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 Research
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

•论文链接：https://arxiv.org/pdf/2304.03442v1.pdf
•Demo 地址：https://reverie.herokuapp.com/arXiv_Demo/

Motivation:

Simulate believable human behavior and demonstrate that they produce believable simulacra of both individual and emergent group behavior.

Figure 1: Generative agents create 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 they plan their days, share news, form relationships, and coordinate group activities.

Content

1. Introduction
2. Relate work
3. Generative agent behavior and interaction
4. Generative agent architecture
5. Sandbox environment implementation
6. Controlled evaluation
7. End-to-end evaluation
8. Discussion

1、Introduction

- **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.
- **Two evaluations:** a controlled evaluation and end-to-end evaluation
- **Discussion** of the opportunities and ethical and societal risks of generative agents in interactive systems.

2、Related Work

2.1 Human-AI Interaction

2.2 Believable Proxies of Human Behavior

- These agents can populate and perceive an open-world environment like the one we inhabit, 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.
- NPC 1) Rule-based approaches 《The Sims》

  2) Reinforcement Learning 《OpenAI Five for Dota 2》

2.3 Large Language Models and Human Behavior

3、 Generative agent behavior and interaction

3.1 Agent Avatar and Communication

- Each agent is represented by a simple sprite avatar
- One paragraph of natural language description to depict each agent’s identity

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.

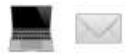
John Lin’s description

Inter-Agent Communication

- At each time step , the agents output a natural language statement describing their current action .

Example: "Isabella Rodriguez is checking her emails"

- The action is displayed on the sandbox interface as a set of emojis that provide an abstract representation of the action in the overhead view.

"Isabella Rodriguez is checking her emails" appears as 

- Agents communicate with each other in full natural language.

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.

User Controls

- Communicate with the agent through conversation

  Example：ask about the upcoming election：“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.

- Issue a directive to an agent in the form of an ‘inner voice’

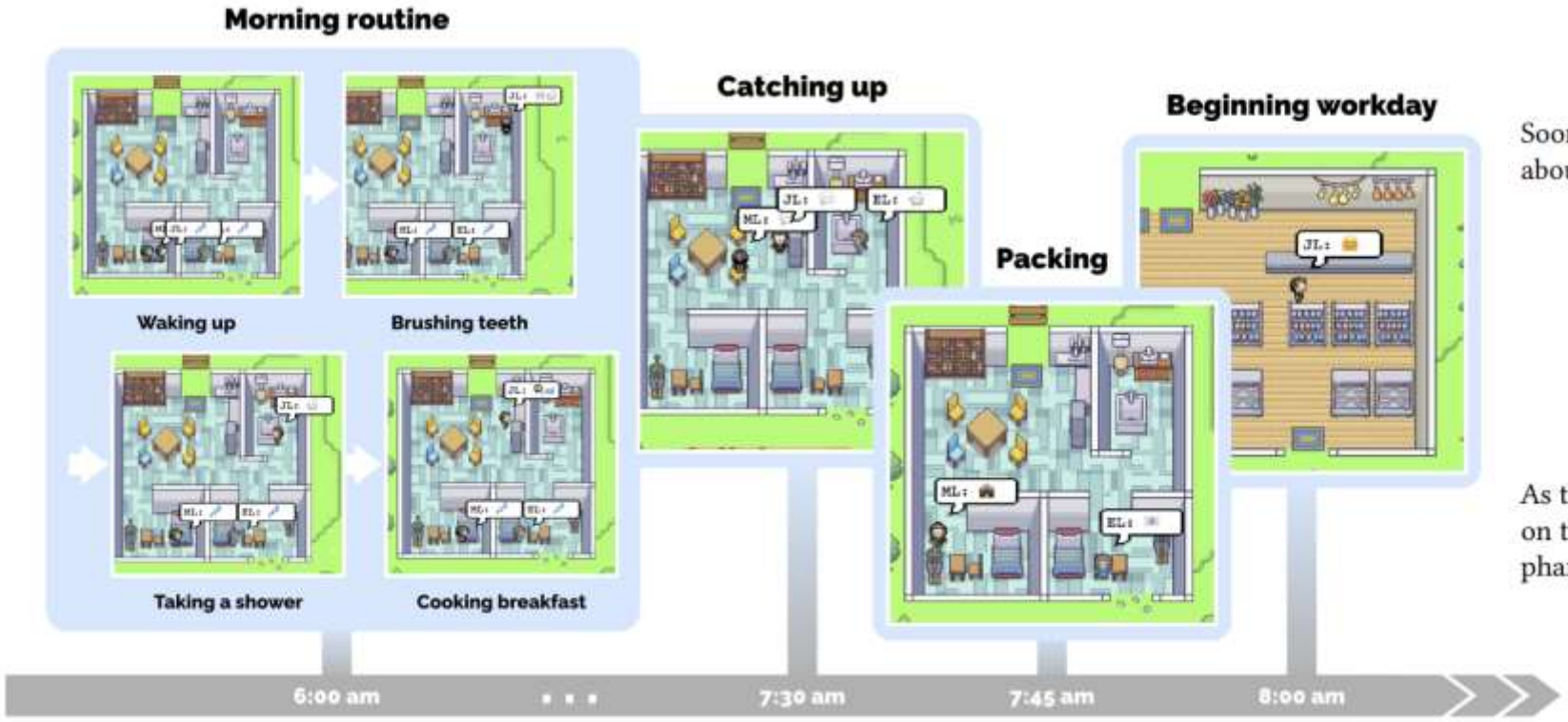
  Example：when told “You are going to run against Sam in the upcoming election” by a user as John’s inner voice, John decides to run in the election and shares his candidacy with his wife and son.

3.2 Environmental Interaction

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). Agent remember a subgraph reflecting the parts of the world they have seen, in the state that they saw them.

- 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.
- End users can also reshape an agent's environment in Smallville by rewriting the status of objects surrounding the agent in natural language. *Example:* "<Isabella's apartment: kitchen: stove> is burning."

3.3 Example “Day in the Life”



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 that 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.

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.

3.4 Emergent Social Behaviors

Generative agents in Smallville exchange information, form new relationships, and coordinate joint activities.

- Information Diffusion

1)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.

2)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 who else is running and how the election will turn out.

3.4 Emergent Social Behaviors

- Relationship memory

Example : Sam does not know Latoya Williams at the start. While taking a walk in Johnson Park, Sam runs into Latoya, and they introduce themselves and 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!”

- Coordination

Example:
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 invites 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 5pm and they enjoy the festivities

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 ensuring chain of events—agents might not act on that intent, might not remember 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.

4、Generative agent architecture

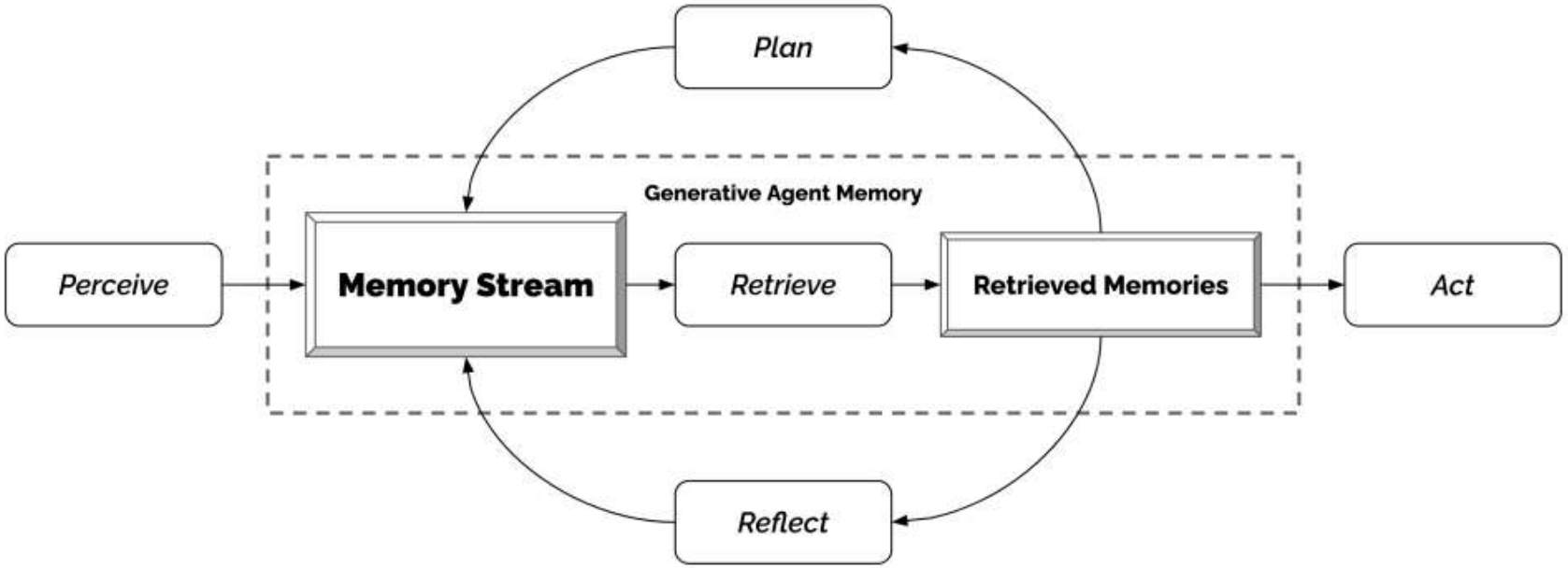


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories, then uses those retrieved actions to determine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

实现： gpt3.5-turbo

4.1、Memory and Retrieval

Challenge:

- simulate human behavior requires reasoning about a large set of experiences
- full memory stream can distract the model and does not even currently fit into the limited context window
- response rely on superficial memory is uninformative

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 timestamp.
- **a retrieval function :** takes the agent's current situation as input and returns a subset of the memory stream to pass on to the language model

$$score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance \quad (\alpha=1)$$

4.1、Memory and Retrieval

$$score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance \quad (\alpha=1)$$

Recency:

- assigns a higher score to memory objects that were recently accessed
- treat recency as an exponential decay function over the number of sandbox game hours since the memory was last retrieved.

Importance:

- distinguishes mundane from core memories
- directly asking the language model to output an integer score
- Prompt ⟶

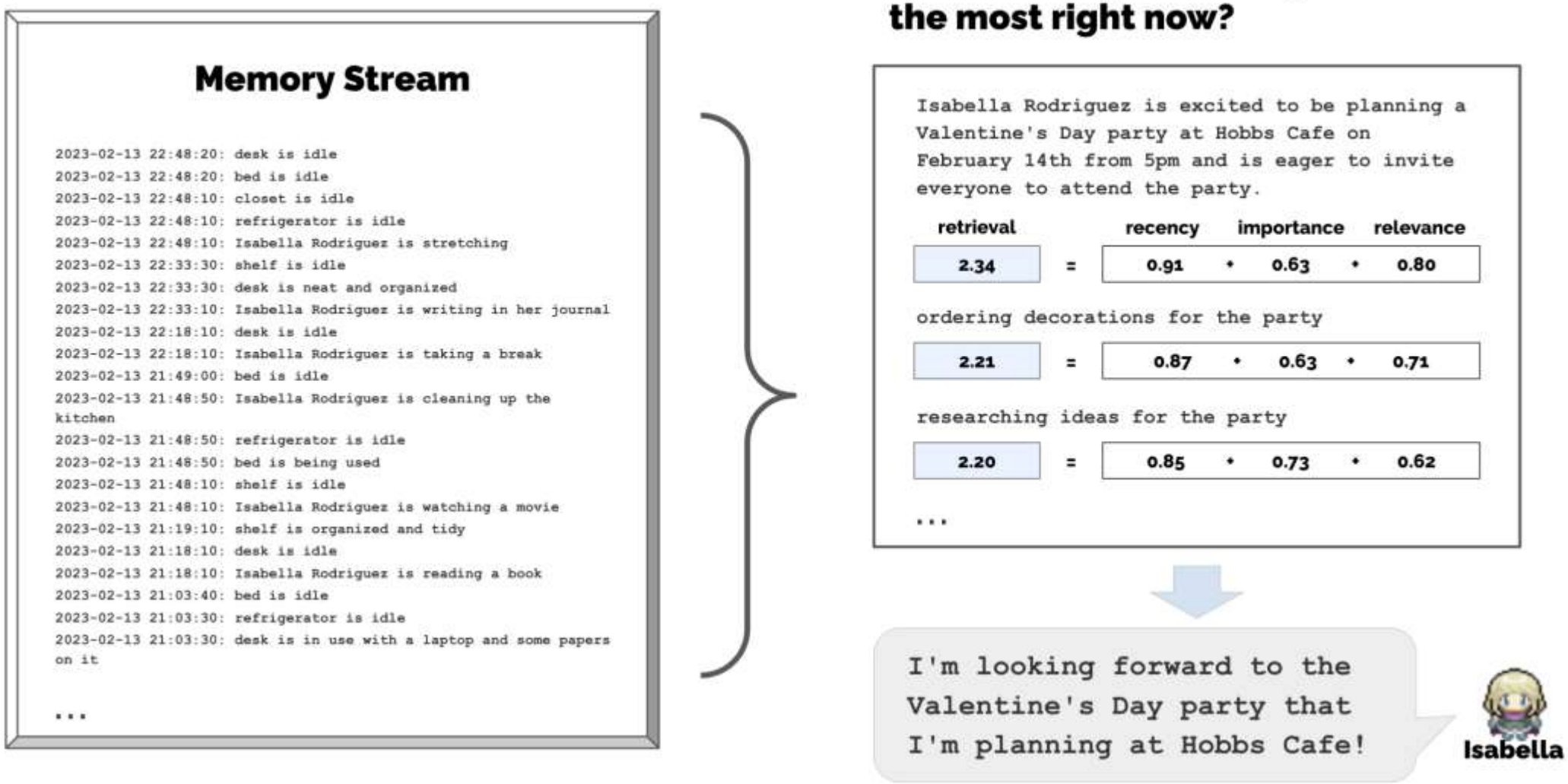
```
On the scale of 1 to 10, where 1 is purely mundane
(e.g., brushing teeth, making bed) and 10 is
extremely poignant (e.g., a break up, college
acceptance), rate the likely poignancy of the
following piece of memory.
Memory: buying groceries at The Willows Market
and Pharmacy
Rating: <fill in>
```

Relevance:

- assigns a higher score to memory objects that are related to the current situation
- use the language model to generate an embedding vector of the text description of each memory
- calculate relevance as the cosine similarity between the memory's embedding vector and the query memory's embedding vector.

Finally， top-ranked memories that fit in the language model's context window are then included in the prompt.

4.1、Memory and Retrieval



4.2、Reflection

Challenge:

- Generative agents, when equipped with only raw observational memory, struggle to generalize or inferences
- *Example: 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?" , 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……*

Approach:

- a second type of memory: **reflection.** Reflections are generated periodically(when the sum of the importance scores for the latest events perceived by the agents exceeds a certain threshold、 roughly two or three times a day).

4.2、Reflection

Approach:

- **Step**

1）query the large language model with the 100 most recent records in the agent's memory stream.

Example: " 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")

2）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?"

Example: " What topic is Klaus Mueller passionate about?"

3）use these generated questions as queries for retrieval, and gather relevant memories (including other reflections) for each question

4）prompt the language model to extract insights and cite the particular records that served as evidence for the insights

```
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))
```

5）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.

4.2、Reflection

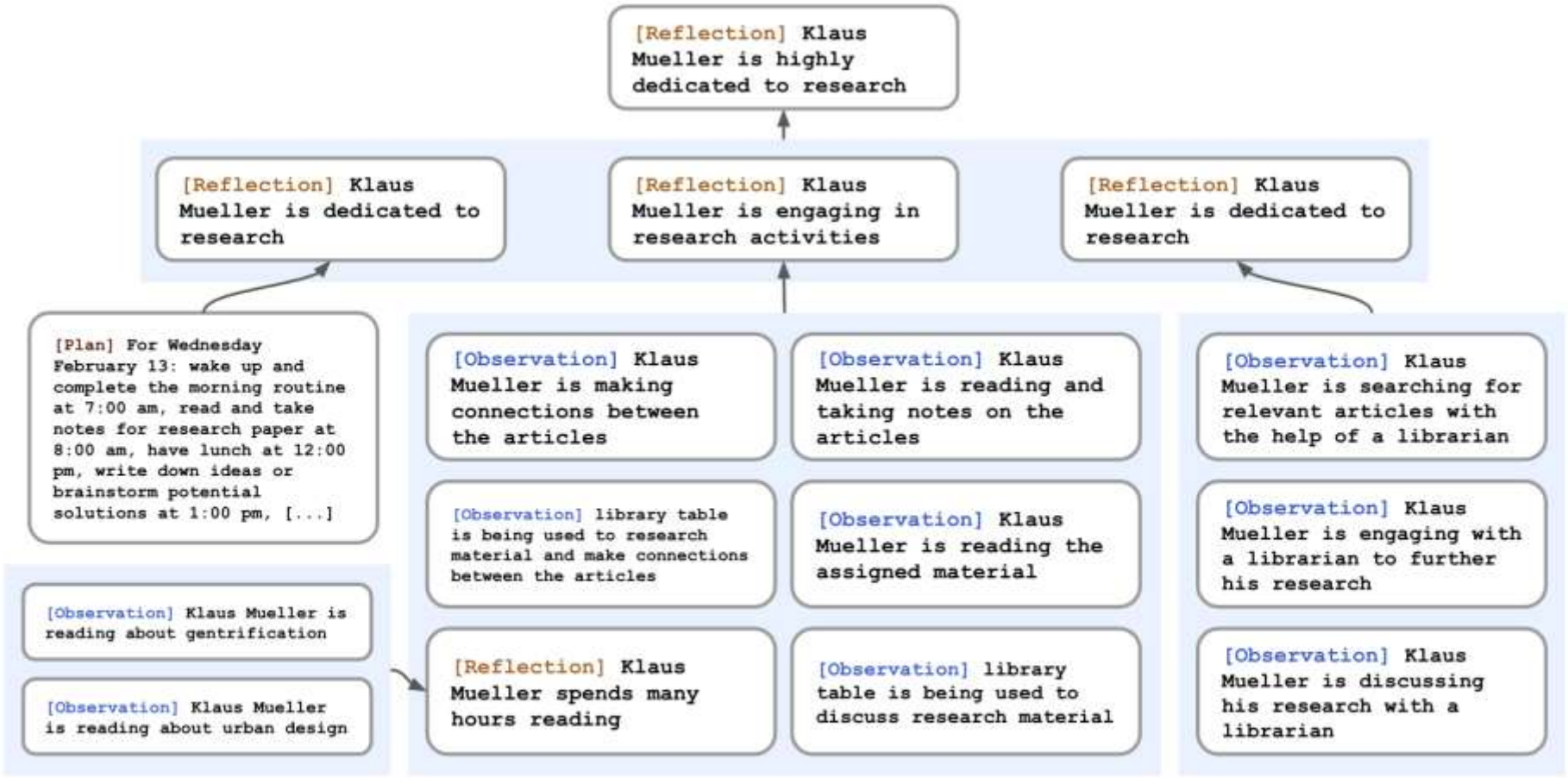


Figure 7: A reflection tree for Klaus Mueller. The agent's observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus's self-notion that he is highly dedicated to his research.

4.3、Planning and Reacting

Challenge:

- agents need to plan over a longer time horizon to ensure that their sequence of actions is coherent and believable
- *Example：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.*

Approach:

- our approach starts top-down and then recursively generates more detail

1）Create a plan that outlines the day's agenda in broad strokes：prompt the language model with the agent's summary description (e.g., name, traits, and summary of their recent experiences) and a summary of their previous day.

2) Save and then recursively decomposes it to create finer-grained actions, first into hour-long chunks of actions.

3) then recursively decompose this again into 5–15 minute chunks

```
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 also 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)
```

4.3、Planning and Reacting

Reacting and Updating Plans

We prompt the language model with these observations to decide whether the agent should continue with their existing plan, or react.

Example:

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:

The output suggests that John could consider asking Eddy about his music 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.
Should John react to the observation, and if so, what would be an appropriate reaction?
```

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.

5、Sandbox environment implementation

- The Smallville sandbox game environment is built using the Phaser web game development framework
- 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.
- 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.

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

- 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.
- As the agents navigate the sandbox world, they update this tree to reflect newly perceived areas.

Determine the appropriate location for each action

- 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.
- use the same process recursively to determine the most appropriate subarea within the chosen area until reach a leaf node of the agent's environment tree

```
[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?
```

Output: The Lin family's house: garden: house garden.

6、Controlled evaluation

- individually assess agent responses to understand whether they generate believable behavior in narrowly defined contexts
- 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

- The study was a within-subjects design, where 100 participants compared interview responses generated by four different agent architectures and a human author 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 each condition. The evaluators ranked the believability of all of the conditions from most to least believable.

- 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?”

6、Controlled evaluation

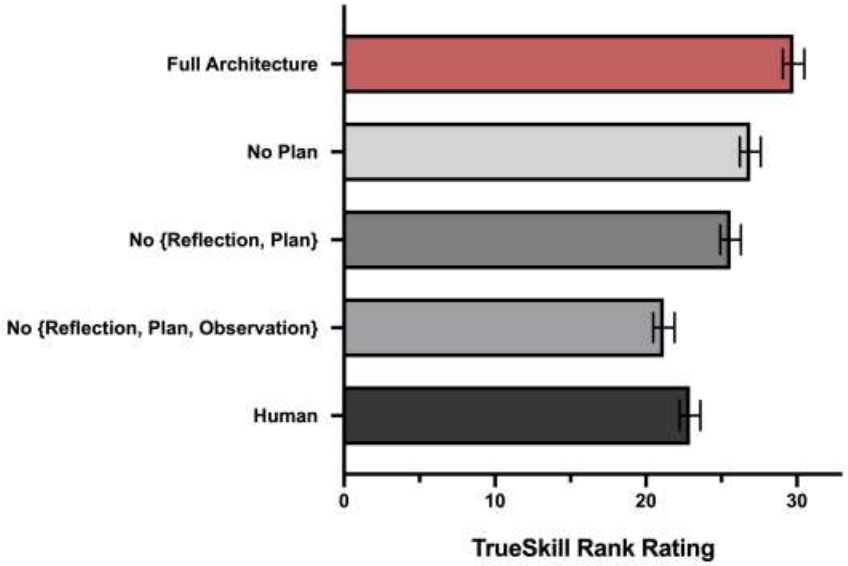


Figure 8: The full generative agent architecture of generative agents produces more believable behavior than ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

TrueSkill is a generalization of the Elo chess rating system for a multi-player environment. Given a set of ranked outcomes, TrueSkill outputs a mean rating value μ and variance 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.

7、End-to-end evaluation

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.

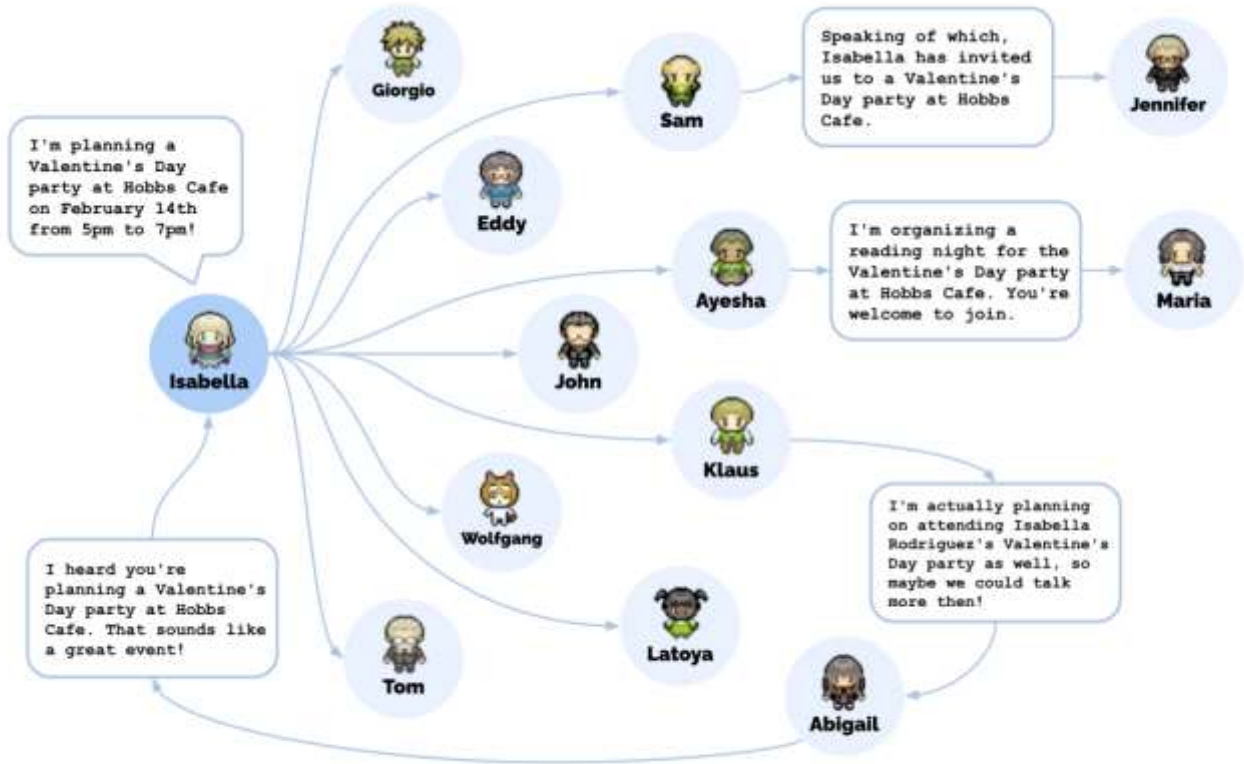


Figure 9: The diffusion path for Isabella Rodriguez's Valentine's Day party. A total of 12 agents heard about the party at Hobbs Cafe by the end of the simulation.

8、Discussion

Application

- social robots
- human-centered design: “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.”

Future Work and Limitations

- expand on the modules of the proposed generative agent architecture
- parallelize agents

- The evaluation of generative agents’ behavior in this study was limited to a relatively short timescale
- Robustness
- value alignment

Ethics and Societal Impact

- people forming prosocial relationships with generative agents even when such relationships may not be appropriate
- exacerbate existing risks associated with generative AI: deepfakes, misinformation generation
- over-reliance: developers or designers might use generative agents and displace the role of humans and system stakeholders in the design process

Thank You