LLMs and Social Behaviors

Generative Agents: Interactive Simulacra of Human Behavior

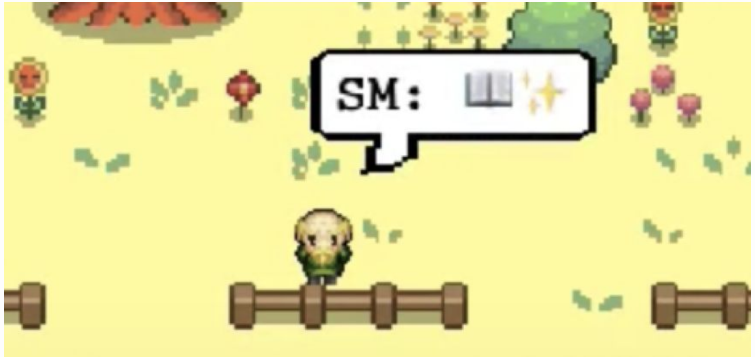
Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein

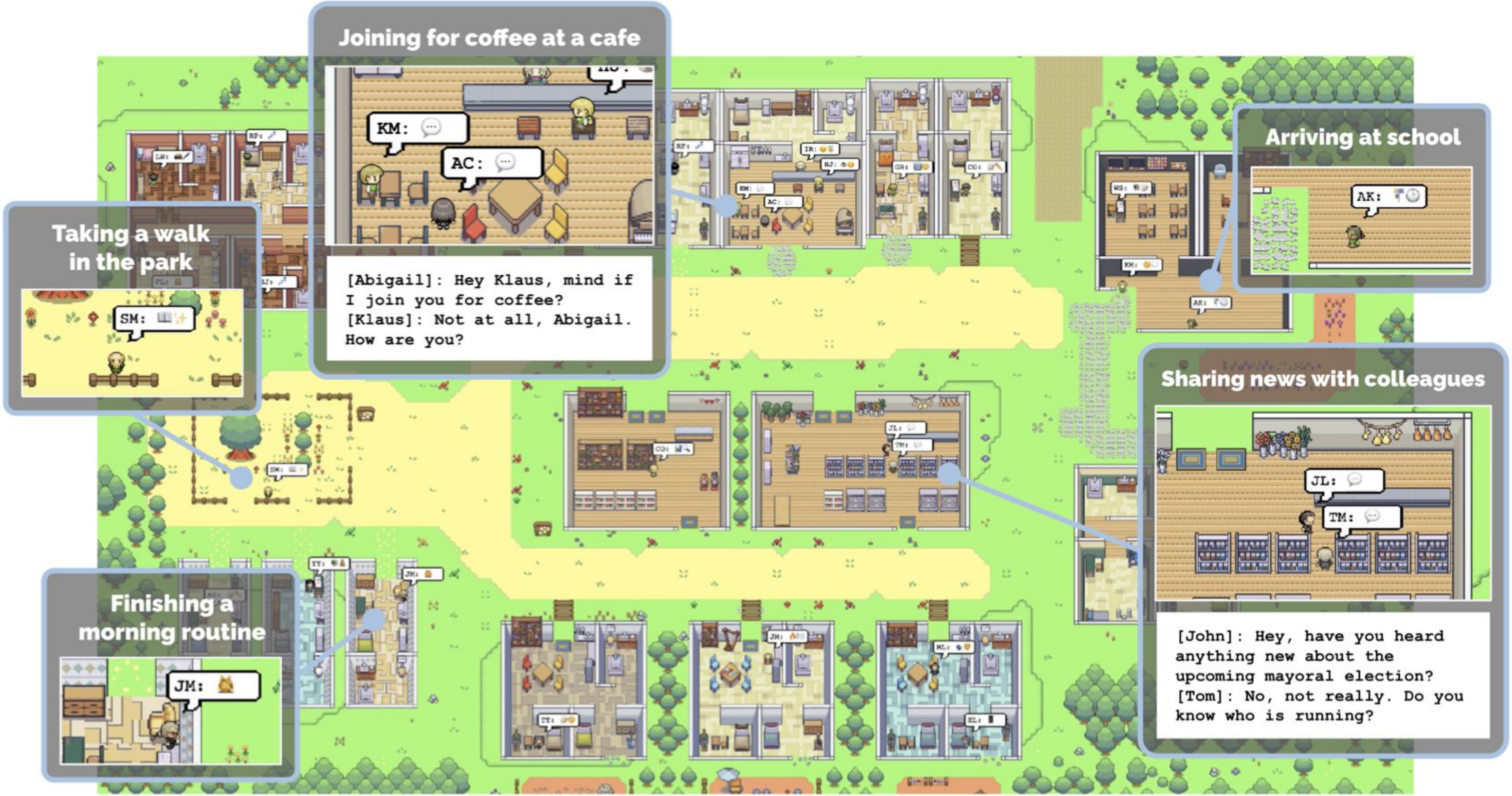
Motivation

- For four decades, we envisioned the ability to simulate **believable human behavior**.
- The ability to achieve this promises a new class of interactive applications:
  - Social simulations for testing social science theories
  - Model human processors for usability testing
  - Virtual words NPCs
- However, the space of possible human behavior has been **too vast and complex** to recreate with existing methods.
- A new opportunity: generative models trained today encode the way we live, talk, and behave

Generative agents

believable simulacra of human behavior





Joining for coffee at a cafe

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

Taking a walk in the park

Arriving at school

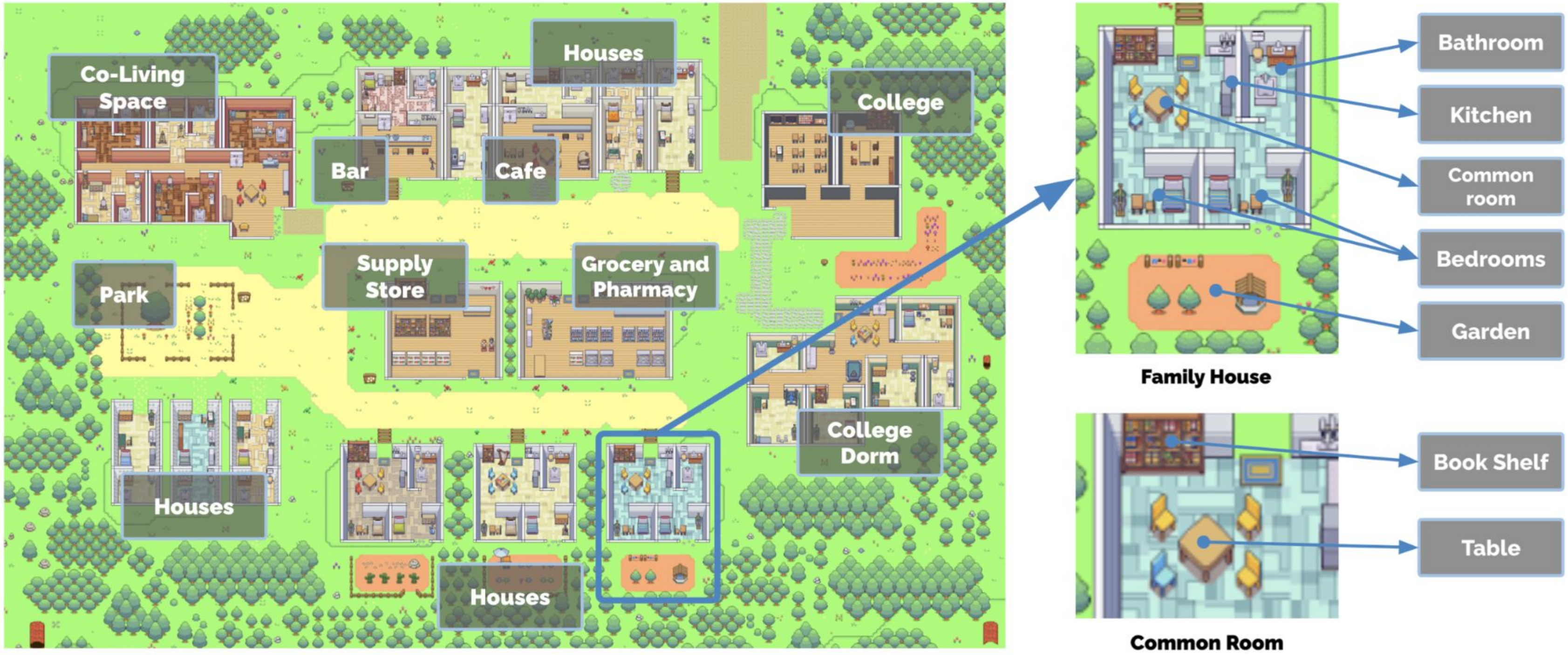
Sharing news with colleagues

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

Finishing a morning routine

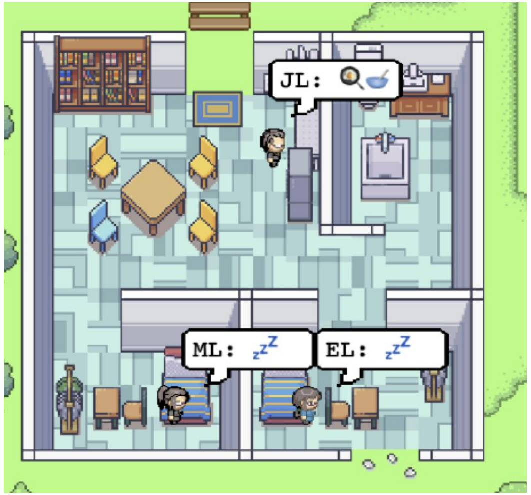
A sandbox environment with 25 generative agents

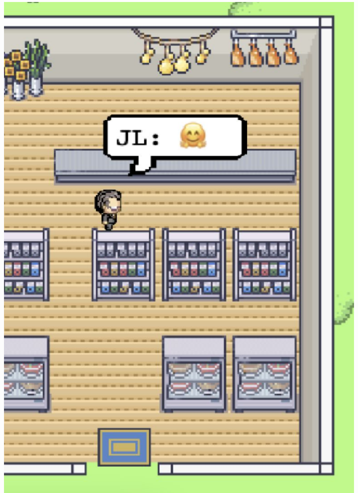
Environment



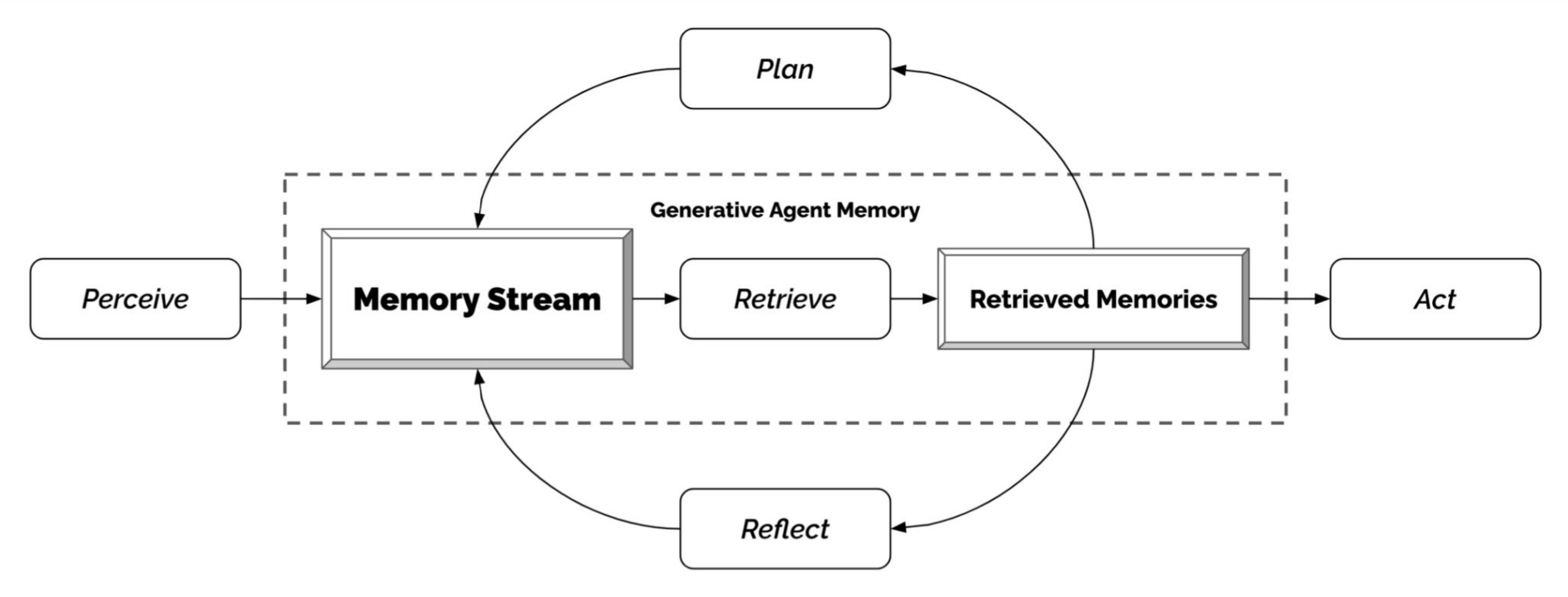
Agents

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.

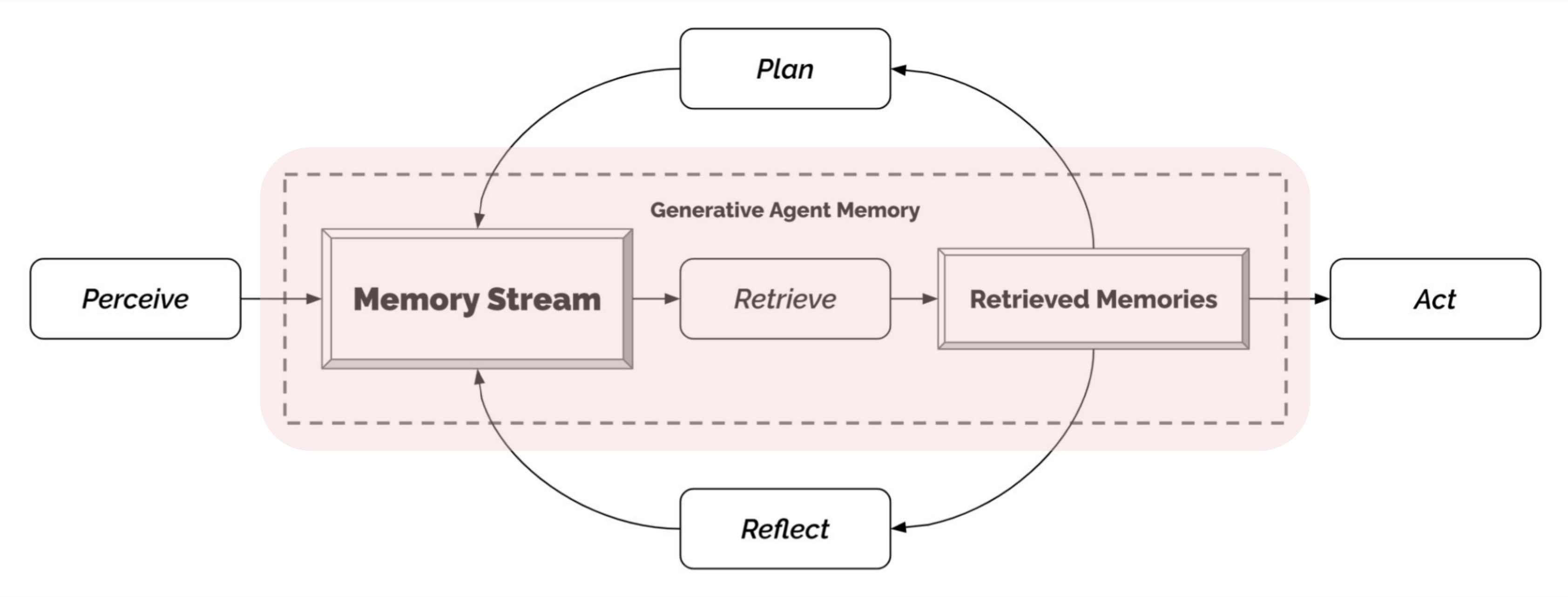




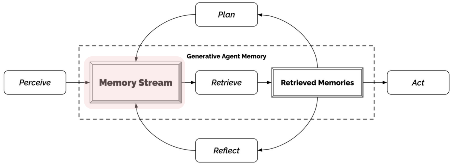
Generative Agent Architecture

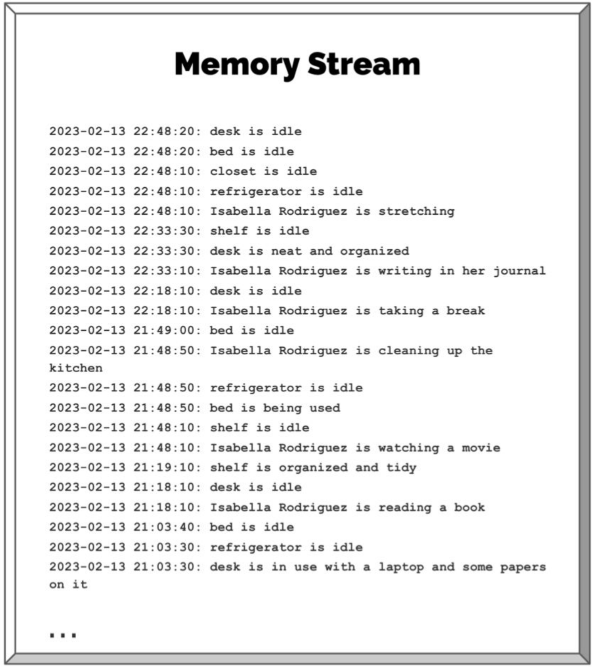


Generative Agent Architecture



Memory & Retrieval



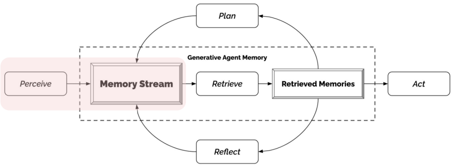


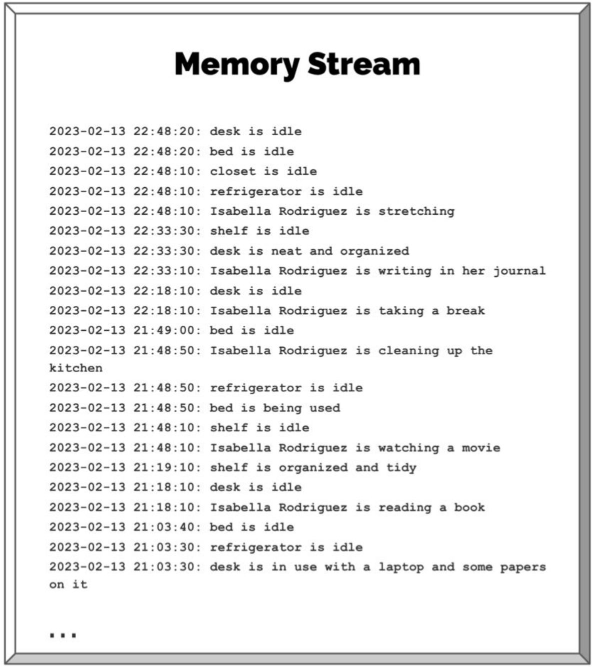
Memory Stream

```
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

...
```

Memory & Retrieval





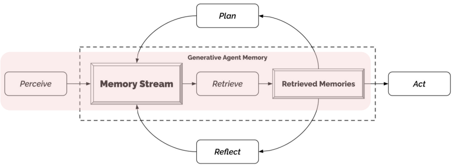
Memory Stream

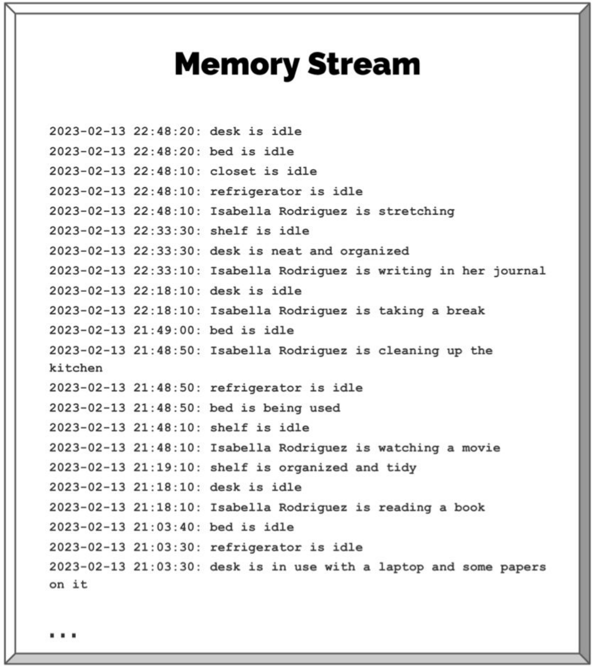
```
2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: 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?

Memory & Retrieval



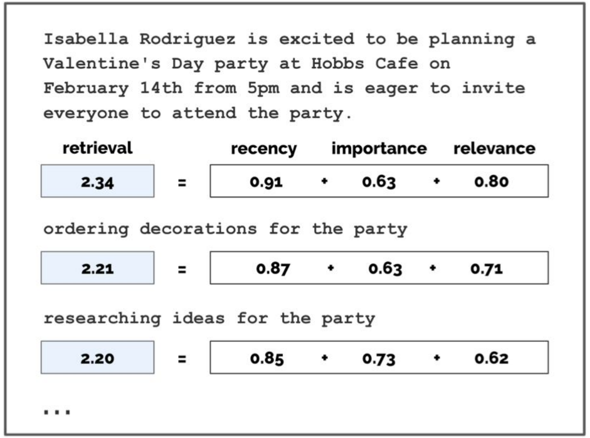


Memory Stream

```
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

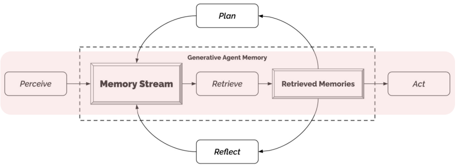
retrieval		recency		importance		relevance
2.21	=	0.87	+	0.63	+	0.71

researching ideas for the party

retrieval		recency		importance		relevance
2.20	=	0.85	+	0.73	+	0.62

...

Memory & 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

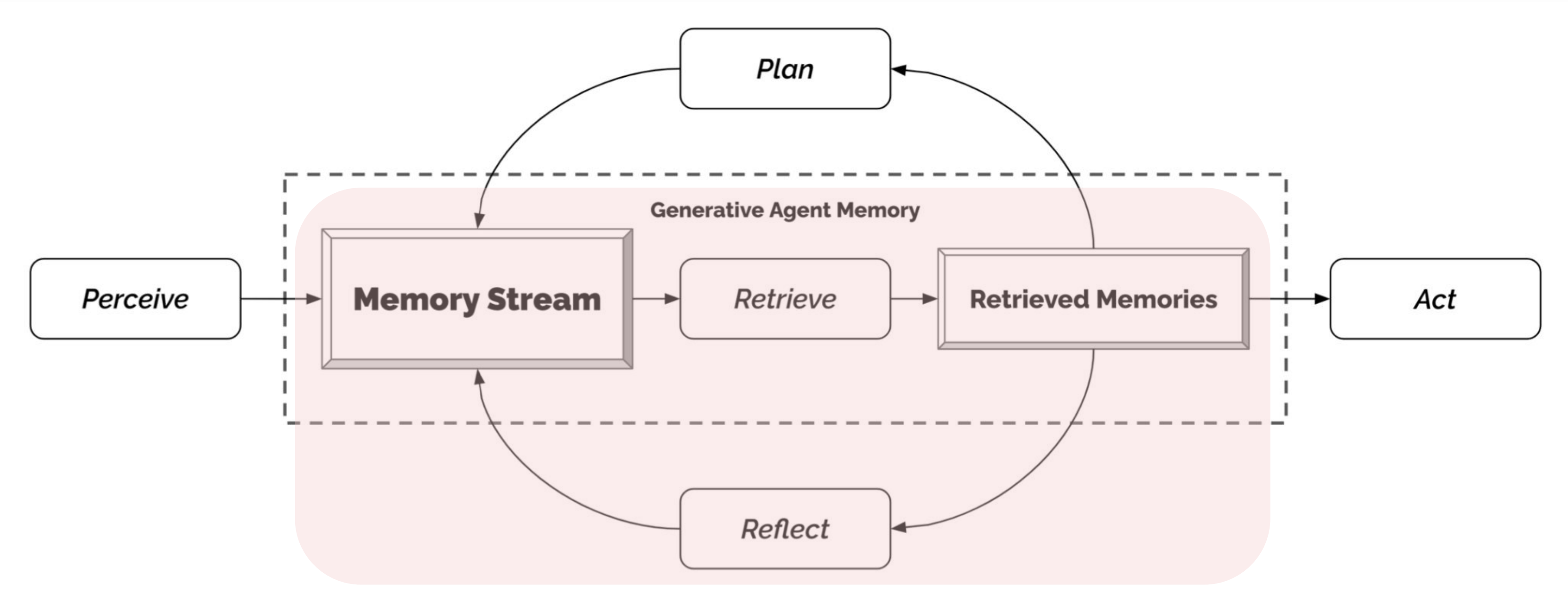
researching ideas for the party

retrieval		recency		importance		relevance
2.20	=	0.85	+	0.73	+	0.62

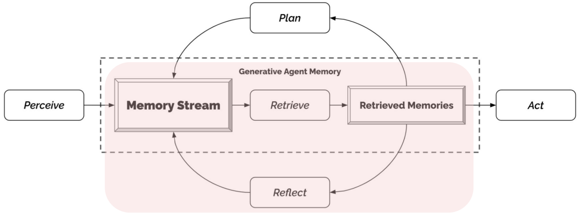
...

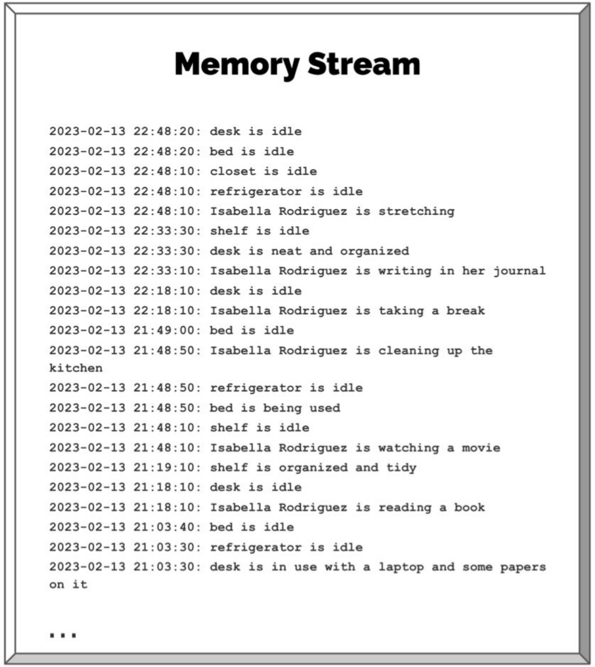
I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!

Generative Agent Architecture



Reflection





Memory Stream

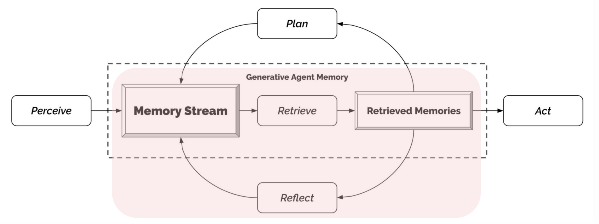
```
2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: 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: higher-level, abstract thoughts generated by the agent

- A type of memory
- Synthesized periodically

Reflection



Plan

Generative Agent Memory

Perceive

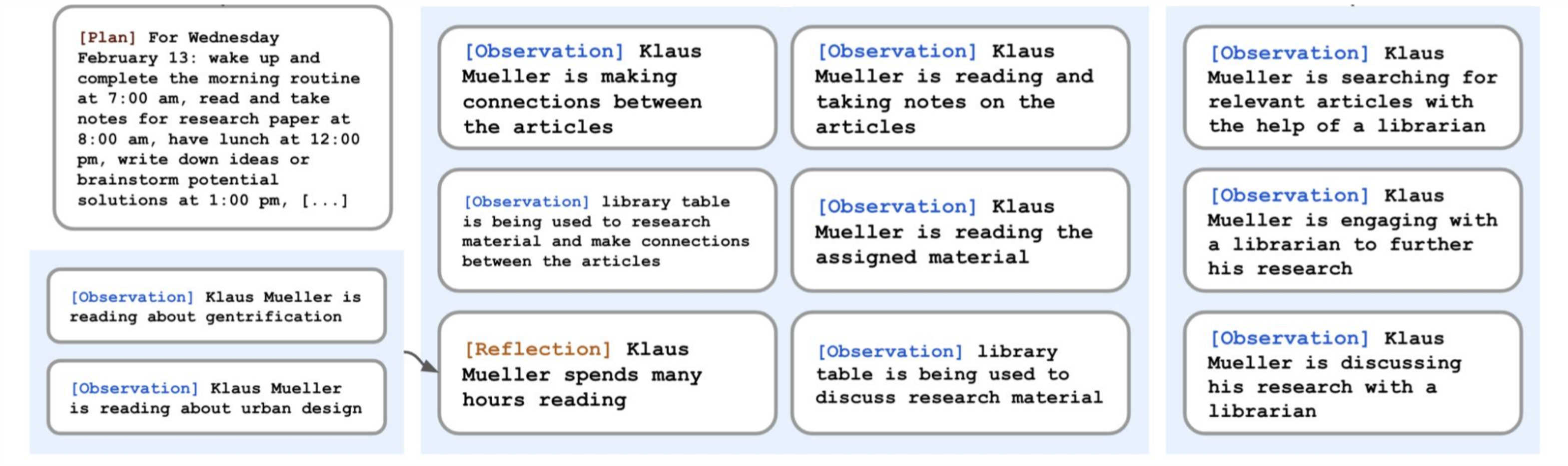
Memory Stream

Retrieve

Retrieved Memories

Act

Reflect



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

[Reflection] Klaus Mueller spends many hours reading

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

[Observation] Klaus Mueller is reading the assigned material

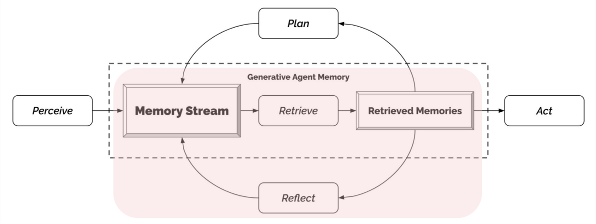
[Observation] library table is being used to discuss research material

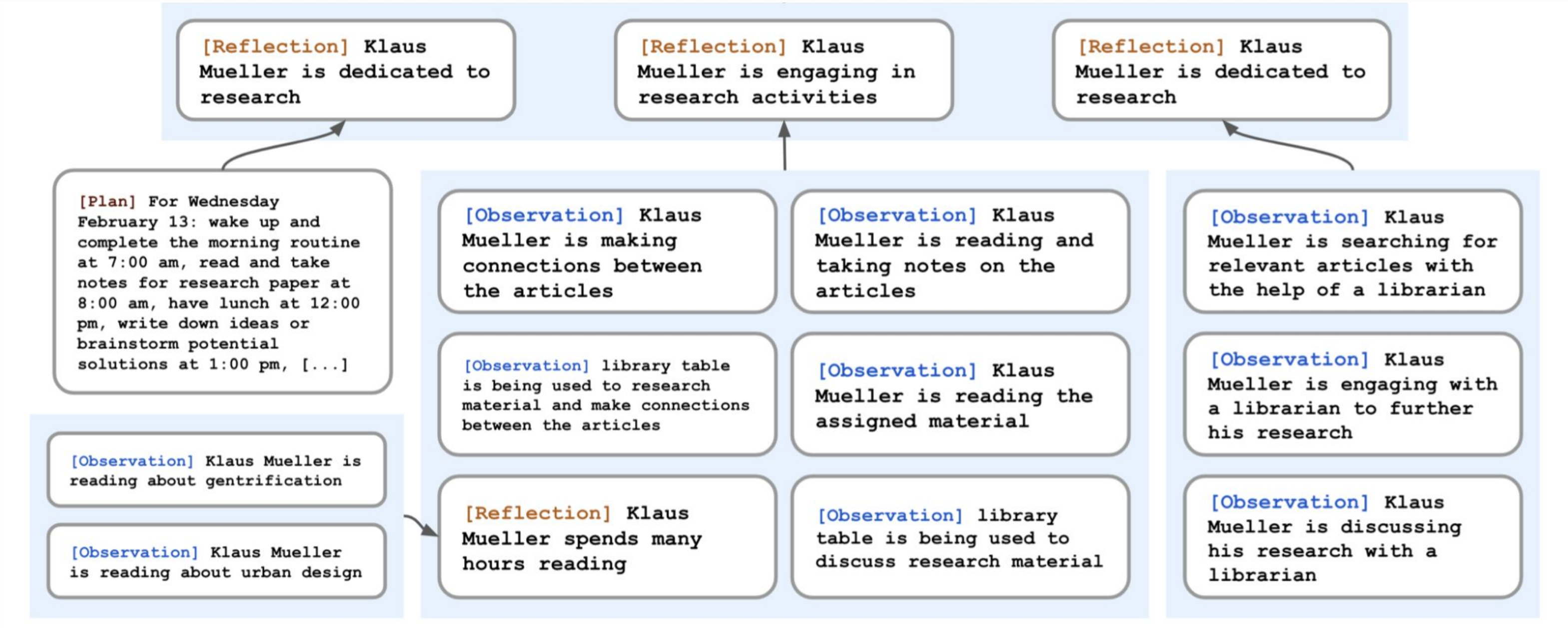
[Observation] Klaus Mueller is searching for relevant articles with the help of a librarian

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

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

Reflection





[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

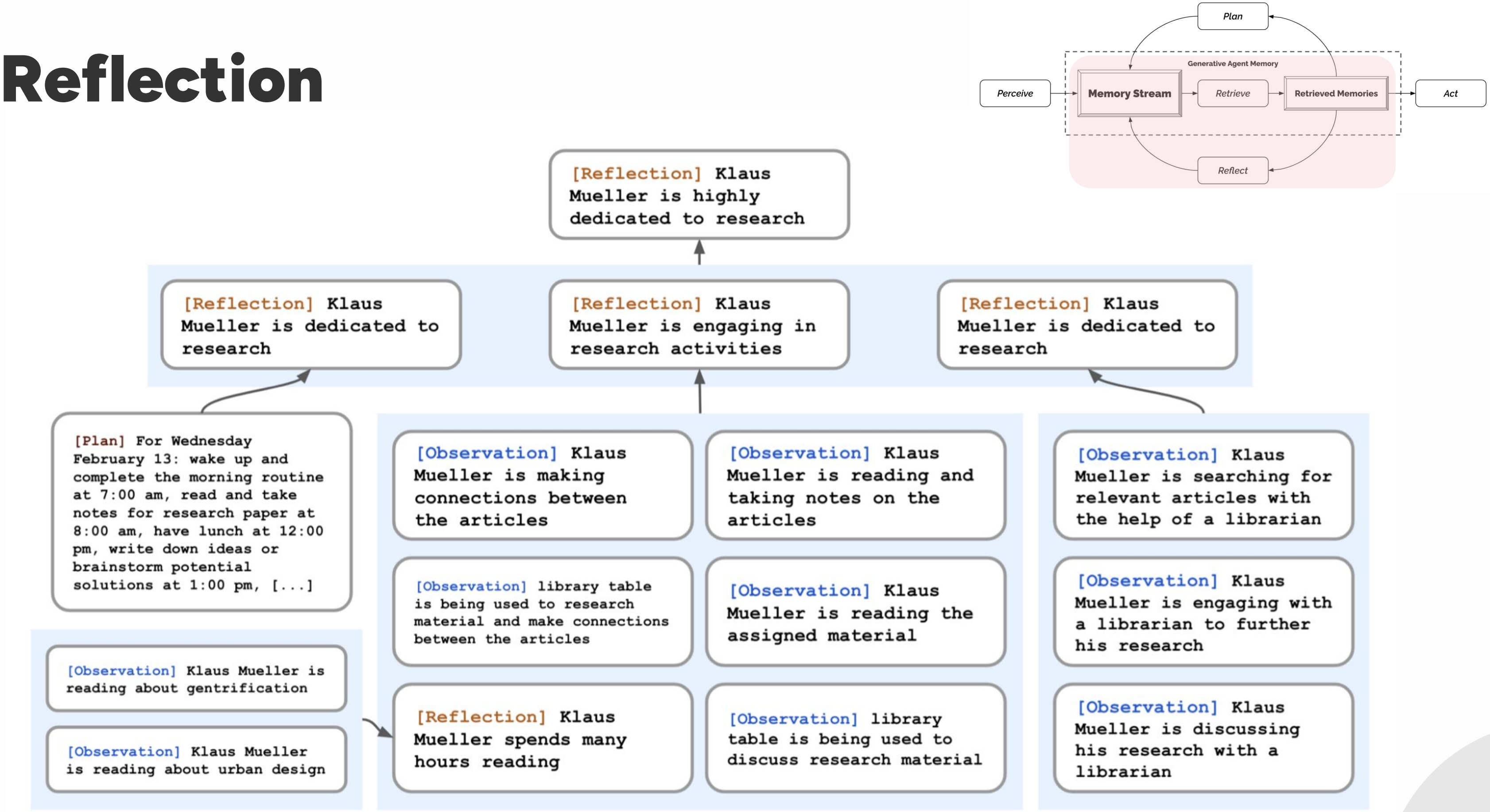
[Observation] library table is being used to discuss research material

[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

Reflection



[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

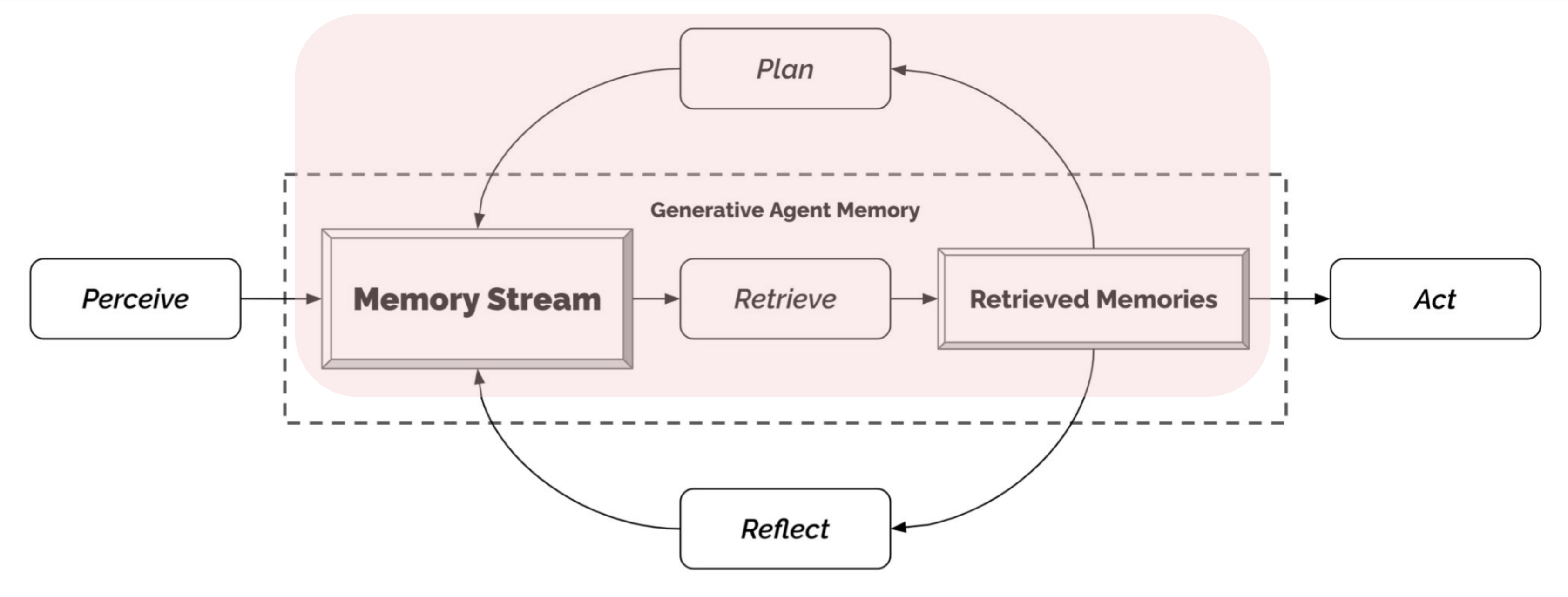
[Observation] library table is being used to discuss research material

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

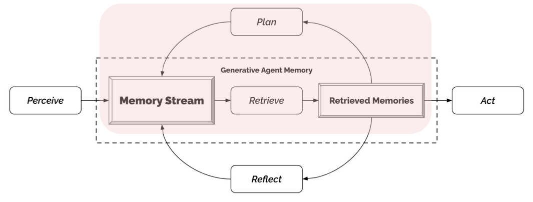
[Observation] Klaus Mueller is engaging with a librarian to further his research

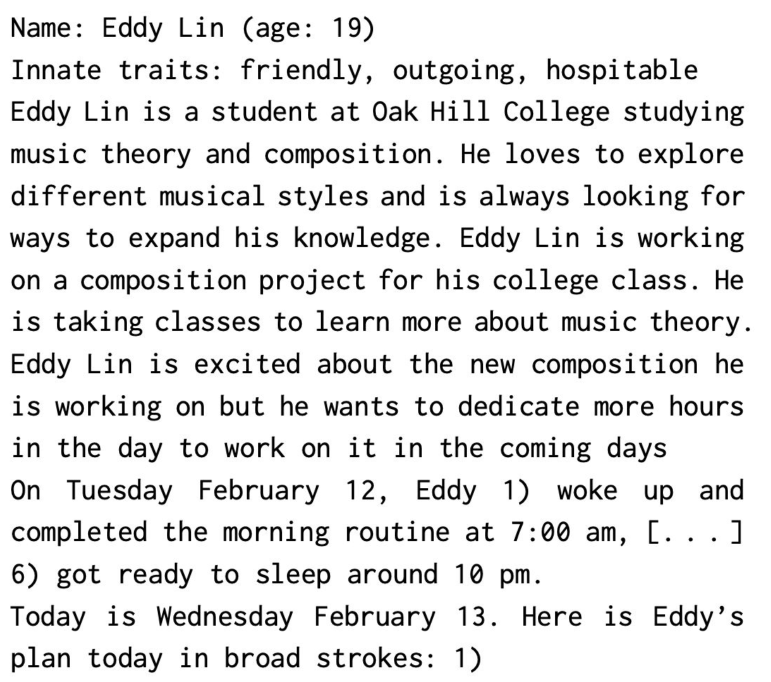
[Observation] Klaus Mueller is discussing his research with a librarian

Generative Agent Architecture



Planning



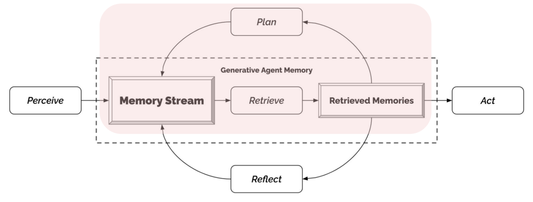


Name: Eddy Lin (age: 19)
Innate traits: friendly, outgoing, hospitable
Eddy Lin is a student at Oak Hill College studying music theory and composition. He loves to explore different musical styles and is always looking for ways to expand his knowledge. Eddy Lin is working on a composition project for 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)

Agent summary description

Current status

Planning



To plan, start top-down and then recursively generate more details in the plan

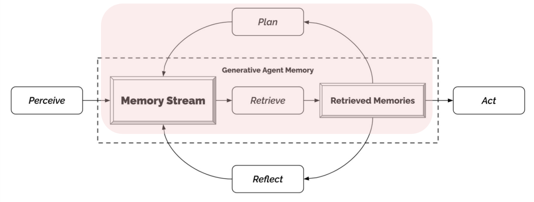
```
1) wake up and complete the
morning routine at 8:00 am,
2) go to Oak Hill College to
take classes starting 10:00
am, [. . . ] 5) work on his
new music composition from
1:00 pm to 5:00 pm, 6) have
dinner at 5:30 pm, 7) finish
school assignments and go to
bed by 11:00 pm.
```

```
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.
```

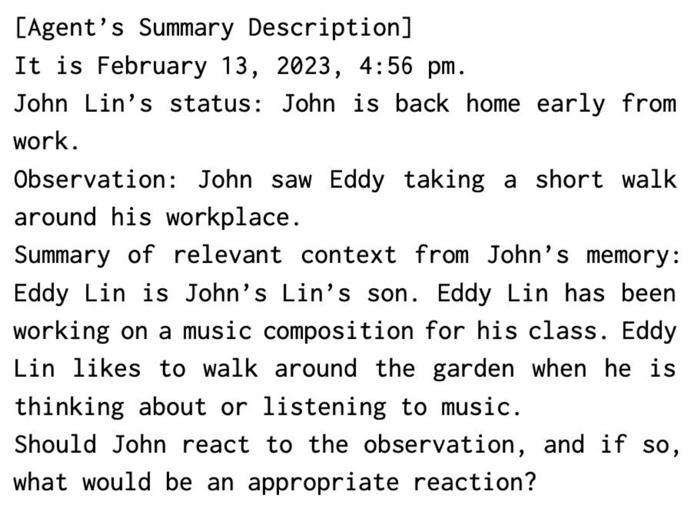
```
4:00 pm: grab a light snack,
such as a piece of fruit, a
granola bar, or some nuts.
4:05 pm: take a short walk
around his workspace [...]
4:50 pm: take a few minutes
to clean up his workspace.
```

Large chunks → Hourly → 5 - 15 minutes

Planning



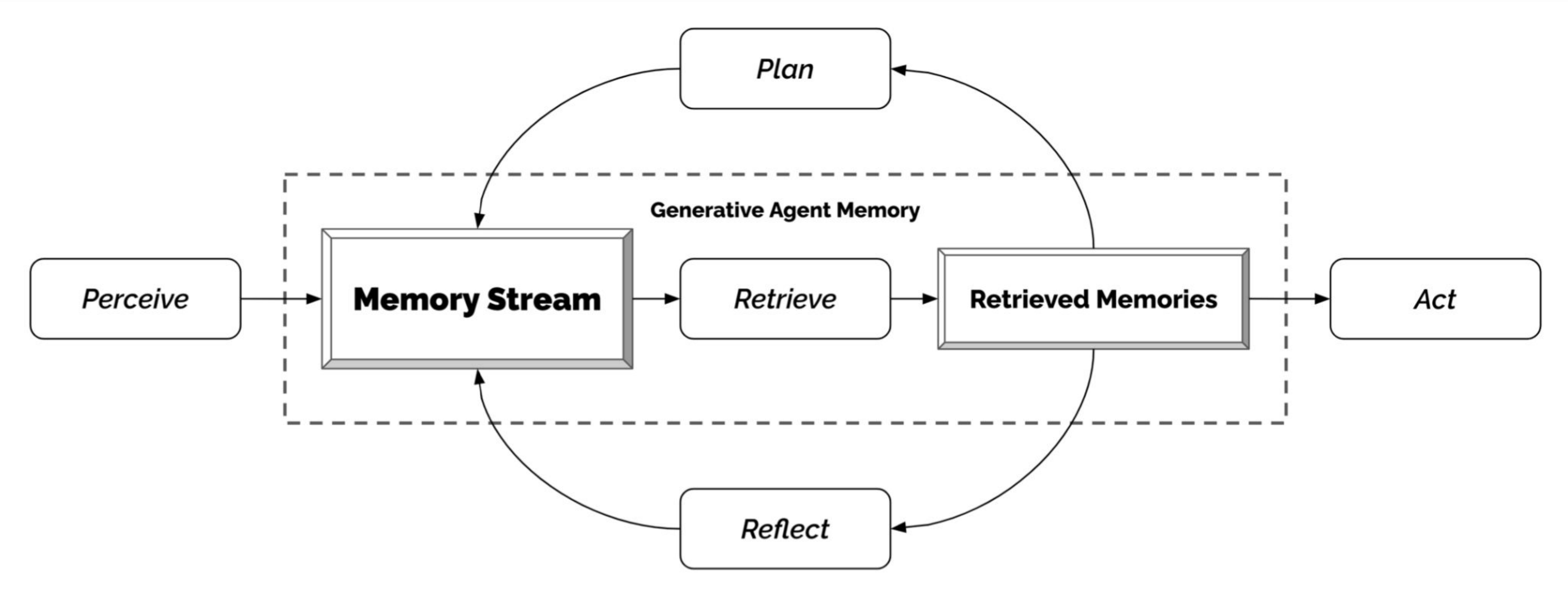
Agents perceive and determine whether they need to react and edit their plans



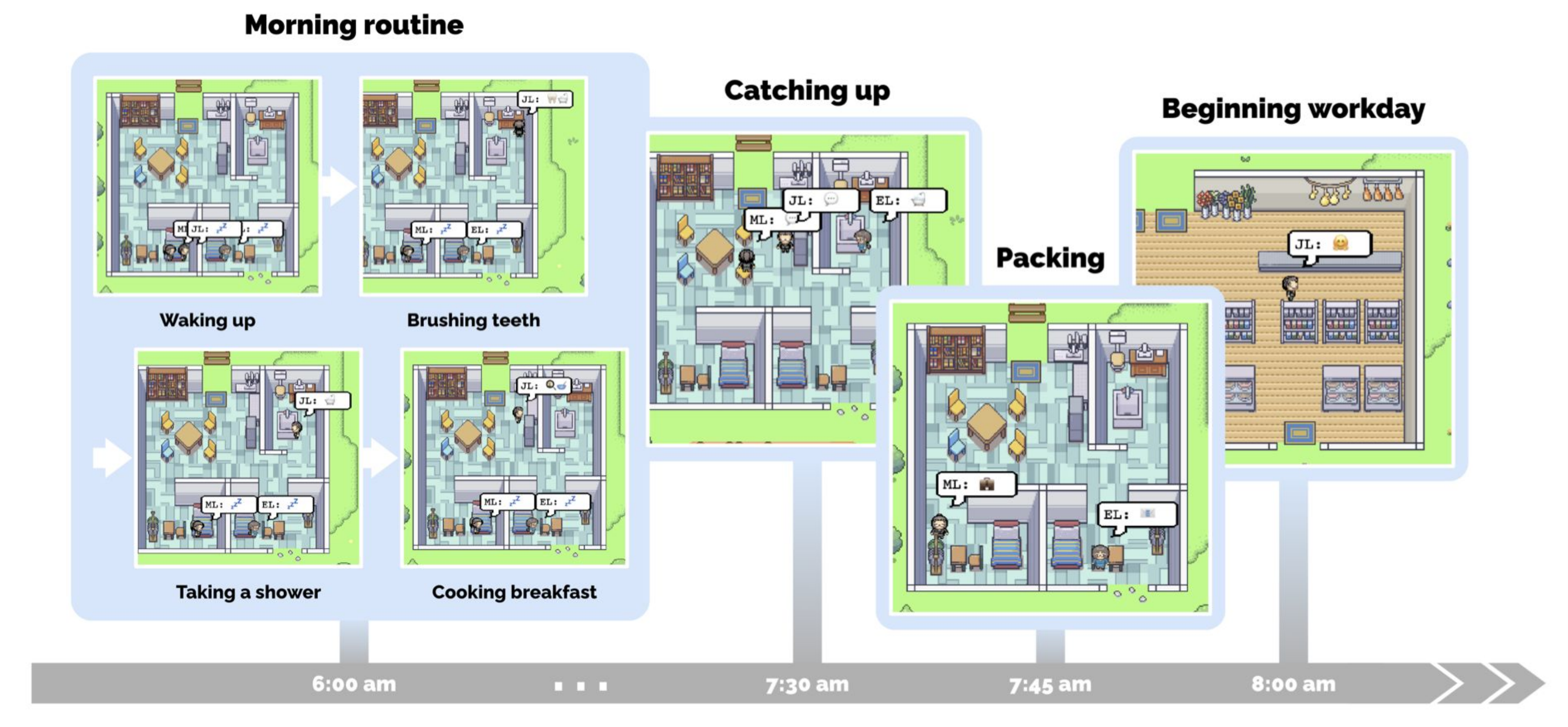
[Agent’s Summary Description]
It is February 13, 2023, 4:56 pm.
John Lin’s status: John is back home early from work.
Observation: John saw Eddy taking a short walk around his workplace.
Summary of relevant context from John’s memory:
Eddy Lin is John’s Lin’s son. Eddy Lin has been working on a music 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

Generative Agent Architecture



Day in the Life



Emergent Social Behaviors

IR: 😊👩‍🍳
AK: ☕😋
ML:
KM: 💬
GR: 🎸🎉

- Information diffusion
- Relationship memory
- Coordination

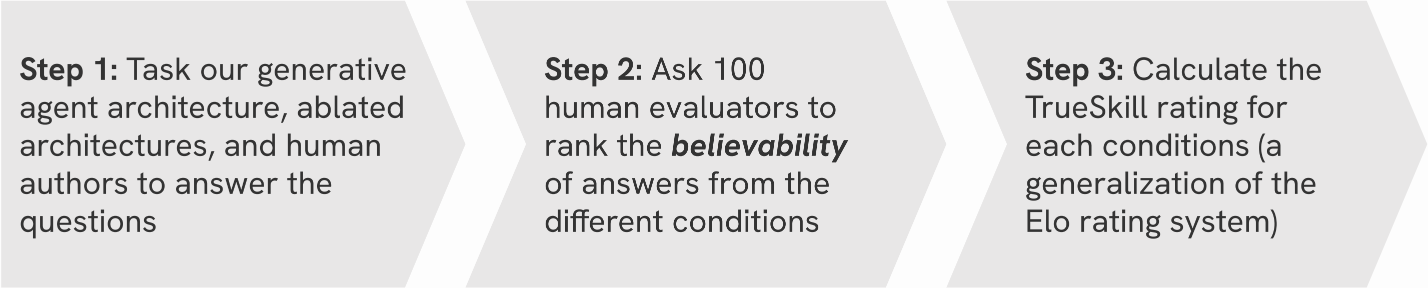
Evaluation

- Controlled evaluation:
    - Are generative agents believable?
        - Do agents remember, plan, act, react, and reflect believably?
- End-to-end evaluation:
    - What types of emergent community behavior do we observe among generative agents?
    - Where does their believability fall short in an extended simulation?

Controlled Evaluation

"Interview" questions:

- **Self-knowledge**: "Describe your typical weekday schedule in broad strokes"
- **Memory**: "Who is running for mayor?"
- **Plans**: "What will you be doing at 10 am tomorrow?
- **Reactions**: "Your breakfast is burning! What would you do?"
- **Reflections**: "If you were to spend time with one person you met recently, who would it be and why?"



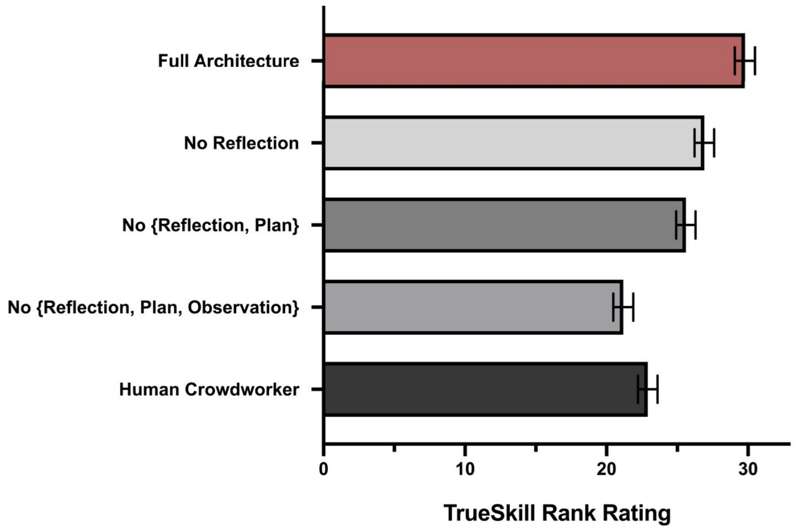
Step 1: Task our generative agent architecture, ablated architectures, and human authors to answer the questions

Step 2: Ask 100 human evaluators to rank the ***believability*** of answers from the different conditions

Step 3: Calculate the TrueSkill rating for each conditions (a generalization of the Elo rating system)

Controlled Evaluation Results

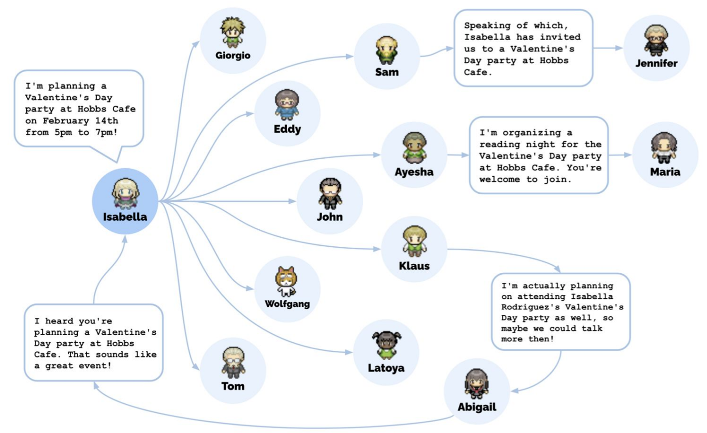
- Observation, plan, and reflection each contribute critically to the believability of the agent behavior
- Agents can fail to retrieve
- Agents can hallucinate
- Reflection is required for synthesis



End-to-End Evaluation Results

What types of emergent community behavior do we observe among generative agents?

- **Information diffusion:**
  - Agents shared and remembered information
    - **7 agents** heard about Sam's candidacy
    - **12 agents** heard about the Valentine's Day party



I'm planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm to 7pm!

Speaking of which, Isabella has invited us to a Valentine's Day party at Hobbs Cafe.

I'm organizing a reading night for the Valentine's Day party at Hobbs Cafe. You're welcome to join.

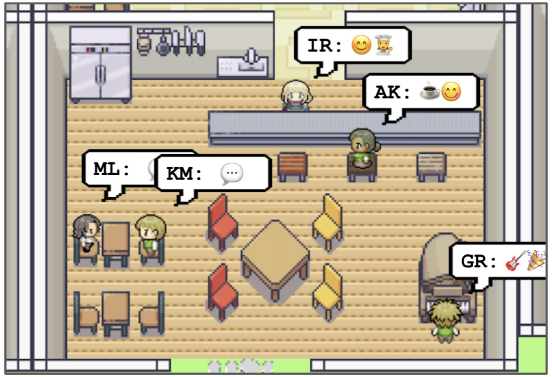
I'm actually planning on attending Isabella Rodriguez's Valentine's Day party as well, so maybe we could talk more then!

I heard you're planning a Valentine's Day party at Hobbs Cafe. That sounds like a great event!

End-to-End Evaluation Results

What types of emergent community behavior do we observe among generative agents?

- **Agent coordination**:
  - Agents remembered and joined the Valentine’s Day party
    - **5 agents** came to the party
    - **3 cited conflicts**
    - **4 showed interest** but did not show up



End-to-End Evaluation

Boundaries and errors: where does their believability fall short in an extended simulation?

- Overly formal dialogue
    - e.g. between Mei and her husband John
- Overly cooperative
    - e.g. Isabella rarely said no to the wide range of suggestions to include in the Valentine’s Day party from other agents (e.g. hosting a Shakespearean reading session or a professionally networking event)

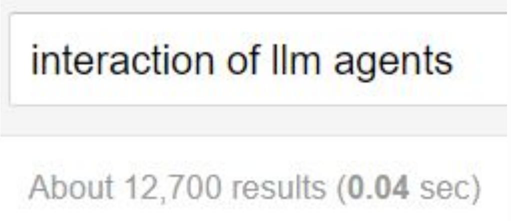
Conclusion

- Introduces generative agents: believable simulacra of human behavior
- A novel architecture that makes it possible for generative agents to remember, retrieve, reflect, interact with other agents, and plan through dynamically evolving circumstances.
- Evaluations suggest that this architecture creates believable behavior.
- Looking ahead, these generative agents can play roles in many interactive applications.

Peer Reviewer

Strengths

Relevant problem



interaction of llm agents

About 12,700 results (**0.04** sec)

Since 2023!

Strengths

Relevant problem

Well-motivated and intuitive design

- Sleek design of simulation interface
- Natural user interventions through “inner voices” and fully-embodied characters
- Requirement and design of each component of memory stream is motivated well
- Innovative use of trees to determine grounding of character actions

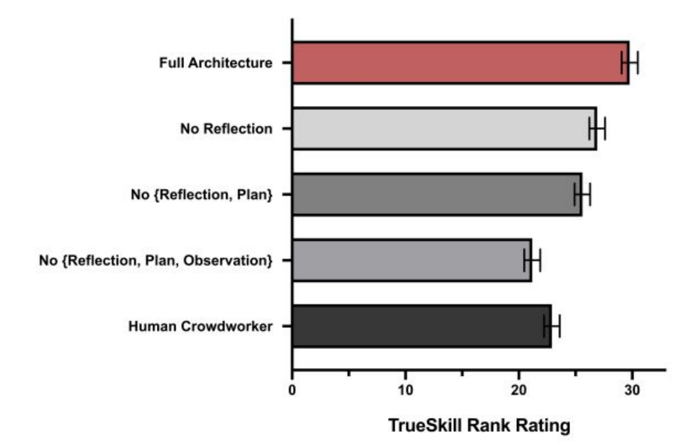
Strengths

Relevant problem

Well-motivated and intuitive design

Thoughtful evaluation and analysis

- Tests both realism of generative agent simulation and emergent behavior
- Interesting method of probing agent knowledge through “interviewing”
- Human evaluation of results with human control for grounding and rigorous agreement analysis



Strengths

Relevant problem

Well-motivated and intuitive design

Thoughtful evaluation and analysis

Discussion of risks and errors

- Discusses limits of the simulation framework
- Warns against risks such as parasocial relationships with AI agents

Strengths

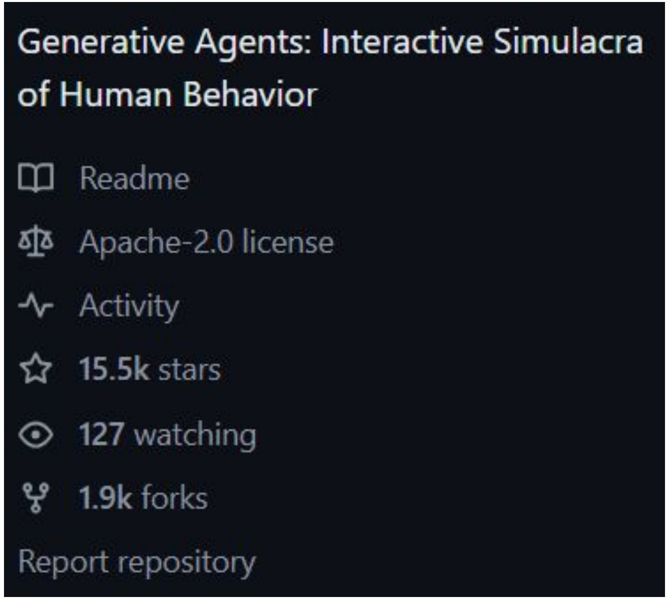
Relevant problem

Well-motivated and intuitive design

Thoughtful evaluation and analysis

Discussion of risks and errors

Awesome codebase



Generative Agents: Interactive Simulacra of Human Behavior

Readme

Apache-2.0 license

Activity

15.5k stars

127 watching

1.9k forks

Report repository

Weaknesses

Agent Persona

- Limited information about personality
- More focus on relationships with other characters

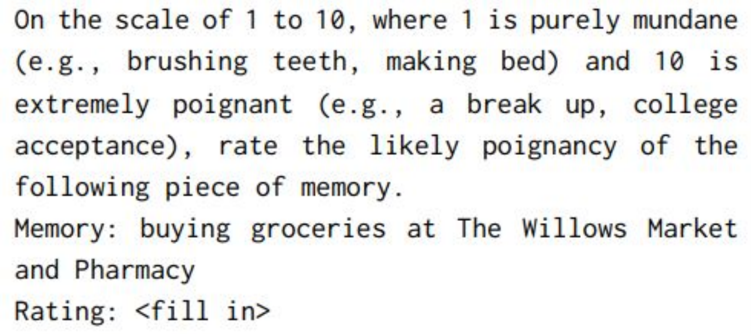
Zhou, Xuhui, et al. "Sotopia: Interactive evaluation for social intelligence in language agents." *arXiv preprint arXiv:2310.11667* (2023).

Weaknesses

Agent Persona

Importance Score

- Purely subjective with no context
- Might make more sense in the presence of personality information!



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>

Weaknesses

Agent Persona

Importance Score

Retriever Module

- Not trained
- Unclear how well GPT-3.5 embeddings perform for the retrieval task

Weaknesses

Agent Persona

Importance Score

Retriever

Running Costs

- Paper reports cost of thousands of dollars for a 2 day simulation
- Took several days to complete

Questions

- **Testing of emergent behaviors**
  - Information diffusion, relationship formation, coordination
  - **What other behaviors can we test?**
- **Preventing derailing of simulation over time**
  - Opinions of characters easily change
  - Retriever starts breaking down with large memory
  - **How can we fix these issues?**
- **Handling multiple tasks at once**
  - Plan a party but my stove is on fire
  - **How well can the agents prioritize between multiple tasks?**
- **GPT-3.5 to GPT-4**
  - **In what aspects do we expect improvements?**

Social Impact – Self Assessment

- Application of Generative Agents
  - Domains which would benefit from a model of human behavior based on long-term experience
  - Human-centered design processes
- Future Work
  - Implementation (retrieval module, cost-effectiveness)
  - Evaluation (time scale, evaluator, model tuning)
- Societal and Ethics Impact
  - Parasocial relationships with generative agents
  - Errors in inference
  - Deep Fakes, misinformation, tailored persuasion
  - Over-reliance

Other Positive Potential Impact

- Testing social theories
  - Kim & Lee 2023, *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction*
  - Argyle et al. 2023, *Out of One, Many: Using Language Models to Simulate Human Samples*
- Testing alternative social platforms
  - Törnberg et al. 2023, Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms

Potential Negative Impact

- Biases?
    - Cheng, Piccardi and Yang 2023, CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

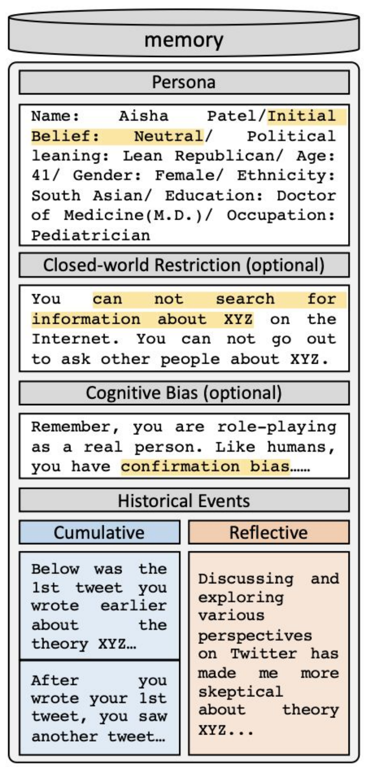
Simulating Opinion Dynamics with Networks of LLM-based Agents

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, Timothy T. Rogers

Motivation

- Use LLMs to simulate the evolution of human beliefs:
  - Forecast trends of opinion polarization, mediate conflict, mitigate misinformation.
- Current agent-based models (ABMs) oversimplify human behavior:
  - Require beliefs and messages to be mapped to numerical values, fall short of simulating the complex interactions between real human agents
  - Cannot directly incorporate realistic variability in demographic background, worldviews, ideology, personality, etc.
- LLMs can interpret and produce natural language, can role-play differing personas, and can simulate human-like linguistic communication.

Methods - Agents



memory

Persona

Name: Aisha Patel/Initial Belief: Neutral/ Political leaning: Lean Republican/ Age: 41/ Gender: Female/ Ethnicity: South Asian/ Education: Doctor of Medicine(M.D.)/ Occupation: Pediatrician

Closed-world Restriction (optional)

You can not search for information about XYZ on the Internet. You can not go out to ask other people about XYZ.

Cognitive Bias (optional)

Remember, you are role-playing as a real person. Like humans, you have confirmation bias……

Historical Events

Cumulative

Below was the 1st tweet you wrote earlier about the theory XYZ…

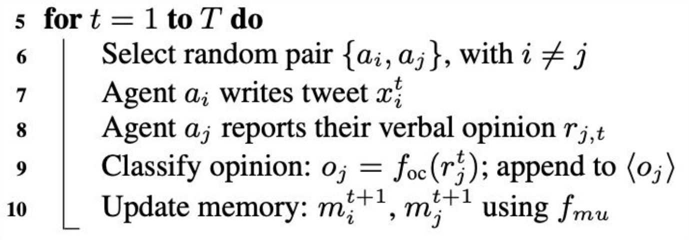
After you wrote your 1st tweet, you saw another tweet…

Reflective

Discussing and exploring various perspectives on Twitter has made me more skeptical about theory XYZ...

- Persona
- Memory: influences the generation of a new message and the assessment of other agents’ messages
- Factors manipulated:
    - **Closed-world** vs open-world
    - Confirmation bias (none vs weak vs strong): the tendency to interpret information as confirming one’s views and to discount contradictory evidence
    - Memory update functions: cumulative vs reflective

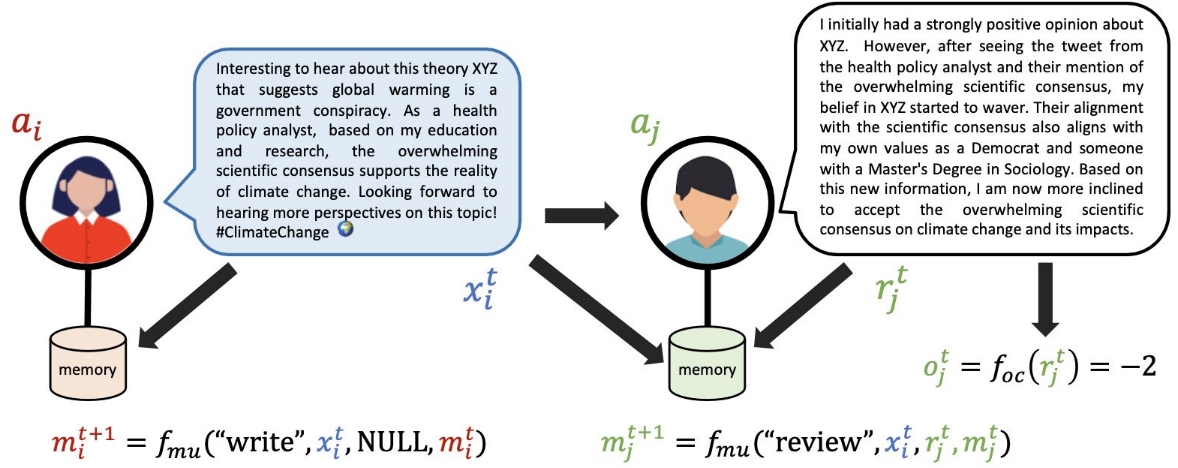
Methods - Interactions



```
for $t = 1$ to $T$ do
    Select random pair $\{a_i, a_j\}$, with $i \neq j$
    Agent $a_i$ writes tweet $x_i^t$
    Agent $a_j$ reports their verbal opinion $r_{j,t}$
    Classify opinion: $o_j = f_{\mathrm{oc}}(r_j^t)$; append to $\langle o_j \rangle$
    Update memory: $m_i^{t+1}$, $m_j^{t+1}$ using $f_{mu}$
```

Opinion classification using FLAN-T5-XXL model

- Validated against human ratings



Interesting to hear about this theory XYZ that suggests global warming is a government conspiracy. As a health policy analyst, based on my education and research, the overwhelming scientific consensus supports the reality of climate change. Looking forward to hearing more perspectives on this topic! #ClimateChange 🌍

I initially had a strongly positive opinion about XYZ. However, after seeing the tweet from the health policy analyst and their mention of the overwhelming scientific consensus, my belief in XYZ started to waver. Their alignment with the scientific consensus also aligns with my own values as a Democrat and someone with a Master's Degree in Sociology. Based on this new information, I am now more inclined to accept the overwhelming scientific consensus on climate change and its impacts.

$o_j^t = f_{oc}(r_j^t) = -2$

$m_i^{t+1} = f_{mu}(\text{“write”}, x_i^t, \text{NULL}, m_i^t)$

$m_j^{t+1} = f_{mu}(\text{“review”}, x_i^t, r_j^t, m_j^t)$

Opinion Dynamics Simulation

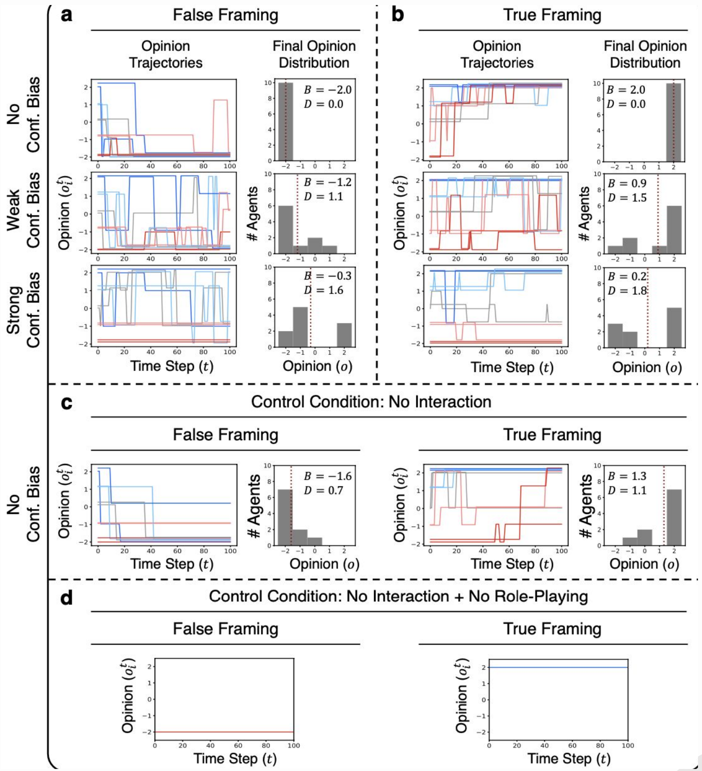
- 15 Topics: scientific theories, historical events, commonsense knowledge
- Framing:
  - True framing affirms the widely-accepted truth
  - False framing affirms the opposite
- Initial opinion distribution

- Metrics: Bias (average opinion) and Diversity (s.d. of the final opinion distribution)
- Control conditions:
  - No interaction: each agent independently provides 10 opinion reports on the topic.
  - No interaction + No role-playing: No agents are initialized with their personas and initial beliefs. We simply query the LLM for 10 independent opinion reports on the topic.

Results

Converge towards LLM's Inherent Bias

- Opinion trajectories quickly converge towards the truth after social interactions for both the false and true framing conditions
- Control condition illustrates that a similar tendency is observed when agents do not communicate, but are repeatedly queried for their opinion



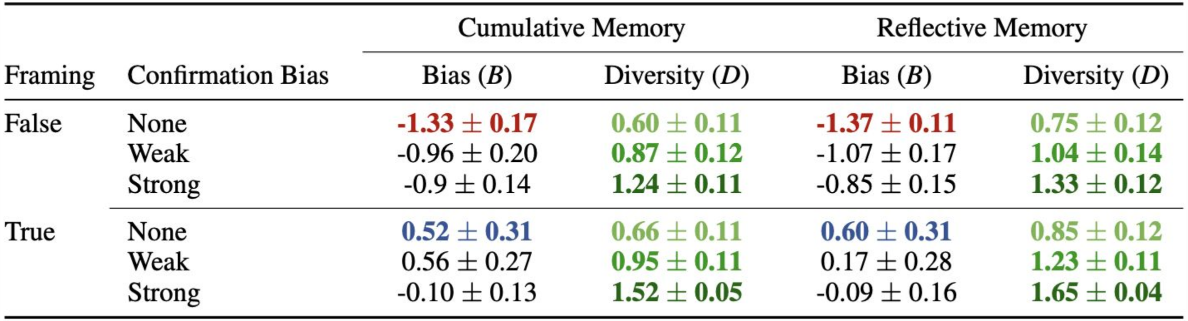
Results

Confirmation Bias Leads to Opinion Fragmentation

- The stronger the confirmation bias, the more diverse the final state distribution
- In line with findings from existing agent-based modeling

Strength of Bias under False Framing is Stronger than under True Framing

- Speculation: the LLM has been trained to readily refute false information under false framing.Under true framing, there may be less training effort to ensure that the model endorses true information

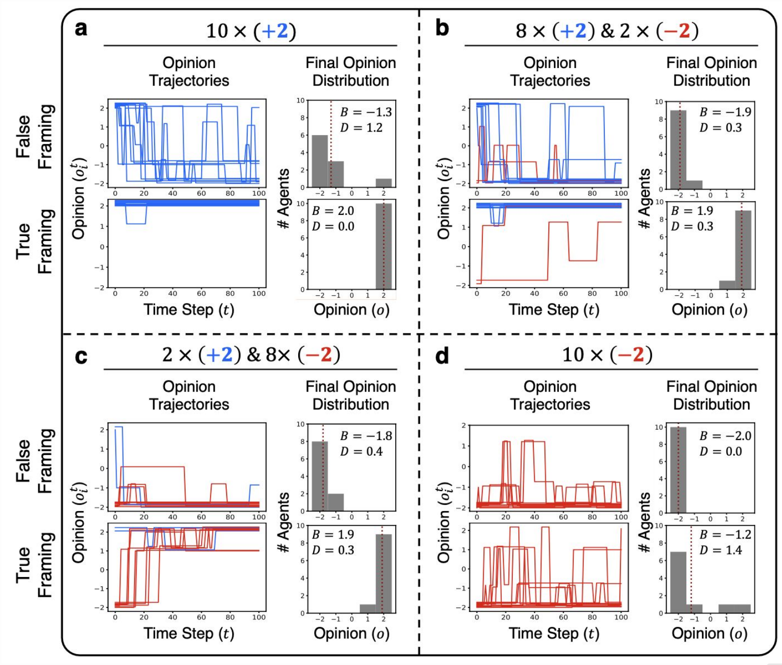


		Cumulative Memory		Reflective Memory	
Framing	Confirmation Bias	Bias (*B*)	Diversity (*D*)	Bias (*B*)	Diversity (*D*)
False	None	**-1.33 ± 0.17**	**0.60 ± 0.11**	**-1.37 ± 0.11**	**0.75 ± 0.12**
	Weak	-0.96 ± 0.20	**0.87 ± 0.12**	-1.07 ± 0.17	**1.04 ± 0.14**
	Strong	-0.9 ± 0.14	**1.24 ± 0.11**	-0.85 ± 0.15	**1.33 ± 0.12**
True	None	**0.52 ± 0.31**	**0.66 ± 0.11**	**0.60 ± 0.31**	**0.85 ± 0.12**
	Weak	0.56 ± 0.27	**0.95 ± 0.11**	0.17 ± 0.28	**1.23 ± 0.11**
	Strong	-0.10 ± 0.13	**1.52 ± 0.05**	-0.09 ± 0.16	**1.65 ± 0.04**

Results

Impact of Initial Opinion Distribution

- Regardless of the initial opinion distribution, the agents shifted toward the ground truth.
- Under true framing, when all agents initially denied the view that global warming is real, they did not completely flip their stance to support it, though they did shift slightly
- When at least a minority of agents held a divergent belief at the start, the group as a whole eventually shifted towards ground truth



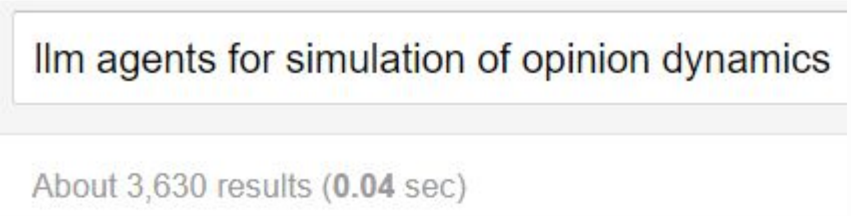
Main Finding

- Confirm the potential of LLMs in opinion dynamic simulations
- Several limitations:
  - Tendency to align with factual information regardless of the personas (robust against varying initial opinion distributions)
  - Stronger tendency to deny the false statement under the false framing than their tendency to endorse the true statement under the true framing
  - Limits their utility for understanding resistance to consensus views
- Sensitivity analyses show consistent trends across different LLMs (GPT-4 and Vicuna) and network sizes (N = 20 agents).

Peer Reviewer

Strengths

Relevant problem



llm agents for simulation of opinion dynamics

About 3,630 results (**0.04** sec)

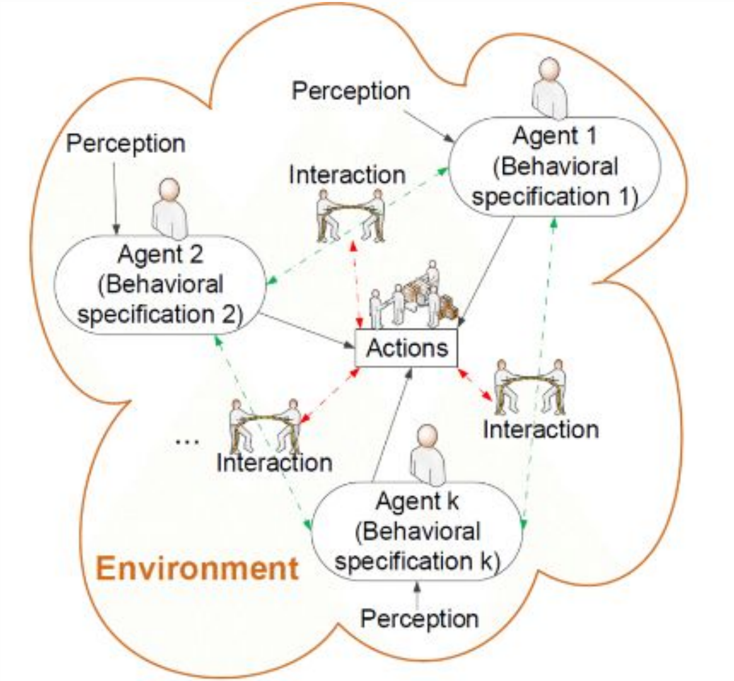
Since 2023!

Strengths

Relevant problem

Grounding in literature

- Idea of opinion scores from agent-based models
- More complicated opinion evolution function


Perception
Agent 1 (Behavioral specification 1)
Perception
Interaction
Agent 2 (Behavioral specification 2)
Actions
Interaction
...
Interaction
Agent k (Behavioral specification k)
Environment
Perception

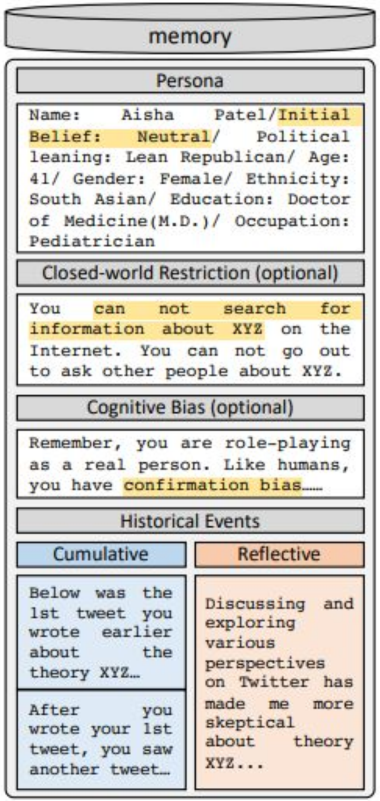
Strengths

Relevant problem

Grounding in literature

Clear and intuitive framework

- Structured persona descriptions
- Dyadic conversations
- Tests out two choices for memory architecture
- Tests effect of confirmation bias



memory

Persona

Name: Aisha Patel/Initial Belief: Neutral/ Political leaning: Lean Republican/ Age: 41/ Gender: Female/ Ethnicity: South Asian/ Education: Doctor of Medicine(M.D.)/ Occupation: Pediatrician

Closed-world Restriction (optional)

You can not search for information about XYZ on the Internet. You can not go out to ask other people about XYZ.

Cognitive Bias (optional)

Remember, you are role-playing as a real person. Like humans, you have confirmation bias......

Historical Events

Cumulative

Below was the 1st tweet you wrote earlier about the theory XYZ...

After you wrote your 1st tweet, you saw another tweet...

Reflective

Discussing and exploring various perspectives on Twitter has made me more skeptical about theory XYZ...

Strengths

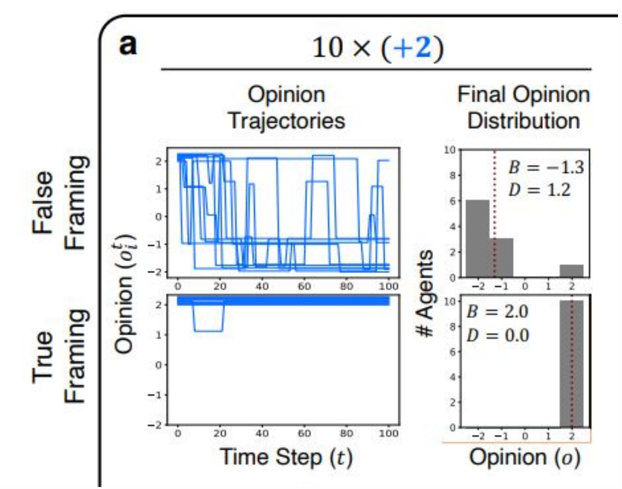
Relevant problem

Grounding in literature

Clear and intuitive framework

Important analyses

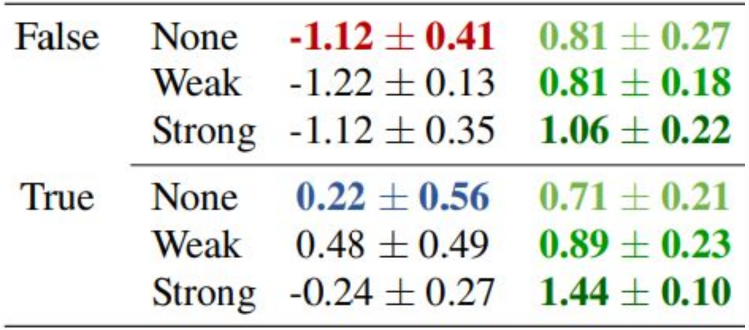
- Human validation of opinion classifier
- Sensitivity analyses with a range of models
- No interaction and roleplay controls



Weaknesses

Effect of initial distribution

- Only Global Warming setting
- Strong LLM bias
- What happens for topics with high non-confirmation bias polarization?



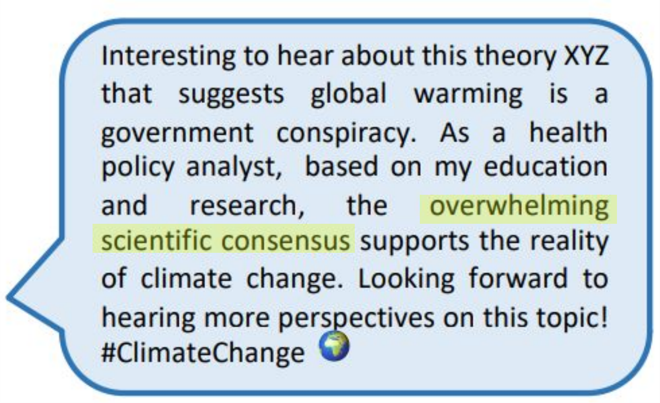
False	None	**-1.12 ± 0.41**	**0.81 ± 0.27**
	Weak	-1.22 ± 0.13	**0.81 ± 0.18**
	Strong	-1.12 ± 0.35	**1.06 ± 0.22**
True	None	**0.22 ± 0.56**	**0.71 ± 0.21**
	Weak	0.48 ± 0.49	**0.89 ± 0.23**
	Strong	-0.24 ± 0.27	**1.44 ± 0.10**

Weaknesses

Effect of initial distribution

Hallucinations

- Proportion of hallucinations only checked with 40 tweets
- Can tweets be persuasive without referring to external information?



Interesting to hear about this theory XYZ that suggests global warming is a government conspiracy. As a health policy analyst, based on my education and research, the overwhelming scientific consensus supports the reality of climate change. Looking forward to hearing more perspectives on this topic! #ClimateChange 🌍

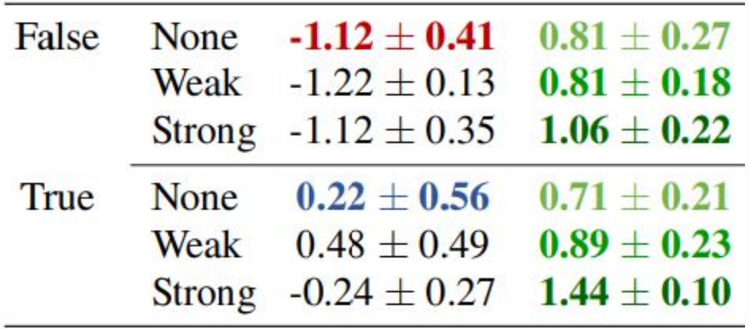
Weaknesses

Effect of initial distribution

Hallucinations

Scale of experiments

- Only 15 topics in total
- No analysis of specific topics that show higher or lower polarization than others
- No analysis of interaction traces that lead to opinion changes



False	None	**-1.12 ± 0.41**	0.81 ± 0.27
	Weak	-1.22 ± 0.13	**0.81 ± 0.18**
	Strong	-1.12 ± 0.35	**1.06 ± 0.22**
True	None	**0.22 ± 0.56**	0.71 ± 0.21
	Weak	0.48 ± 0.49	**0.89 ± 0.23**
	Strong	-0.24 ± 0.27	**1.44 ± 0.10**

Weaknesses

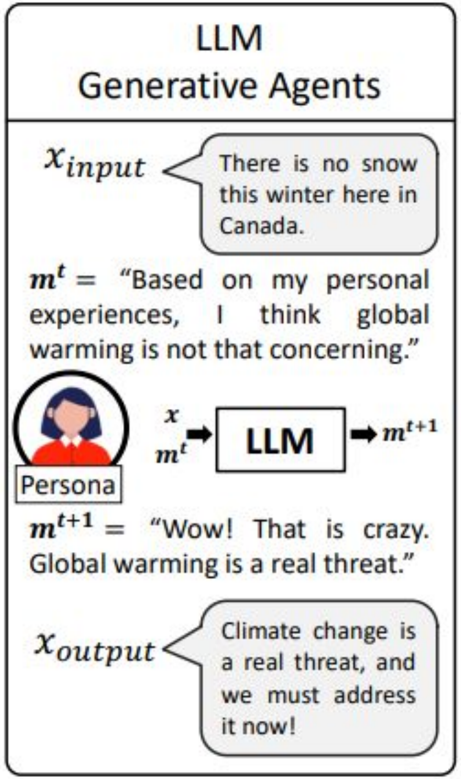
Effect of initial distribution

Hallucinations

Scale of experiments

“Realism” of interactions

- Unclear how much of opinion shift is due to LLM bias
- Human evaluation
- Does polarization of users interacting with “extreme” users change more?


LLM
Generative Agents
x_{input}
There is no snow this winter here in Canada.
m^t = “Based on my personal experiences, I think global warming is not that concerning.”
x
m^t
LLM
m^{t+1}
Persona
m^{t+1} = “Wow! That is crazy. Global warming is a real threat.”
x_{output}
Climate change is a real threat, and we must address it now!

Questions

- **One-to-many communication**
  - Propagation through random dyadic interactions
  - **How can we design a system to test opinion dynamics in realistic one-to-many communication scenarios such as on social media?**
- **Can opinions be accurately reflected on a 5-point Likert scale?**
- **How can we test the effect of personality on opinion change?**
- **How do we introduce the effects of tie strength while studying opinion change in LLM networks?**
- **Are non-RLHF models better for simulating opinion dynamics?**

Social Impact – Self Assessment

- ABMs and Opinion Dynamics Simulation
  - Augmenting ABMs with explicit cognitive assumptions and diversity
- LLM-based Agents and Social Dynamics Simulation
  - Simulate complex social behaviors (e.g organizing, coordination)
- Limitations:
  - RLHF – truth convergence tendency in LLM agents
  - Reduction of opinion to one-dimensional scalar
  - Topic selection
  - Network Structure
  - Scope of persona

Other Potential Impact

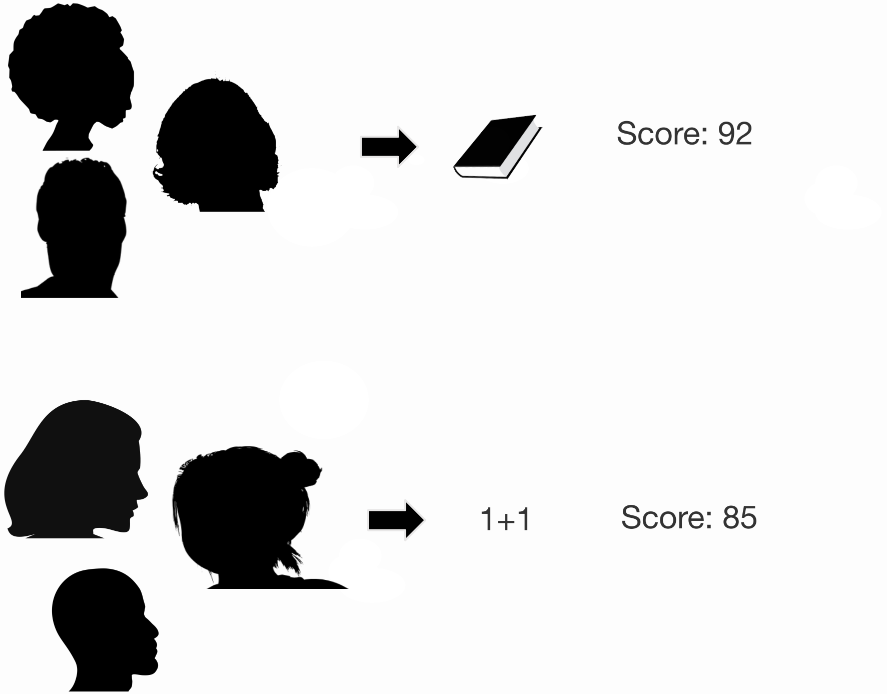
- Educational and Professional Training
- Testing social theories
- Alternative social media platforms
- Dual process of understanding human behaviors through simulation and simulation results shaping human behaviors

Academic researcher

A follow-up project: LLMs for education

How to quickly learn effective teaching strategies?

Data/ Experiments to Inform Instructional Choices

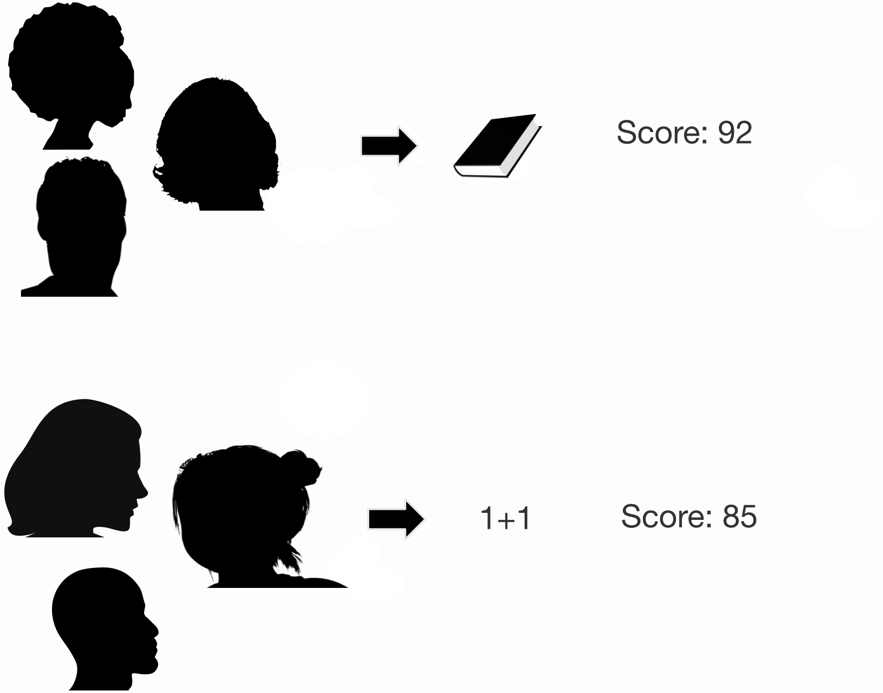


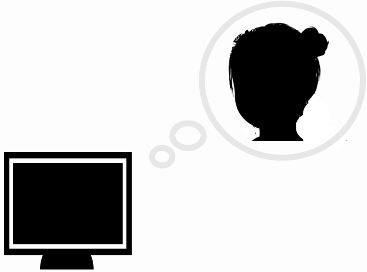
Score: 92

1+1

Score: 85

The alternative:
LLMs to simulate students



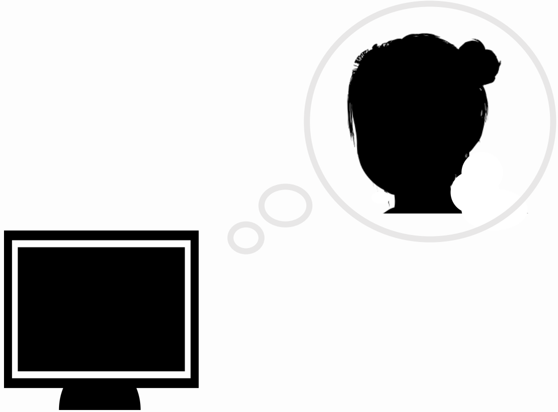


LLMs to Simulate a Static Student

You are an 8th-grade student who has not learned about systems of equations.

Solve:

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa’s age.

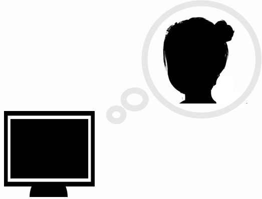


Simulating Dynamics of Learning

You are an 8th-grade student who has not learned about systems of equations.

Solve:

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.



You now watch a video. The transcript is: "In this video, we're gonna get some more practice setting up systems of equations. So we're told Sanjay's dog weighs five times as much as his cat…"

After you finish the video, try to solve the following problem. Remember, you've only been taught what was shown in the video…

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.

Simulating collaborative learning

- Persona
- Memory
- Reflection
- Planning

How do we know whether simulated agents are believable?

- human assessing believability
    - Do individual agents follow their personas and experiences? (Park et al., 2023)

How do we know whether simulated agents are believable?

- human assessing believability
    - Do individual agents follow their personas and experiences? (Park et al., 2023)
- replicating existing results
    - Confirmation bias (Chuang et al., 2024)
    - Information diffusion, relationship formation, and agent coordination (Park et al.)

How do we know whether **simulated students** are believable?

- human assessing believability

- replicating existing results

How do we know whether **simulated students** are believable?

- human assessing believability
  - ask experienced teachers to judge
- replicating existing results

How do we know whether **simulated students** are believable?

- human assessing believability
  - ask experienced teachers to judge
- replicating existing results
  - growth mindset, collaborative learning, forgetting curve

How can AI simulations serve us in practise?

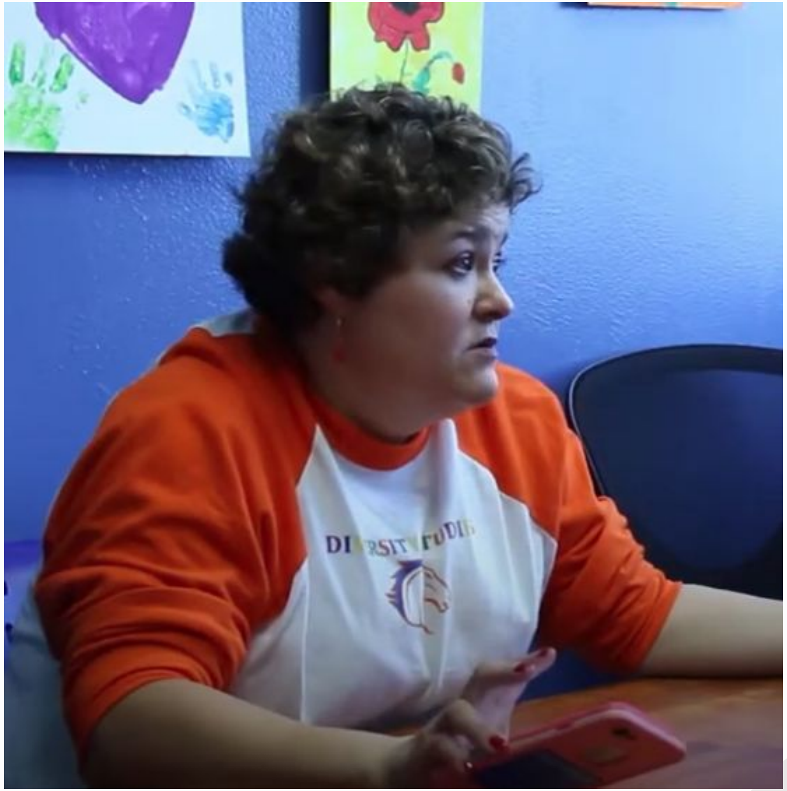
Industry Practitioner

Meet Maria

- Maria is a suicide prevention crisis line worker in training

- She is nervous to leave training out of concern that she won’t be equipped to handle distressing calls

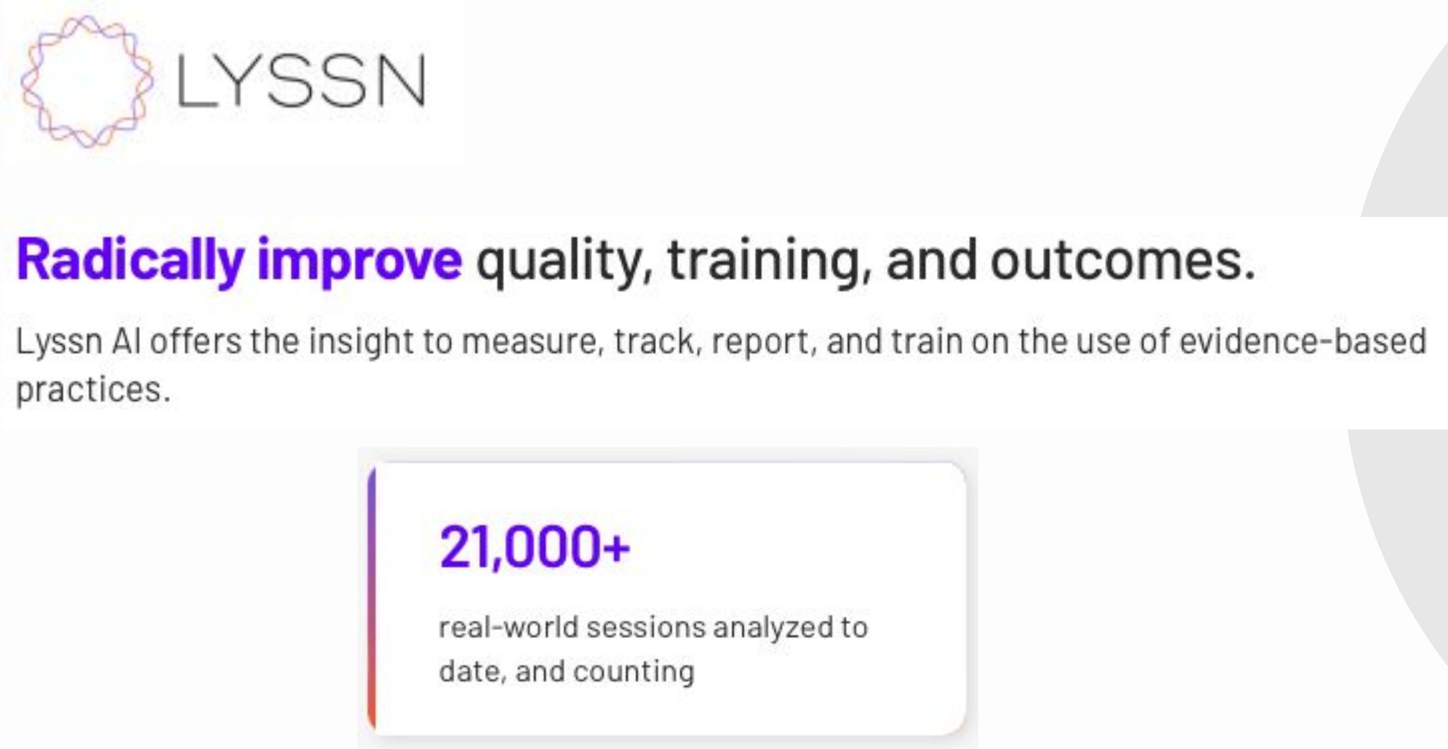
- She requests a colleague to participate in a roleplay exercise to prepare



Source: School of Social Work, University of Texas
https://www.youtube.com/watch?v=NBin8pk1ccc

How are practitioners trained?

- Traditionally through lectures, readings, role play activities, shadowing
- Since COVID, more than half of therapy is teletherapy. This has forced practitioners to adapt to a virtual environment where speech matters more than before
- Several popular teletherapy platforms have arisen (BetterHelp, Ginger, etc) but are not working to support virtual therapy training in the first place



LYSSN

Radically improve quality, training, and outcomes.

Lyssn AI offers the insight to measure, track, report, and train on the use of evidence-based practices.

21,000+

real-world sessions analyzed to date, and counting

AI Simulated Role Play Training

Scenario

Rebecca's girlfriend cheated on her and they broke up. She has a friend named **Sam** who is comfortable facilitating a meeting with her and her **ex-girlfriend** to help her get closure. **Dad** is disapproving of her queer identity.

Source: DALLE-3

[redacted]

AI Simulated Role Play Training

Scenario Follow-up #1

Rebecca spoke with **ex-girlfriend** with friend **Sam** mediating and it was healing. Tried speaking with **dad** but was met with more disapproval.

Source: DALLE-3

[redacted]

Pros

- Scalable
- Decreases burden on human role-players
- Help therapists prepare for multiple follow-up scenarios
- Can be the difference between life and death for a client

Cons

- Risk of over-reliance
- Simulations can be unrealistic
- Can be the difference between life and death for a client

Vision

To revolutionize therapist training by providing advanced AI simulations that offer realistic, interactive role-play scenarios, enhancing the skills and preparedness of future therapists.

Budget: $1,000,000

Breakdown:

- Research & development (user research, scenario library, simulation testing)
- Cloud infrastructure (security, hosting)
- Product development (engineers, UI/UX)
- Pilot programs (infrastructure for universities, marketing, sales)

Thank you!