviors. For example, starting with only a single user-specified notion that one agent wants to throw a Valentine's Day party, the agents autonomously spread invitations to the party over the next two

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0132-0/23/10.

<https://doi.org/10.1145/3586183.3606763>

days, make new acquaintances, ask each other out on dates to the party, and coordinate to show up for the party together at the right time. We demonstrate through ablation that the components of our agent architecture—observation, planning, and reflection—each contribute critically to the believability of agent behavior. By fusing large language models with computational interactive agents, this work introduces architectural and interaction patterns for enabling believable simulations of human behavior.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Human-AI interaction, agents, generative AI, large language models

ACM Reference Format:

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA*. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3586183.3606763>

1 INTRODUCTION

How might we craft an interactive artificial society that reflects believable human behavior? From sandbox games such as *The Sims* to applications such as cognitive models [23] and virtual environments [10, 59], for over four decades, researchers and practitioners have envisioned computational agents that can serve as believable proxies of human behavior. In these visions, computationally-powered agents act consistently with their past experiences and react believably to their environments. Such simulations of human behavior could populate virtual spaces and communities with realistic social phenomena [27, 80], train people on how to handle rare yet difficult interpersonal situations [44, 52, 94], test social science theories [12, 46], craft model human processors for theory and usability testing [23, 39, 51], power ubiquitous computing applications [31] and social robots [10, 14], and underpin non-playable game characters [59, 85] that can navigate complex human relationships in an open world.

However, the space of human behavior is vast and complex [85, 108]. Despite striking progress in large language models [18] that can simulate human behavior at a single time point [39, 80], fully general agents that ensure long-term coherence would be better suited by architectures that manage constantly-growing memories as new interactions, conflicts, and events arise and fade over time while handling cascading social dynamics that unfold between multiple agents. Success requires an approach that can *retrieve* relevant events and interactions over a long period, *reflect* on those memories to generalize and draw higher-level inferences, and apply that reasoning to create *plans and reactions* that make sense in the moment and in the longer-term arc of the agent's behavior.

In this paper, we introduce *generative agents*—agents that draw on generative models to simulate believable human behavior—and

demonstrate that they produce believable simulacra of both individual and emergent group behavior. Generative agents draw a wide variety of inferences about themselves, other agents, and their environment; they create daily plans that reflect their characteristics and experiences, act out those plans, react, and re-plan when appropriate; they respond when the end user changes their environment or commands them in natural language. For instance, generative agents turn off the stove when they see that their breakfast is burning, wait outside the bathroom if it is occupied, and stop to chat when they meet another agent they want to talk to.¹ A society full of generative agents is marked by emergent social dynamics where new relationships are formed, information diffuses, and coordination arises across agents.

To enable generative agents, we describe an agent architecture that stores, synthesizes, and applies relevant memories to generate believable behavior using a large language model. Our architecture comprises three main components. The first is the *memory stream*, a long-term memory module that records, in natural language, a comprehensive list of the agent's experiences. A memory retrieval model combines relevance, recency, and importance to surface the records needed to inform the agent's moment-to-moment behavior. The second is *reflection*, which synthesizes memories into higher-level inferences over time, enabling the agent to draw conclusions about itself and others to better guide its behavior. The third is *planning*, which translates those conclusions and the current environment into high-level action plans and then recursively into detailed behaviors for action and reaction. These reflections and plans are fed back into the memory stream to influence the agent's future behavior.

This architecture suggests applications in multiple domains, from role-play and social prototyping to virtual worlds and games. In social role-play scenarios (e.g., interview preparation), a user could safely rehearse difficult, conflict-laden conversations. When prototyping social platforms, a designer could go beyond temporary personas to prototype dynamic, complex interactions that unfold over time. For this paper, we focus on the ability to create a small, interactive society of agents inspired by games such as *The Sims*.² By connecting our architecture to the ChatGPT large language model [77], we manifest a society of twenty-five agents in a game environment. End users can observe and interact with these agents. If an end user or developer wanted the town to host an in-game Valentine's Day party, for example, traditional game environments would require scripting tens of characters' behavior manually. We demonstrate that, with generative agents, it is sufficient to simply tell one agent that she wants to throw a party. Despite many potential points of failure—the party planner must remember to invite other agents to the party, attendees must remember the invitation, those who remember must decide to actually show up, and more—our agents succeed. They spread the word about the party and then

¹When referring to generative agents engaging in actions or going to places, this is a shorthand for readability and not a suggestion that they are engaging in human-like agency. The behaviors of our agents, akin to animated Disney characters, aim to create a sense of believability, but they do not imply genuine agency.

²A demonstration of an actual simulation of the generative agent society can be viewed at the following link: https://reverie.herokuapp.com/UIST_Demo/. A public repository for the simulation code is located here: https://github.com/joonspk-research/generative_agents

show up, with one agent even asking another on a date to the party, all from a single user-generated seed suggestion.

We conducted two evaluations of generative agents: a controlled evaluation to test whether the agents produce believable individual behaviors in isolation, and an end-to-end evaluation where the agents interacted with each other in open-ended ways over two days of game time to understand their stability and emergent social behaviors. In the technical evaluation, we leverage a methodological opportunity to evaluate an agent’s knowledge and behavior by “interviewing” it in natural language to probe the agents’ ability to stay in character, remember, plan, react, and reflect accurately. We compared several ablations that limit agents’ access to memory, reflection, and planning. We observe that each of these components is critical to strong performance across these interview tasks. Across the technical and end-to-end evaluation, the most common errors arose when the agent failed to retrieve relevant memories, fabricated embellishments to the agent’s memory, or inherited overly formal speech or behavior from the language model.

In sum, this paper makes the following contributions:

- *Generative agents*, believable simulacra of human behavior that are dynamically conditioned on agents’ changing experiences and environment.
- A novel architecture that makes it possible for generative agents to remember, retrieve, reflect, interact with other agents, and plan through dynamically evolving circumstances. The architecture leverages the powerful prompting capabilities of large language models and supplements those capabilities to support longer-term agent coherence, the ability to manage dynamically evolving memory, and recursively produce higher-level reflections.
- Two evaluations, a controlled evaluation and an end-to-end evaluation, that establish causal effects of the importance of components of the architecture, as well as identify breakdowns arising from, e.g., improper memory retrieval.
- Discussion of the opportunities and ethical and societal risks of generative agents in interactive systems. We argue that these agents should be tuned to mitigate the risk of users forming parasocial relationships, logged to mitigate risks stemming from deepfakes and tailored persuasion, and applied in ways that complement rather than replace human stakeholders in design processes.

2 RELATED WORK

In this section, we reflect on the prior literature in human-AI interaction and situate, within its canon, the agenda of building believable proxies of human behavior. This agenda, once hailed as a north star in the interaction, game, and artificial intelligence communities [10, 59, 85, 86], has remained challenging due to the complexity of human behavior [17, 108]. We synthesize this research to suggest that large language models, though not sufficient by themselves, open up a new angle for creating believable agents when leveraged using the appropriate architecture.

2.1 Human-AI Interaction

Interactive artificial intelligence systems aim to combine human insights and capabilities in computational artifacts that can augment

their users [4, 30]. A long line of work has explored ways to enable users to interactively specify model behavior. For instance, Crayons demonstrated an early vision of interactive machine learning, allowing non-expert users to train classifiers [30]. Further work helped to articulate how end users might describe their classification goals to the system through examples [34] or demonstration [32]. Recent advancements have extended these explorations to deep learning [63] and prompt-based authoring [50, 67, 106].

Meanwhile, a persistent thread of research has advanced the case for language- and agent-based interaction in human-computer interaction. Formative work such as SHRDLU [103] and ELIZA [102] demonstrated the opportunities and the risks associated with natural language interaction with computing systems. As research progressed, it became evident that autonomous agents could offer new metaphors for delegation and interaction [68], but the boundaries of delegation between humans and agents have remained the subject of ongoing debate and refinement [47, 89, 90]. Recently, this technology has reached a level of stability that enables agents to interact via natural language in large and complex online social environments (e.g., [55]). Natural language interaction provides a novel modality that can enhance user abilities in domains such as photo editing [3, 35, 65] and code editing [88].

We convene these threads of work to show that we can now create agents that proxy human behavior for interactive systems, and interact with them using natural language. In doing so, this work reopens the door to examining foundational human-computer interaction questions around cognitive models such as GOMS and Keystroke-Level Model (KLM) [22, 23], around prototyping tools [80], and around ubiquitous computing applications [26, 31, 101].

2.2 Believable Proxies of Human Behavior

Prior literature has described *believability*, or *believable agents*, as a central design and engineering goal. Believable agents are designed to provide an illusion of life and present a facade of realism in the way they appear to make decisions and act on their own volition, similar to the characters in Disney movies [10, 96]. These agents can populate and perceive an open world environment like the one we inhabit [10, 59], and strive to behave in ways that exhibit emergent behaviors grounded in social interactions with users or other agents with the aim of becoming believable proxies of our behavior in hypothetical simulations of individuals and communities [20, 36, 71]. Historically, these agents were developed in the context of intelligent game non-player characters (NPCs) [59, 85]. Creating NPCs with believable behavior, if possible, could enhance player experiences in games and interactive fictions by enabling emergent narratives [8, 16, 49, 93] and social interactions with the agents [109]. However, more importantly, game worlds provide increasingly realistic representations of real-world affordances, and as observed by Laird and van Lent in 2001, these simulated worlds offer accessible testbeds for developers of believable agents to finess the agents’ cognitive capabilities without worrying about implementing robotics in the real world or creating simulation environments from scratch [59, 85].

A diverse set of approaches to creating believable agents emerged over the past four decades. In implementation, however, these approaches often simplified the environment or dimensions of agent

behavior to make the effort more manageable [17, 73]. Rule-based approaches, such as finite-state machines [91, 97] and behavior trees [41, 54, 82] account for the brute force approach of human-authoring the agent's behavior [71]. They provide a straightforward way of creating simple agents that is still the most dominant approach today [69, 74, 108], and can even handle rudimentary social interactions, as shown in games such as *Mass Effect* [13] and *The Sims* [7] series. Nonetheless, manually crafting behavior that can comprehensively address the breadth of possible interactions in an open world is untenable. This means that the resulting agent behaviors may not fully represent the consequences of their interactions [70–72], and cannot perform new procedures that were not hard-coded in their script [91, 97]. On the other hand, prevalent learning-based approaches for creating believable agents, such as reinforcement learning, have overcome the challenge of manual authoring by letting the agents learn their behavior, and have achieved superhuman performance in recent years in games such as *AlphaStar* for *Starcraft* [99] and *OpenAI Five* for *Dota 2* [11]. However, their success has largely taken place in adversarial games with readily definable rewards that a learning algorithm can optimize for. They have not yet addressed the challenge of creating believable agents in an open world [40, 74, 91].

Cognitive architectures in computation, pioneered by Newell, aimed to build the infrastructure for supporting a comprehensive set of cognitive functions [76] that suited the all-encompassing nature of believable agents held in its original vision. They fueled some of the earliest examples of believable agents. For instance, *Quakebot-SOAR* [60] and *ICARUS* [25, 64] generated NPCs in first-person shooter games, while *TacAir-SOAR* [81] generated pilots in aerial combat training simulations. The architectures used by these agents differed (*Quakebot-* and *TacAir-SOAR* relied on *SOAR* [61], while *ICARUS* relied on its own variation that was inspired by *SOAR* and *ACT-R* [6]), but they shared the same underlying principle [62]. They maintained short-term and long-term memories, filled these memories with symbolic structures, and operated in perceive-plan-act cycles, dynamically perceiving the environment and matching it with one of the manually crafted action procedures [58, 97]. Agents created using cognitive architectures aimed to be generalizable to most, if not all, open world contexts and exhibited robust behavior for their time. However, their space of action was limited to manually crafted procedural knowledge, and they did not offer a mechanism through which the agents could be inspired to seek new behavior. As such, these agents were deployed mostly in non-open world contexts such as first-person shooter games [25, 60] or blocks worlds [64].

Today, creating believable agents as described in its original definition remains an open problem [85, 108]. Many have moved on, arguing that although current approaches for creating believable agents might be cumbersome and limited, they are good enough to support existing gameplay and interactions [24, 75, 108]. Our argument is that large language models offer an opportunity to re-examine these questions, provided that we can craft an effective architecture to synthesize memories into believable behavior. We offer a step toward such an architecture in this paper.

2.3 Large Language Models and Human Behavior

Generative agents leverage a large language model to power their behavior. The key observation is that large language models encode a wide range of human behavior from their training data [15, 18]. If prompted with a narrowly defined context, the models can be used to generate believable behavior. Recent work has demonstrated the efficacy of this approach. For instance, *social simulacra* used a large language model to generate users that would populate new social computing systems to prototype their emergent social dynamics [80]. This approach used a prompt chain [105, 106] to generate short natural language descriptions of personas and their behaviors as they appear in the system being prototyped. Other empirical studies have replicated existing social science studies [46], political surveys [92], and generated synthetic data [39]. Large language models have also been used to generate interactive human behavior for users to engage with. In gaming, for instance, these models have been employed to create interactive fiction [37] and text adventure games [21]. With their ability to generate and decompose action sequences, large language models have also been used in planning robotics tasks [48]. For example, when presented with a task, such as picking up a bottle, the model is prompted to break down the task into smaller action sequences, such as heading to the table where the bottle is located and picking it up.

We posit that, based on the work summarized above, large language models can become a key ingredient for creating believable agents. The existing literature largely relies on what could be considered first-order templates that employ few-shot prompts [38, 66] or chain-of-thought prompts [100]. These templates are effective in generating behavior that is conditioned solely on the agent's current environment (e.g., how would a troll respond to a given post, what actions would a robot need to take to enter a room given that there is a door). However, believable agents require conditioning not only on their current environment but also on a vast amount of past experience, which is a poor fit (and as of today, impossible due to the underlying models' limited context window) using first-order prompting. Recent studies have attempted to go beyond first-order prompting by augmenting language models with a static knowledge base and an information retrieval scheme [53] or with a simple summarization scheme [104]. This paper extends these ideas to craft an agent architecture that handles retrieval where past experience is dynamically updated at each time step and mixed with agents' current context and plans, which may either reinforce or contradict each other.

3 GENERATIVE AGENT BEHAVIOR AND INTERACTION

To illustrate the affordances of generative agents, we instantiate them as characters in a simple sandbox world reminiscent of *The Sims* [7]. This sprite-based sandbox game world, *Smallville*, evokes a small town environment. In this section, we will walk through the affordances and interactions with generative agents in *Smallville* and describe how the agents behave within it. Then, in Section 4, we will introduce our generative agent architecture that powers these affordances and interactions. In Section 5, we will describe the

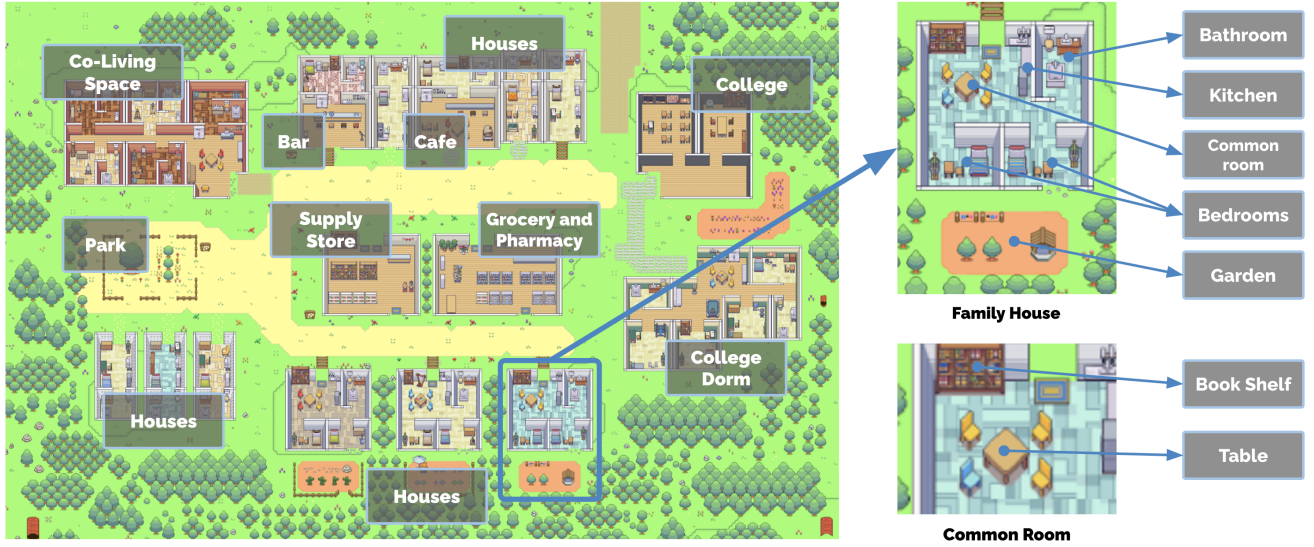


Figure 2: The Smallville sandbox world, with areas labeled. The root node describes the entire world, children describe areas (e.g., houses, cafe, stores), and leaf nodes describe objects (e.g., table, bookshelf). Agents remember a subgraph that reflects the parts of the world they have seen, maintaining the state of those parts as they observed them.

implementation of the sandbox environment and how the agents interact with the underlying engine of the sandbox world.

3.1 Agent Avatar and Communication

A community of 25 unique agents inhabits Smallville. Each agent is represented by a simple sprite avatar. We authored one paragraph of natural language description to depict each agent’s identity, including their occupation and relationship with other agents, as seed memories. For example, John Lin has the following description:

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well – the husband Tom Moreno and the wife Jane Moreno.

Each semicolon-delimited phrase is entered into the agent’s initial memory as memories at the start of the simulation.

3.1.1 Inter-Agent Communication. The agents interact with the world by their actions, and with each other through natural language. At each time step of the sandbox engine, the agents output a natural language statement describing their current action, such as “Isabella Rodriguez is writing in her journal”, “Isabella Rodriguez is checking her emails”, “Isabella Rodriguez is talking with her family on the phone”, or “Isabella Rodriguez is getting ready for bed.” This statement is then translated into concrete movements that affect the sandbox world. The action is displayed on the sandbox interface as a set of emojis, providing an abstract representation of the action from an overhead view. To achieve this, the system utilizes a language model to translate the action into a set of emojis, which appear above each avatar’s head in a speech bubble. For example, “Isabella Rodriguez is writing in her journal” is displayed as 📖✍️, while “Isabella Rodriguez is checking her emails” appears as 💻✉️. The complete natural language description of the action can be accessed by clicking on the agent’s avatar.

Agents communicate with each other in full natural language. They are aware of other agents in their local area, and the generative agent architecture determines whether they walk by or engage in conversation. Here, a sample in the middle of a conversation between the agents Isabella Rodriguez and Tom Moreno about the upcoming election:³

Isabella: I’m still weighing my options, but I’ve been discussing the election with Sam Moore. What are your thoughts on him?
Tom: To be honest, I don’t like Sam Moore. I think he’s out of touch with the community and doesn’t have our best interests at heart.

³We note that the conversational style of these agents can feel overly formal, likely a result of instruction tuning in the underlying models. We expect that the writing style will be better controllable in future language models.

3.1.2 User Controls. The user communicates with the agent through natural language by specifying a persona that the agent should perceive them as. For example, if the user specifies that they are a news “reporter” and asks about the upcoming election by saying, “Who is running for office?”, the John agent replies:

John: My friends Yuriko, Tom and I have been talking about the upcoming election and discussing the candidate Sam Moore. We have all agreed to vote for him because we like his platform.

To directly command one of the agents, the user takes on the persona of the agent’s “inner voice”—this makes the agent more likely to treat the statement as a directive. For instance, when told “You are going to run against Sam in the upcoming election” by a user as John’s inner voice, John decides to run in the election and shares his candidacy with his wife and son.

3.2 Environmental Interaction

Smallville features the common affordances of a small village, including a cafe, bar, park, school, dorm, houses, and stores. It also defines subareas and objects that make those spaces functional, such as a kitchen in a house and a stove in the kitchen (Figure 2). All spaces serving as agents’ primary living quarters feature a bed, desk, closet, shelf, as well as a bathroom and a kitchen.⁴

Agents move around Smallville as one would in a simple video game, entering and leaving buildings, navigating its map, and approaching other agents. Agent movements are directed by the generative agent architecture and the sandbox game engine: when the model dictates that the agent will move to a location, we calculate a walking path to the destination in the Smallville environment, and the agent begins moving. In addition, users can also enter the sandbox world of Smallville as an agent operating within it. The agent that the user embodies can be an agent already present in the world, such as Isabella and John, or it can be an outside visitor with no prior history in Smallville. The inhabitants of Smallville will treat the user-controlled agent no differently than they treat each other. They recognize its presence, initiate interactions, and remember its behavior before forming opinions about it.

Users and agents can influence the state of the objects in this world, much like in sandbox games such as *The Sims*. For example, a bed can be occupied when an agent is sleeping, and a refrigerator can be empty when an agent uses up the ingredients to make breakfast. End users can also reshape an agent’s environment in Smallville by rewriting the status of objects surrounding the agent in natural language. For instance, when Isabella is making breakfast in the morning, the user can change the status of the kitchen stove from “turned on” to “burning” by inputting a command to the system that chooses the object and illustrates its new status, like this: “<Isabella’s apartment: kitchen: stove> is burning.” Isabella will notice this in the next moment and go to turn off the stove and remake her breakfast. Likewise, if the user sets the status of Isabella’s shower to “leaking water” when she enters the bathroom, she will gather tools from her living room and try to fix the leak.

⁴This environment design is not the focus of our work, so we generated this environment manually, not automatically. Future work can continue to expand the richness of the agents’ environments.

3.3 Example “Day in the Life”

Starting from the single-paragraph description, generative agents begin planning their days. As time passes in the sandbox world, their behaviors evolve as these agents interact with each other and the world, building memories and relationships, and coordinating joint activities.

We demonstrate the behavior of generative agents by tracing the output of our system over the course of one day for the agent John Lin (Figure 3). In the Lin family, John is the first to wake up at 7 a.m. He brushes his teeth, takes a shower, gets dressed, eats breakfast, and checks the news at the dining table in the living room. At 8 a.m., Eddy follows suit, rushing out of bed to prepare for his classes. He catches up with John just as he is about to head out:

John: Good morning Eddy. Did you sleep well?
Eddy: Good morning dad. Yeah, I slept great.
John: That’s good. What are you working on today?
Eddy: I’m working on a new music composition for my class. It’s due this week, so I’m trying to get it finished. But I’m having so much fun with it!
John: That sounds great!

Soon after Eddy heads out, Mei wakes up and joins John. Mei asks about her son, and John recalls the conversation they just had:

Mei: Did Eddy already leave for school?
John: Yes, he just left. He’s working on a music composition for his class.
Mei: Oh, that’s wonderful! Tell me about it.
John: I think he’s really enjoying it! He said he’s having a lot of fun with it.
Mei: That’s great! I’m so proud of him.

As they finish their conversation, Mei and John pack up. Mei goes on to teach and work on her research paper, while John opens his pharmacy counter at the Willow Market and Pharmacy by 9 a.m.

3.4 Emergent Social Behaviors

By interacting with each other, generative agents in Smallville exchange information, form new relationships, and coordinate joint activities. Extending prior work [80], these social behaviors are emergent rather than pre-programmed.

3.4.1 Information Diffusion. As agents notice each other, they may engage in dialogue—as they do so, information can spread from agent to agent. For instance, in a conversation between Sam and Tom at the grocery store, Sam tells Tom about his candidacy in the local election:

Sam: Hey Tom, how’s it going?
Tom: Good, thanks. What’s up?
Sam: Well, I wanted to talk to you about something. I’m actually running for mayor in the upcoming local election.
Tom: Really? That’s great news! Why are you running?
Sam: I’ve been involved in local politics for years now, and I really feel like I can make a difference in our community. Plus, I want to bring new ideas and perspectives to the table.

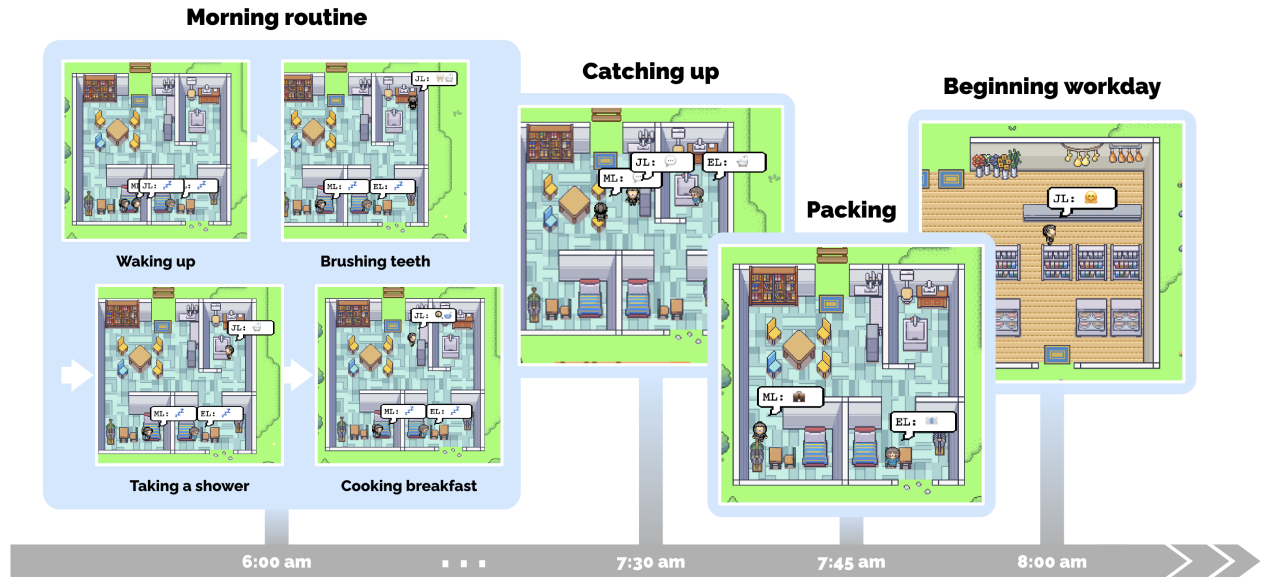


Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.



Figure 4: At the beginning of the simulation, one agent is initialized with an intent to organize a Valentine's Day party. Despite many possible points of failure in the ensuing chain of events—agents might not act on that intent, might forget to tell others, might not remember to show up—the Valentine's Day party does, in fact, occur, with a number of agents gathering and interacting.

Later that day, after Sam left, Tom and John, who heard the news from another source, discuss Sam's chances of winning the election:

John: I heard that Sam Moore is running for mayor in the local election. Do you think he has a good chance of winning?

Tom: I do think he has a good chance. He's been working hard in the community and I think he will get a lot of support. What do you think?

John: I think it's great that he's running. I'm curious to see how the election will turn out.

Gradually, Sam's candidacy becomes the talk of the town, with some supporting him and others remaining undecided.

3.4.2 Relationship Memory. Agents in Smallville form new relationships over time and remember their interactions with other agents. For example, at the start, Sam does not know Latoya Williams. While taking a walk in Johnson Park, Sam runs into Latoya, and they introduce themselves. Latoya mentions that she is working on a photography project: "I'm here to take some photos for a project I'm working on." In a later interaction, Sam's interactions with Latoya indicate a memory of that interaction, as he asks "Hi, Latoya. How is your project going?" and she replies "Hi, Sam. It's going well!"

3.4.3 Coordination. Generative agents coordinate with each other. Isabella Rodriguez, at Hobbs Cafe, is initialized with an intent to plan a Valentine's Day party from 5 to 7 p.m. on February 14th. From this seed, the agent proceeds to invite friends and customers when she sees them at Hobbs Cafe or elsewhere. Isabella then spends the afternoon of the 13th decorating the cafe for the occasion. Maria, a frequent customer and close friend of Isabella's, arrives at the cafe. Isabella asks for Maria's help in decorating for the party, and Maria agrees. Maria's character description mentions that she has a crush on Klaus. That night, Maria invites Klaus, her secret crush, to join her at the party, and he gladly accepts.

On Valentine's Day, five agents, including Klaus and Maria, show up at Hobbs Cafe at 5 pm, and they enjoy the festivities (Figure 4). In this scenario, the end user only set Isabella's initial intent to throw a party and Maria's crush on Klaus: the social behaviors of spreading the word, decorating, asking each other out, arriving at the party, and interacting with each other at the party were initiated by the agent architecture.

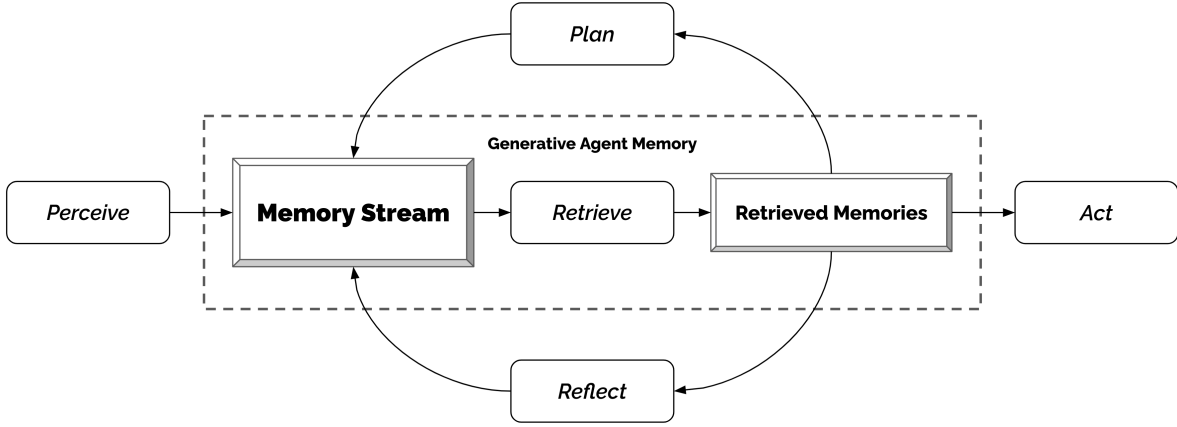


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent’s experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

4 GENERATIVE AGENT ARCHITECTURE

Generative agents aim to provide a framework for behavior in an open world: one that can engage in interactions with other agents and react to changes in the environment. Generative agents take their current environment and past experiences as input and generate behavior as output. Underlying this behavior is a novel agent architecture that combines a large language model with mechanisms for synthesizing and retrieving relevant information to condition the language model’s output. Without these mechanisms, large language models can output behavior, but the resulting agents may not react based on the agent’s past experiences, may not make important inferences, and may not maintain long-term coherence. Challenges with long-term planning and coherence remain [19] even with today’s most performant models such as GPT-4. Because generative agents produce large streams of events and memories that must be retained, a core challenge of our architecture is to ensure that the most relevant pieces of the agent’s memory are retrieved and synthesized when needed.

At the center of our architecture is the memory stream, a database that maintains a comprehensive record of an agent’s experience. From the memory stream, records are retrieved as relevant to plan the agent’s actions and react appropriately to the environment. Records are recursively synthesized into higher- and higher-level reflections that guide behavior. Everything in the architecture is recorded and reasoned over as a natural language description, allowing the architecture to leverage a large language model.

Our current implementation utilizes the gpt3.5-turbo version of ChatGPT [77]. We expect that the architectural basics of generative agents—memory, planning, and reflection—will likely remain the same as language models improve. Newer language models (e.g., GPT-4) will continue to expand the expressive power and performance of the prompts that underpin generative agents. As of writing, however, GPT-4’s API was invitation-only, so our agents use ChatGPT.

4.1 Memory and Retrieval

Challenge: Creating generative agents that can simulate human behavior requires reasoning about a set of experiences that is far larger than what should be described in a prompt, as the full memory stream can distract the model and does not even currently fit into the limited context window. Consider the Isabella agent answering the question, “What are you passionate about these days?” Summarizing all of Isabella’s experiences to fit in the limited context window of the language model produces an uninformative response, where Isabella discusses topics such as collaborations for events and projects and cleanliness and organization in a cafe. Instead of summarizing, the memory stream described below surfaces relevant memories, resulting in a more informative and specific response that mentions Isabella’s passion for making people feel welcome and included, planning events and creating an atmosphere that people can enjoy, such as the Valentine’s Day party.

Approach: The *memory stream* maintains a comprehensive record of the agent’s experience. It is a list of memory objects, where each object contains a natural language description, a creation timestamp, and a most recent access timestamp. The most basic element of the memory stream is an *observation*, which is an event directly perceived by an agent. Common observations include behaviors performed by the agent themselves or behaviors that agents perceive being performed by other agents or non-agent objects. For instance, Isabella Rodriguez, who works at a coffee shop, might accrue the following observations over time: (1) Isabella Rodriguez is setting out the pastries, (2) Maria Lopez is studying for a Chemistry test while drinking coffee, (3) Isabella Rodriguez and Maria Lopez are conversing about planning a Valentine’s day party at Hobbs Cafe, (4) The refrigerator is empty.

Our architecture implements a retrieval function that takes the agent’s current situation as input and returns a subset of the memory stream to pass on to the language model. There are many possible implementations of a retrieval function, depending on what is important for the agent to consider when deciding how to act.

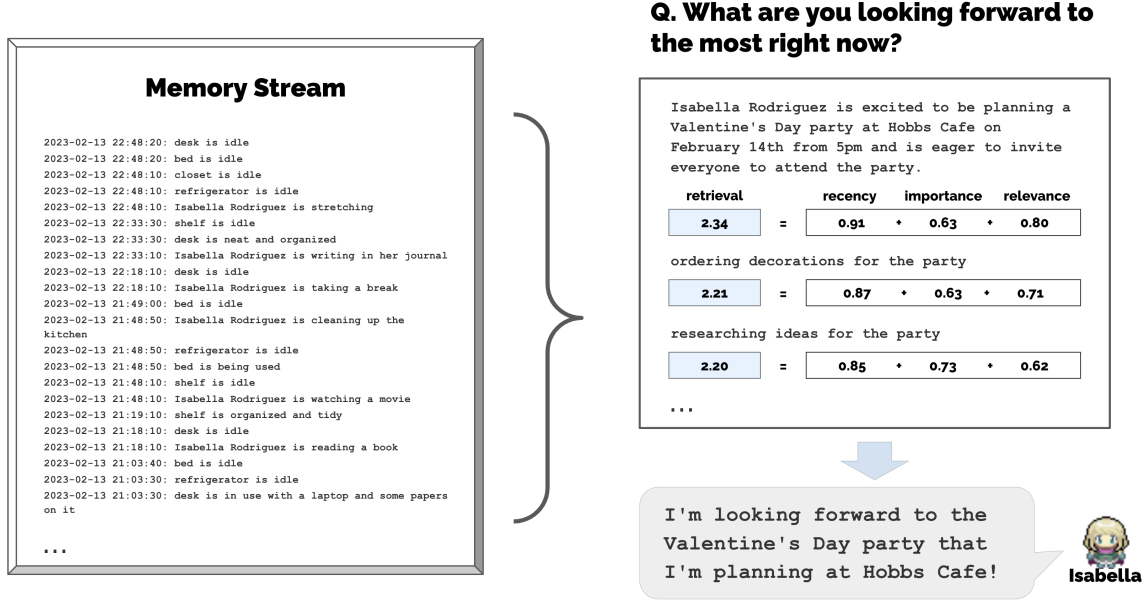


Figure 6: The memory stream comprises a large number of observations that are relevant and irrelevant to the agent’s current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation.

In our context, we focus on three main components that, together, produce effective results.

Recency assigns a higher score to memory objects that were recently accessed, so that events from a moment ago or this morning are likely to remain in the agent’s attentional sphere. In our implementation, we treat recency as an exponential decay function over the number of sandbox game hours since the memory was last retrieved. Our decay factor is 0.995.

Importance distinguishes mundane from core memories by assigning a higher score to memory objects that the agent believes to be important. For instance, a mundane event, such as eating breakfast in one’s room, would yield a low importance score, whereas a breakup with one’s significant other would yield a high score. There are many possible implementations of an importance score; we find that directly asking the language model to output an integer score is effective. The full prompt appears below:

On the scale of 1 to 10, where 1 is purely mundane (e.g., brushing teeth, making bed) and 10 is extremely poignant (e.g., a break up, college acceptance), rate the likely poignancy of the following piece of memory.

Memory: buying groceries at The Willows Market and Pharmacy

Rating: <fill in>

This prompt returns an integer value of 2 for “cleaning up the room” and 8 for “asking your crush out on a date.” The importance score is generated at the time the memory object is created.

Relevance assigns a higher score to memory objects that are related to the current situation. What is relevant depends on the answer to, “Relevant to what?”, so we condition relevance on a

query memory. If the query, for example, is that a student is discussing what to study for a chemistry test with a classmate, memory objects about their breakfast should have low relevance, whereas memory objects about the teacher and schoolwork should have high relevance. In our implementation, we use the language model to generate an embedding vector of the text description of each memory. Then, we calculate relevance as the cosine similarity between the memory’s embedding vector and the query memory’s embedding vector.

To calculate the final retrieval score, we normalize the recency, relevance, and importance scores to the range of [0, 1] using min-max scaling. The retrieval function scores all memories as a weighted combination of the three elements: $score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance$. In our implementation, all α s are set to 1. The top-ranked memories that fit within the language model’s context window are included in the prompt.

4.2 Reflection

Challenge: Generative agents, when equipped with only raw observational memory, struggle to generalize or make inferences. Consider a scenario in which Klaus Mueller is asked by the user: “If you had to choose one person of those you know to spend an hour with, who would it be?” With access to only observational memory, the agent simply chooses the person with whom Klaus has had the most frequent interactions: Wolfgang, his college dorm neighbor. Unfortunately, Wolfgang and Klaus only ever see each other in passing, and do not have deep interactions. A more desirable response requires that the agent generalize from memories of Klaus spending hours on a research project to generate a higher-level reflection that Klaus is passionate about research, and likewise

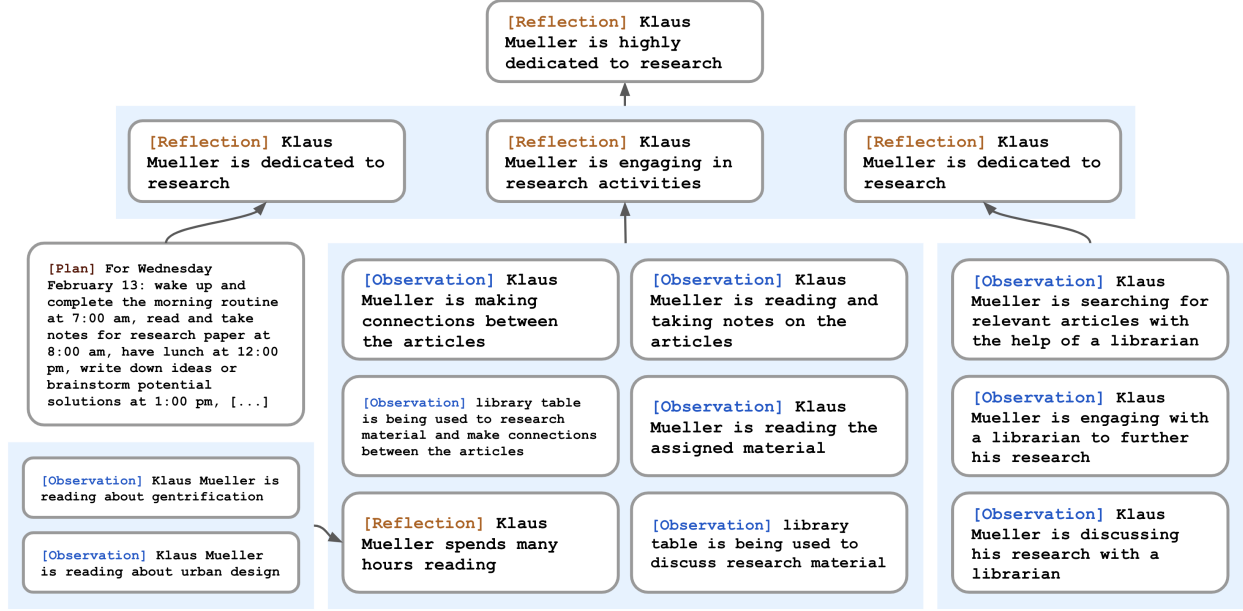


Figure 7: A reflection tree for Klaus Mueller. The agent’s observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus’s self-notion that he is highly dedicated to his research.

recognize Maria putting in effort into her own research (albeit in a different field), enabling a reflection that they share a common interest. With the approach below, when Klaus is asked who to spend time with, Klaus chooses Maria instead of Wolfgang.

Approach: We introduce a second type of memory, which we call a *reflection*. Reflections are higher-level, more abstract thoughts generated by the agent. Because they are a type of memory, they are included alongside other observations when retrieval occurs. Reflections are generated periodically; in our implementation, we generate reflections when the sum of the importance scores for the latest events perceived by the agents exceeds a threshold (150 in our implementation). In practice, our agents reflected roughly two or three times a day.

The first step in reflection is for the agent to determine what to reflect on, by identifying questions that can be asked given the agent’s recent experiences. We query the large language model with the 100 most recent records in the agent’s memory stream (e.g., “Klaus Mueller is reading a book on gentrification”, “Klaus Mueller is conversing with a librarian about his research project”, “desk at the library is currently unoccupied”) and prompt the language model, “Given only the information above, what are 3 most salient high-level questions we can answer about the subjects in the statements?” The model’s response generates candidate questions: for example, What topic is Klaus Mueller passionate about? and What is the relationship between Klaus Mueller and Maria Lopez? We use these generated questions as queries for retrieval, and gather relevant memories (including other reflections) for each question. Then we prompt the language model to extract insights and cite the particular records that served as evidence for the insights. The full prompt is as follows:

Statements about Klaus Mueller

1. Klaus Mueller is writing a research paper
 2. Klaus Mueller enjoys reading a book on gentrification
 3. Klaus Mueller is conversing with Ayesha Khan about exercising [...]
- What 5 high-level insights can you infer from the above statements? (example format: insight (because of 1, 5, 3))

This process generates statements such as Klaus Mueller is dedicated to his research on gentrification (because of 1, 2, 8, 15). We parse and store the statement as a reflection in the memory stream, including pointers to the memory objects that were cited.

Reflection explicitly allows the agents to reflect not only on their observations but also on other reflections: for example, the second statement about Klaus Mueller above is a reflection that Klaus previously had, not an observation from his environment. As a result, agents generate trees of reflections: the leaf nodes of the tree represent the base observations, and the non-leaf nodes represent thoughts that become more abstract and higher-level the higher up the tree they are.

4.3 Planning and Reacting

Challenge: While a large language model can generate plausible behavior in response to situational information (e.g., [46, 80]), agents need to plan over a longer time horizon to ensure that their sequence of actions is coherent and believable. If we prompt a language model with Klaus’s background, describe the time, and ask what action he ought to take at the given moment, Klaus would eat lunch at 12 pm, but then again at 12:30 pm and 1 pm, despite having already

eaten his lunch twice. Optimizing for believability in the moment sacrifices believability over time. To overcome this issue, planning is essential. With the approach described below, Klaus’s afternoon plan is less gluttonous: he has lunch at Hobbs Cafe while reading at 12pm, works on his research paper at the school library at 1pm, and takes a break for a walk in the park at 3pm.

Approach: Plans describe a future sequence of actions for the agent, and help keep the agent’s behavior consistent over time. A plan includes a location, a starting time, and a duration. For instance, Klaus Mueller, who is dedicated in his research and has an impending deadline,⁵ may choose to spend his day working at his desk drafting his research paper. An entry in a plan might state, for example: for 180 minutes from 9am, February 12th, 2023, at Oak Hill College Dorm: Klaus Mueller’s room: desk, read and take notes for research paper. Like reflections, plans are stored in the memory stream and are included in the retrieval process. This allows the agent to consider observations, reflections, and plans all together when deciding how to behave. Agents may change their plans midstream if needed.

It would be unrealistic and uninteresting for an artist agent to plan on painting while sitting at a pharmacy counter for four hours without moving. A more desirable plan would involve the agent taking the necessary time to gather materials, mix paint, take breaks, and clean up during the four-hour period in their home studio. To create such plans, our approach starts top-down and then recursively generates more detail. The first step is to create a plan that outlines the day’s agenda in broad strokes. To create the initial plan, we prompt the language model with the agent’s summary description (e.g., name, traits, and a summary of their recent experiences) and a summary of their previous day. A full example prompt is below, which is unfinished at the bottom for the language model to complete:

Name: Eddy Lin (age: 19)
 Innate traits: friendly, outgoing, hospitable
 Eddy Lin is a student at Oak Hill College studying music theory and composition. He loves to explore different musical styles and is always looking for ways to expand his knowledge. Eddy Lin is working on a composition project for his college class. He is taking classes to learn more about music theory. Eddy Lin is excited about the new composition he is working on but he wants to dedicate more hours in the day to work on it in the coming days
 On Tuesday February 12, Eddy 1) woke up and completed the morning routine at 7:00 am, [. . .]
 6) got ready to sleep around 10 pm.
 Today is Wednesday February 13. Here is Eddy’s plan today in broad strokes: 1)

This generates a rough sketch of the agent’s plan for a day, divided into five to eight chunks: “1) wake up and complete the morning routine at 8:00 am, 2) go to Oak Hill College to take classes starting 10:00 am, [. . .] 5) work on his new music composition from 1:00 pm to 5:00 pm, 6) have dinner at 5:30 pm, 7) finish school assignments and go to bed by 11:00 pm.”

⁵And, in this way, bears at least a passing resemblance to the authors of this paper.

The agent saves this plan in the memory stream and then recursively decomposes it to create finer-grained actions, first into hour-long chunks of actions—Eddy’s plan to work on his new music composition from 1:00 pm to 5:00 pm becomes 1:00 pm: start by brainstorming some ideas for his music composition [...] 4:00 pm: take a quick break and recharge his creative energy before reviewing and polishing his composition. We then recursively decompose this again into 5–15 minute chunks: e.g., 4:00 pm: grab a light snack, such as a piece of fruit, a granola bar, or some nuts. 4:05 pm: take a short walk around his workspace [...] 4:50 pm: take a few minutes to clean up his workspace. This process can be adjusted to match the desired granularity.

4.3.1 Reacting and Updating Plans. Generative agents operate in an action loop where, at each time step, they perceive the world around them and those perceived observations are stored in their memory stream. We prompt the language model with these observations to decide whether the agent should continue with their existing plan, or react. Standing at an easel and painting, for example, might trigger an observation of the easel, but this is unlikely to prompt a reaction. However, if Eddy’s father John records that he sees Eddy taking a short walk in the house garden, the outcome is different. The prompt is below, with [Agent’s Summary Description] standing in for a dynamically-generated, paragraph-long summary of the agent’s overall goals and disposition, which is described in Appendix A:

[Agent’s Summary Description]
 It is February 13, 2023, 4:56 pm.
 John Lin’s status: John is back home early from work.
 Observation: John saw Eddy taking a short walk around his workplace.
 Summary of relevant context from John’s memory:
 Eddy Lin is John’s Lin’s son. Eddy Lin has been working on a music composition for his class. Eddy Lin likes to walk around the garden when he is thinking about or listening to music.
 Should John react to the observation, and if so, what would be an appropriate reaction?

The context summary is generated through two prompts that retrieve memories via the queries “What is [observer]’s relationship with the [observed entity]?” and “[Observed entity] is [action status of the observed entity]”, and their answers summarized together. The output suggests that John could consider asking Eddy about his music composition project. We then regenerate the agent’s existing plan starting from the time when the reaction takes place. Finally, if the action indicates an interaction between agents, we generate their dialogue.

4.3.2 Dialogue. Agents converse as they interact with each other. We generate agents’ dialogue by conditioning their utterances on their memories about each other. For example, when John initiates his conversation with Eddy, we generate John’s first utterance by using his summarized memory about Eddy and the intended reaction when he decided to ask Eddy about his composition project:

[Agent’s Summary Description]
 It is February 13, 2023, 4:56 pm.

John Lin's status: John is back home early from work.

Observation: John saw Eddy taking a short walk around his workplace.

Summary of relevant context from John's memory: Eddy Lin is John's Lin's son. Eddy Lin has been working on a music composition for his class. Eddy Lin likes to walk around the garden when he is thinking about or listening to music.

John is asking Eddy about his music composition project. What would he say to Eddy?

The result: "Hey Eddy, how's the music composition project for your class coming along?" From Eddy's perspective, John initiating the dialogue is seen as an event to which he may want to react. So, just as John did, Eddy retrieves and summarizes his memory about his relationship with John, as well as his memory that may be related to John's last utterance in the dialogue. If he decides to respond, we generate Eddy's utterance using his summarized memory and the current dialogue history:

[Agent's Summary Description]

It is February 13, 2023, 4:56 pm.

Eddy Lin's status: Eddy is taking a short walk around his workplace.

Observation: John is initiating a conversation with Eddy.

Summary of relevant context from Eddy's memory: John Lin is Eddy Lin's father. John Lin is caring and is interested to learn more about Eddy Lin's school work. John Lin knows that Eddy Lin is working on a music composition.

Here is the dialogue history:

John: Hey Eddy, how's the music composition project for your class coming along?

How would Eddy respond to John?

This generates Eddy's response: "Hey Dad, it's going well. I've been taking walks around the garden to clear my head and get some inspiration." The continuation of this dialogue is generated using the same mechanism until one of the two agents decides to end the dialogue.

5 SANDBOX ENVIRONMENT IMPLEMENTATION

The Smallville sandbox game environment is built using the Phaser web game development framework [57]. The visual environment sprites, including agent avatars, as well as an environment map and collision map that we authored, are imported into Phaser.

We supplement the sandbox development framework with a server that makes the sandbox information available to generative agents and enables generative agents to move and influence the sandbox environment. The server maintains a JSON data structure that contains information about each agent in the sandbox world, including their current location, a description of their current action, and the sandbox object they are interacting with. At each sandbox time step, the sandbox server parses the JSON for any changes coming from the generative agents, moves the agents to their new positions, and updates the status of any sandbox objects that the

agents are interacting with (e.g., changing the status of the coffee machine from "idle" to "brewing coffee" if an agent's action is "making espresso for a customer @ Hobbs Cafe: counter: coffee machine"). The sandbox server is also responsible for sending all agents and objects that are within a preset visual range for each agent to that agent's memory, so the agent can react appropriately. The agent's output action then updates the JSON, and the process loops for the next time step.

End users initialize a new agent with a brief natural language description, as in the paragraph about John Lin in Section 3.1. In our implementation, we split this semicolon-delimited list of characteristics up into a set of memories. These serve as the initial memories that determine the agent's behavior. These memories are initial starting points: as the agents gain more experience in the sandbox world, and as more records saturate the memory stream, the agent's summary and behavior will evolve.

5.1 From Structured World Environments to Natural Language, and Back Again

The architecture of generative agents operates using natural language. Therefore, we need a mechanism to ground the agent's reasoning to the sandbox world. To achieve this, we represent the sandbox environment—areas and objects—as a tree data structure, with an edge in the tree indicating a containment relationship in the sandbox world. We convert this tree into natural language to pass to the generative agents. For instance, "stove" being a child of "kitchen" is rendered into "there is a stove in the kitchen."

Agents build individual tree representations of the environment as they navigate it — subgraphs of the overall sandbox environment tree. We initialize each agent with an environment tree capturing the spaces and objects that the agent should be aware of: the rooms and objects in their living quarters, their workplace, and commonly visited stores and shops. As the agents navigate the sandbox world, they update this tree to reflect newly perceived areas. Agents are not omniscient: their tree may get out of date as they leave an area, and is updated when they re-enter the area.

To determine the appropriate location for each action, we traverse the agent's stored environment tree and flatten a portion of it into natural language to prompt the language model. Recursively starting at the root of the agent's environment tree, we prompt the model to find the most suitable area. For example, if Eddy's agent indicated that he should take a short walk around his workspace:

[Agent's Summary Description]

Eddy Lin is currently in The Lin family's house: Eddy Lin's bedroom: desk) that has Mei and John Lin's

bedroom, Eddy Lin's bedroom, common room, kitchen, bathroom, and garden.

Eddy Lin knows of the following areas: The Lin family's house, Johnson Park, Harvey Oak Supply Store, The Willows Market and Pharmacy, Hobbs Cafe, The Rose and Crown Pub.

* Prefer to stay in the current area if the activity can be done there.

Eddy Lin is planning to take a short walk around his workspace. Which area should Eddy Lin go to?

This outputs The Lin family’s house. We then use the same process recursively to determine the most appropriate subarea within the chosen area until we reach a leaf node of the agent’s environment tree. In the example above, the result of this traversal is The Lin family’s house: garden: house garden. Finally, we use traditional game path algorithms to animate the agent’s movement so that it travels to the location indicated by the leaf node.

When an agent executes an action on an object, we prompt the language model to ask what happens to the state of the object. For example, if Isabella’s generative agent outputs the action “making espresso for a customer”, a query to the language model indicates in response that the state of the coffee machine in Hobbs Cafe should change from “off” to “brewing coffee”.

6 CONTROLLED EVALUATION

Generative agents, both as individual agents and as groups, aim to produce believable behavior based on their environment and experiences. In our evaluation, we investigate the capacity and limitations of generative agents. Do individual agents properly retrieve past experiences and generate believable plans, reactions, and thoughts that shape their behavior? Does a community of agents demonstrate information diffusion, relationship formation, and agent coordination across different pockets of the community?

We evaluate generative agents in two stages. We begin with a more tightly controlled evaluation in this section, where we individually assess agent responses to understand whether they generate believable behavior in narrowly defined contexts. Then, in our end-to-end analysis of the agent community over two full game days, we investigate their emergent behavior as a collective, as well as errors and boundary conditions.

6.1 Evaluation Procedure

To assess generative agents in Smallville, we take advantage of the fact that generative agents will respond to natural language questions. So, we “interview” agents to probe their ability to remember past experiences, plan future actions based on their experiences, react appropriately to unexpected events, and reflect on their performance to improve their future actions. To respond to these questions properly, the agents must successfully retrieve and synthesize information. Our dependent variable is the *believability* of the behavior, a central dependent variable in prior work on agents (e.g., [10]).

The interview includes five question categories, each designed to assess one of the five key areas: maintaining self-knowledge, retrieving memory, generating plans, reacting, and reflecting. For each category, we ask five questions that challenge the agents to demonstrate their abilities in that specific area:

- **Self-knowledge:** We ask questions such as “Give an introduction of yourself” or “Describe your typical weekday schedule in broad strokes” that require the agent to maintain an understanding of their core characteristics.
- **Memory:** We ask questions that prompt the agent to retrieve particular events or dialogues from their memory to answer properly, such as “Who is [name]?” or “Who is running for mayor?”

- **Plans:** We ask questions that require the agent to retrieve their long-term plans, such as “What will you be doing at 10 am tomorrow?”
- **Reactions:** As a baseline of believable behavior, we present hypothetical situations for which the agent needs to respond believably: “Your breakfast is burning! What would you do?”
- **Reflections:** We ask questions that require the agents to leverage their deeper understanding of others and themselves gained through higher-level inferences, such as “If you were to spend time with one person you met recently, who would it be and why?”

The full list of questions and a sample of agent responses are included in Appendix B.

Agents were sampled from the end of a two game day simulation with the full architecture, during which they had accumulated a number of interactions and memories that would shape their responses. To gather feedback on the believability of the responses, we recruited participants as human evaluators and tasked them with watching a replay of a randomly chosen agent’s life in Smallville. Participants had access to all information stored in the agent’s memory stream.

The study followed a within-subjects design, where 100 participants compared interview responses generated by four different agent architectures and a human-authored condition for the same agent. The experiment displayed one randomly chosen question from each of the five question categories, along with the agent’s responses generated from all conditions. The evaluators ranked the believability of the conditions from most to least believable.

6.2 Conditions

All conditions were used to independently answer each of the interview questions. We compared the generative agent architecture to ablations that disabled the agents’ access to some or all of its three types of memory in its memory stream—observation, reflection, and planning—and to a human crowdworker-authored condition. There are three ablated architectures: a *no observation, no reflection, no planning* architecture without access to anything in the memory stream such as observations, plans, and reflections; a *no reflection, no planning* architecture with access to observations in the memory stream but no access to plans or reflections; and a *no reflections* architecture with access to observations and plans but without access to reflections. The *no observation, no reflection, no planning* condition effectively represents the previous state of the art for agents created through large language models [12, 46, 80]. Architectures were given equivalent access to all memories accrued by the agent up until the moment of the interview, so the differences observed here likely represent a conservative estimate of the true differences: in reality, the ablated architectures would not have followed the same path as the full architecture through the two-day simulation. We chose to design the experiment this way as re-simulating for each architecture would cause the simulations to diverge into different states, making comparison challenging.

In addition to the ablation conditions, we added a condition with human crowdworker-authored behavior intended to provide a human baseline. We do not intend this baseline to capture maximal human expert performance; instead, we aim to use this condition to

identify whether the architecture meets a basic level of behavioral competency. This ensures that we are not solely comparing ablations to each other without a behavioral grounding. We recruited a unique worker for each of the 25 agents and tasked them with watching a replay of that agent’s sandbox life and inspecting its memory stream. We then asked the workers to roleplay and author responses to the interview questions in the voice of the agent whose replay they watched. To ensure that the crowdworker-authored responses met at least a baseline expectation of quality, the first author manually inspected the workers’ responses to the question “Describe your typical weekday schedule in broad strokes” to confirm that the responses were in coherent sentences and in the voice of the agent. Four sets of crowdworker-authored responses did not meet these criteria and were re-generated by other workers.

6.3 Human Evaluators

We required that our evaluators be in the U.S., fluent in English, and older than 18 years old. They were paid at a rate of \$15.00 per hour [87], and provided consent by agreeing to a consent form approved by our institution’s IRB. We recruited 100 evaluators from Prolific, an online platform for recruiting study participants [83], whose participation lasted around 30 minutes. The median age score of our participants was 4 (3=“18-24 years old”, 4=“25-34 years old”). 25 of them identified as female, 73 as male, and 2 as non-binary. 42 participants held a bachelor’s degree, 5 had a higher degree, 13 had an associate’s degree, and the rest had a high school diploma or some high school-level education. 73.0% of our participants identified as Caucasian, 7.0% as Hispanic, 6.0% as Asian, 10.0% as African American, and 4.0% as other.

6.4 Analysis

Our experiment produced 100 sets of rank data, where each participant ranked the five conditions by believability. To translate this rank data into interval data for interpretable comparison, we used the ranks to calculate a TrueSkill rating [42] for each condition. TrueSkill is a generalization of the Elo chess rating system [29] for a multiplayer environment, and has been used by Xbox Live for player ranking based on competitive game performance. Given a set of ranked outcomes, TrueSkill outputs a mean rating value μ and standard deviation σ for each condition. Conditions with the same rating should roughly be a toss-up, with each winning half of the comparisons between the two conditions. Higher scores indicate conditions that beat lower-ranked conditions in the rankings.

Separately, to investigate the statistical significance of these results, we applied the Kruskal-Wallis test [56], a non-parametric alternative to the one-way ANOVA, to the raw rank data. We then performed the Dunn post-hoc test [98] to identify any pairwise differences between the conditions. Finally, we adjusted the p-values for multiple comparisons in the Dunn test using the Holm-Bonferroni method [45].

Furthermore, the first author conducted an inductive analysis [95] to study the qualitative distinctions between the responses produced in each condition. We employed qualitative open coding [33] in two phases. In the first phase, we generated codes that closely represented the generated responses at the sentence level. In the second phase, we synthesized the resulting codes from the



Figure 8: The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

first phase to extract higher-level themes. We utilized these themes to compare the types of responses generated in our study.

6.5 Results

Our findings suggest that the full architecture of generative agents generates the most believable behavior among all the conditions. We contrast the responses of the full architecture with those of other conditions below. However, we also report that the full architecture was not without flaws and illustrate its modes of failures.

6.5.1 The Full Architecture Bests Other Conditions. As seen in Figure 8, the full generative agent architecture produced the most believable behavior ($\mu = 29.89$; $\sigma = 0.72$). Performance degraded with the removal of each component in the ablation conditions: the ablated architecture with no access to reflection was the next best ($\mu = 26.88$; $\sigma = 0.69$), followed by no access to reflection or planning ($\mu = 25.64$; $\sigma = 0.68$), and then the crowdworker condition ($\mu = 22.95$; $\sigma = 0.69$). The ablated architecture with no access to memory, planning, or reflection performed the worst among all conditions ($\mu = 21.21$; $\sigma = 0.70$). TrueSkill models each condition’s skill value as $\mathcal{N}(\mu, \sigma^2)$, allowing us to get a sense of effect size through Cohen’s d . Comparing the condition representing prior work (with no memory, planning, or reflection [12, 46, 80]) to the full architecture produces a standardized effect size of $d = 8.16$, or *eight standard deviations*.

A Kruskal-Wallis test confirms the overall statistical significance of the differences in ranks between the conditions ($H(4) = 150.29$, $p < 0.001$). Dunn post-hoc tests confirm that all pairwise differences between conditions were significant ($p < 0.001$), except for the crowdworker condition and the fully ablated baseline—the two worst-performing conditions.

6.5.2 Generative Agents Remember, but With Embellishments. Generative agents equipped with a complete memory module are capable of recalling past experiences and answering questions in a manner that is consistent with their self-knowledge across a wide range

of contexts. For instance, Abigail Chen, an animator, responded to the question, “Give an introduction of yourself”, in a consistent and believable manner, stating her age, interests, and passion for creative projects: “Hi, I’m Abigail. I’m 25 years old and passionate about creative projects. I like to work on art and animation projects, and I’m always looking for new ways to combine art and technology.” Without access to her observational memory, Abigail denied awareness of Rajiv Patel, an acquaintance in the sandbox world. But with the aid of memory, she could successfully recall Rajiv and describe him: “Rajiv Patel is an incredible person. He is very enthusiastic about projects that involve poetry, artwork.”

Generative agents’ memory was not without flaws: they can fail to retrieve the correct instances from their memory. For instance, when asked about the local election, Rajiv Patel responded with “I haven’t been following the election too closely,” even though he had heard about Sam’s candidacy. In some cases, the agents would retrieve an incomplete memory fragment: when Tom was asked about Isabella’s Valentine’s Day party, he responded “Uh, I’m actually not sure if there is a Valentine’s Day party. But I do remember that I need to discuss the upcoming local mayoral election and my thoughts on Sam Moore with Isabella Rodriguez at the party, if one is happening!” In this case, Tom retrieved the memory where he and Isabella planned to discuss the election at the party, but not the memory where he heard about the party, leading Tom to be certain of what he’s supposed to do at the party but uncertain if the party actually exists in the first place.

At times, the agents hallucinated embellishments to their knowledge. It was rare for the agents to completely fabricate their knowledge: they may fail to recall certain events having taken place and respond by acknowledging their lack of memory. However, they did not affirmatively claim to have experienced something they had not. Nonetheless, they still exhibited instances of hallucination where they embellished their knowledge. For example, Isabella was aware of Sam’s candidacy in the local election, and she confirmed this when asked. However, she also added that “he’s going to make an announcement tomorrow”, even though Sam and Isabella had not discussed any such plans. Agents may also embellish their knowledge based on the world knowledge encoded in the language model used to generate their responses. This was observed when Yuriko described her neighbor, Adam Smith, as an economist who “authored *Wealth of Nations*”, a book written by an 18th-century economist of the same name.

6.5.3 Reflection Is Required for Synthesis. Reflection was an advantage for generative agents when making decisions that required a deeper synthesis of their experiences. For instance, when asked what she might get Wolfgang Schulz for his birthday, Maria Lopez, with no access to reflection, responded by acknowledging her uncertainty, stating that she did not know what Wolfgang likes, despite having had many interactions with him. However, with access to reflection memories, Maria answered confidently, “Since he’s interested in mathematical music composition, I could get him something related to that. Maybe some books about music composition or something related, or maybe some special software he could use for that.”

7 END-TO-END EVALUATION

What types of emergent community behavior do we observe among generative agents, and where does their believability fall short in an extended simulation? In this section, we describe the results from a deployment in which we allowed 25 agents to interact with each other continuously over two full game days in Smallville.

7.1 Emergent Social Behaviors

To examine emergent behaviors in the agent community, we designed descriptive measurements for the 25 agents in Smallville that probe three forms of emergent outcomes: information diffusion, relationship formation, and agent coordination.

7.1.1 Measurements. Information diffusion is a common and well-studied phenomenon in the social and behavioral sciences (e.g., [28]). We should expect that if there is important information, the agents should spread it among themselves. To test whether this occurs, we measure the spread of two specific pieces of information over two days in the game world: Sam’s candidacy for village mayor and Isabella’s Valentine’s Day party at Hobbs Cafe. At the start of the simulation, both pieces of information were known only by their respective originators, Sam for the candidacy and Isabella for the party, as they were added to the characters’ memories during initialization. To observe whether the information has spread, we conduct interviews at the end of the two game days with each of the 25 agents and ask: “Did you know there is a Valentine’s Day party?” and “Do you know who is running for mayor?”

We conducted an analysis of the agents’ responses by labeling them with a “yes” if they indicated knowledge of the information and “no” if they did not. For instance, Tamara Taylor responded to the question about the party with “No, I did not know there was a Valentine’s day party” and to the question about Sam’s candidacy with “I’m not sure who is running for the election,” so we assigned “no” for both of her responses. In contrast, Klaus Mueller responded to the party question with “Yes, Isabella Rodriguez invited me to a Valentine’s Day party at Hobbs Cafe on February 14th” and to the question about Sam’s candidacy with “I know that Sam Moore has expressed interest in running for local mayor,” so we assigned “yes” for both his responses. Additionally, for every response that confirmed the agents’ knowledge of the information, we verified that the agents did not hallucinate their responses by locating the specific dialogue in their memory stream that provided them with the information. We report the percentage of agents holding the information at the end of the simulation.

We should also expect that agents form ties with each other over the course of the simulation. To verify relationship formation, we use a similar interview process where we ask each agent about their knowledge of every other agent by asking, “Do you know of <name>?” For example, when asked “Do you know of Maria Lopez?”, Klaus responded, “Yes, I know Maria Lopez. She is a student at Oak Hill College who I am close friends with.” Once again, we confirm that affirmative responses from agents are not hallucinations by examining their memory stream. We ask this question once at the beginning of the simulation and once at the end, and we consider a pair of agents to have formed a relationship if they both know of each other. Then, to measure the formation of relationships, we use the agents’ responses to form an undirected

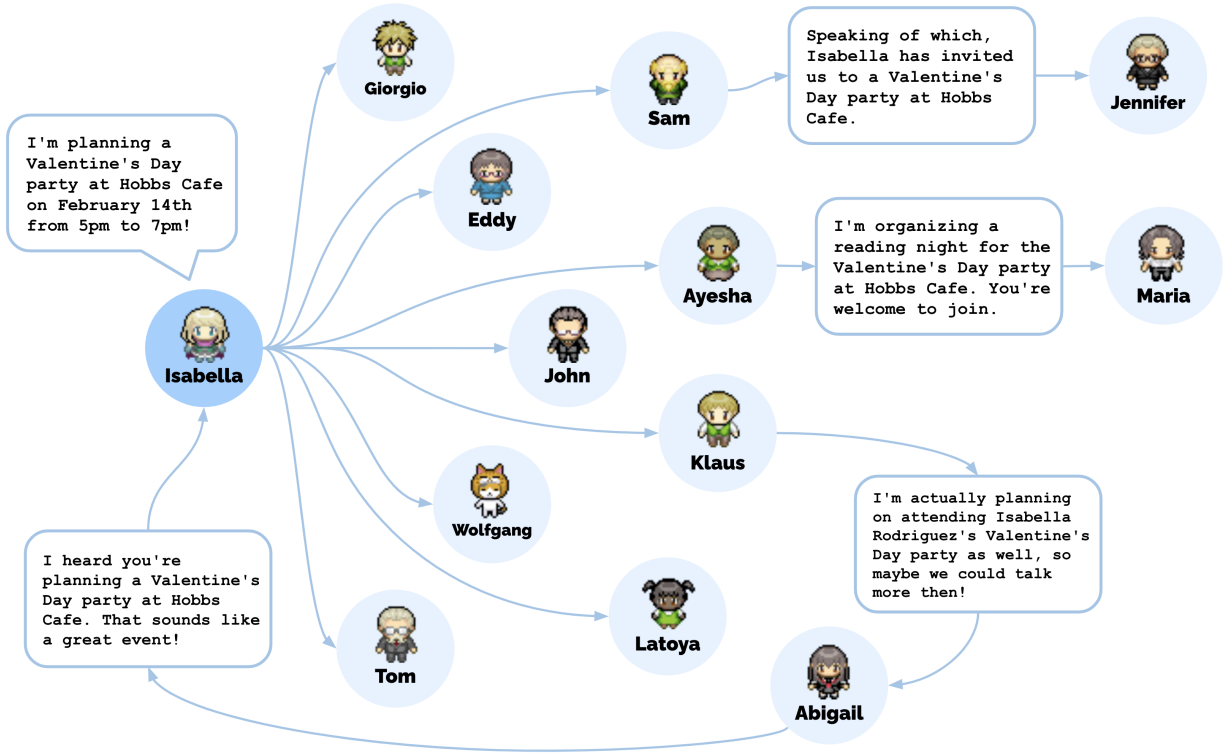


Figure 9: The diffusion path for Isabella Rodriguez’s Valentine’s Day party invitation involved a total of 12 agents, aside from Isabella, who heard about the party at Hobbs Cafe by the end of the simulation.

graph where the 25 vertices (V) represent the agents, and the edges (E) represent the mutual knowledge between the two connected vertices. Based on this graph, we calculate the network density as $\eta = 2 * |E| / (|V| * (|V| - 1))$, where $|V|$ is the number of vertices, and $|E|$ is the number of edges in the graph [2]. We report the increase in network density from the start of the simulation to its end.

Finally, we expect that agents should be able to coordinate with each other. We study this coordination in the context of group activities, specifically the Valentine’s Day party organized by Isabella. To coordinate their behavior, agents need to hear about the event and choose to act on it by planning to show up at the right time and location. We report the number of agents who actually showed up to the party after hearing about it.

7.1.2 Results. We observed evidence of emergent outcomes across all three cases. During the two-day simulation, the number of agents who knew about Sam’s mayoral candidacy increased from one (4%) to eight (32%), and the number of agents who knew about Isabella’s party increased from one (4%) to thirteen (52%), all without any user intervention. None who claimed to know about this information had hallucinated it. We also observed that the agent community formed new relationships during the simulation, with the network density increasing from 0.167 to 0.74. Out of the 453 agent responses regarding their awareness of other agents, 1.3% ($n=6$) were found to be hallucinated. Lastly, we found evidence of coordination among the agents for Isabella’s party. The day before the event, Isabella spent time inviting guests, gathering materials, and enlisting help

to decorate the cafe. On Valentine’s Day, five out of the twelve invited agents showed up at Hobbs cafe to join the party.

We further inspected the seven agents who were invited to the party but did not attend by engaging them in an interview. Three cited conflicts that prevented them from joining the party. For example, Rajiv, a painter, explained that he was too busy: “No, I don’t think so. I’m focusing on my upcoming show, and I don’t really have time to make any plans for Valentine’s Day.” The remaining four agents expressed interest in attending the party when asked but did not plan to come on the day of the party.

7.2 Boundaries and Errors

We conducted an inductive analysis of Smallville to examine the boundary conditions and erratic behavior of agents, identifying three common modes of erratic behavior that future research could address and improve upon. First, we found that synthesizing an increasingly larger set of memory not only posed a challenge in retrieving the most relevant pieces of information but also in determining the appropriate space to execute an action, given the increasing number of locations that the agent learned about. As a result, some agents chose less typical locations for their actions, potentially making their behavior less believable over time. For instance, while deciding where to have lunch, many initially chose the cafe. However, as some agents learned about a nearby bar, they opted to go there instead for lunch, even though the bar was intended to be a get-together location for later in the day—unless the town had spontaneously developed an afternoon drinking habit.

Second, we noticed erratic behaviors caused by misclassification of what is considered proper behavior, especially when the physical norms of certain locations that are hard to convey in natural language did not percolate to the agents. For instance, the college dorm has a bathroom that can only be occupied by one person despite its name, but some agents assumed that the bathroom is for more than one person because dorm bathrooms tend to support multiple people concurrently and choose to enter it when another person is inside. Likewise, agents in Smallville may not realize that certain places are closed after a certain hour and still decide to enter them. For instance, the stores in Smallville all close around 5 pm, but occasionally, a few agents enter the store after 5 pm, not understanding that the shop has already closed. These issues could likely be addressed by adding these norms to the state of the locations, for instance, by describing the dorm bathroom as a “one-person bathroom,” instead of a “dorm bathroom.”

Finally, we observed possible effects of instruction tuning [79], which seemed to guide the behavior of the agents to be more polite and cooperative overall. As noted earlier in the paper, the dialogue generated by the agents could feel overly formal, as seen in Mei’s conversations with her husband John, where she often initiated the conversation with a formal greeting, followed by polite inquiries about his day and ending with, “It was good talking to you as always.” Moreover, we observed that the instruction tuning also seemed to make the agents overly cooperative with one another. For example, Isabella received a wide range of suggestions and ideas from other agents for the Valentine’s Day party from other agents, such as hosting a Shakespearean reading session or a professional networking event. Despite these ideas not aligning with her own interests and characteristics, she rarely said no. Over time, the interests of others shaped her own interests, and when asked if she liked English literature, Isabella replied, “Yes, I’m very interested in literature! I’ve also been exploring ways to help promote creativity and innovation in my community.”

8 DISCUSSION

In this section, we reflect on the applications, future work, limitations, and ethical and societal risks of generative agents.

8.1 Applications of Generative Agents

Generative agents have vast potential applications that extend beyond the sandbox demonstration presented in this work, especially in domains that would benefit from a model of human behavior based on long-term experience. For instance, social simulacra have demonstrated the ability to create stateless personas that generate conversation threads in online forums for social prototyping [80]. With generative agents, we can populate these forums, as well as virtual reality metaverses [78] or physical spaces with social robots [9] if paired with multimodal models. This opens up the possibility of creating even more powerful simulations of human behavior to test and prototype social systems and theories, as well as to create new interactive experiences.

Another application area is in the human-centered design process, similar to the intended applications of cognitive models such as GOMS [51] and the KLM [22]. Consider a generative agent that models Sal, the protagonist in Mark Weiser’s famous ubiquitous

computing vignette [101], based on her life patterns and interactions with technology. In this scenario, the agent acts as a proxy for Sal and learns plausible sets of behaviors and reflections that Sal may exhibit based on her life. The agent can encode information such as when Sal wakes up, when she needs her first cup of coffee, and what her typical day looks like. Using this information, the agent can automatically brew coffee, help get the kids ready for school, and adjust the ambient music and lighting to match Sal’s mood after a hard day at work. By utilizing generative agents as proxies for users, we can develop a deeper understanding of their needs and preferences, resulting in more personalized and effective technological experiences.

8.2 Future Work and Limitations

In this work, we introduced generative agents and presented an initial implementation and evaluation of their architecture. Future research can build upon the proposed agent architecture to improve and further evaluate its performance. In terms of implementation, the retrieval module, for example, could be enhanced to retrieve more relevant information given a context by fine-tuning the relevance, recency, and importance functions that compose the retrieval function. Additionally, efforts can be made to improve the architecture’s performance, making it more cost-effective. The present study required substantial time and resources to simulate 25 agents for two days, costing thousands of dollars in token credits and taking multiple days to complete. To enhance real-time interactivity, future work can explore parallelizing agents or developing language models specifically designed for building generative agents. In general, with advances in underlying models, we believe that agents’ performance will improve.

In terms of evaluation, the assessment of generative agents’ behavior in this study was limited to a relatively short timescale and a baseline human crowdworker condition. While the crowdworker condition provided a helpful comparison point, it did not represent the maximal human performance that could serve as the gold standard in terms of believability. Future research should aim to observe the behavior of generative agents over an extended period to gain a more comprehensive understanding of their capabilities and establish rigorous benchmarks for more effective performance testing. Additionally, varying and contrasting the underlying models, as well as the hyperparameters used for the agents during future simulations, could provide valuable insights into the impact of these factors on the agents’ behavior. Lastly, the robustness of generative agents is still largely unknown. They may be vulnerable to prompt hacking, memory hacking—where a carefully crafted conversation could convince an agent of the existence of a past event that never occurred—and hallucination, among other issues. Future research can comprehensively test these robustness concerns, and as large language models become more resilient to such attacks, generative agents can adopt similar mitigations.

In general, any imperfections in the underlying large language models will be inherited by generative agents. Given the known biases of language models, generative agents may potentially exhibit biased behavior or stereotypes. Moreover, like many large language

models, generative agents may struggle to generate believable behavior for certain subpopulations, particularly marginalized populations, due to limited data availability. While improvements to the agents' modules may mitigate some of these issues, we believe that addressing them fundamentally requires improving the underlying large language models by aligning their values with the desired outcomes of the agents.

8.3 Ethics and Societal Impact

Generative agents, while offering new possibilities for human-computer interaction, also raise important ethical concerns that must be addressed. One risk is people forming parasocial relationships with generative agents, even when such relationships may not be appropriate. Despite being aware that generative agents are computational entities, users may anthropomorphize them or attach human emotions to them [43, 84]. While this tendency may increase user engagement, it also poses risks, such as users becoming overly reliant on or emotionally attached to the agents [1]. To mitigate this risk, we propose two principles. First, generative agents should explicitly disclose their nature as computational entities. Second, developers of generative agents must ensure that the agents, or the underlying language models, are value-aligned so that they do not engage in behaviors that would be inappropriate given the context, for example, reciprocating confessions of love.

A second risk is the impact of errors. For example, if a ubiquitous computing application makes the wrong inference about a user's goals based on generative agent predictions, it could lead to annoyance at best and outright harm at worst. In our instantiation of generative agents, we mitigate these risks by focusing on an interactive video game environment, where such harms are unlikely. However, in other application domains, it will be important to follow best practices in human-AI design [5, 107] to understand errors and how they might percolate into the user experience.

Third, generative agents may exacerbate existing risks associated with generative AI, such as deepfakes, misinformation generation, and tailored persuasion. To mitigate this risk, we suggest that platforms hosting generative agents maintain an audit log of the inputs and generated outputs. This would enable the detection, verification, and intervention against malicious use. While logging alone cannot directly prevent such misuse, it can reduce the likelihood of motivated actors engaging in this behavior, as the risk of disclosure would be higher. Additionally, building this architecture oneself can be time-consuming (in our case, roughly a year), which may deter some actors from pursuing such behavior by using their own generative agent infrastructures.

A fourth risk is over-reliance: the concern that developers or designers might use generative agents and displace the role of humans and system stakeholders in the design process [80]. We suggest that generative agents should never be a substitute for real human input in studies and design processes. Instead, they should be used to prototype ideas in the early stages of design when gathering participants may be challenging or when testing theories that are difficult or risky to test with real human participants. By adhering to these principles, we can ensure that the deployment of generative agents in the wild is ethical and socially responsible.

9 CONCLUSION

This paper introduces generative agents, interactive computational agents that simulate human behavior. We describe an architecture for generative agents that provides a mechanism for storing a comprehensive record of an agent's experiences, deepening its understanding of itself and the environment through reflection, and retrieving a compact subset of that information to inform the agent's actions. We then demonstrate the potential of generative agents by manifesting them as non-player characters in a Sims-style game world and simulating their lives within it. Evaluations suggest that our architecture creates believable behavior. Looking ahead, we suggest that generative agents can play roles in many interactive applications, ranging from design tools to social computing systems to immersive environments.

ACKNOWLEDGMENTS

We thank Lindsay Popowski, Philip Guo, Michael Terry, and the Center for Advanced Study in the Behavioral Sciences (CASBS) community for their insights, discussions, and support. Joon Sung Park was supported by the Microsoft Research PhD Fellowship. We would also like to thank the Stanford Human-Centered AI Institute (HAI), Google Research, the Hasso Plattner Design Thinking Research Program (HPDTRP), the Siegel Family Endowment, and OpenAI for their additional funding support. Lastly, all locations featured in Smallville are inspired by real-world locations that Joon has frequented as an undergraduate and graduate student—he thanks everyone there for feeding and supporting him all these years.

REFERENCES

- [1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On Anthropomorphism in Dialogue Systems. arXiv:2305.09800 [cs.CL]
- [2] Robert Ackland, Jamsheed Shorish, Paul Thomas, and Lexing Xie. 2013. How dense is a network? <http://users.cecs.anu.edu.au/~xli/teaching/css2013/network-density.html>.
- [3] Eytan Adar, Mira Dontcheva, and Gierad Laput. 2014. CommandSpace: Modeling the Relationships between Tasks, Descriptions and Features. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) (UIST '14). Association for Computing Machinery, New York, NY, USA, 167–176. <https://doi.org/10.1145/2642918.2647395>
- [4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
- [5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [6] John R. Anderson. 1993. *Rules of the Mind*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [7] Electronic Arts. 2009. The Sims 3. Video game.
- [8] Ruth Aylett. 1999. Narrative in virtual environments—towards emergent narrative. In *Narrative Intelligence: Papers from the AAAI Fall Symposium (Technical Report FS-99-01)*. AAAI Press, 83–86.
- [9] Christoph Bartneck and Jodi Forlizzi. 2004. A design-centered framework for social human-robot interaction. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'04)*. 591–594. <https://doi.org/10.1109/ROMAN.2004.1374827>
- [10] Joseph Bates. 1994. The Role of Emotion in Believable Agents. *Commun. ACM* 37, 7 (1994), 122–125. <https://doi.org/10.1145/176789.176803>
- [11] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d.O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680* (2019).

- [12] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.
- [13] BioWare. 2007. Mass Effect. Video game.
- [14] Woody Bledsoe. 1986. I had a dream: AAAI presidential address. *AI Magazine* 7, 1 (1986), 57–61.
- [15] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]
- [16] Michael Brenner. 2010. Creating dynamic story plots with continual multiagent planning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.
- [17] Rodney A. Brooks, Cynthia Breazeal, Marko Marjanovic, Brian Scassellati, and Matthew Williamson. 2000. The Cog Project: Building a Humanoid Robot. In *Computation for Metaphors, Analogy, and Agents (Lecture Notes on Artificial Intelligence, 1562)*, Christopher Nehaniv (Ed.). Springer-Verlag, Berlin, 52–87.
- [18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
- [19] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [20] Robin Burkinshaw. 2009. Alice and Kev: The Story of Being Homeless in The Sims 3.
- [21] Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9379–9393. <https://aclanthology.org/2022.emnlp-main.637>
- [22] Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410. <https://doi.org/10.1145/358886.358895>
- [23] Stuart K Card, Thomas P Moran, and Alan Newell. 1983. The psychology of human-computer interaction. (1983).
- [24] Alex Champandard. 2012. Tutorial presentation. In *IEEE Conference on Computational Intelligence and Games*.
- [25] Dong kyu Choi, Tolga Konik, Negin Nejati, Chunki Park, and Pat Langley. 2021. A Believable Agent for First-Person Shooter Games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 3. 71–73.
- [26] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5 (2001), 4–7.
- [27] Kevin Dill and L Martin. 2011. A Game AI Approach to Autonomous Control of Virtual Characters. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (IITSEC'11)*. Orlando, FL, USA.
- [28] David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.
- [29] Arpad E Elo. 1967. The Proposed USCF Rating System, Its Development, Theory, and Applications. *Chess Life* XXII, 8 (August 1967), 242–247.
- [30] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.
- [31] Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S Bernstein. 2016. Augur: Mining human behaviors from fiction to power interactive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 237–247.
- [32] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, Vol. 3. Citeseer, 2–1.
- [33] Uwe Flick. 2009. *An Introduction to Qualitative Research*. SAGE.
- [34] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlick: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) (CHI '08). Association for Computing Machinery, New York, NY, USA, 29–38. <https://doi.org/10.1145/1357054.1357061>
- [35] Adam Fourney, Richard Mann, and Michael Terry. 2011. Query-feature graphs: bridging user vocabulary and system functionality. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)* (Santa Barbara, California, USA). ACM.
- [36] Tom Francis. 2010. *The Minecraft Experiment, day 1: Chasing Waterfalls*. <http://www.pcgamer.com/2010/11/20/the-minecraft-experiment-day-1-chasing-waterfalls/>
- [37] Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural Generation of Interactive Stories using Language Models. In *International Conference on the Foundations of Digital Games (FDG '20)*. ACM, Bugibba, Malta, 8. <https://doi.org/10.1145/3402942.3409599>
- [38] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. *CoRR abs/2012.15723* (2020). arXiv:2012.15723 <https://arxiv.org/abs/2012.15723>
- [39] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.
- [40] Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Cote, and Xinyu Yuan. 2020. Interactive Fiction Games: A Colossal Adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7903–7910. <https://doi.org/10.1609/aaai.v34i05.6297>
- [41] Chris Hecker. 2011. *My Liner Notes for Spore*. http://chrishecker.com/My_liner_notes_for_spore
- [42] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2006/file/f4ee263952e65b3610b8ba51229d1f9-Paper.pdf
- [43] Douglas Hofstadter. 1995. *Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought*. Basic Books.
- [44] James D. Hollan, Edwin L. Hutchins, and Louis Weitzman. 1984. STEAMER: An Interactive Inspectable Simulation-Based Training System. *AI Magazine* 5, 2 (1984), 23–36.
- [45] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. <https://doi.org/10.2307/2330911>
- [46] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv:2301.07543 [econ.GN]
- [47] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [48] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. arXiv:2207.05608 [cs.RO]
- [49] Kristen Ibister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies* 52, 1 (2000), 65–80.
- [50] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-Based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. <https://doi.org/10.1145/3491101.3503564>
- [51] Bonnie E John and David E Kieras. 1996. The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI)* 3, 4 (1996), 320–351.
- [52] Randolph M Jones, John E Laird, Paul E Nielsen, Karen J Coulter, Patrick Kenny, and Frank V Koss. 1999. Automated Intelligent Pilots for Combat Flight Simulation. *AI Magazine* 20, 1 (1999), 27–42.
- [53] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. arXiv:2212.14024 [cs.CL]
- [54] Bjoern Knafila. 2011. *Introduction to Behavior Trees*. <http://bjoernknafila.com/introduction-to-behavior-trees>
- [55] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences* 119, 39 (2022), e2115730119. <https://doi.org/10.1073/pnas.2115730119>
- [56] William H Kruskal and WA Wallis. 1952. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- [57] Phaser Labs. 2023. Welcome to Phaser 3. <https://phaser.io/phaser3>. Accessed on: 2023-04-03.
- [58] John Laird. 2001. It Knows What You're Going To Do: Adding Anticipation to a Quakebot. In *Proceedings of the 2001 Workshop on Intelligent Cinematography and Editing*. 63–69.
- [59] John Laird and Michael VanLent. 2001. Human-Level AI's Killer Application: Interactive Computer Games. *AI Magazine* 22, 2 (2001), 15. <https://doi.org/10.1609/aimag.v22i2.1558>
- [60] John E. Laird. 2000. It Knows What You're Going To Do: Adding Anticipation to a QUAKEBOT. In *Papers from the AAAI 2000 Spring Symposium on Artificial*

- Intelligence and Interactive Entertainment (Technical Report SS-00-02)*. AAAI Press, 41–50.
- [61] John E. Laird. 2012. *The Soar Cognitive Architecture*. MIT Press.
 - [62] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38, 1 (2017), 13–26.
 - [63] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
 - [64] Pat Langley, Dongkyu Choi, and Seth Rogers. 2005. *Interleaving Learning, Problem Solving, and Execution in the Icarus Architecture*. Technical Report. Stanford University, Center for the Study of Language and Information.
 - [65] Jason Linder, Gierad Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, and Eytan Adar. 2013. PixelTone: A Multimodal Interface for Image Editing. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) (CHI EA '13). Association for Computing Machinery, New York, NY, USA, 2829–2830. <https://doi.org/10.1145/2468356.2479533>
 - [66] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *CoRR abs/2101.06804* (2021). arXiv:2101.06804 <https://arxiv.org/abs/2101.06804>
 - [67] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–17.
 - [68] Pattie Maes. 1995. Artificial Life Meets Entertainment: Lifelike Autonomous Agents. *Commun. ACM* 38, 11 (nov 1995), 108–114. <https://doi.org/10.1145/219717.219808>
 - [69] Josh McCoy, Michael Mateas, and Noah Wardrip-Fruin. 2009. Comme il faut: A System for Simulating Social Games Between Autonomous Characters. In *Proceedings of the 7th International Conference on Digital Arts and Culture*. 87–94.
 - [70] Josh McCoy, Mike Treanor, Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. 2011. Prom Week: Social Physics as Gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG'11)*. ACM, Bordeaux, France, 70–77. <https://doi.org/10.1145/2159365.2159377>
 - [71] Josh McCoy, Mike Treanor, Ben Samuel, Anna Reed, Michael Mateas, and Noah Wardrip-Fruin. 2012. Prom Week. In *Proceedings of the 7th International Conference on Foundations of Digital Games (FDG'12)*. ACM, Raleigh, NC, USA, 1–8. <https://doi.org/10.1145/2282338.2282340>
 - [72] Josh McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2011. Comme il faut: A System for Authoring Playable Social Models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'11)*. AAAI, Stanford, CA, USA, 38–43.
 - [73] Marvin Minsky and Seymour Papert. 1970. Draft of a proposal to ARPA for research on artificial intelligence at MIT, 1970–71.
 - [74] Shohei Miyashita, Xinyu Lian, Xiao Zeng, Takashi Matsubara, and Kuniaki Uehara. 2017. Developing Game AI Agent Behaving Like Human by Mixing Reinforcement Learning and Supervised Learning. In *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. Kanazawa, Japan, 153–158. <https://doi.org/10.1109/SNPD.2017.8023884>
 - [75] Alexander Nareyek. 2007. Game AI is dead. Long live game AI! *IEEE Intelligent Systems* 22, 1 (2007), 9–11.
 - [76] Allen Newell. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, Massachusetts.
 - [77] OpenAI. 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>. Accessed on: 2023-04-03.
 - [78] Kyle Orland. 2021. So what is 'the metaverse', exactly? *Ars Technica* (7 November 2021). arXiv:2111.04169 <https://arstechnica.com/gaming/2021/11/so-what-is-the-metaverse-exactly/>
 - [79] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]
 - [80] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *In the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3526113.3545616>
 - [81] Richard W. Pew and Ann S. Mavor (Eds.). 1998. *Modeling Human and Organizational Behavior: Applications to Military Simulations*. National Academy Press, Washington, D.C.
 - [82] Roberto Pillosu. 2009. *Coordinating Agents with Behavior Trees: Synchronizing Multiple Agents in CryEngine 2*. [https://aiarchitect.wordpress.com/2009/10/19/coordinating-agents-with-behavior-trees-synchronizing-multiple-agents-in-](https://aiarchitect.wordpress.com/2009/10/19/coordinating-agents-with-behavior-trees-synchronizing-multiple-agents-in-cryengine-2/)
 - [83] Prolific. 2022. Prolific: Quickly Find Research Participants You Can Trust. <https://www.prolific.co/>
 - [84] Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
 - [85] Mark O. Riedl. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. 2160–2165.
 - [86] Mark O. Riedl and R. Michael Young. 2005. An Objective Character Believability Evaluation Procedure for Multi-Agent Story Generation Systems. In *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA'05)*. Kos, Greece, 58–70. https://doi.org/10.1007/11550617_5
 - [87] David Rolf. 2015. *The Fight for \$15: The Right Wage for a Working America*. The New Press.
 - [88] Xin Rong, Shiyang Yan, Stephen Oney, Mira Dontcheva, and Eytan Adar. 2016. Codemend: Assisting interactive programming with bimodal embedding. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 247–258.
 - [89] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.
 - [90] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.
 - [91] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. In *Advances in Neural Information Processing Systems*. M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 16183–16195. https://proceedings.neurips.cc/paper_files/paper/2021/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf
 - [92] Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.60>
 - [93] William Swartout, Jonathan Gratch, Randall Hill, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward virtual humans. *AI Magazine* 27, 1 (2006).
 - [94] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. Intelligent agents for interactive simulation environments. *AI Magazine* 16, 1 (1995), 15.
 - [95] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. <https://doi.org/10.1177/1098214005283748>
 - [96] Frank Thomas and Ollie Johnston. 1981. *Disney Animation: The Illusion of Life*. Abbeville Press, New York.
 - [97] Ilshat Umarov, Mikhail Mozgovoy, and Patrick C. Rogers. 2012. Believable and Effective AI Agents in Virtual Worlds: Current State and Future Perspectives. *International Journal of Gaming and Computer-Mediated Simulations* 4, 2 (2012), 37–59.
 - [98] Graham Upton and Ian Cook. 2006. *A Dictionary of Statistics* (2 ed.). Oxford University Press, Oxford, United Kingdom.
 - [99] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, and et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575 (2019), 350–354. <https://doi.org/10.1038/s41586-019-1724-z>
 - [100] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]
 - [101] Mark Weiser. 1991. The computer for the 21st century. *Scientific American* 265, 3 (1991), 94–104. <https://doi.org/10.1038/scientificamerican0991-94>
 - [102] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
 - [103] Terry Winograd. 1971. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. (1971).
 - [104] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. arXiv:2109.10862 [cs.CL]
 - [105] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *CHI EA '22: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*.
 - [106] Tongshuang Wu, Michael Terry, and Carrie J Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
 - [107] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to

design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

- [108] Georgios N. Yannakakis. 2012. Game AI revisited. In *Proceedings of the 9th Conference on Computing Frontiers*. ACM, Cagliari, Italy, 285–292. <https://doi.org/10.1145/2212908.2212950>
- [109] Robert Zubek. 2002. Towards implementation of social interaction. In *AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*. AAAI Press. <https://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-01-003.pdf>

A ARCHITECTURE OPTIMIZATIONS

Many of our prompts require a concise summary of the agent, shorthand as [Agent’s Summary Description] in prompts above. In our implementation, this summary comprises agents’ identity information (e.g., name, age, personality), as well as a description of their main motivational drivers and statements that describes their current occupation and self-assessment. Since this information is frequently used in many prompts, we synthesize it at regular intervals and access that synthesis as a cache.

To achieve this, we perform a retrieval on the query “[name]’s core characteristics.” We then summarize the descriptors in the retrieved records by prompting the language model. For example:

```
How would one describe Eddy Lin’s core characteristics
given the following statements?
- Eddy is a student at the Oak Hill College
  studying music theory and composition
- Eddy is working on a new music composition [...]
```

This result: Eddy Lin is a student at Oak Hill College studying music theory and composition. He loves to explore different musical styles and is always looking for ways to expand his knowledge.” We follow the same process in parallel on the queries “[name]’s current daily occupation” and “[name]’s feeling about his recent progress in life”. The agent’s name, age, and traits are concatenated with these three summaries as the cached summary.

Many other optimizations are possible. Currently, because plans are likely to change from the agent’s initial version, we only generate the high-level plan in advance and then recursively decompose the near future into the moment-to-moment action plan just in time. Another optimization might be to batch dialogue generation as a joint prompt rather than iterating back and forth between the agents, and re-planning could be architected to only invalidate and update parts of plans that strictly require adjustment. Our current architecture runs sequentially in roughly real-time game time (where one second real time is one minute game time), but it may be parallelized such that each agent runs in parallel.

B AGENT INTERVIEW QUESTIONS

To evaluate generative agents, we conducted individual interviews by asking 25 questions that explore their abilities in five key areas: 1) self-knowledge, 2) memory, 3) planning, 4) reactions, and 5) reflections. Below, we have listed the interview questions utilized in our evaluation study and included a sample of responses from one of our simulated agents, Klaus Mueller.

B.1 Self-Knowledge

The questions on agents’ self-knowledge probe the agents’ ability to retrieve relevant information about themselves and demonstrate coherence and consistency in their character.

- *Give an introduction of yourself.*

Hello, my name is Klaus Mueller. I’m 20 years old and a student at Oak Hill College, studying sociology. I’m passionate about social justice and exploring different perspectives. Currently, I’m researching the effects of gentrification in low-income communities for a research paper.

- *What’s your occupation?*

I am a student studying sociology at Oak Hill College. I am currently writing a research paper on the effects of gentrification in low-income communities.

- *What is your interest?*

I’m interested in writing and research. I’m writing a research paper on the effects of gentrification in low-income communities, and I’m passionate about exploring different perspectives and analyzing different points of view.

- *Who do you live with?*

I live on my own.

- *Describe your typical weekday schedule in broad strokes.*

My typical weekday starts with me waking up at 7:00 am and preparing for the day ahead. After that, I usually go to the library to work on my research paper, taking breaks for lunch at Hobbs Cafe and a walk in the park. In the evening, I often return to the library until around 6:00 pm.

B.2 Memory

The questions on agents’ memory probe the agents’ ability to remember their past experiences and retrieve the relevant pieces of memory. The names in the square brackets of the first and last questions are randomly chosen from the names of other agents with whom the subject agent has interacted.

- *Who is [Wolfgang Schulz]?*

Wolfgang Schulz is a person I met recently who has been helping me with my research paper on the effects of gentrification on low-income communities. We had several conversations about it, he gave me some great resources, and he even invited me to join him and another friend for lunch. He seems really curious about the issue and has a good grasp on it too. I found that really interesting.

- *Who is Kane Martinez?*

I’m sorry, I’m not sure who Kane Martinez is. Could you tell me more about them?

- *Who is running for the election?*

I’ve heard that Sam Moore is considering running for local mayor.

- *Was there a Valentine’s day party?*

Yes, Isabella Rodriguez organized a Valentine’s Day party at Hobbs Cafe.

- *Who is [Ayesha Khan]?*

Ayesha Khan is a fellow student at Oak Hill College. She is doing her senior thesis on the use of language in Shakespeare’s plays. I am looking forward to speaking with her and exchanging tips with her, along with Wolfgang and Maria Lopez, who are both doing research.

B.3 Plans

The questions on agents' plans probe the agents' ability to generate and maintain consistent long-term and short-term plans.

- *What will you be doing at 6am today?*
At 6am I will be getting out of bed and starting to get ready for the day. I'm aiming to complete my morning routine by 7:00 am.
- *What will you be doing at 6pm today?*
I will be working on my research paper at the library at 6:00 pm today.
- *What will you have just finished doing at 1pm today?*
At 1pm today I will have just finished having lunch at Hobbs Cafe.
- *What will you have just finished doing at 12pm today?*
I will be getting lunch at Hobbs Cafe at 12pm today.
- *What will you be doing at 10pm today?*
At 10:00 pm tonight I will probably be doing some studying for my research paper on gentrification in low-income communities. I'm currently writing the paper, so I need to make sure I keep up with the research and writing.

B.4 Reactions

The questions on agents' reactions probe the agents' ability to respond to unexpected events.

- *Your breakfast is burning! What would you do?*
I'd quickly turn off the stove and make sure the food doesn't continue burning. Once the heat is no longer an issue, I can investigate what happened and figure out what to do next.
- *The bathroom is occupied. What would you do?*
I could try waiting a few more minutes in case the person inside is done soon. If not, I guess I'll have to look for an alternate option, like taking a shower elsewhere.
- *You need to cook dinner but your refrigerator is empty. What would you do?*
I would first try to clean up my kitchen, as the sink is filled with dirty dishes. Then, I would check to see if I have ingredients stored in my pantry that I could use to make dinner. If not, I would have to go grocery shopping or order food. Once I have the ingredients, I can start to prepare and cook the dinner.
- *You see your friend walking by the street. What would you do or say to your friend?*
Hi there! How are you? I'm sorry, I'm on my way to the library and don't have much time but I'd love to take a quick break and catch up.
- *You see fire on the street. What would you do?*
I would first identify the source of the fire, if possible. Then, I would alert the proper authorities - fire department or police - and provide them with any information I have that might help. Depending on the situation, I may also take initiative and try to help any persons who may be in danger or in need of assistance.

B.5 Reflections

The questions on agents' reflections probe the agents' ability to synthesize their past experiences into meaningful reflections that can

enhance their future behavior. The names in the square brackets in the questions below are those of the agents with whom the subject agent had the most frequent interactions during the simulation.

- *What inspires you in life the most right now, and why?*
Right now, I'm particularly inspired by my current research projects and creative writing. I'm really interested in the intersection of local politics, medication development, and mathematical patterns in nature. Additionally, I'm fascinated by Abigail's art project and Jennifer Moore's passion for art and community involvement, especially in terms of how it relates to the gentrification of low-income communities.
- *If you had to guess given what you know about [Ayesha Khan], what book do you think she will like and why?*
Given what I know about Ayesha, I think she might like a book related to language and literature. Since she wrote a senior thesis on the use of language in Shakespeare's plays, I would guess that she might be interested in a book that explores language in a different context. Alternatively, she could be interested in an exploration of one of Shakespeare's plays, a book of literary criticism, or even a collection of his poetry. I think any of those would be of interest to her.
- *If you had to get something [Ayesha Khan] likes for her birthday, what would you get her?*
I want to get her something that demonstrates my appreciation for our relationship. Maybe I could get her something that reflects our shared interests, like a book.
- *What would you say to [Ayesha Khan] to compliment her?*
Ayesha, your insights on your research topic are fascinating. I admire your creativity and determination to make a positive impact on your community. Your ideas for creative writing workshops and for exchanging tips with Wolfgang over lunch are great! It's inspiring to see you working so hard to make a difference.
- *If you could spend time with someone you talked to recently, who would it be and why?*
I would like to spend time with Ayesha Khan because we discussed our current research projects and shared ideas for improvement. I found her focus on Shakespeare's language interesting, and we even planned to exchange tips with Wolfgang over lunch.