



中國人民大學
RENMIN UNIVERSITY OF CHINA



高瓴人工智能学院
Gaoling School of Artificial Intelligence

Generative Agents: Interactive Simulacra of Human Behavior

**Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris,
Percy Liang and Michael S. Bernstein**

Stanford University, Google Research, Google DeepMind

Shen Yuan

2023-10-12

Outline

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- Discussion
- Conclusion

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- Discussion
- Conclusion





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.



Contributions

- **Generative agents**
- 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.

- Introduction
- **Generative agent behavior and interaction**
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- Discussion
- Conclusion



Agent Avatar and Communication

- A community of 25 unique agents inhabits **Smallville**.
- Each agent is represented by a simple sprite avatar.
- For instance, John Lin has the following description including their occupation and relationship with other agents:



Isabella



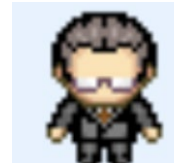
Abigail



Klaus



Eddy







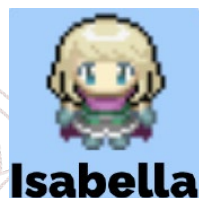
John

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.

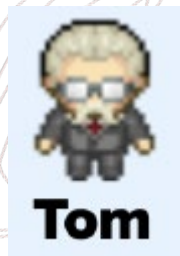


Inter-Agent Communication

- The agents interact with the world by their **actions**, and with each other through **natural language**.
 - “Isabella Rodriguez is writing in her journal” is displayed as  
 - “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

- The user communicates with the agent through natural language by specifying a persona that the agent should perceive them as.
- 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.



Example “Day in the Life”

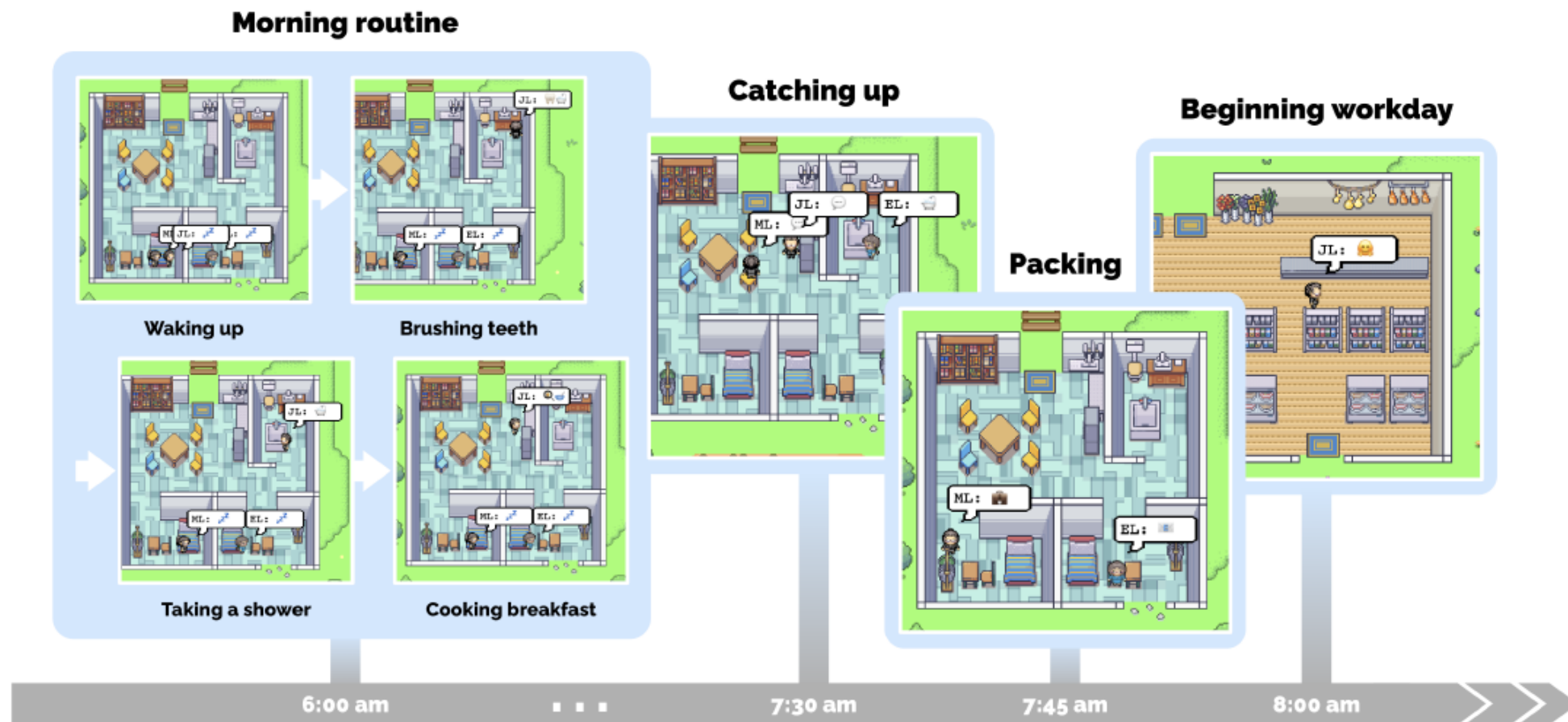


Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.



Emergent Social Behaviors

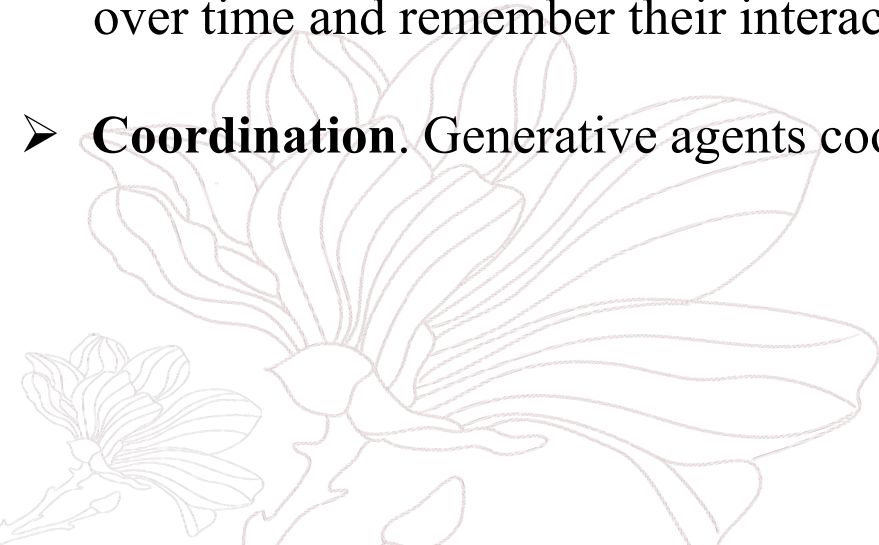


➤ By interacting with each other, generative agents in Smallville could do:

- **Information Diffusion.** Information can spread from agent to agent.
- **Relationship Memory.** Agents in Smallville form new relationships over time and remember their interactions with other agents.
- **Coordination.** Generative agents coordinate with each other.



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.



- Introduction
- Generative agent behavior and interaction
- **Generative agent architecture**
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- Discussion
- Conclusion





Generative Agent Architecture

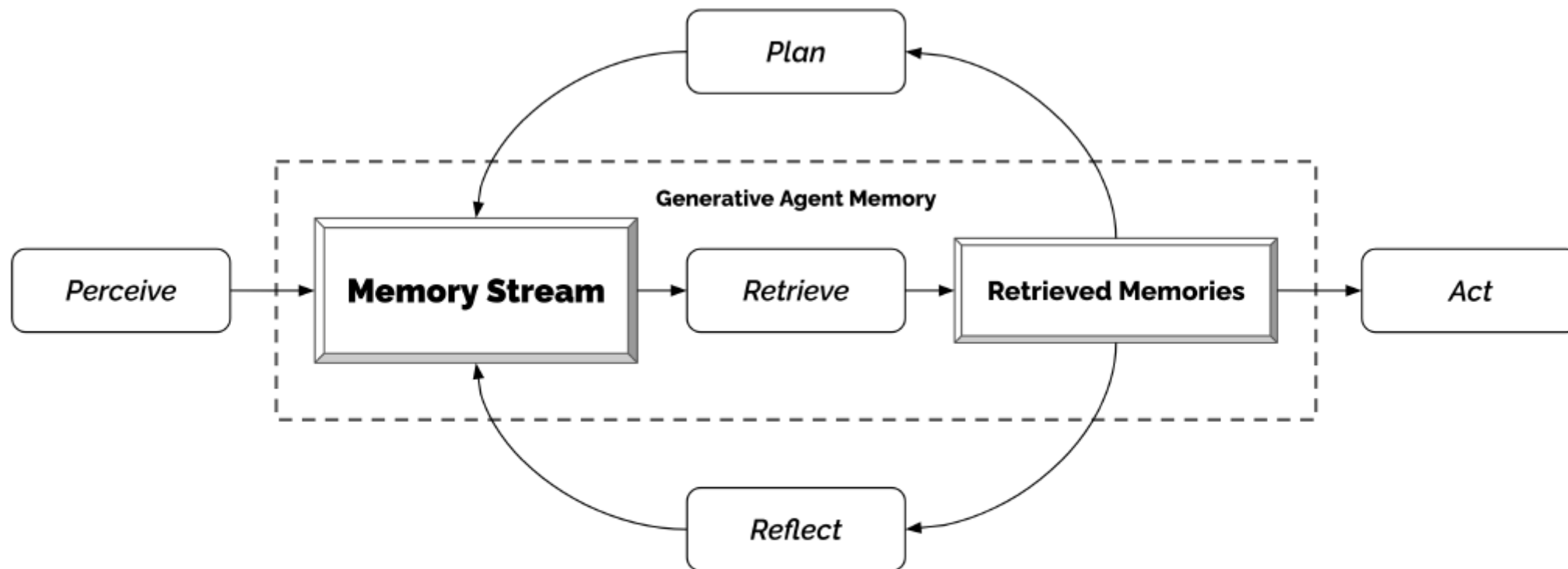


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine 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.



Memory and Retrieval

- **Challenge:** the full memory stream is far larger than what should be described in a prompt, and does not even currently fit into the limited context window.
- **Approach:**
 - **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.
 - **Observation** is the most basic element of the memory stream, which is an event directly perceived by an agent.
 - **Retrieval** that takes the agent's current situation as input and returns **the most relevant** subset of the memory stream to pass on to the language model.

Memory Stream

```

2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10: desk is idle
2023-02-13 22:18:10: Isabella Rodriguez is taking a break
2023-02-13 21:49:00: bed is idle
2023-02-13 21:48:50: Isabella Rodriguez is cleaning up the kitchen
2023-02-13 21:48:50: refrigerator is idle
2023-02-13 21:48:50: bed is being used
2023-02-13 21:48:10: shelf is idle
2023-02-13 21:48:10: Isabella Rodriguez is watching a movie
2023-02-13 21:19:10: shelf is organized and tidy
2023-02-13 21:18:10: desk is idle
2023-02-13 21:18:10: Isabella Rodriguez is reading a book
2023-02-13 21:03:40: bed is idle
2023-02-13 21:03:30: refrigerator is idle
2023-02-13 21:03:30: desk is in use with a laptop and some papers on it
...

```

Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

retrieval	=	recency	importance	relevance
2.34	=	0.91	0.63	0.80

ordering decorations for the party


2.21	=	0.87	0.63	0.71
------	---	------	------	------

researching ideas for the party

2.20	=	0.85	0.73	0.62
------	---	------	------	------

...

I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!



Isabella

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.



Retrieval

- **Recency** assigns a higher score to memory objects that were recently accessed.
- **Importance** distinguishes mundane from core memories by assigning a higher score to memory objects that the agent believes to be important.

- **Relevance** 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. ; that are related

Memory: buying groceries at The Willows Market and Pharmacy
Rating: <fill in>



Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

retrieval		recency	importance	relevance
2.34	=	0.91	• 0.63	• 0.80

ordering decorations for the party

2.21	=	0.87	• 0.63	• 0.71
------	---	------	--------	--------

researching ideas for the party

2.20	=	0.85	• 0.73	• 0.62
------	---	------	--------	--------

...



I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!



Isabella



Reflection



- **Challenge:** Generative agents, when equipped with only raw observational memory, struggle to generalize or make inferences.
- **Approach:**
 - **Reflection**, the second type of memory.
 - Reflections are **higher-level, more abstract** thoughts generated by the agent.
 - Reflections can be included alongside other **observations** when **retrieval** occurs.
 - Reflections are generated **periodically** when the sum of the importance scores for the latest events perceived by the agents exceeds a threshold (150 in our implementation).



Reflection

➤ Steps:

- **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"

- **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.

e.g., What topic is Klaus Mueller passionate about? and What is the relationship between Klaus Mueller and Maria Lopez?

- Use these generated questions to **prompt** the language model to extract 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))

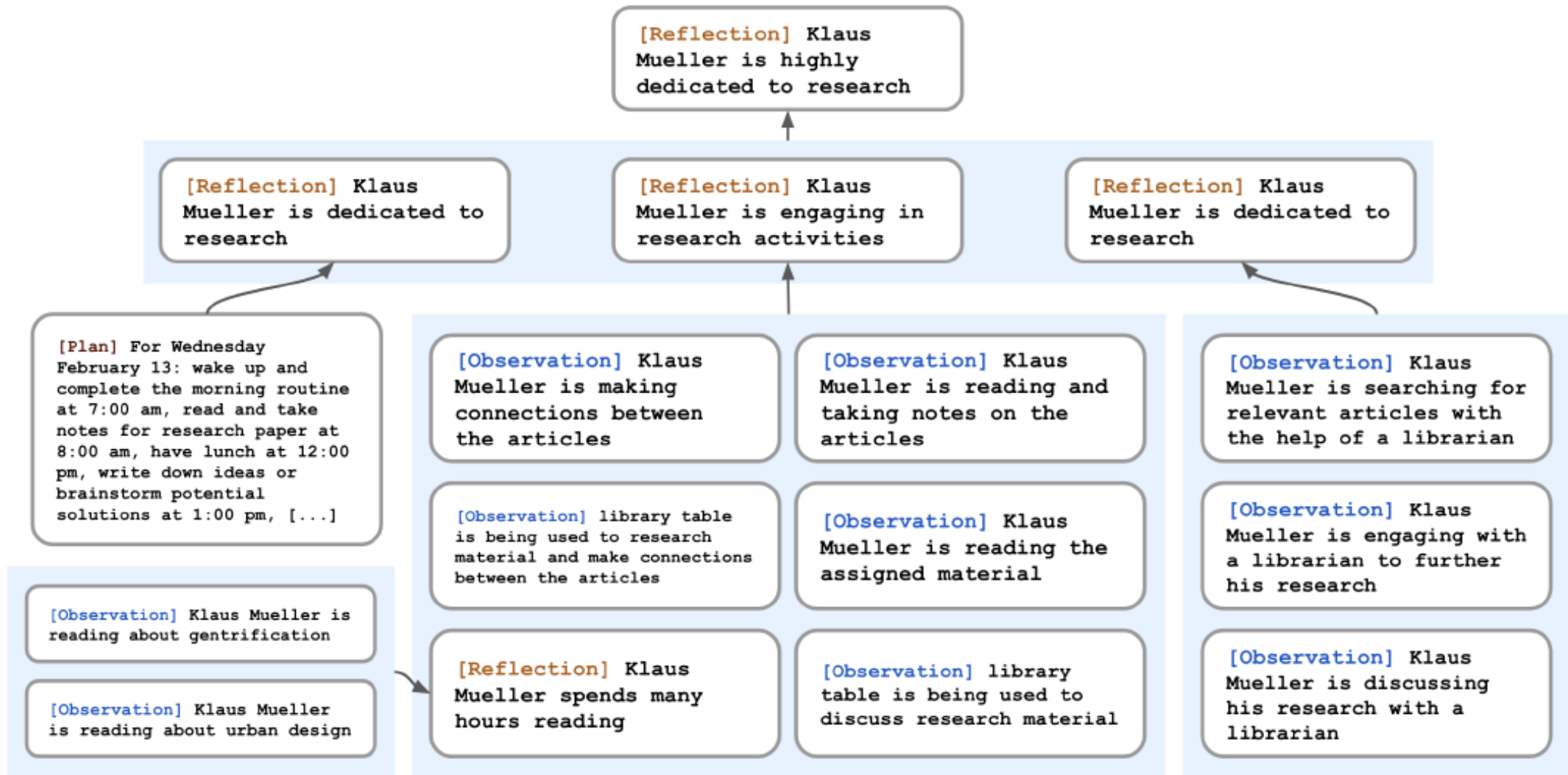


Figure 7: A reflection tree for Klaus Mueller. The agent's observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus's self-notion that he is highly dedicated to his research.



Planning and Reacting

- **Challenge:** agents need to plan over a longer time horizon to ensure that their sequence of actions is coherent and believable.
- **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**.
 - Like reflections, **plans** are stored in the **memory stream** and are included in the **retrieval** process.
 - To create such plans, the approach starts **top-down** and then **recursively** generates more detail.



Planning and Reacting

- To create such plans, our approach starts **top-down** and then **recursively** generates more detail.

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)



Work on his new music composition from 1:00 pm to 5:00 pm



*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.

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- **Sandbox environment implementation**
- Controlled evaluation
- End-to-end evaluation
- Discussion
- Conclusion



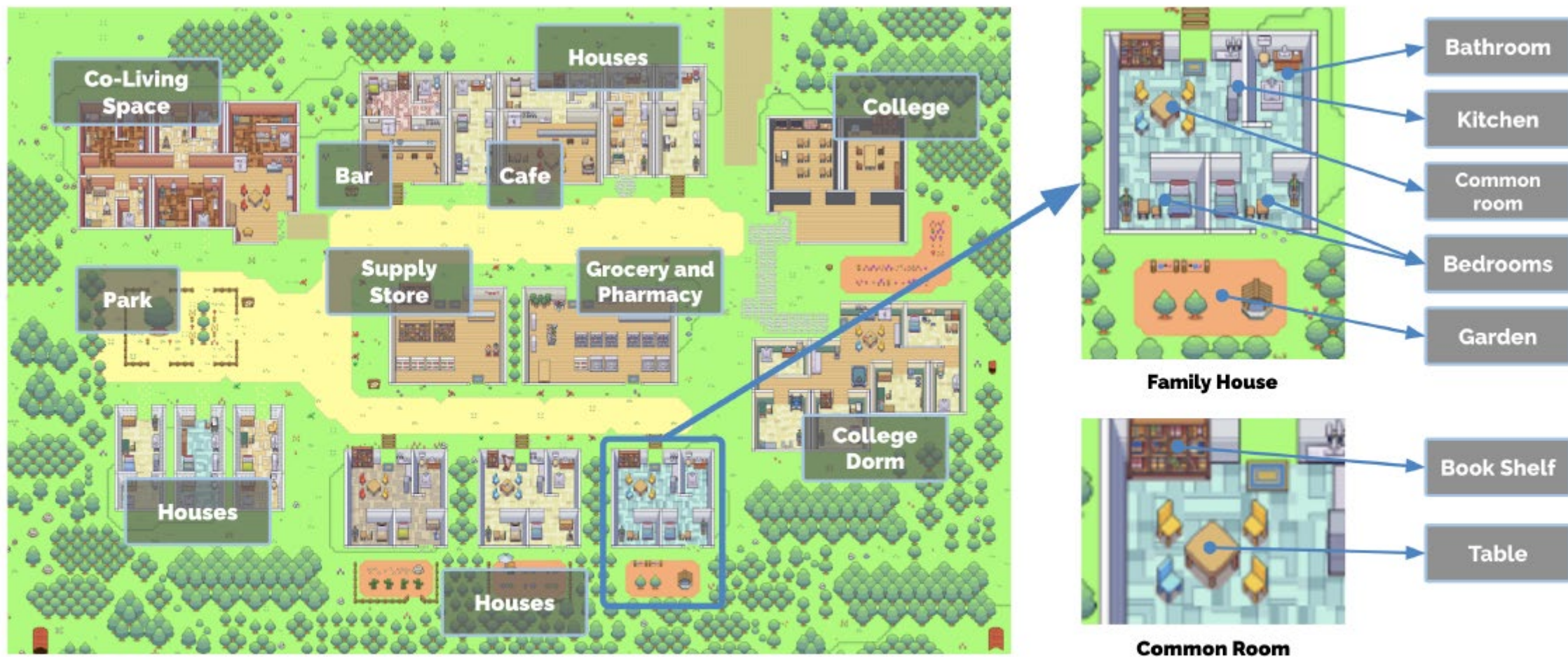




Figure 2: The Smallville sandbox world, with areas labeled. The root node describes the entire world, children describe areas (e.g., houses, cafe, stores), and leaf nodes describe objects (e.g., table, bookshelf). Agents remember a subgraph that reflects the parts of the world they have seen, maintaining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

- 
- 
- Introduction
 - Generative agent behavior and interaction
 - Generative agent architecture
 - Sandbox environment implementation
 - **Controlled evaluation**
 - End-to-end evaluation
 - Discussion
 - Conclusion



Evaluation Procedure

- 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 interview includes five question categories:
 - **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?*”



Results

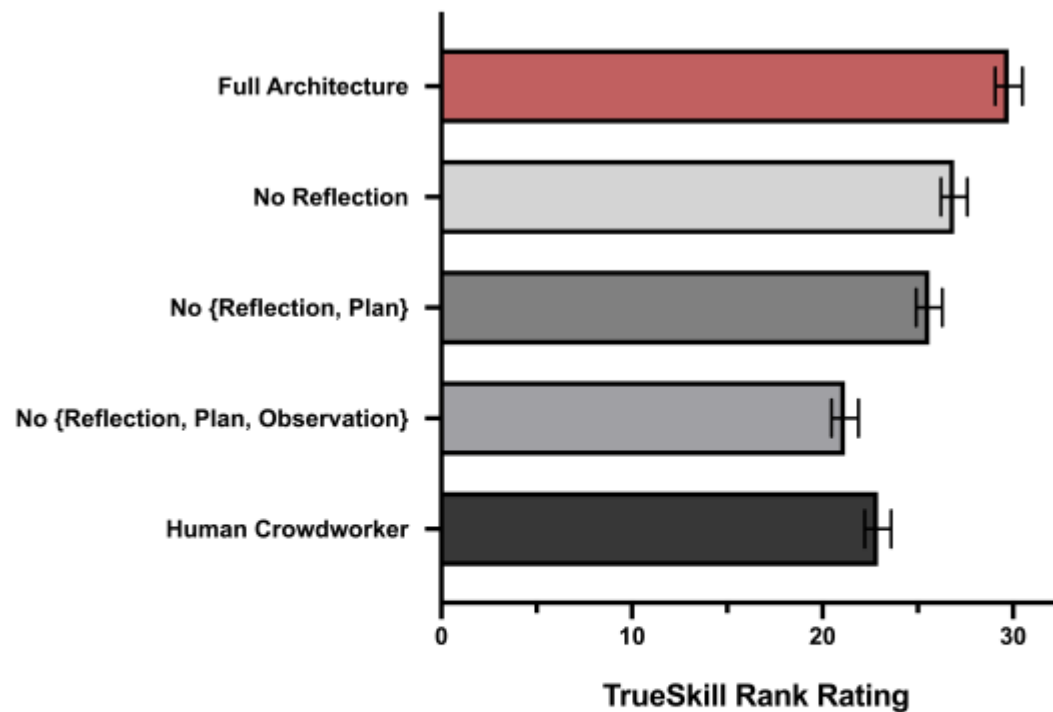


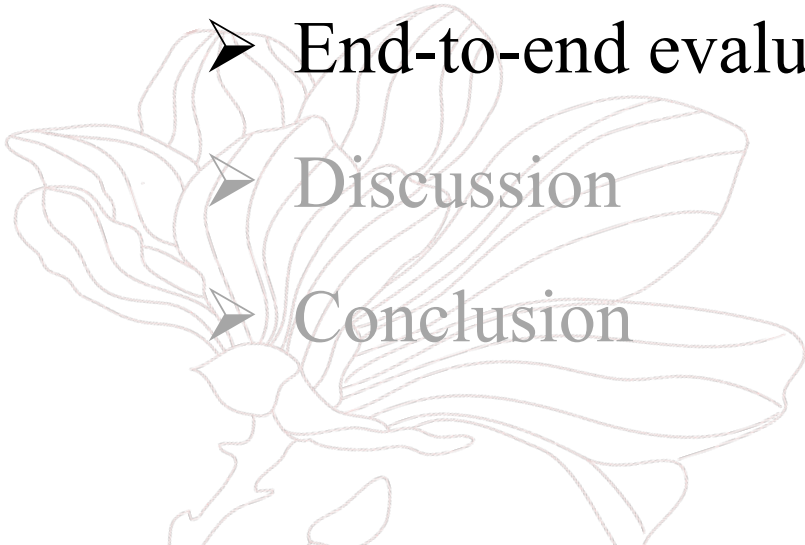
Figure 8: The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.



Problems

- Generative agents maybe **fail to retrieve the correct instances** from their memory.
 - *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!”*
- At times, the agents **hallucinated embellishments** to their knowledge.
 - *Isabella said that “Sam’s going to make an announcement tomorrow”, however, 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.*

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- **End-to-end evaluation**
- Discussion
- Conclusion



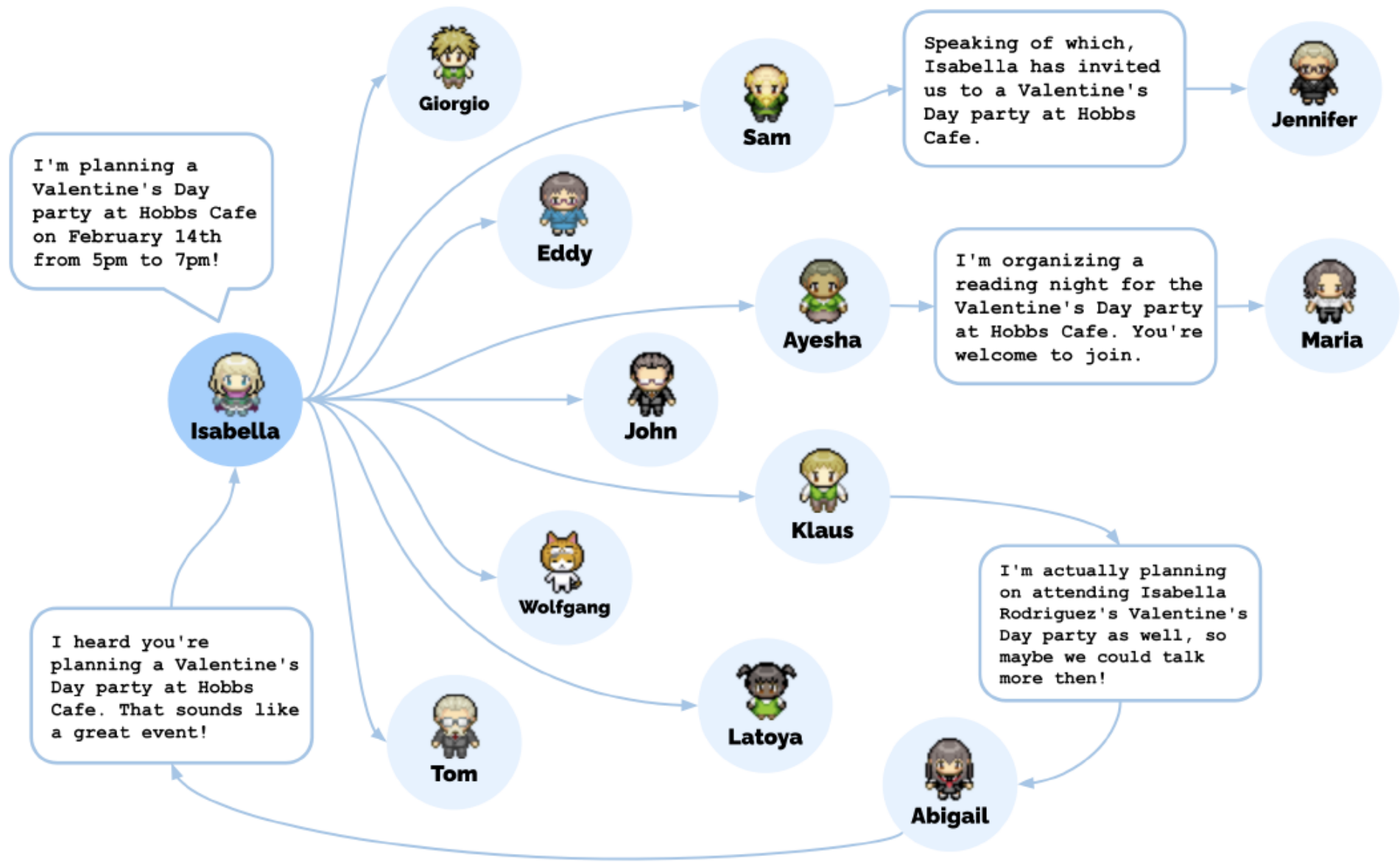


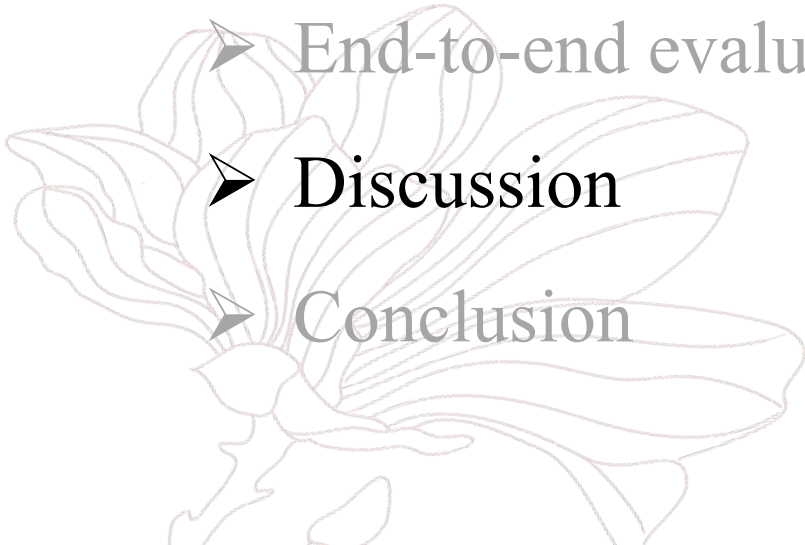
Figure 9: The diffusion path for Isabella Rodriguez's Valentine's Day party invitation involved a total of 12 agents, aside from Isabella, who heard about the party at Hobbs Cafe by the end of the simulation.



Problems

- we noticed erratic behaviors caused by misclassification of what is considered proper behavior, especially when the certain locations that are **hard to describe in natural language**.
 - *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.*
- we observed possible effects of **instruction tuning**, which seemed to guide the behavior of the agents to be **too polite and cooperative** overall.
 - *When Mei talked with her husband John, she often ended the conversation with a **too formal** greeting, “It was good talking to you as always.”*

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- **Discussion**
- Conclusion

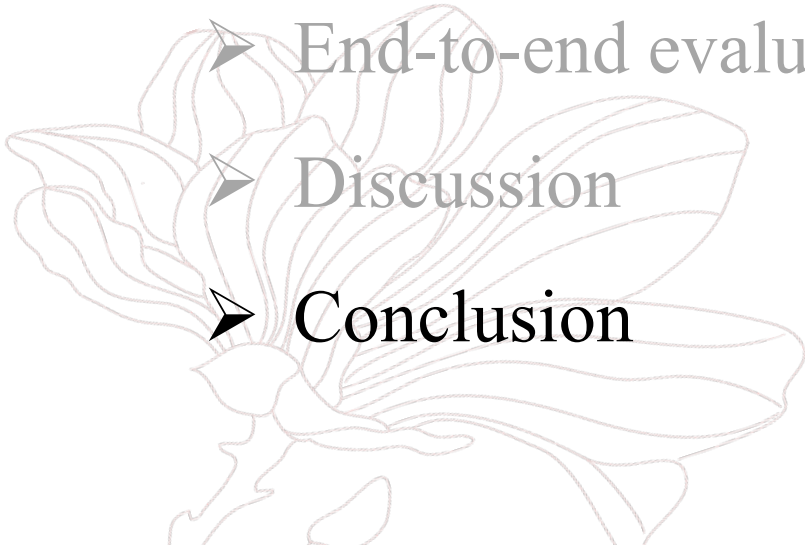




Future Work and Limitations

- **Cost-effective.** The present study simulated 25 agents for two days, costing thousands of dollars in token credits and taking multiple days to complete.
- **More effective performance testing.** 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 benchmark.
- **Robustness.** The robustness of generative agents is still largely unknown. They may be vulnerable to prompt hacking, memory hacking, and hallucination, among other issues.
- **Biases of LLM.** Any imperfections in the underlying LLM will be inherited by generative agents. Given the known biases of language models, generative agents may potentially exhibit biased behavior or stereotypes.
- **OOD.** Generative agents may struggle to generate believable behavior for certain subpopulations, particularly marginalized populations, due to limited data availability.

- Introduction
- Generative agent behavior and interaction
- Generative agent architecture
- Sandbox environment implementation
- Controlled evaluation
- End-to-end evaluation
- Discussion
- **Conclusion**





Conclusion

- This paper introduces generative agents, interactive computational agents that simulate human behavior.
- The authors 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.
- Evaluations suggest that the architecture creates believable behavior.





中國人民大學
RENMIN UNIVERSITY OF CHINA

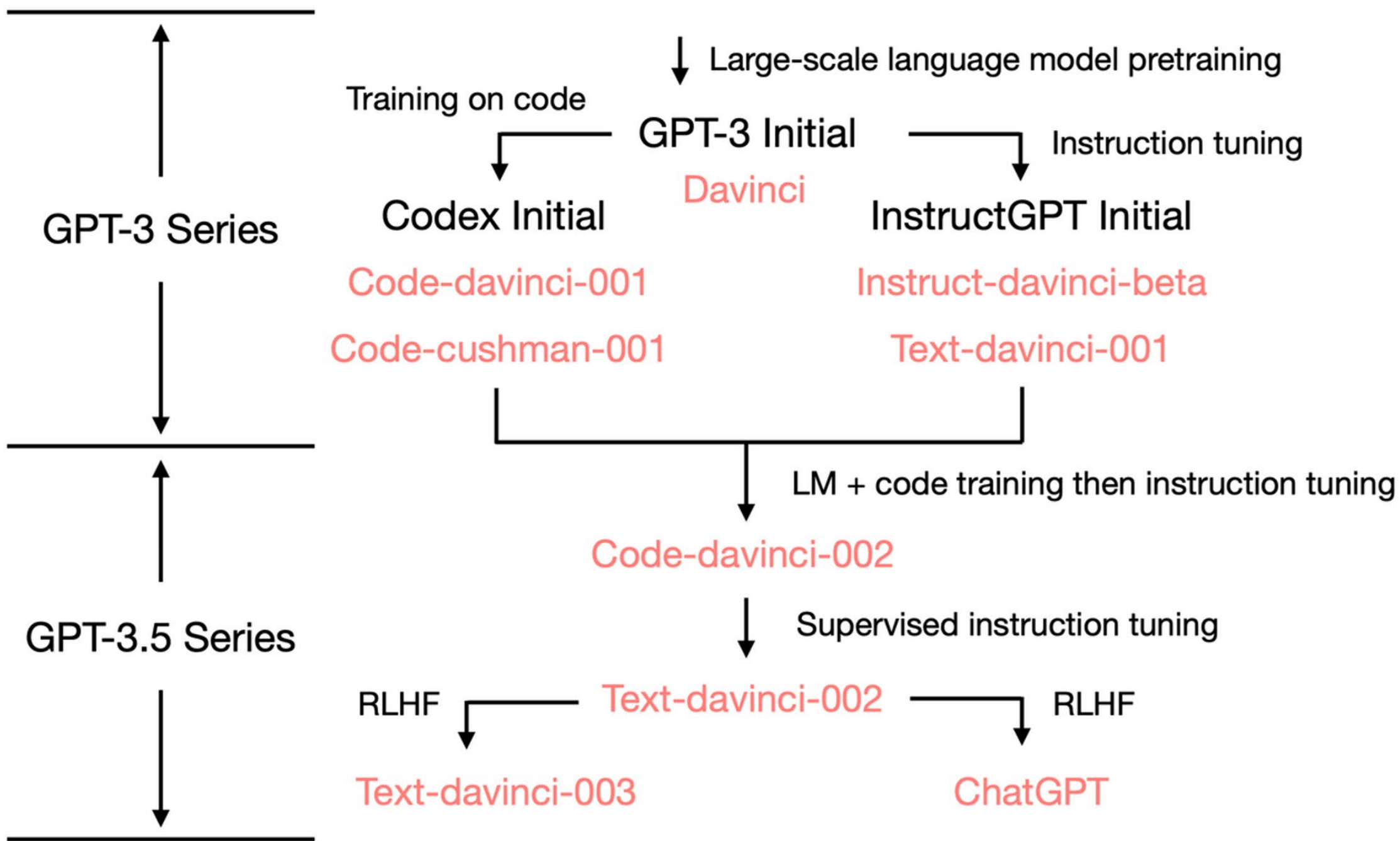


高瓴人工智能学院
Gaoling School of Artificial Intelligence

Thank You for listening!

Shen Yuan

2023-10-12





英文	中文	释义
Emergent Ability	突现能力	小模型没有，只在模型大到一定程度才会出现的能力
Prompt	提示词	把 prompt 输入给大模型，大模型给出 completion
In-Context Learning	上下文学习	在 prompt 里面写几个例子，模型就可以照着这些例子做生成
Instruction Tuning	指令微调	用 instruction 来 fine-tune 大模型
Code Tuning	在代码上微调	用代码来 fine-tune 大模型
Reinforcement Learning with Human Feedback (RLHF)	基于人类反馈的强化学习	让人给模型生成的结果打分，用人打的分来调整模型
Chain-of-Thought	思维链	在写 prompt 的时候，不仅给出结果，还要一步一步地写结果是怎么推出来的
Scaling Laws	缩放法则	模型的效果的线性增长要求模型的大小指数增长
Alignment	与人类对齐	让机器生成符合人类期望的，符合人类价值观的句子

