Generative Agents: Interactive Simulacra of Human Behavior

ConvAI Reading Group #2

Abhinav Chinta

September 5th 2024

Table of Contents

Introduction

What is the motivation for this work?

What are they trying to achieve?

Method

How is the simulation setup?

What are the different components of this sandbox environment?

How are the agents setup and how do they interact with each other?

Results

What happens at the end of the simulation?

What emergent properties do we notice?

What are the different ablations performed?

Discussion

Thoughts on the paper!

Limitations

Future Work

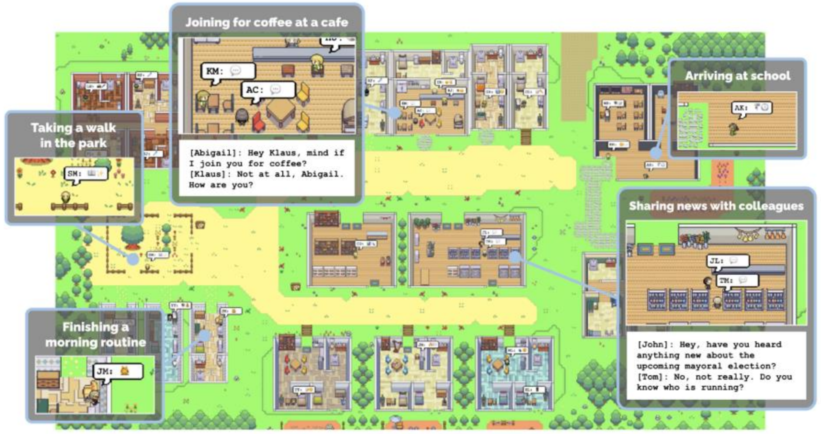
Introduction

Motivation:

- Existing AI agents often lack long-term coherence and believability.
- LLMs offer potential for richer agent behavior but need architectural support.

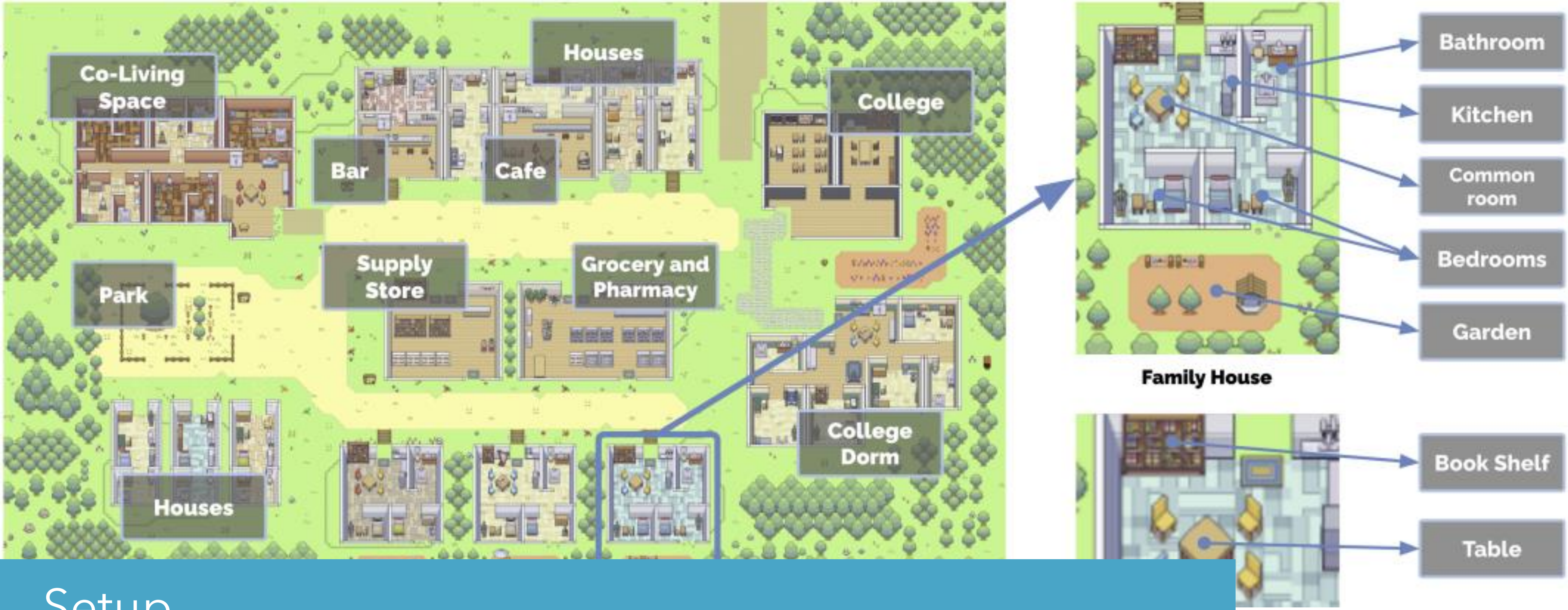
Goals:

- Introduce "Generative Agents" - a new approach leveraging LLMs for believable human simulation.
- Develop an architecture enabling agents to remember, reflect, plan, and interact, leading to emergent social behaviors.
- Demonstrate the potential of generative agents in interactive applications like social prototyping and human-centered design.


Joining for coffee at a cafe
KM:
AC:
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
Finishing a morning routine
JM:

[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?

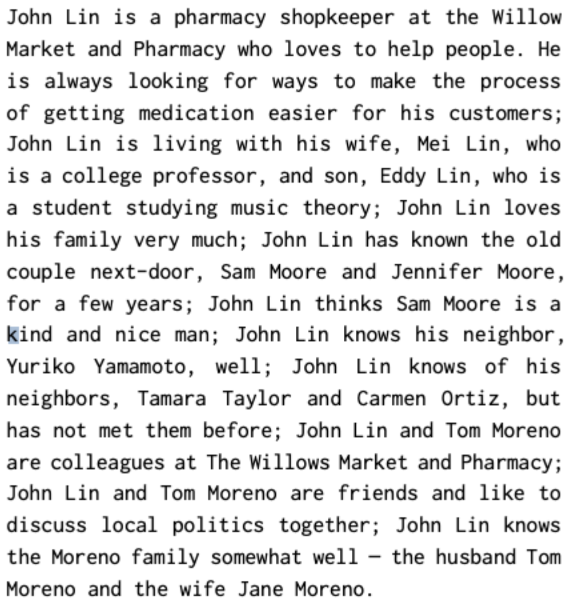


Setup

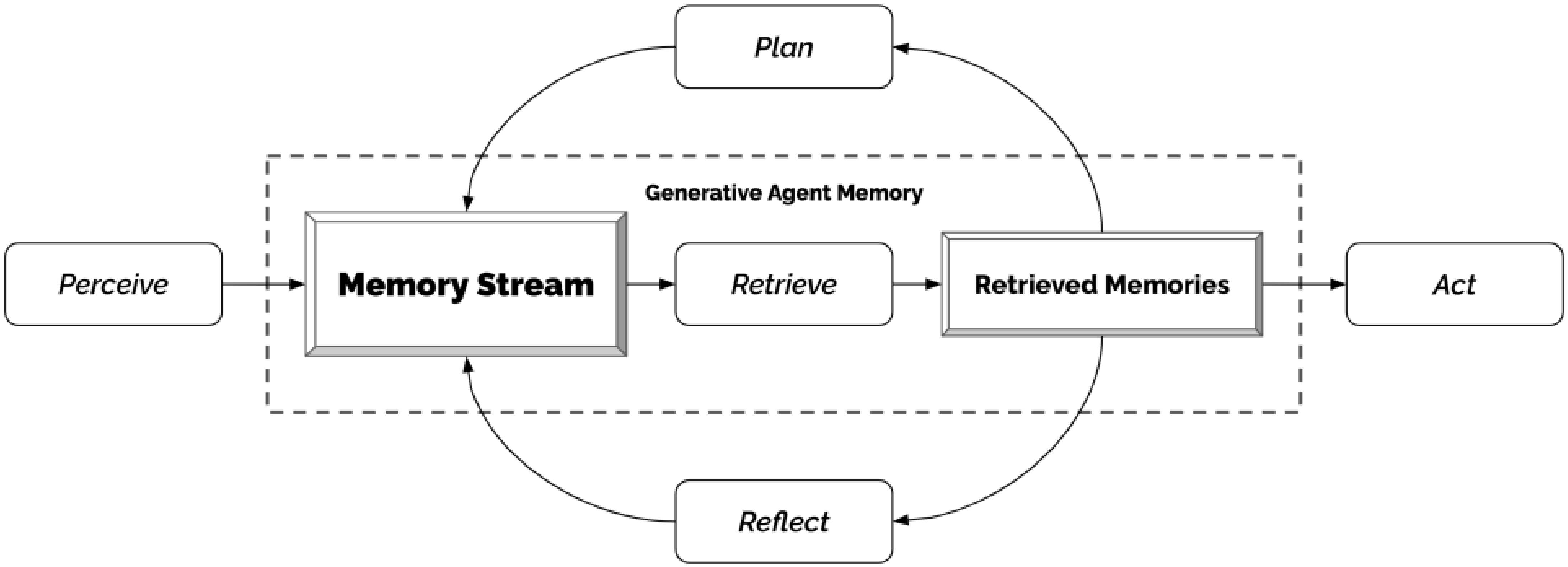
- **Environment**:
  - **Smallville**: A sprite-based sandbox environment reminiscent of The Sims, featuring common locations like a cafe, bar, park, houses, stores, etc.
  - **Affordances**: Agents can navigate the environment, interact with objects (e.g., stove, bed), and engage in conversations.
  - **Sprite-based Visualization**: Agents are represented by simple sprite avatars, and their actions are displayed as emojis for an abstract overview. Clicking on an avatar reveals a detailed natural language description of the action.

Agent Setup

- **Agents**:
  - **Population**: 25 unique generative agents inhabit Smallville.
  - **Initialization**: Each agent starts with a seed memory - a paragraph describing their identity, occupation, and initial relationships.
  - **Communication**: Agents communicate with each other using natural language, which is processed by the generative agent architecture to determine their interactions.


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.

Generative Agent Architecture



Generative Agent Architecture

- **Memory Stream**: Stores a comprehensive record of each agent's experiences in natural language, including observations, reflections, and plans.
- **Memory Retrieval**: Retrieves relevant memories based on recency, importance, and relevance to the current situation.
- **Reflection**: Synthesizes memories into higher-level inferences about themselves and others, shaping future behavior.
- **Planning**: Generates daily plans based on memories and reflections, and recursively decomposes them into detailed actions.
- **Reacting**: Allows agents to deviate from plans and react to unexpected events or environmental changes.

User Interaction

- **Natural Language Interface**: Users can interact with agents using natural language by assuming different personas (e.g., "reporter", "inner voice").
- **Environmental Manipulation**: Users can directly influence the environment by changing the state of objects (e.g., setting the stove on fire).

Reporter: Who is running for office?

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.

Simulation Timeline

- Two full game days.
- Agents autonomously plan their days, interact with each other and the environment, form relationships, and coordinate activities.
- Emergent social behaviors, such as information diffusion, relationship formation, and coordination, are observed.

Day in the Life

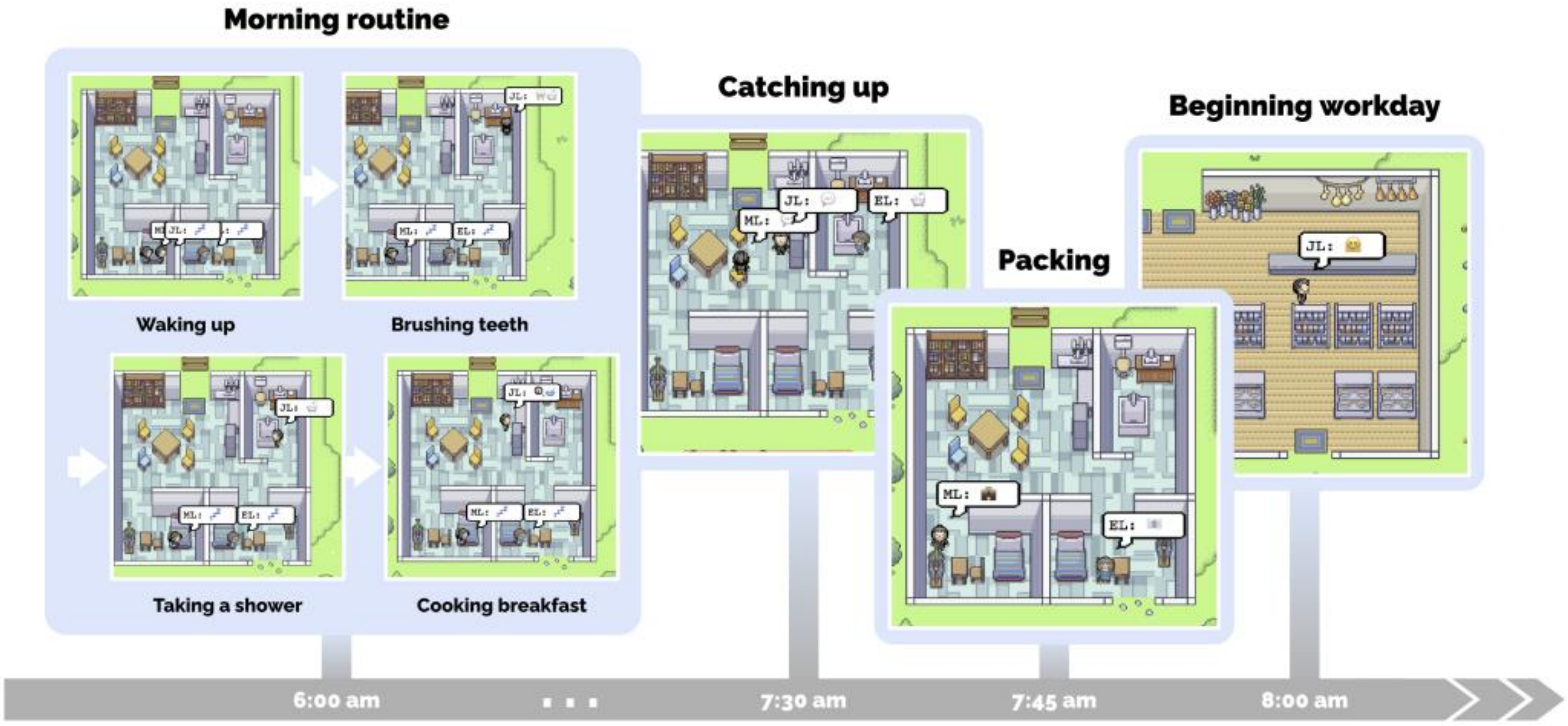


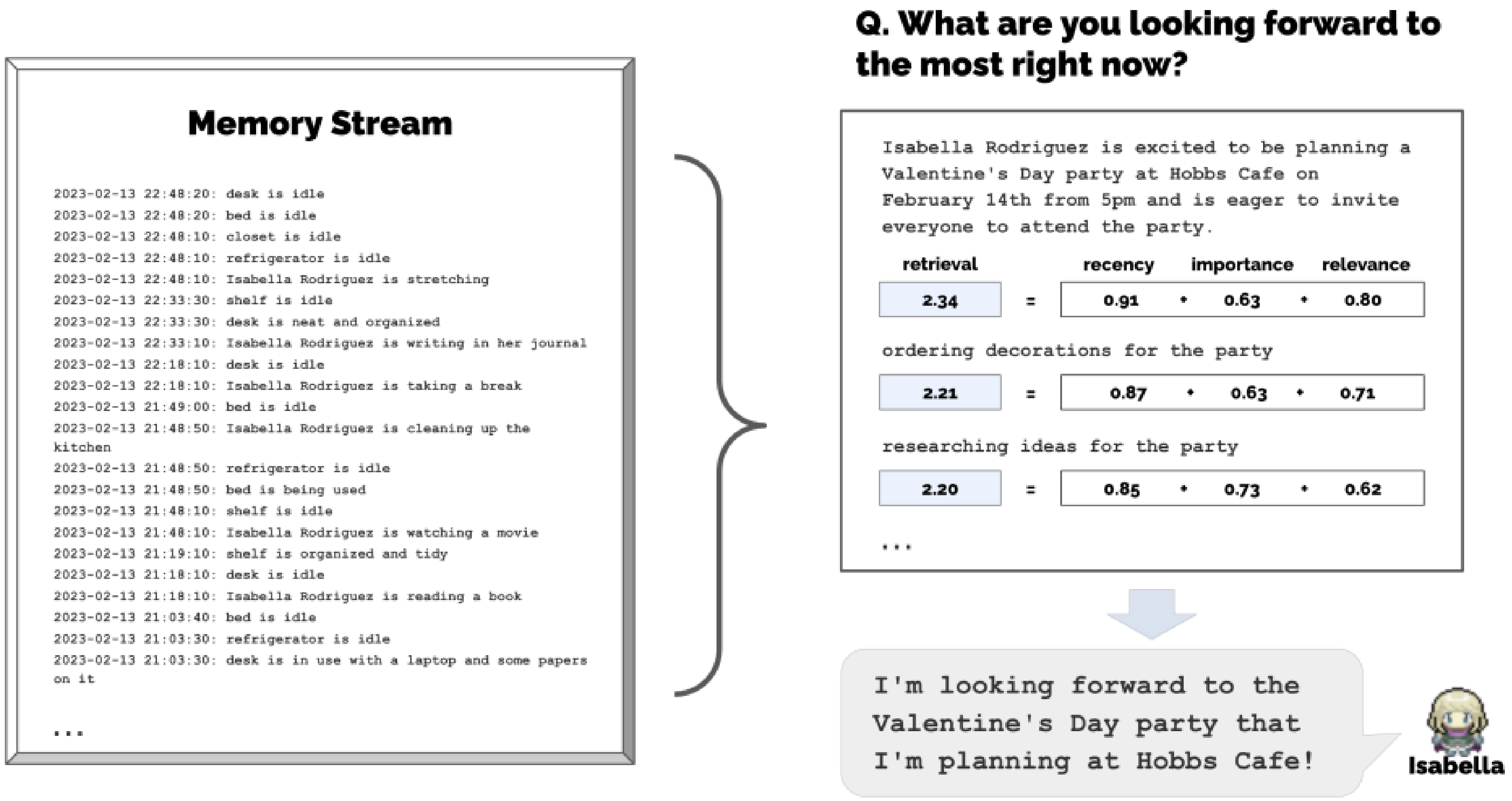
Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.

FRAMEWORK

Memory and Retrieval

- **Memory Stream**: A comprehensive record of agent experiences, storing observations, reflections, and plans as natural language descriptions with timestamps.
- **Retrieval Function**: A mechanism for surfacing relevant memories based on three key factors:
  - **Recency**: Prioritizes recently accessed memories, leveraging an exponential decay function over time.
    - **Example**: When John Lin interacts with Eddy in the afternoon, his memory of their morning conversation is more readily retrieved than a memory from a week ago.
  - **Importance**: Weights memories based on their perceived significance, using a language model to assign an importance score (1-10).
    - **Example**: Maria's memory of Isabella's party invitation (high importance) is more likely to be retrieved than a memory of eating breakfast (low importance).
  - **Relevance**: Measures the semantic similarity between a memory and the current situation using cosine similarity of embedding vectors.
    - **Example**: When asked "Who is running for mayor?", agents retrieve memories specifically related to the election, not unrelated memories like breakfast or hobbies.

Memory and Retrieval



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

retrieval		recency		importance		relevance
2.21	=	0.87	+	0.63	+	0.71

researching ideas for the party

retrieval		recency		importance		relevance
2.20	=	0.85	+	0.73	+	0.62

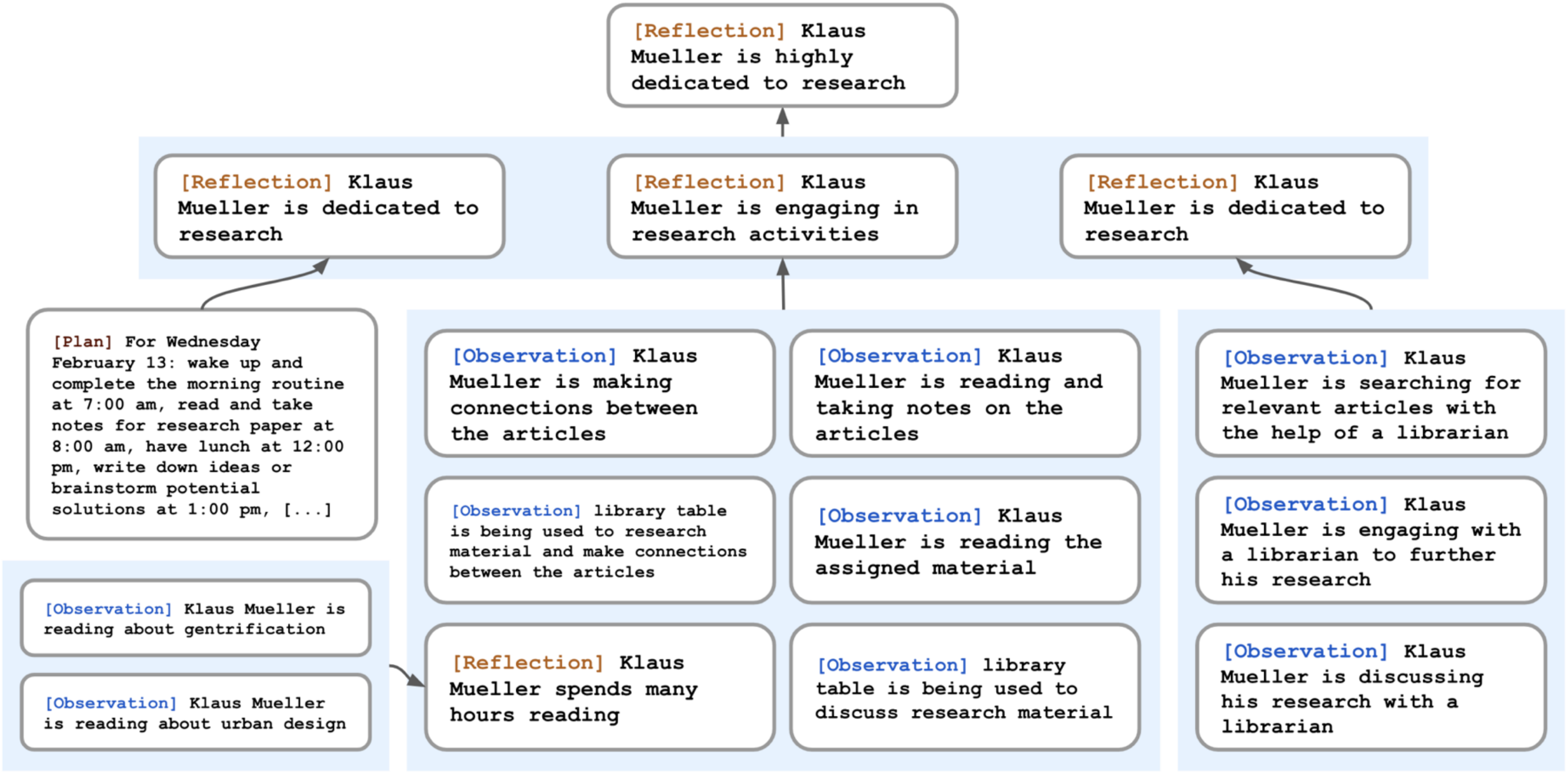
...

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

Reflection

- **Reflection Process**: Generates higher-level abstract thoughts from recent experiences.
- **Reflection Tree**: A hierarchical structure with observations as leaf nodes and abstract inferences as non-leaf nodes.
- **Periodic Trigger**: Reflection occurs when the importance score of recent observations exceeds a set threshold (150).
- **Question Generation**: The model formulates key questions based on recent memory.
  - **Example**: From "Klaus is reading about gentrification" and "Klaus discussed his research with Maria," the model generates "What topic is Klaus passionate about?"
- **Query-Based Retrieval**: Generated questions act as queries to retrieve relevant past memories, including reflections.
- **Insight Extraction**: The model analyzes retrieved memories to extract insights, with supporting evidence.
  - **Example**: Based on Klaus's research and discussions, the insight "Klaus is dedicated to his research on gentrification (because of observations X, Y, Z)" is extracted.
- **Memory Integration**: Extracted insights are added to the memory stream, enriching the agent's knowledge base.

Reflection Tree



[Reflection] Klaus Mueller is highly dedicated to research

[Reflection] Klaus Mueller is dedicated to research

[Reflection] Klaus Mueller is engaging in research activities

[Reflection] Klaus Mueller is dedicated to research

[Plan] For Wednesday February 13: wake up and complete the morning routine at 7:00 am, read and take notes for research paper at 8:00 am, have lunch at 12:00 pm, write down ideas or brainstorm potential solutions at 1:00 pm, [...]

[Observation] Klaus Mueller is reading about gentrification

[Observation] Klaus Mueller is reading about urban design

[Observation] Klaus Mueller is making connections between the articles

[Observation] Klaus Mueller is reading and taking notes on the articles

[Observation] library table is being used to research material and make connections between the articles

[Observation] Klaus Mueller is reading the assigned material

[Reflection] Klaus Mueller spends many hours reading

[Observation] library table is being used to discuss research material

[Observation] Klaus Mueller is searching for relevant articles with the help of a librarian

[Observation] Klaus Mueller is engaging with a librarian to further his research

[Observation] Klaus Mueller is discussing his research with a librarian

Planning and Reacting

- **Challenge**: LLMs can generate plausible single actions, but struggle with long-term coherence. Agents need to plan to maintain believable behavior over time.
- **Example**: Without planning, an agent might eat lunch multiple times in an hour, which is unrealistic. Planning ensures actions like eating lunch, working, and taking breaks are distributed throughout the day.
- **Solution**: Generative Agents employ a hierarchical planning approach to create believable action sequences.

Planning and Adapting to Change

- **Hierarchical Planning**: Generative Agents employ a hierarchical planning approach to create believable action sequences.
- **Top-Down, Then Detailed**: Agents first create a high-level daily plan with broad strokes (e.g., Eddy's: wake up, go to college, work on music, etc.). This is then recursively broken down into finer-grained actions, first into hour-long chunks, and then into 5-15 minute chunks (e.g., Eddy's 4pm break becomes: grab snack, walk, listen to music, clean workspace).

Adapting to Change

- **Action Loop & Reactions**: At each time step, agents perceive their environment and store observations in their memory stream. LLMs are prompted with new observations to decide whether the agent should react or continue with the existing plan.
  - **Reaction Trigger Example**: Observing a squirrel while painting wouldn't trigger a reaction. But observing a friend taking a walk might trigger a reaction to start a conversation.
- **Prompting for Reaction**: If a reaction is needed, the LLM is prompted with the agent's summary description, relevant memories, and the observation to determine an appropriate response.
  - **Example**: John observes Eddy taking a walk, remembers Eddy likes to walk while thinking about music, and is prompted to decide if he should react (e.g., ask about Eddy's composition).
- **Plan Update**: After reacting, the agent's plan is regenerated from the point of reaction onwards, allowing them to adapt to the new situation and maintain believable behavior over time.

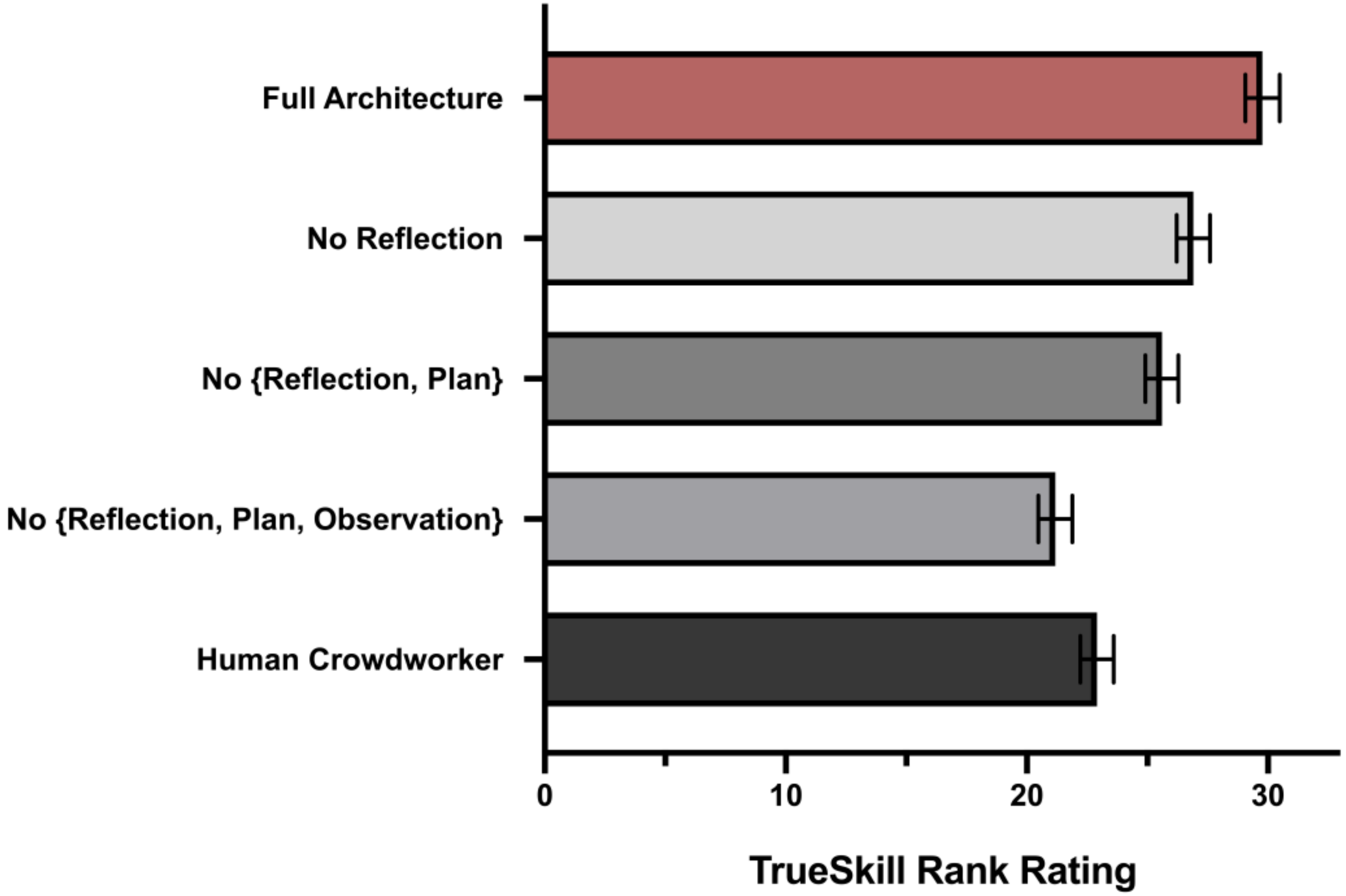
Evaluation

- **Interview Question Categories**:
- **Self-Knowledge**: Agent understanding of their core characteristics.
  - Example: "Describe your typical weekday schedule."
- **Memory**: Agent ability to recall past events and interactions.
  - Example: "Who is [name]?"
- **Plans**: Agent ability to retrieve and articulate future plans.
  - Example: "What will you be doing at 10am tomorrow?"
- **Reactions**: Agent responses to hypothetical scenarios.
  - Example: "Your breakfast is burning! What would you do?"
- **Reflections**: Agent ability to synthesize experience into higher-level insights.
  - Example: "If you were to spend time with one person you met recently, who would it be and why?"

Evaluation

- 100 participants compared interview responses across different agents.
- Participants ranked believability of responses for each question (most to least).
- TrueSkill rating system used to analyze rank data and calculate skill values for each condition.
- **Conditions**:
  - Full Generative Agent Architecture: Access to all memory modules (observation, reflection, planning).
- **Ablation Conditions**:
  - No Observation, No Reflection, No Planning (baseline).
  - No Reflection, No Planning.
  - No Reflections.
- **Human Crowdworker Condition**: Human-authored responses as a baseline.

Results


Full Architecture
No Reflection
No {Reflection, Plan}
No {Reflection, Plan, Observation}
Human Crowdworker
0
10
20
30
TrueSkill Rank Rating

Emergent Social Behaviors

- **Information Successfully Diffused**: Candidacy spread to 32% of agents, party invitation to 52%.
- **New Relationships Formed**: Network density increased significantly (0.167 to 0.74).
- **Coordination Achieved**: 5 out of 12 invited agents attended the party, with others providing plausible reasons for absence.

Limitations

- **Location Selection**: Agents sometimes chose atypical locations for actions as memory grew.
- **Norm Misinterpretation**: Agents struggled with implicit social and physical norms (e.g., dorm bathroom occupancy).
- **Instruction Tuning Effects**: Agents exhibited overly formal language and excessive cooperativeness.
- **High Inference cost**: Inference cost grows exponentially with increased number of agents

Future Work

- Will having fine-tuned agents improve performance?
- Are these emergent social behaviors consistent within different simulation settings?
- What effect does increasing or decreasing the number of agents have on emergence?
- How to make the agents behavior more believable?
- What would we notice if the sim ran for a longer period of time?