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

Presented by:
John Tan Chong Min

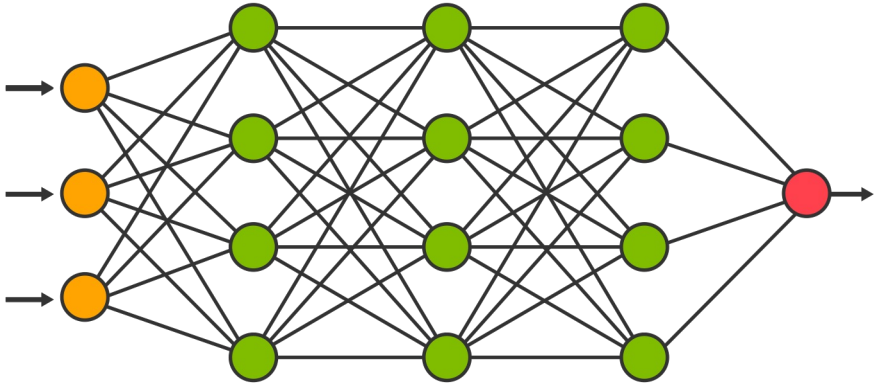
Let’s see it in action!

- https://reverie.herokuapp.com/arXiv_Demo/

Motivation

- GPT-based models can encode a lot of the world via self-supervised learning
- Fine-tuning the model can be difficult (and also impossible right now with ChatGPT / GPT4)
- Use memory as a means of learning
  - Use Natural Language to store and retrieve memories!
- Use multi-agent system to get emergent behaviour!

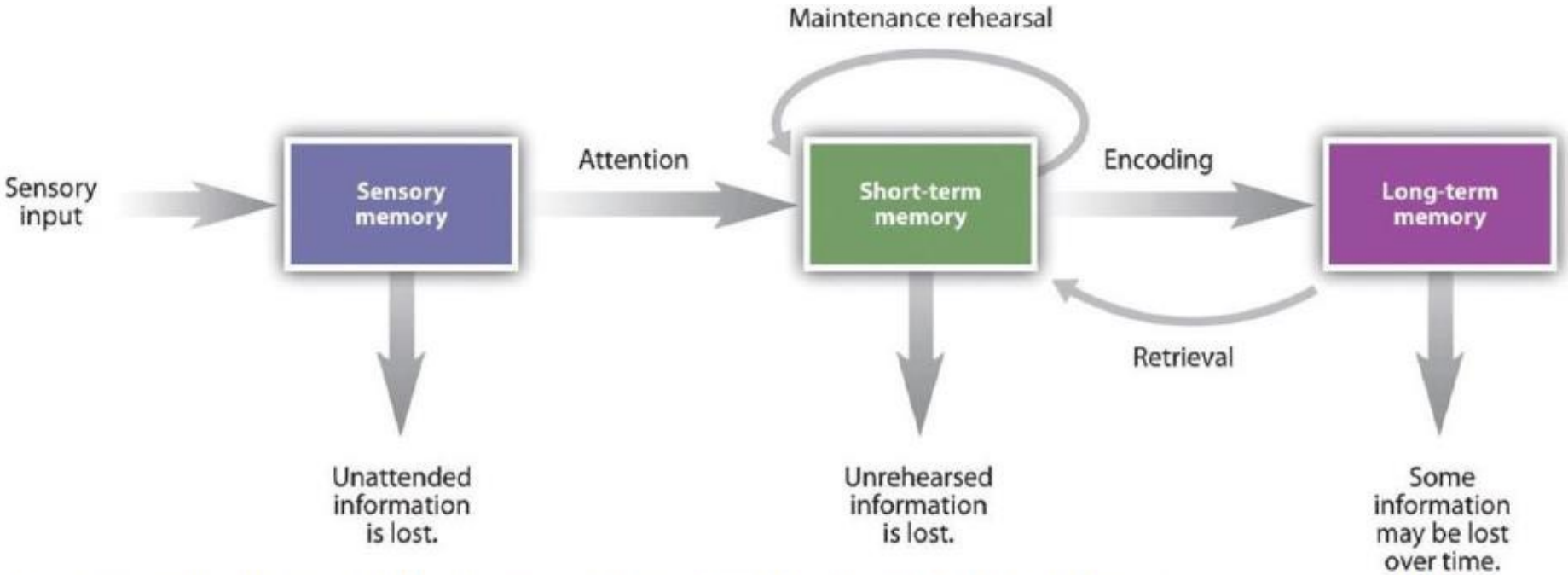
Types of memory



- Fast inference (one pass)
- Slow to learn (requires gradient descent)

- Slow inference unless using approximate techniques
- Can be quick to learn from experience

Can we learn from just memories alone?



Memory can be characterized in terms of stages—the length of time that information remains available to us.

Source: Adapted from Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. Spence (Ed.) *The psychology of learning and motivation* (Vol. 2). Oxford, England: Academic Press.

Is memory related to personality?

- If we wipe out your memory right now, will you still be the same person?

- What are the things retained, what are the things lost?

Overall

- 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**
  - **Retrieve** them dynamically to plan behaviour

Starting Information

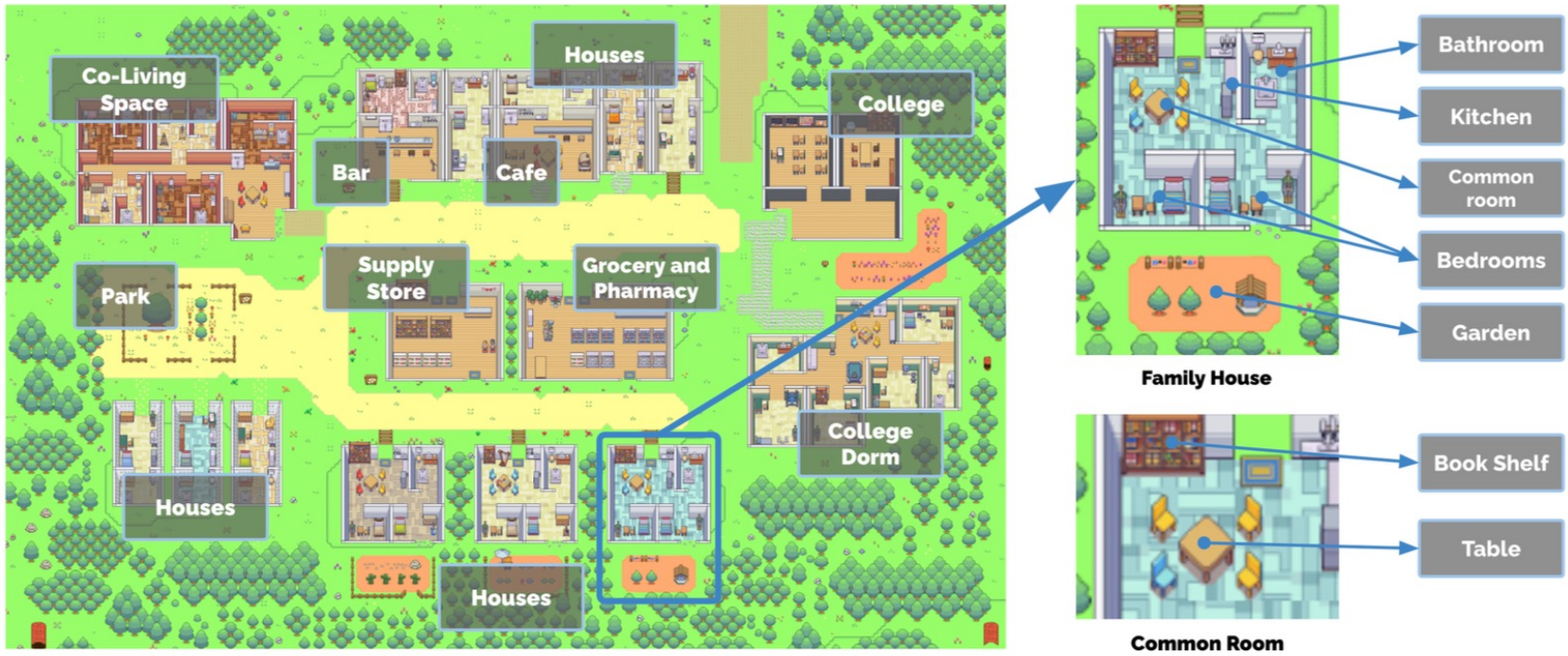


Figure 2: The Smallville sandbox world, with areas labeled. The root node describes the entire world, children describe areas (e.g., houses, cafe, stores), and leaf nodes describe objects (e.g., table, bookshelf). Agent remember a subgraph reflecting the parts of the world they have seen, in the state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

The power of prompting:
Agent character, memories and background

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

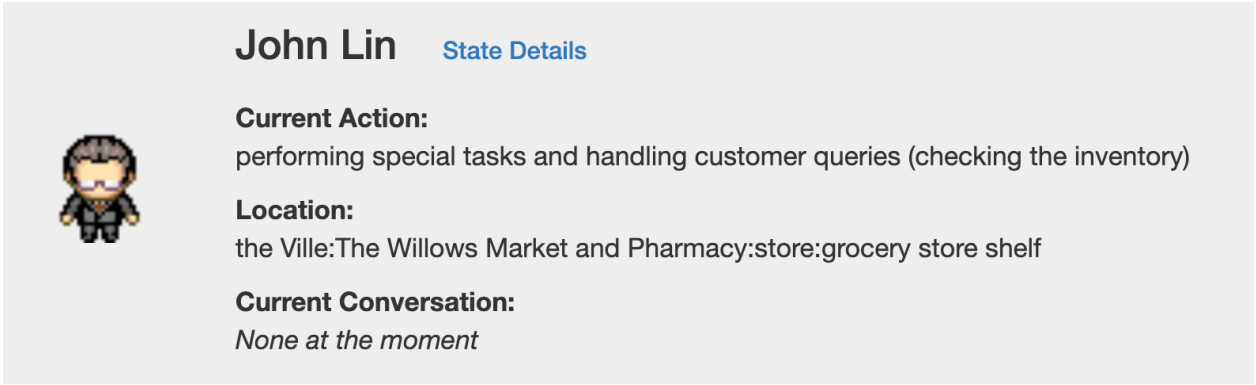
John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmac

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**
- 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”
- Agents communicate with each other in full natural language. Agents are aware of other agents in their local area, and the generative agent architecture determines whether they walk by or engage in conversation



Taking Control of an Agent

- By a new agent
  - Specify characteristic of agent via a prompt (e.g. ”You are a news reporter”)

- Taking over an existing agent via “inner voice”
  - 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

  - Similar to chain-of-thought prompting?

A day in the life of an agent (John Lin)

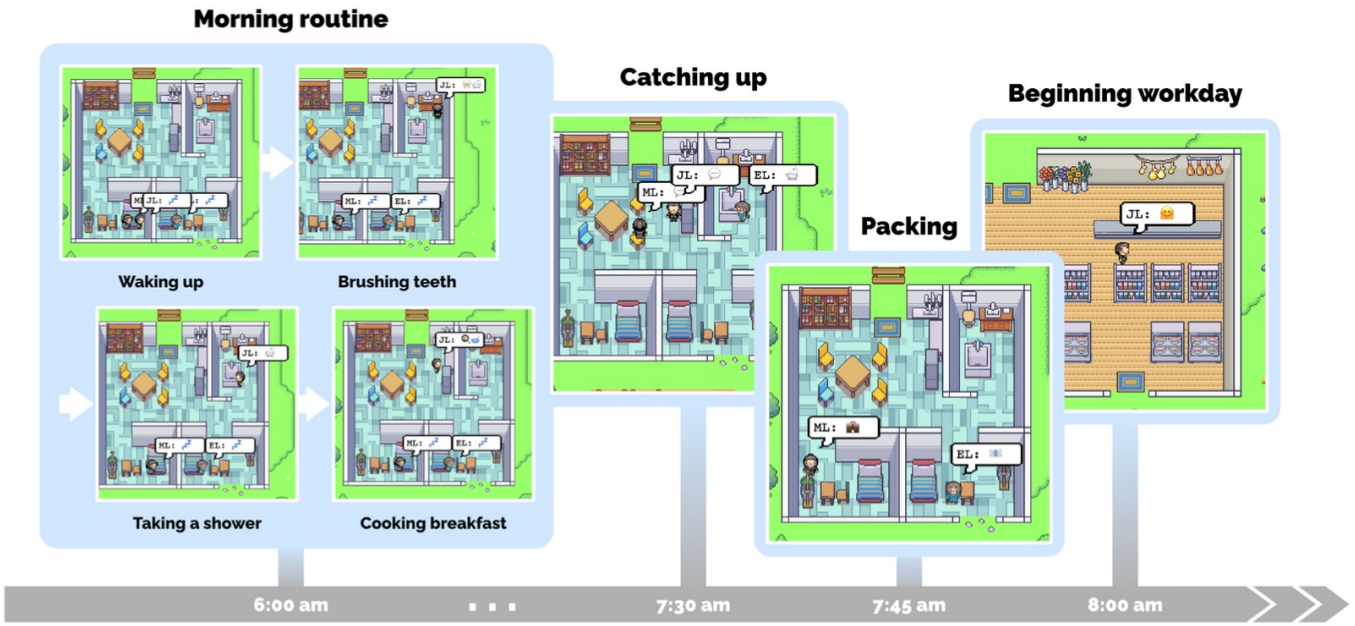


Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.

Initial prompting can lead to cascade of actions

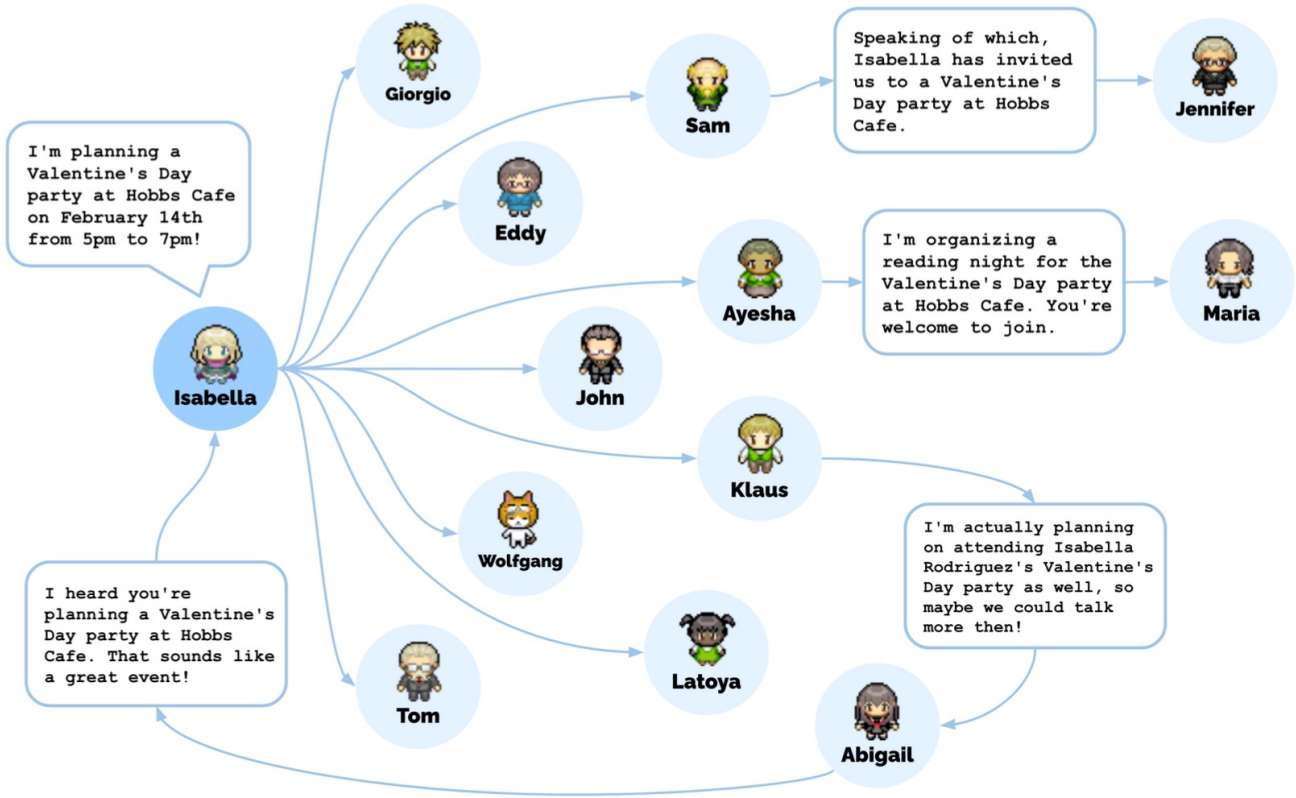


Figure 9: The diffusion path for Isabella Rodriguez's Valentine's Day party. A total of 12 agents heard about the party at Hobbs Cafe by the end of the simulation.

Memory is important

Memory Retrieval for Planning

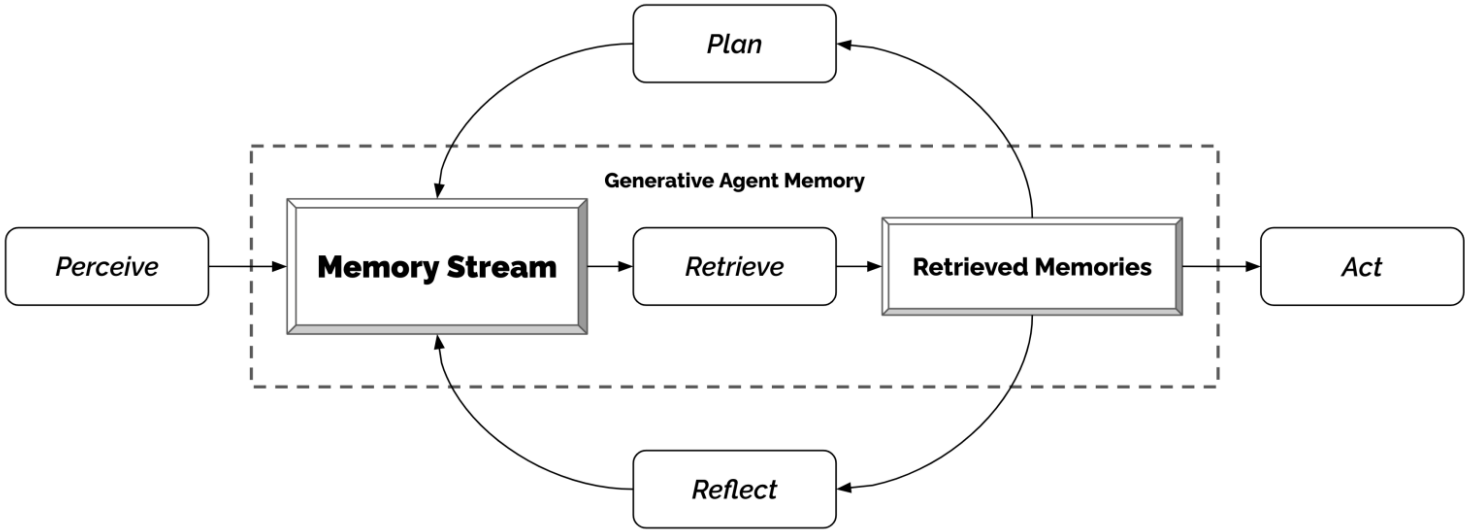


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories, then uses those retrieved actions to determine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

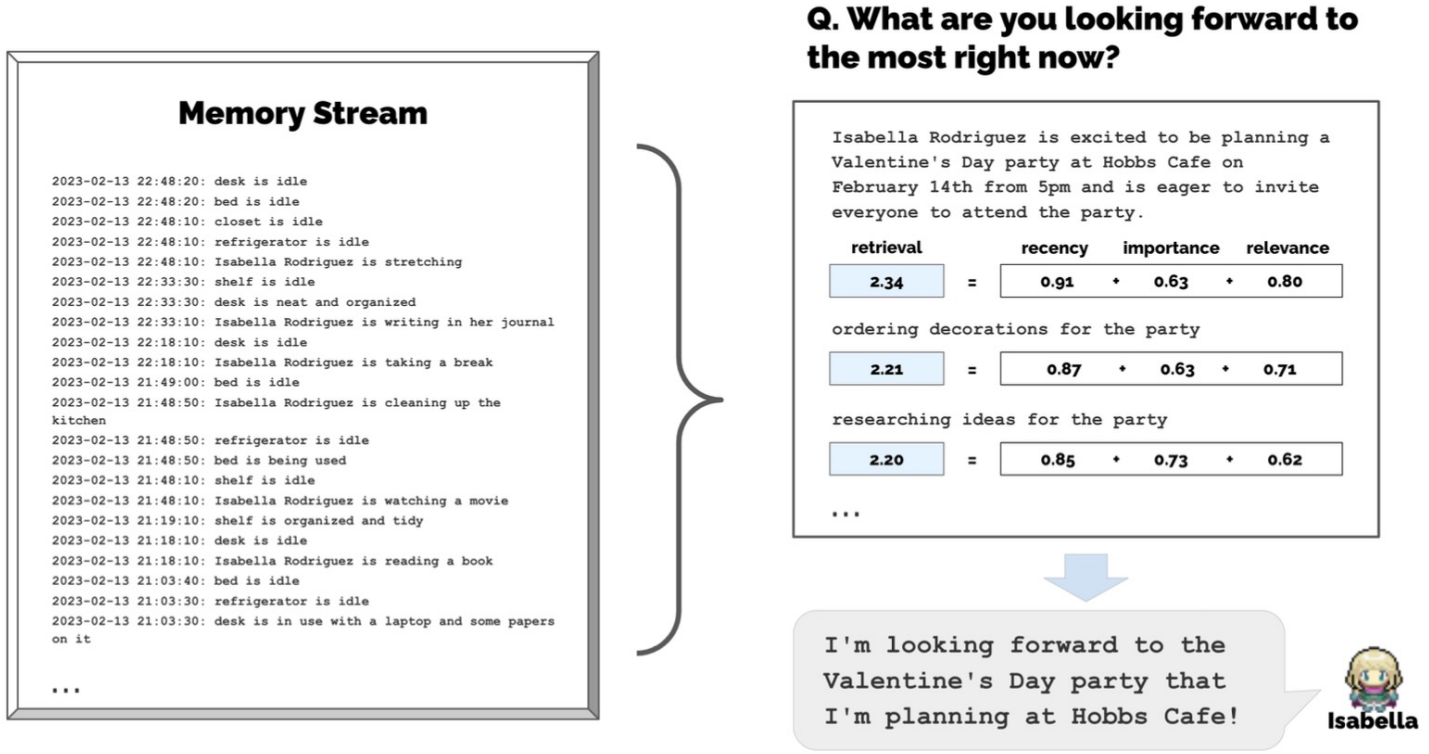
Memory Retrieval Limit

- Current prompting has a token limit
- Need to find a **concise way** to store memories and retrieve them

- Solution
  - Store experience as a memory stream – list of memory objects, with time stamp of creation and last time stamp of access
  - Stores as observation – behaviors by agent, or behaviors by others/environment
  - Retrieves experience based on recency, importance, relevance

Memory Stream

Final score is a weighted sum of all 3 components



- **Recency**
  - Timestamp of last access
- **Importance**
  - LLM rates itself!
- **Relevance**
  - Vector embedding similarity

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.

Importance of memory

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>

- Question to ponder:
  - Is an LLM good for coming up with numbers for a rating?
  - Why not zero-shot classification?

Memory Stream - from observations to reflection

- Observations may be too generic to learn, need some higher level consolidation

- Reflection generation:
  - Generate reflections when the sum of the importance scores for the latest events perceived by the agents exceeds a certain threshold
  - Reflect on average two to three times a day

- Reflection process:
  - Retrieve 100 most recent observations/reflections, then prompt "Given only the information above, what are 3 most salient high-level questions we can answer about the subjects in the statements?”
  - For each question, retrieve most relevant X memories and generate reflections (also state the memories/reflections derived)

Planning

- Plans describe a future sequence of actions for the agent, and help keep the agent's behavior consistent over time (Grounding like Retrieval-Augmented Generation!)

- Without a plan, an agent may keep repeating actions (like keep eating lunch)

- Generation of Plan:
  - At the start of the day, the agent is prompted with his/her innate traits, the full description of the previous day's plans, and asked to generate the plan for the new day
  - Recursively generate even more fine-grained details
  - Plan can be changed if interaction with other agents trigger it (via prompting)
  - Plan is saved in memory stream

Trees of Reflection

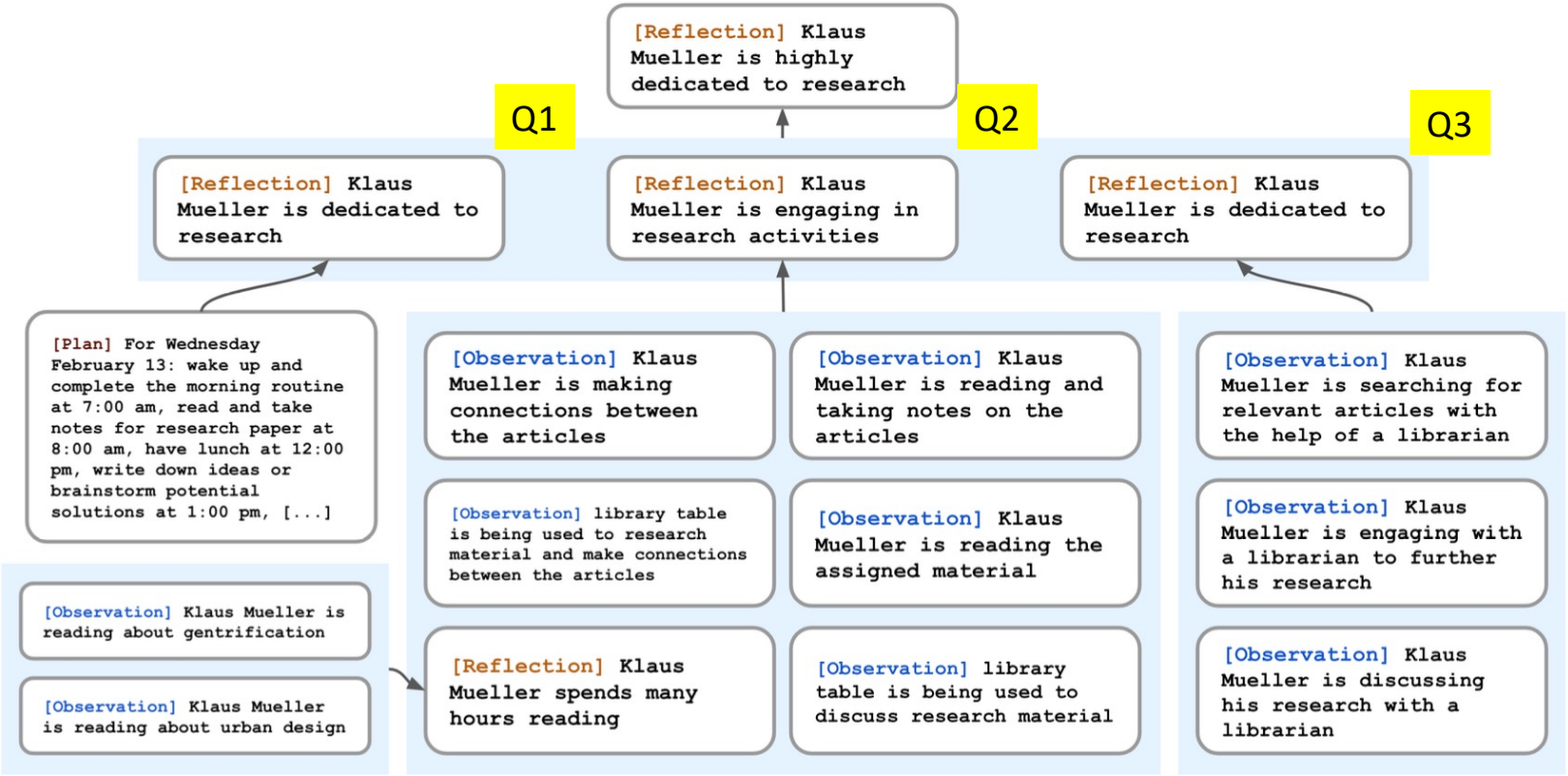


Figure 7: A reflection tree for Klaus Mueller. The agent's observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus's self-notion that he is highly dedicated to his research.

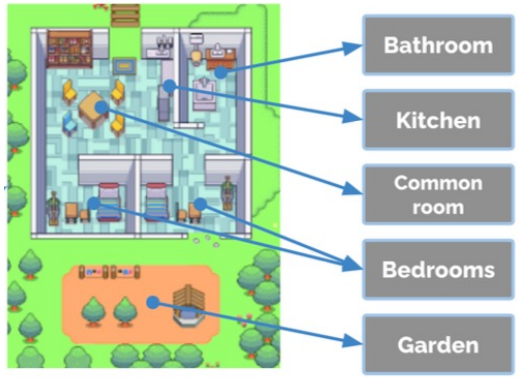
Symbolic Representation: Perceiving and Acting

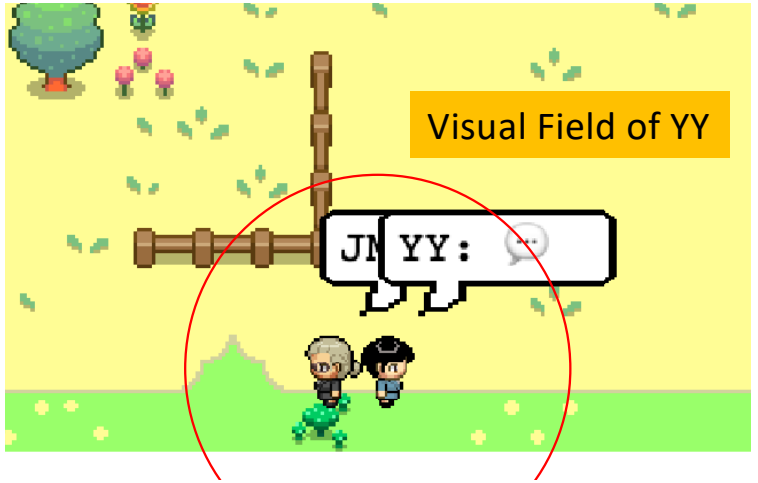
- The server maintains a JSON data structure that contains information about each agent in the sandbox world, including their **current location**, a description of their **current action**, and the **object they are interacting with**

- At each time step, the 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 objects that the agents are interacting with

- The 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.

- Each agent has their own memory of the world, and will only update their memory upon perceiving the changes





Obtaining fine-grained actions via recursive prompting

[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?

- Recursively use prompting to get to lower and lower sub-areas based on the JSON tree
- Ground the agent in feasible actions via prompting

Is the memory retrieval good?

- Generally capable of recalling past experiences and answering questions in a manner consistent with self-knowledge
- Can fail to retrieve correct instances in memory
- Can retrieve an incomplete memory fragment (maybe need episodic memory?)
- Can hallucinate knowledge if they fail to recall (does this sound like humans?)

Are the actions generated plausible?

- Generally yes
- Some areas like college dorm's bathroom can only be occupied by one person at a time, but agents did not know about it from the description and multiple agents entered at the same time
- Some places like shops are closed after certain hours, but agents still enter them
- Can potentially fix by giving prompts of social constructs, or relabelling some locations to give semantic meaning (e.g. one-person bathroom)

Future Work:
Storing and Retrieving Memories

Some Recap of my Memory Framework

Storing memories

- There is limited room in the brain to hold memories

- Need to be very selective as to what can be stored
- Priority of storing is determined by emotion
  - Flashbulb memories are very powerful
- High emotion memories (fear, happy) will be prioritized over low emotion ones (neutral, boredom)
  - In school settings, we largely do not remember the process of studying, but the friends we made and the activities we did with them

Forgetting memories

- Memory retrieval helps to improve the strength of the memory
- Memories that are not accessed often are forgotten and make room for other memories to be stored
- Helps to make sure the memories we have are useful for the current environment
  - Less used memories are those not relevant and forgotten
  - Frequently used memories are those relevant and strengthened

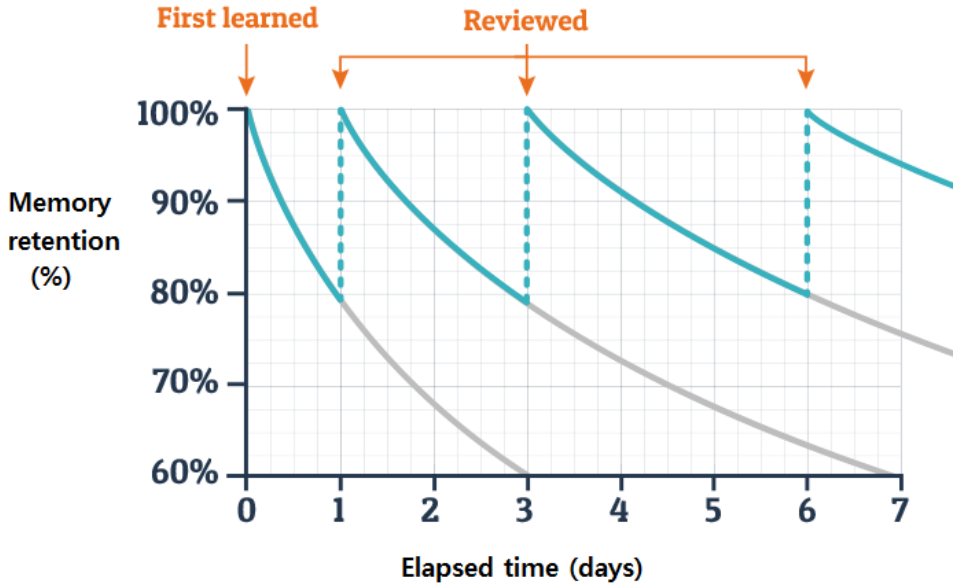
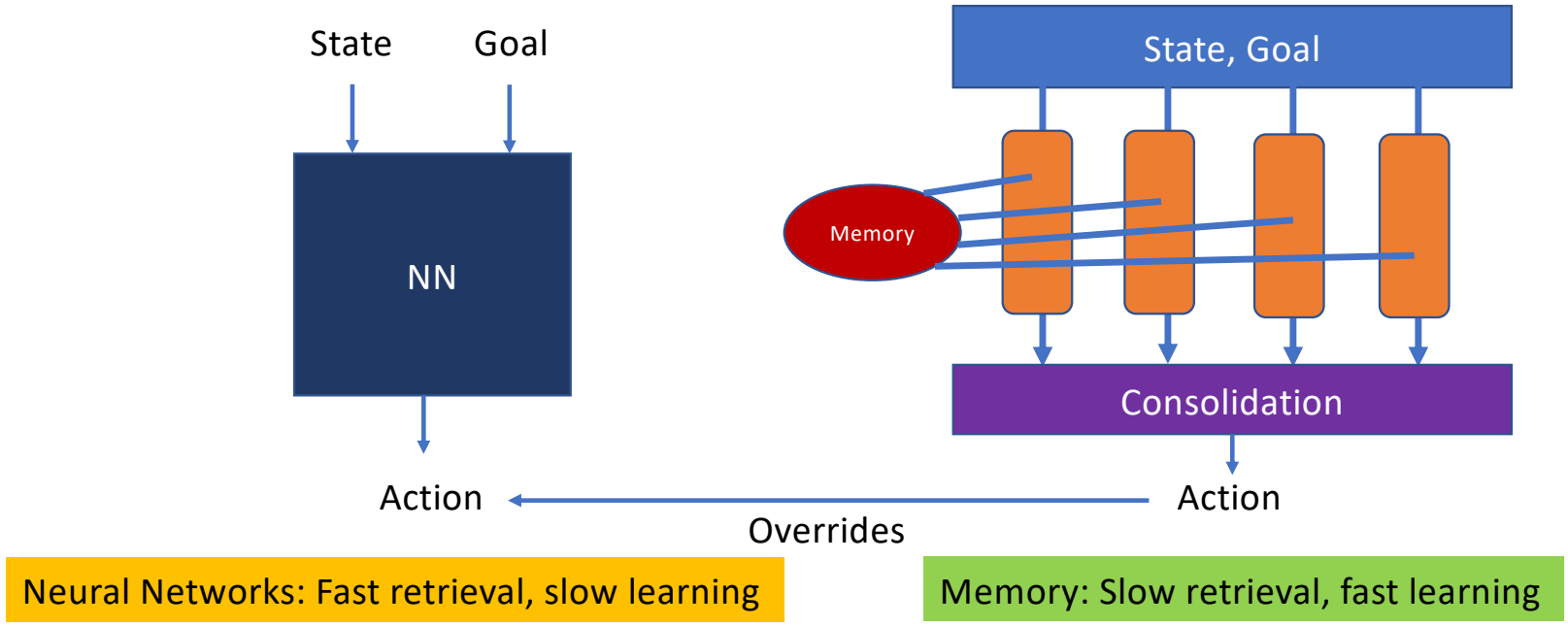


Chart from: Chun, Bo Ae & hae ja, Heo. (2018). The effect of flipped learning on academic performance as an innovative method for overcoming ebbinghaus' forgetting curve. 56-60. 10.1145/3178158.3178206.

Link to my work (Learning, Fast & Slow)

- In this work, memory is stored as a (state, action, next state) tuple to aid planning
- Could memory be more generic and be stored as a timestep with observation?
- Are there different forms of memory that are all useful?



Learning, Fast and Slow: A Goal-Directed Memory-Based Approach for Dynamic Environments. 2023. John and Motani.

Questions to ponder

- Can we encode memory in other forms other than natural language?
- For repetitive actions, could it be that the agent's state is not descriptive enough? If we embellish the agent with information of past actions, can counter repetition?
- What are the types of memory that an agent should have? Is the amount of memory (and prompting) in this simulation too much/too little?
- How can we better store and retrieve memories? How can we implement the emotional aspect of memories, as well as the forgetting aspect into this architecture?
- Instead of retrieval, why not just summarize the memory stream once every X hours?