

Lecture 5.

# Architecting Generative Agents

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park





**Quick housekeeping**

# Announcements

Assignment 1 is available on our course website!

[https://joonspk-research.github.io/cs222-fall24/  
assignment1.html](https://joonspk-research.github.io/cs222-fall24/assignment1.html)

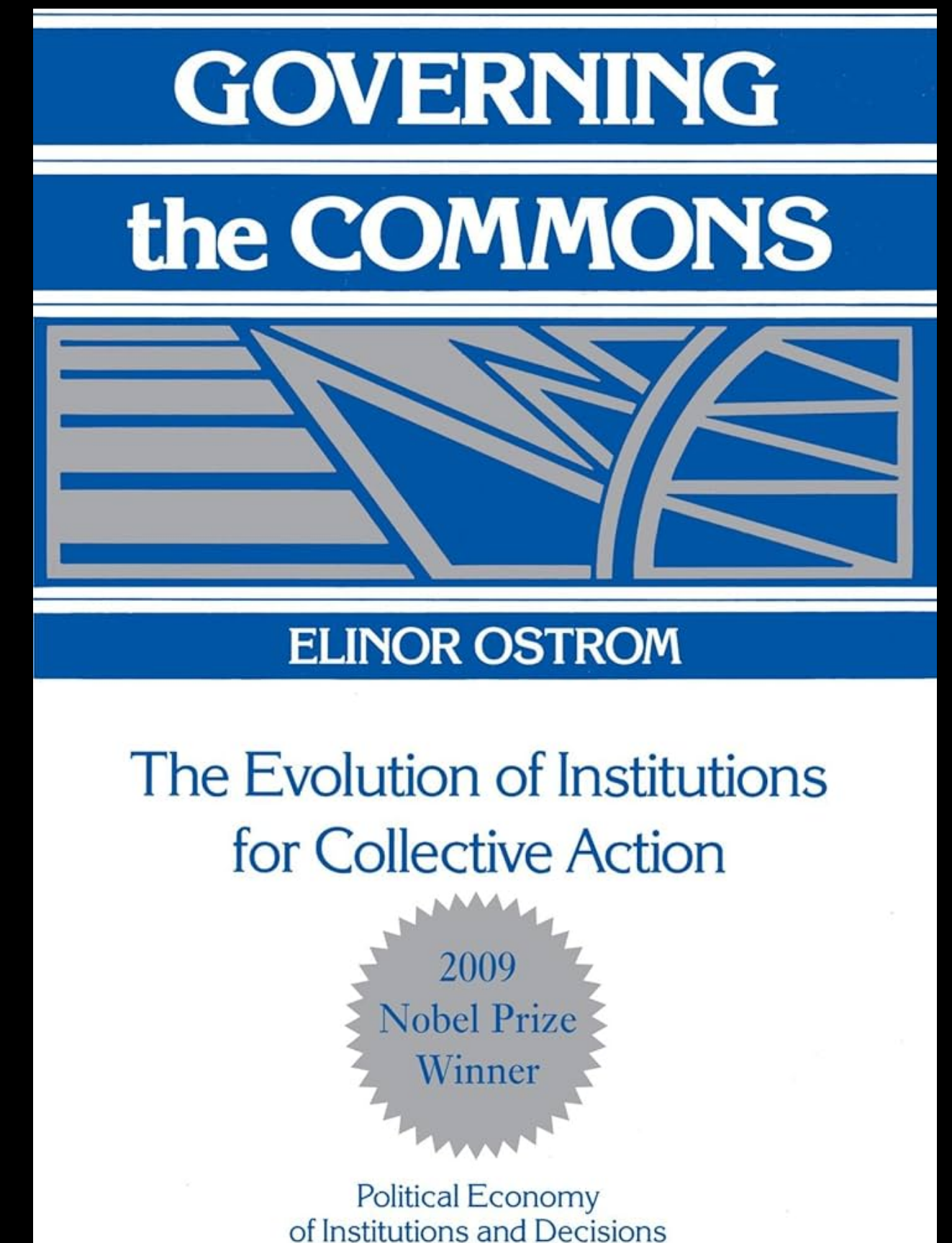
Due Monday, Oct 21!

# Assignment 1 requires OpenAI credit

We created an API key for everyone in the class.

For Assignment 1, we have a collective budget of \$400 (roughly \$8 per student).

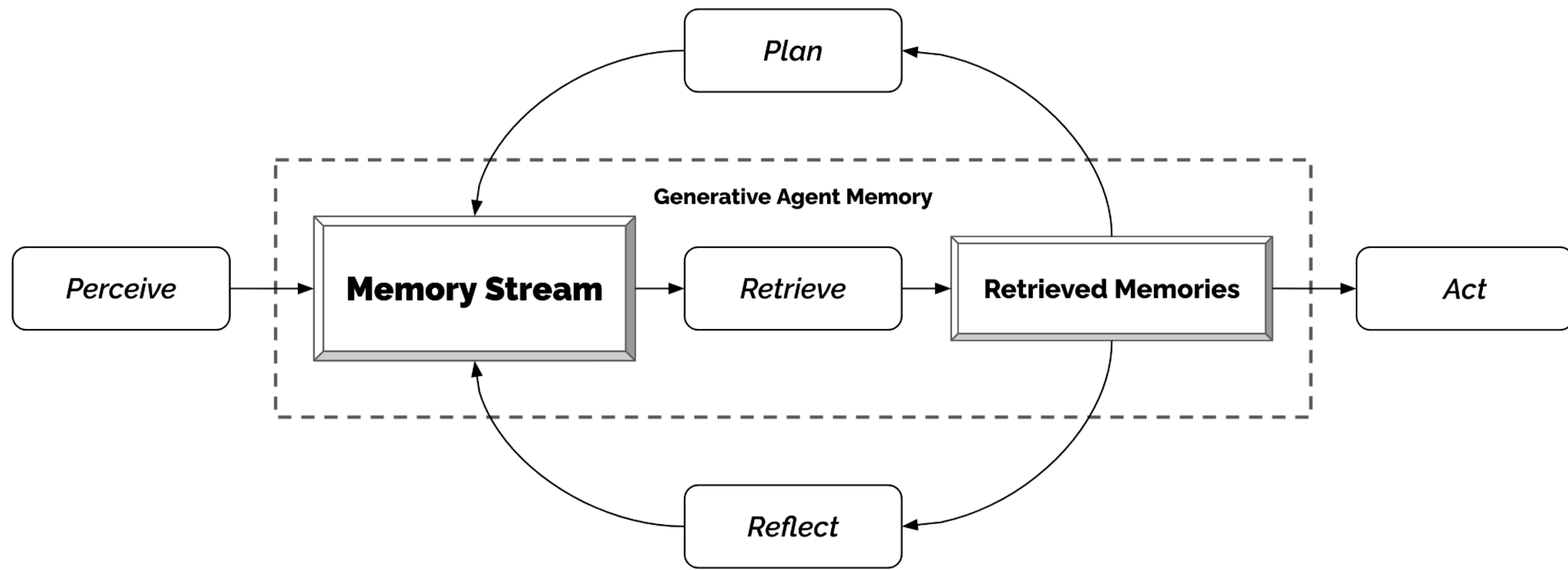
This should be enough for this assignment, but it's important for everyone to keep an eye on the API spending!





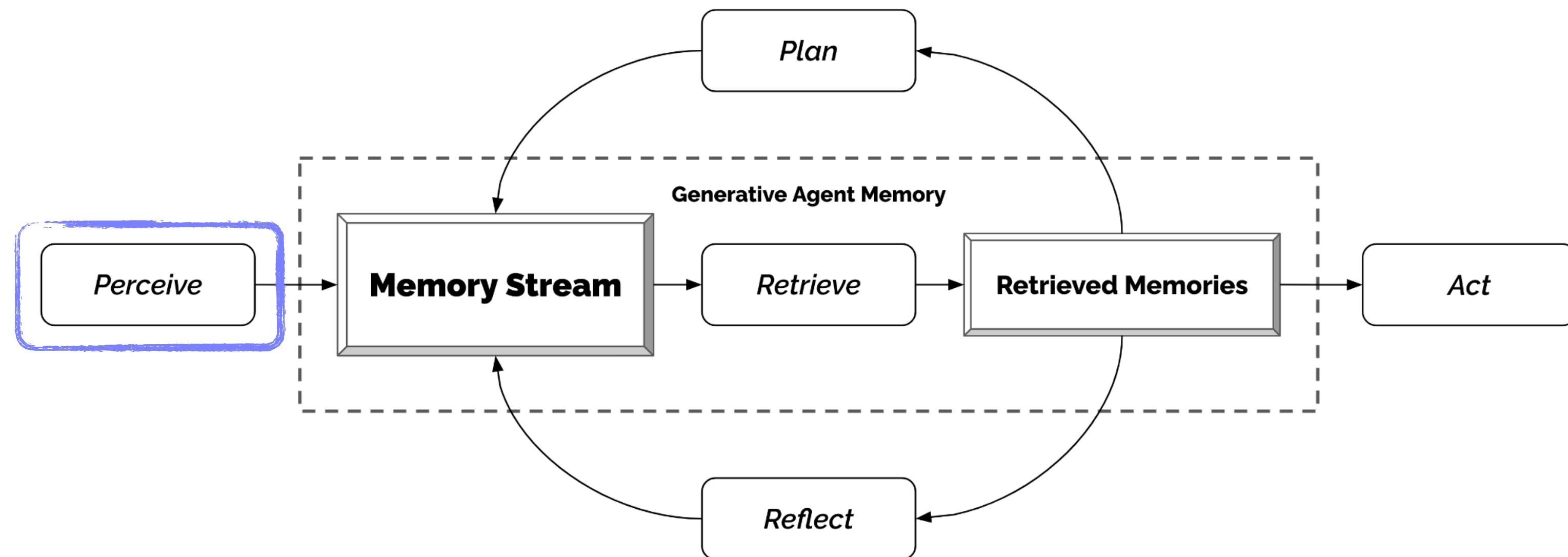
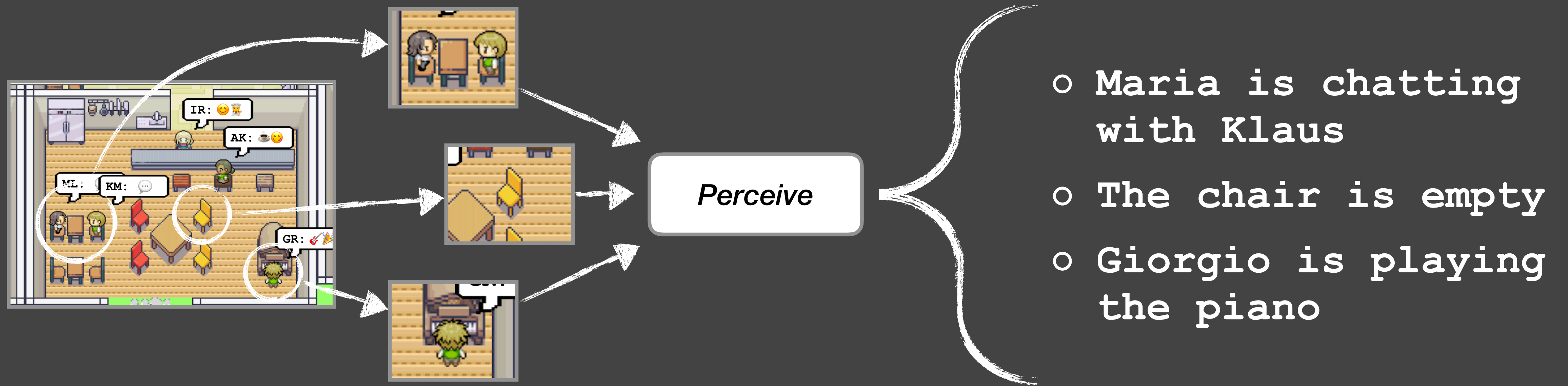
**Today: let's build a generative agent!**





J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).



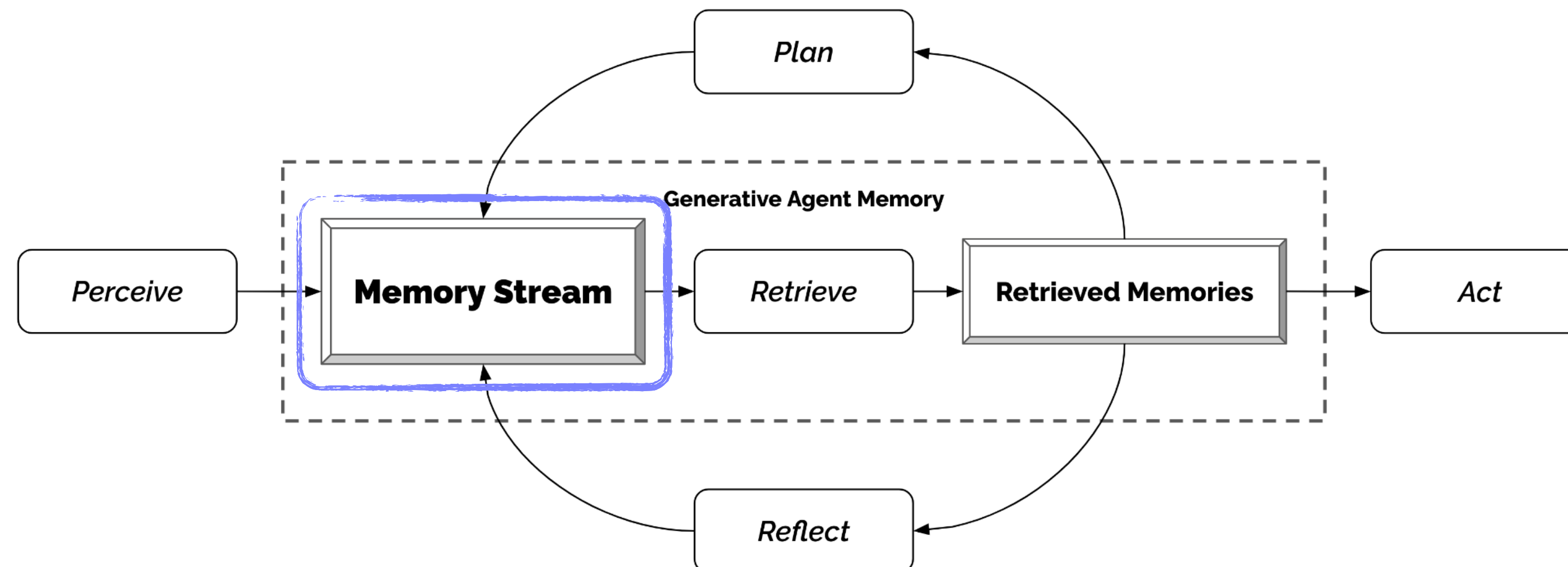




- Maria is chatting with Klaus
- The chair is empty
- Giorgio is playing the piano

## Isabella's Memory Stream

2023-02-13 22:48:20: Maria is chatting with Klaus  
2023-02-13 22:48:20: The chair is empty  
2023-02-13 22:48:20: Giorgio is playing the piano  
2023-02-13 22:48:20: Giorgio is playing the piano  
2023-02-13 22:48:20: Giorgio is playing the piano  
...





**Isabella's Memory Stream**

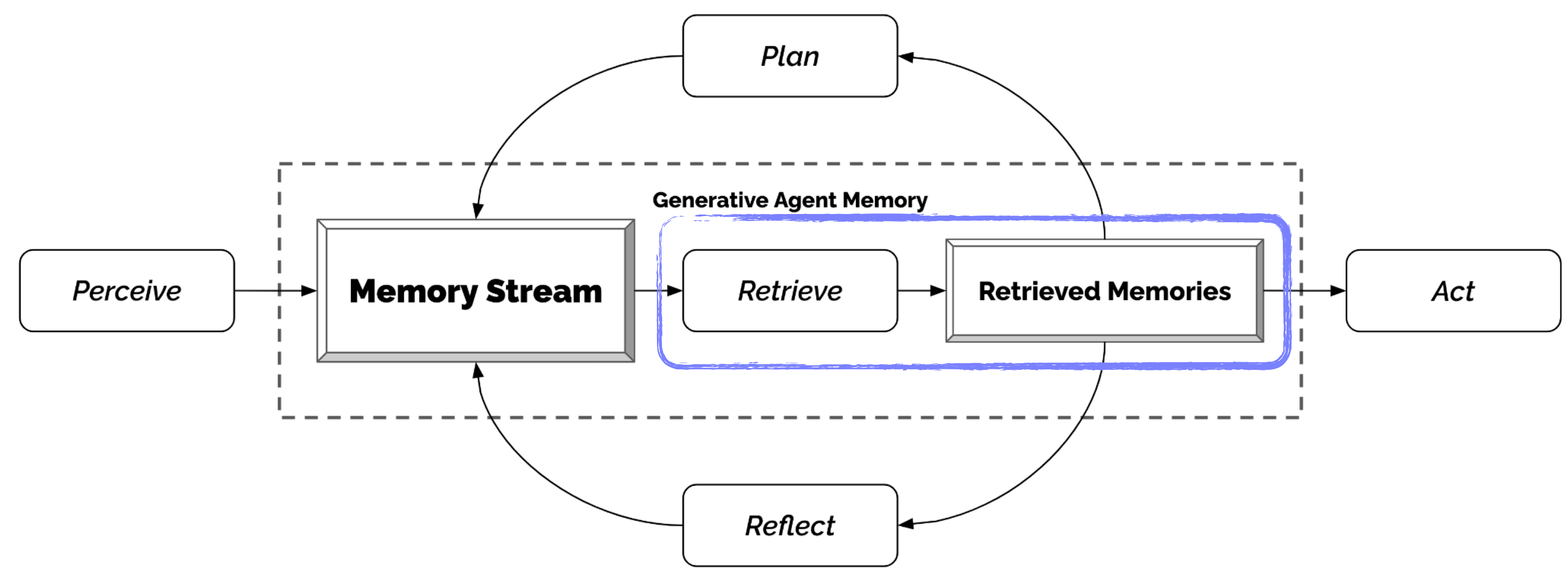


*Retrieve*



### What are you excited about, Isabella?

- Isabella is planning a Valentine's Day party at Hobbs Cafe.
- ordering decorations for the party
- researching ideas for the party





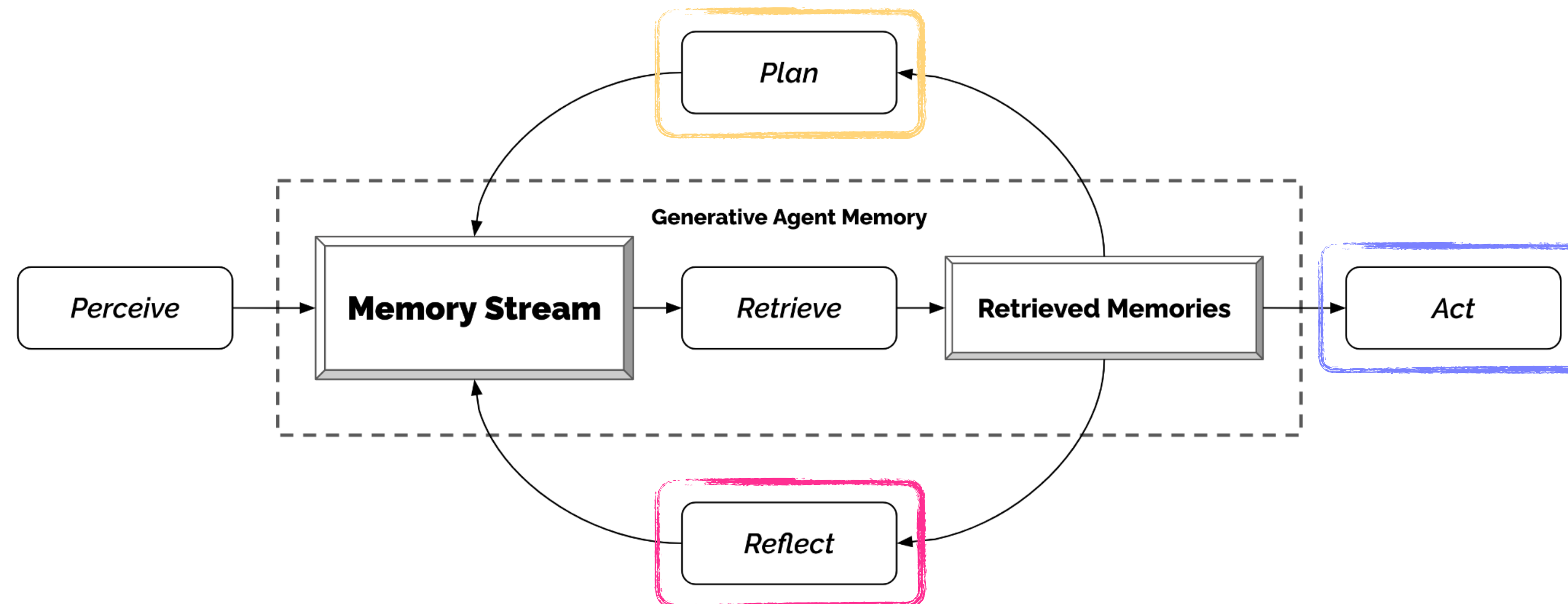
## What are you excited about, Isabella?

- Isabella is planning a Valentine's Day party at Hobbs Cafe.
- ordering decorations for the party
- researching ideas for the party

**[Plan]** Let's decorate the cafe later this afternoon

**[Action]** Heading to the local grocery store to buy supplies for the party

**[Reflection]** I enjoy organizing events and making people feel welcome



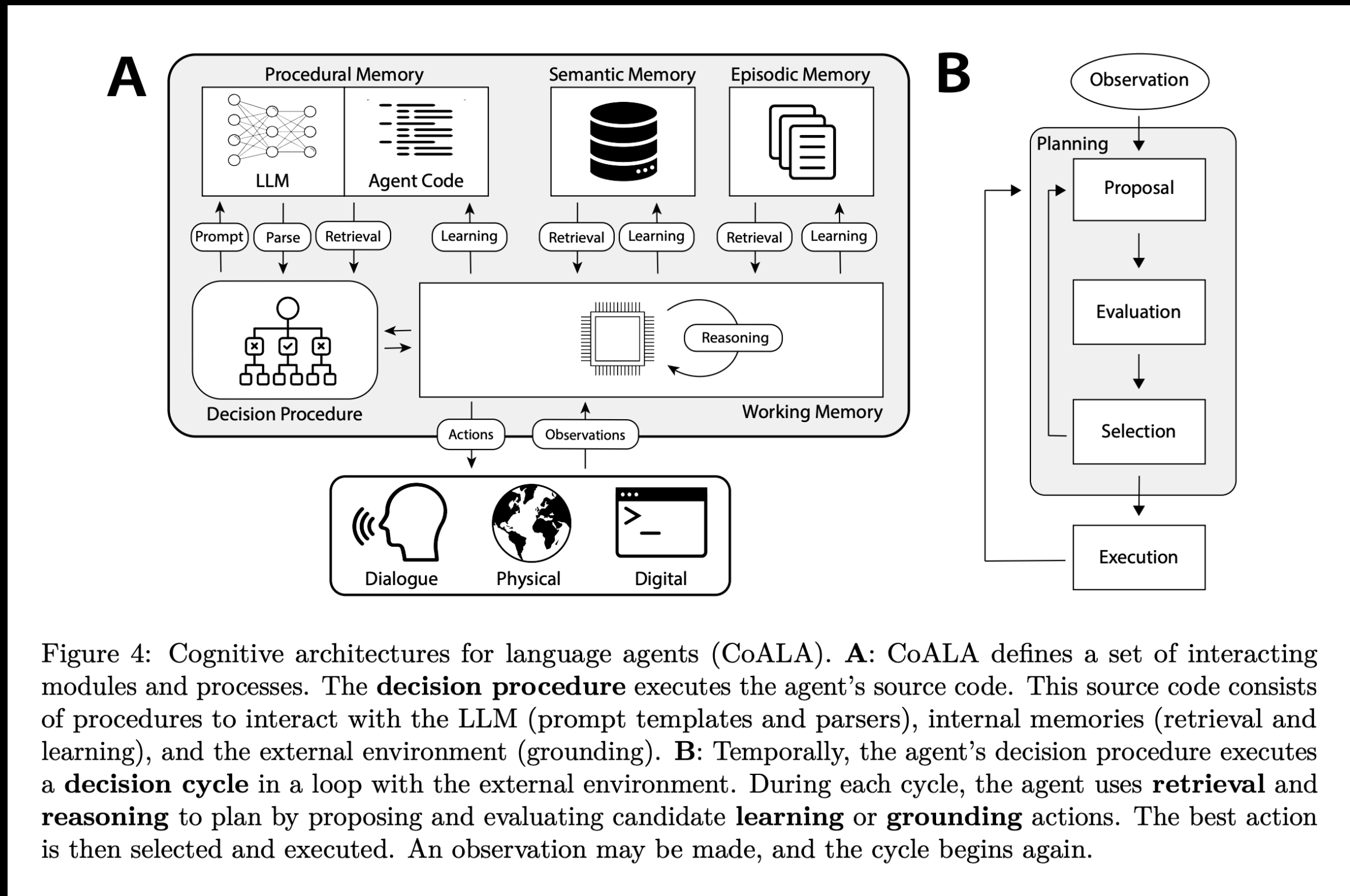
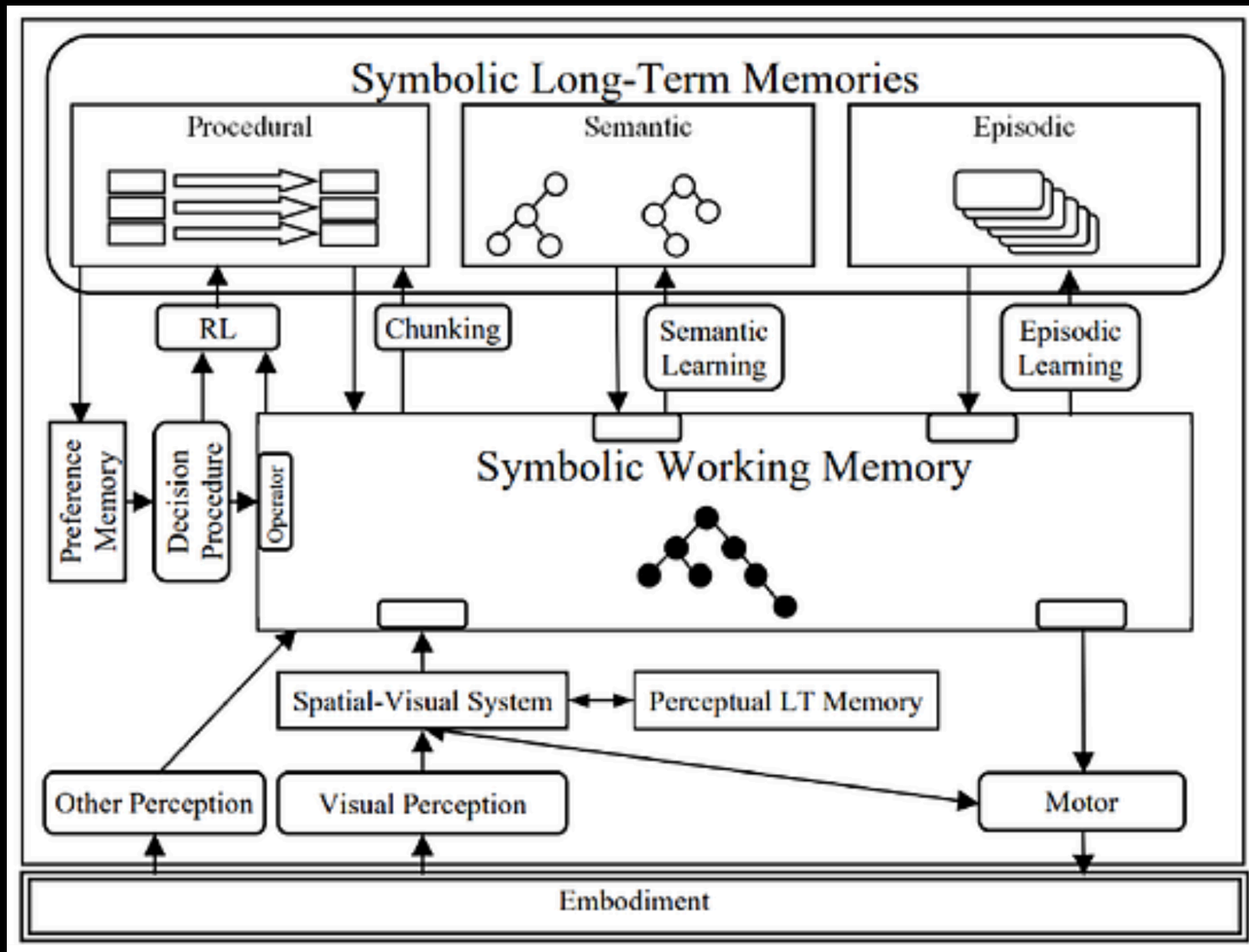


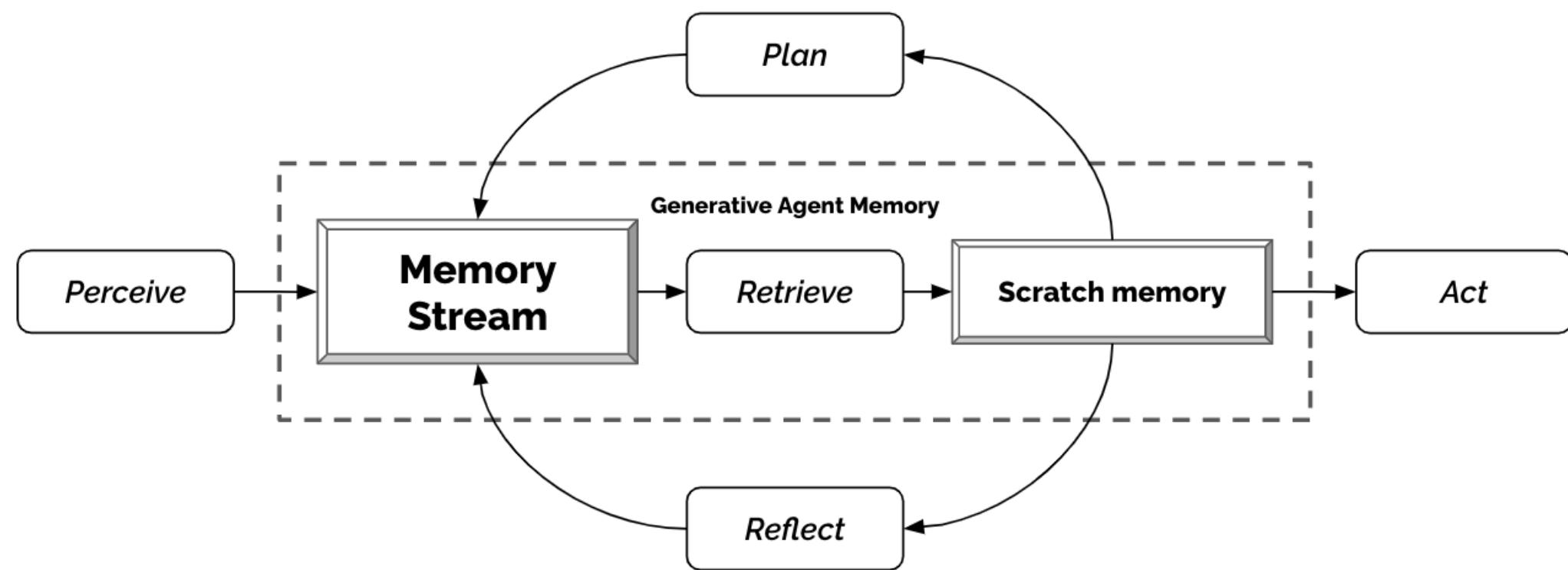
Figure 4: Cognitive architectures for language agents (CoALA). **A:** CoALA defines a set of interacting modules and processes. The **decision procedure** executes the agent’s source code. This source code consists of procedures to interact with the LLM (prompt templates and parsers), internal memories (retrieval and learning), and the external environment (grounding). **B:** Temporally, the agent’s decision procedure executes a **decision cycle** in a loop with the external environment. During each cycle, the agent uses **retrieval** and **reasoning** to plan by proposing and evaluating candidate **learning** or **grounding** actions. The best action is then selected and executed. An observation may be made, and the cycle begins again.



# Generative Agents — Implementation

## Memory stream:

A database that maintains a comprehensive record of an agent's experience in natural language



From the **memory stream**, records are **retrieved** as relevant to **plan the agent's actions** and react appropriately to the environment, and records are recursively synthesized into higher- and higher-level **reflections** that guide behavior.



# **Memory and Retrieval -- Challenges**

**The full memory stream can distract the generative model and does not fit into the limited context window.**

**Q. Imagine a friend asked you to tell them what you are looking forward to the most.**

**What would you tell them?**




# Memory stream stores a comprehensive record of agent experience in natural language

**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

...
```

Each "memory object" contains the timestamp for the creation time.



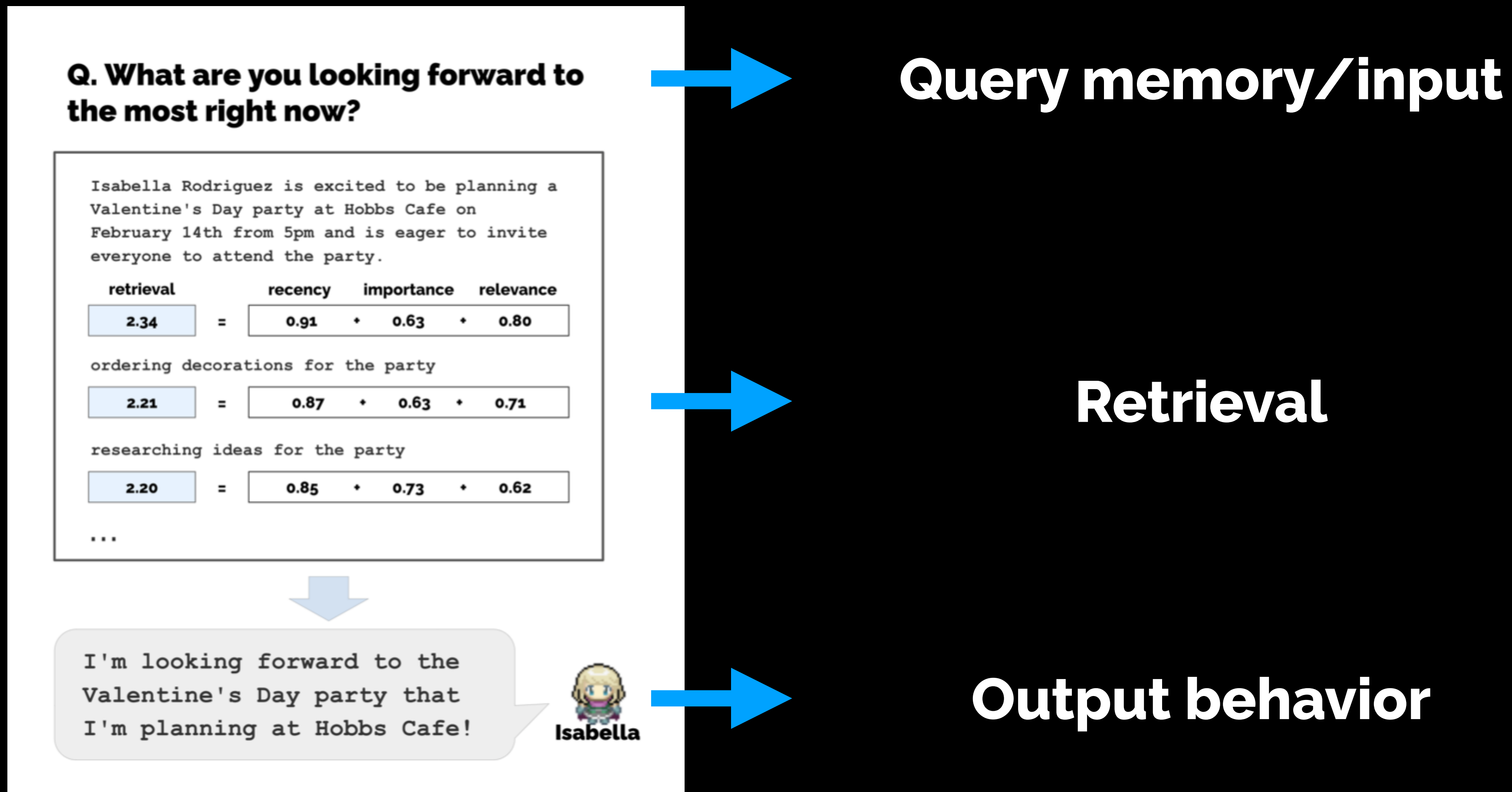
**2023-02-13 22:33:30**  
**Isabella Rodriguez is writing in her journal**

**2023-02-13 22:18:10**  
**Isabella Rodriguez is taking a break**

**2023-02-13 21:48:10**  
**refrigerator is idle**

...

# We retrieve a select portion of the agents' experience using a retrieval function





$$\text{retrieval\_score} = \alpha_1 * \text{recency} + \alpha_2 * \text{importance} + \alpha_3 * \text{relevance}$$

## Recency

Exponential decay

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

Embedding

X

Cosine similarity

# Reflection -- Challenges

**Generative agents, when equipped with only raw observational memory, struggle to generalize or make inferences.**

# Reflections are higher-level, abstract thoughts generated by the agent that are stored in the memory stream

## 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
```

...

Reflections are a type of memory,  
just like the observational memory.  
They are synthesized periodically.



# We synthesize existing records in agents' memory stream to formulate higher-level reflections

Generate what to reflect on by looking at 100 most recent records, then retrieve and reflect.

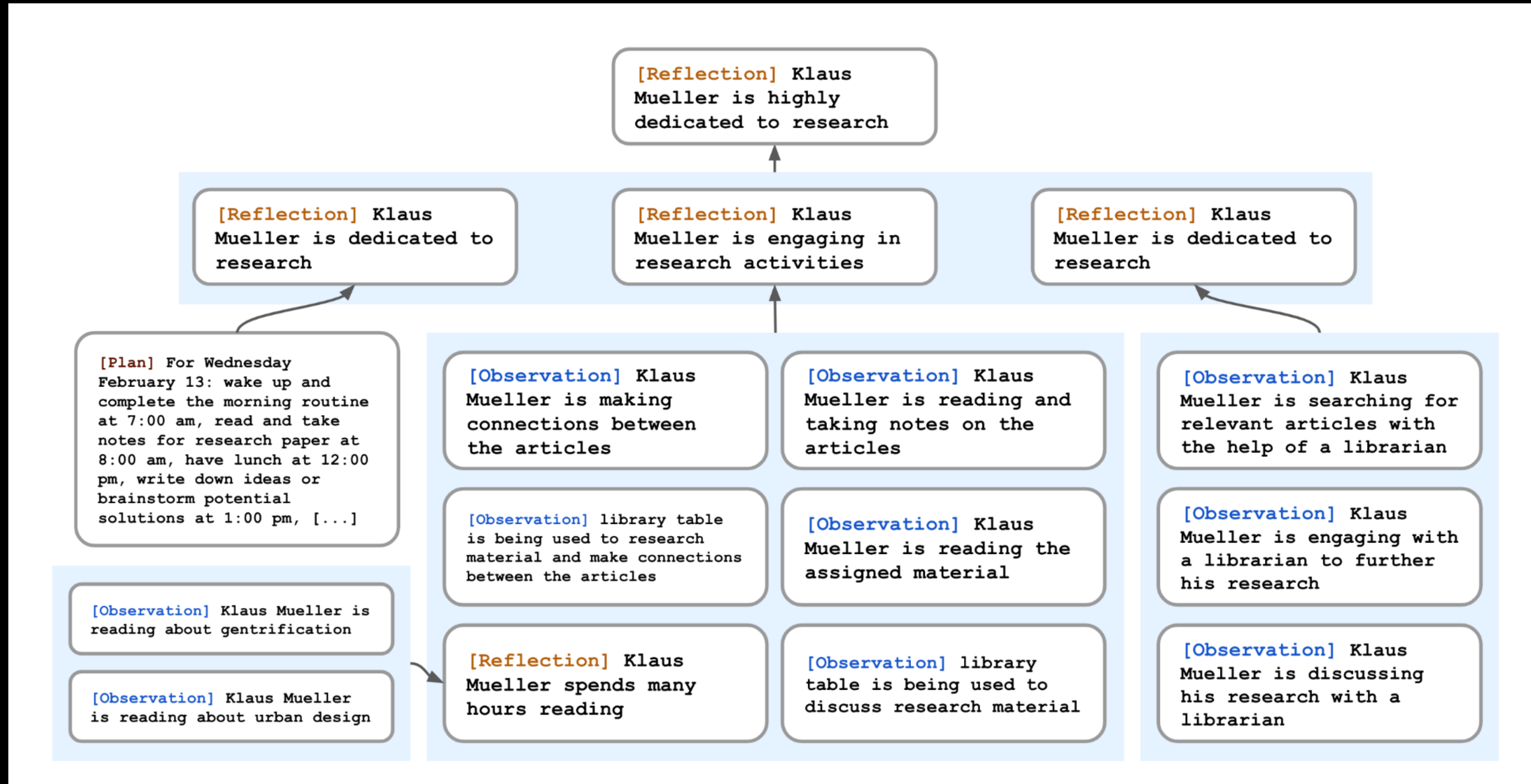
Statements about Klaus Mueller

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))

**Retrieved memory for reflection.**

**Retrieved memory can contain reflections and plans**

Over time, agents generate trees of reflections: the leaf nodes as observations, and the non-leaf nodes as thoughts that become higher-level higher up the tree they are.



## **Planning and Reacting -- Challenges**

**While a large language model can generate plausible behavior in response to situational information, agents need to plan over a longer time horizon.**



**Plans describe a future sequence of actions for the agent that are stored in the memory stream. They include a location, a starting time, and a duration.**

**Example plan for Klaus Mueller, who is dedicated in his research and has an impending deadline:**

**Chooses to spend his day working at his desk drafting his research paper. 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.**

**To generate plans, we prompt a large language model with a prompt that summarizes the agent and the agent's current status.**

**Agent summary description**

**Current status**

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)

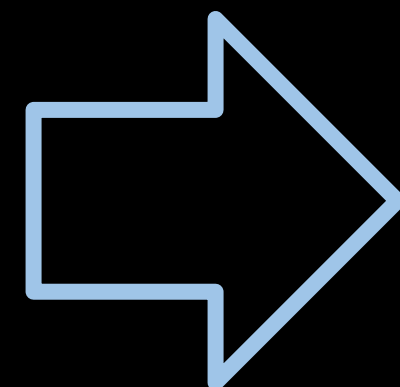
# To plan, our approach starts top-down and then recursively generates more detail in the plan.

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.

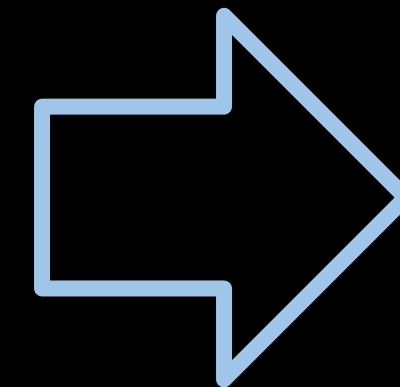
work on his new music composition from 1:00 pm to 5:00 pm becomes 1:00 pm: start by brainstorming some ideas for his music composition [...] 4:00 pm: take a quick break and recharge his creative energy before reviewing and polishing his composition.

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.

Large chunks



Hourly



5 ~ 15 minutes



# Agents perceive, and determines whether they need to react and edit their schedules

[Agent's Summary Description]

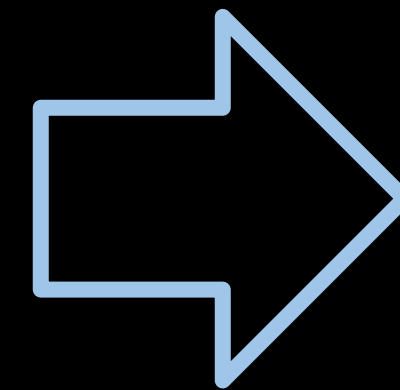
It is February 13, 2023, 4:56 pm.

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?



**Re-plan if the agent needs to react**

**In practice**

**In assignment 1...**

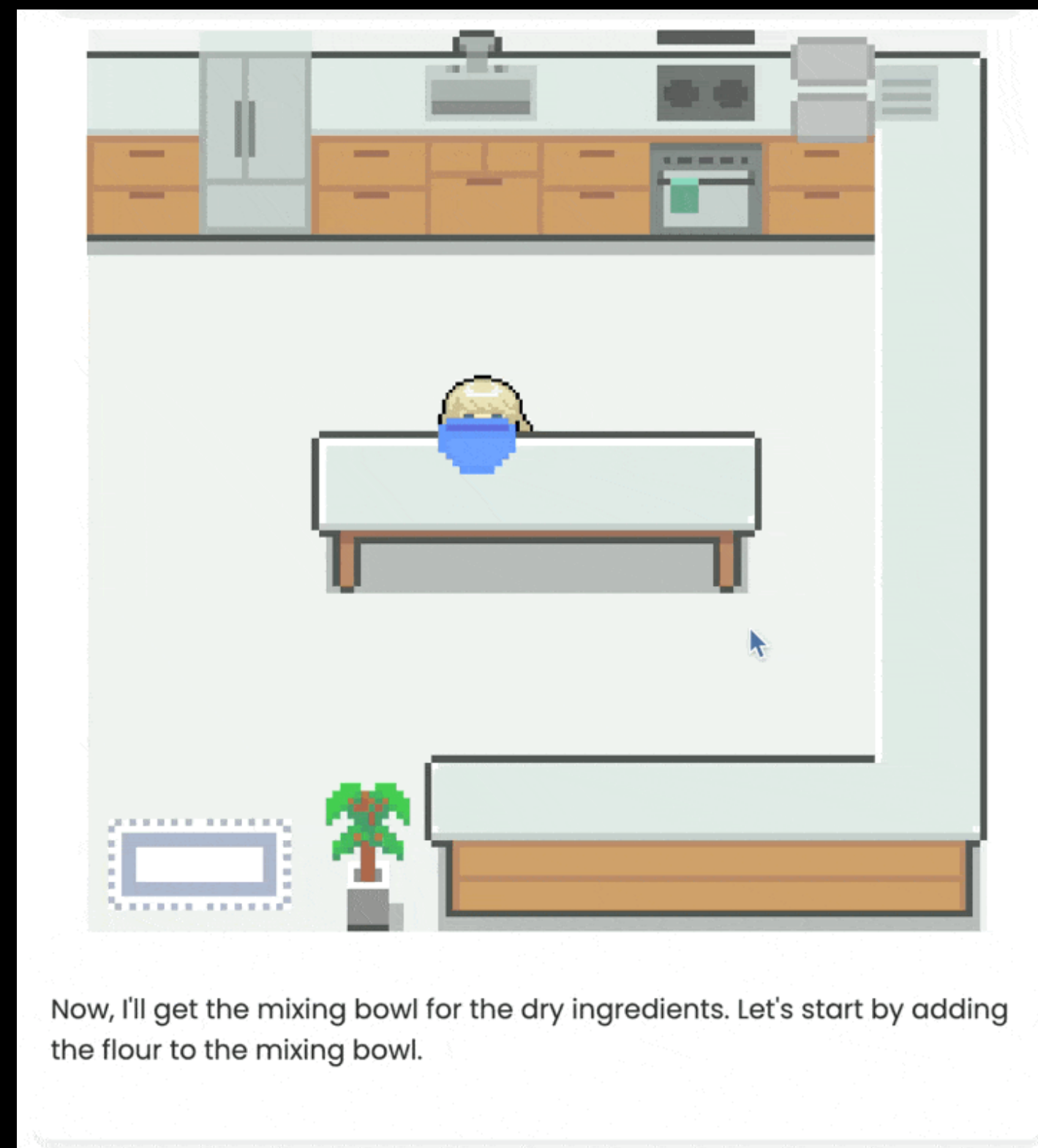
**Implement the core parts of the retrieval function**

**Implement chat function**

**+ some prompt engineering**



# In assignment 1...



**Today...**

**class ConceptNode**  
**class MemoryStream**

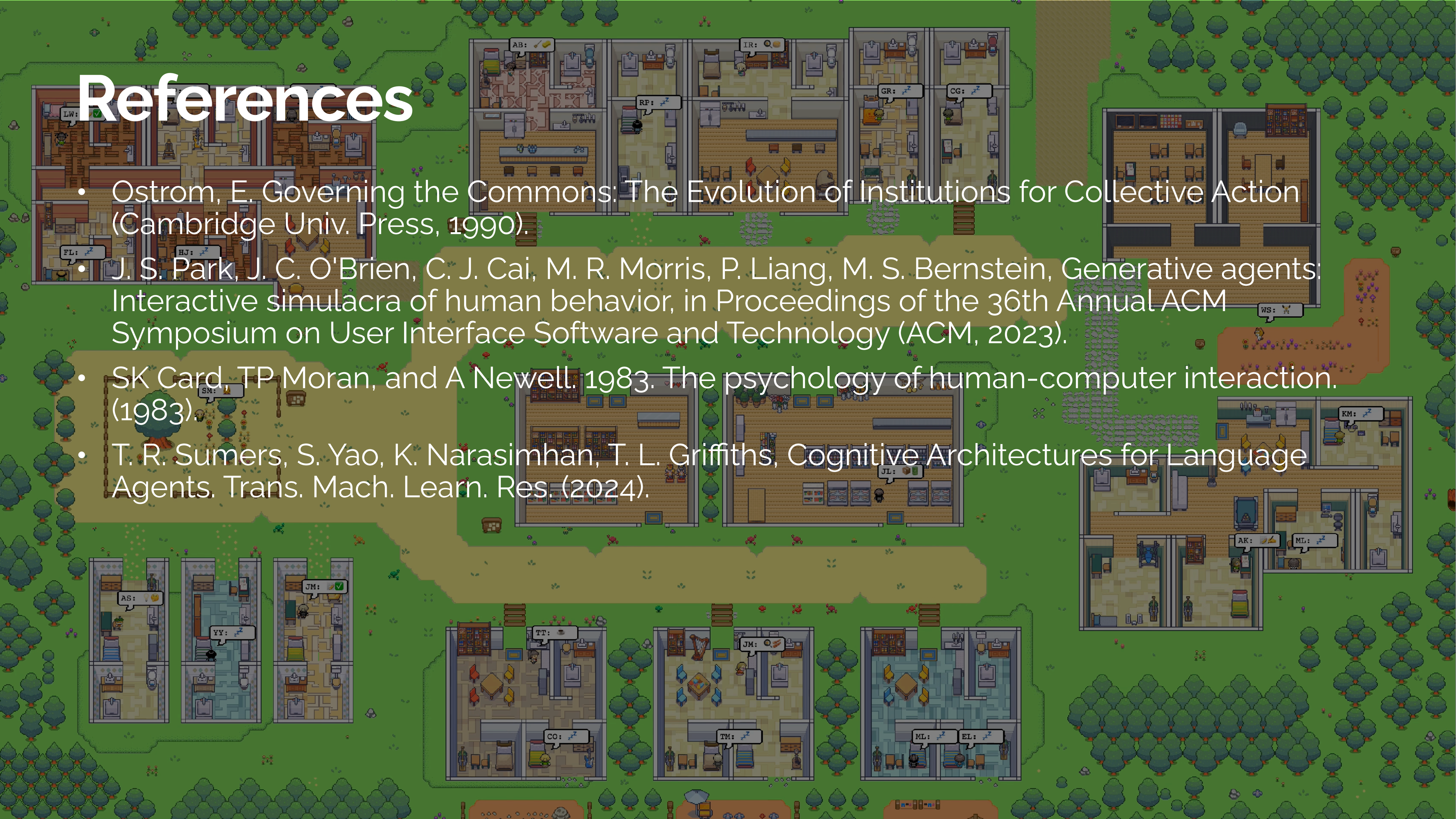
**Generate the importance score**  
**Cosine similarity scores + embeddings**

**Implement “remember”**



# References

- Ostrom, E. Governing the Commons: The Evolution of Institutions for Collective Action (Cambridge Univ. Press, 1990).
- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).
- SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).
- T. R. Summers, S. Yao, K. Narasimhan, T. L. Griffiths, Cognitive Architectures for Language Agents. Trans. Mach. Learn. Res. (2024).





The image shows a top-down view of a simulated environment, likely a campus or park. It features several interconnected buildings with various rooms, including offices, classrooms, a library, and a dining area. Each room contains furniture like desks, chairs, and bookshelves. Numerous small, stylized human figures (agents) are scattered throughout the environment, each with a speech bubble containing a two-letter code (e.g., LW, RP, AC, AB, IR, GR, CG, FL, HJ, WS, JL, KM, AS, YY, JM, TT, CO, TM, ML, EL). The environment is surrounded by green grass, trees, and a central dirt path. The overall style is a colorful, pixelated aesthetic.

# CS 222: AI Agents and Simulations

## Stanford University

### Joon Sung Park