

SIMULATION OF HUMAN BEHAVIOR THROUGH LLM-BASED AUTONOMOUS AGENTS

Raymond Fu and Gabriel Xu

May 21, 2025

Systems Lab, Dr. Yilmaz, Period 1

GOAL

- The goal is to simulate human behavior through computer programming
- This comes in the form of a simulated town where each person will act in a human way



THE GOAL



Park et al.

BACKGROUND – AGENTS

- Agents are the people in the simulated world
- An agent usually takes up a role, such as a teacher, cashier, etc.
- Agents are meant to interact and socialize with each other, forming a realistic society



BACKGROUND – LARGE LANGUAGE MODELS (LLM)

- Large language models are Natural Language Processing AI models that generate text given a prompt
- Input: Text written by a user. The input is also called the prompt
- Output: Text generated by the model that answers the prompt
- In this project, each agent has an LLM attached to it that will make decisions for it

BACKGROUND – LARGE LANGUAGE MODELS (LLM)

Input/Prompt

if $2x + 3 = 7$, what is x ?



To solve the equation $2x + 3 = 7$, follow these steps:

1. Subtract 3 from both sides:

$$2x = 7 - 3$$

$$2x = 4$$

2. Divide both sides by 2:

$$x = \frac{4}{2}$$

$$x = 2$$

AI generated
output

So, $x = 2$.



PROBLEM – MEMORY

- It is difficult to simulate human memory
- People will forget things after it has been too long
- Agents need to be able to forget unimportant things and be able to retrieve memories relevant to their current situation
- For example, an agent about to ask a teacher a question needs to retrieve memories about his question

PROBLEM – COST

- Calling an LLM through an API costs money
- Existing solutions call the API multiple times every second
- For multiple agents making queries with their memories, there would at least tens of thousands of tokens per second
- Even the cheapest API options at around \$0.15 per 1 million tokens would cost around a dollar per minute
- Scaling this to larger games or more users result in thousands of dollars of costs

WHAT MAKES A SUCCESSFUL PROJECT?

- **Memory:** The agent should be able to recall events from the past and forget less important information
- **Society:** Multiple agents should create a functional society.
- **Role-playing ability:** The agents should adhere to their roles and execute on them properly without hallucinations

OTHER SOLUTIONS – GOOGLE RESEARCH

Pros/Cons

- Pros

- Generative Agents: Interactive Simulacra of Human Behavior
- Successful simulation of human behavior

- Cons

- Despite its success, it cost thousands of dollars, which limited the number of simulation days that it was allowed to run
- They were not able to fully test the extent at which the agents stay sane
- Agents never forget anything; not accurate human memory

OTHER SOLUTIONS – GOOGLE RESEARCH

Planning

- At the beginning of each day, every agents generates a plan for the day
- As the day goes on, their behaviors evolve as they interact with other agents and the world



OTHER SOLUTIONS – GOOGLE RESEARCH

Memory

- Memory is main source of cost
- They had a memory stream which took a note of everything that happens every tick (smallest unit of time)
- The memory stream is then called upon and searched through by an LLM whenever memories need to be retrieved

OTHER SOLUTIONS – GOOGLE RESEARCH

Memory

Memory Stream

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

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

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

retrieval		recency		importance		relevance
2.34	=	0.91	+	0.63	+	0.80

ordering decorations for the party

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

researching ideas for the party

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

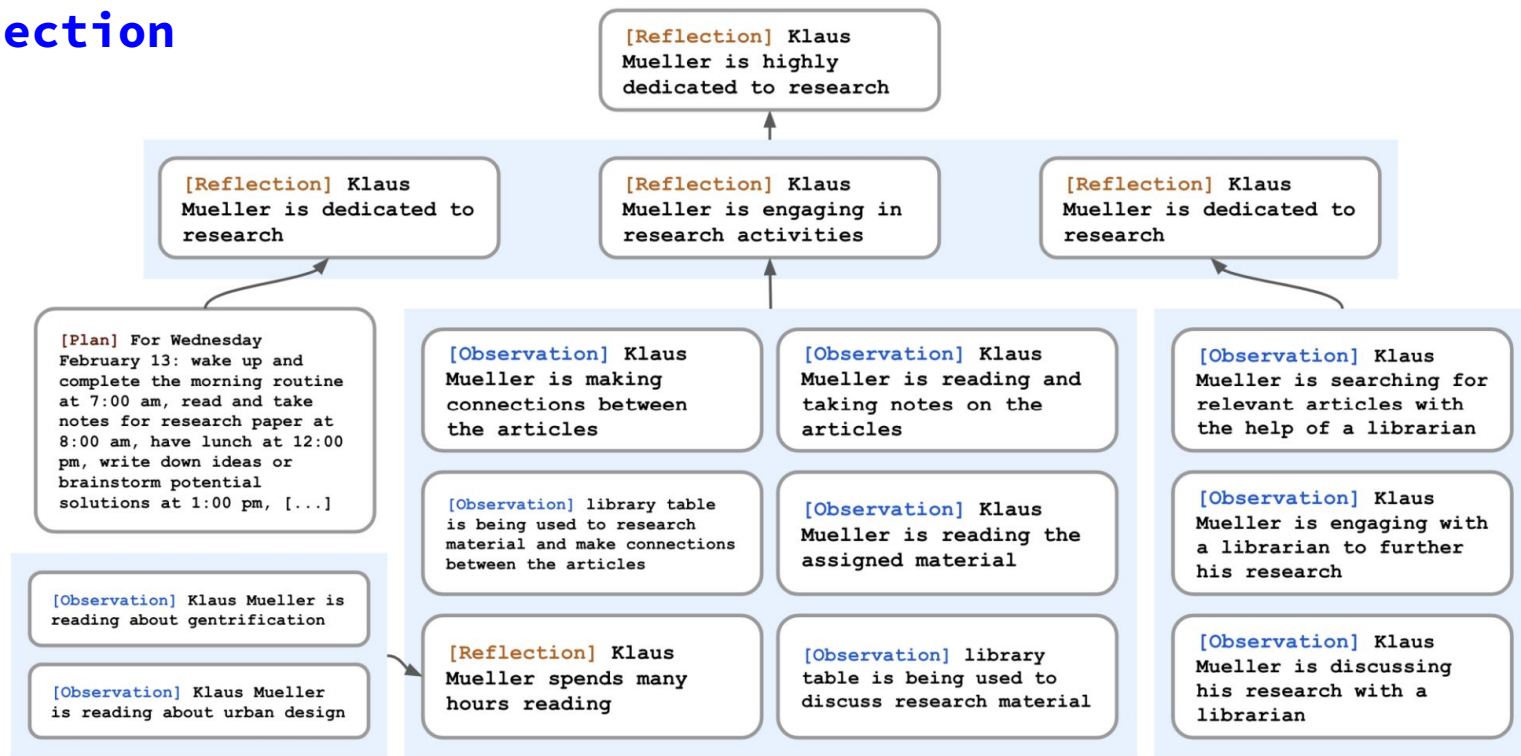
...

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



OTHER SOLUTIONS – GOOGLE RESEARCH

Reflection



OTHER SOLUTIONS – GOOGLE FEEDBACK DEMO



OTHER SOLUTIONS – AGENTSIMS

Overview

- [AgentSims: An Open-Source Sandbox for Large Language Model Evaluation](#)
- Cuts cost relative to Park et al. through developing a vector-based system to retrieve memories
- AgentSims is very extensible by a user
- AgentSims introduces a unique way for agents to interact with objects in their surroundings
- This system condenses memories after each day, so an LLM would still be called for memories within a day

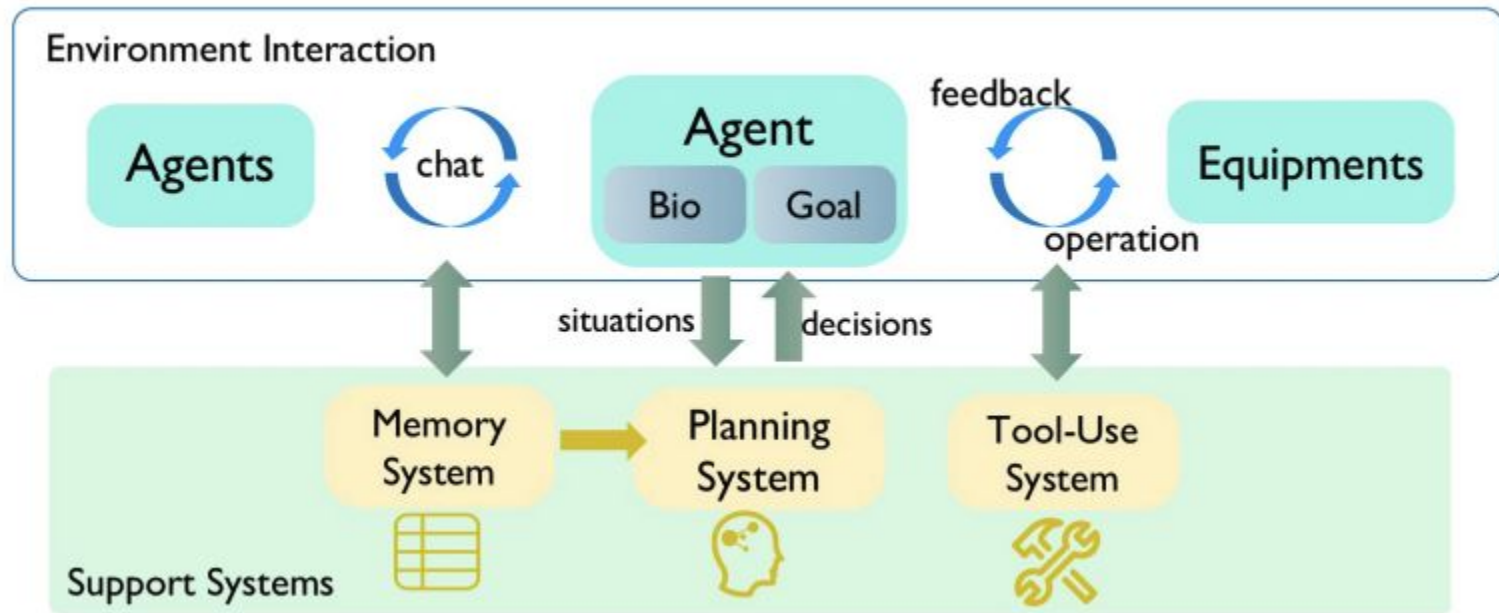
OTHER SOLUTIONS – AGENTSIMS

Feedback System

- For every action, two things need to pass
 1. A hard-coded possibility check
 2. A possibility check generated by the LLM
- The LLM would then learn from these two checks
- Example:
 - An agent wants to get tea from the stove. The first, hard-coded check says “Meaningless operation.” The second, LLM generated check says “You cannot get tea from a stove.” The model then learns from this

OTHER SOLUTIONS – AGENTSIMS

Systems Architecture



Lin et al.

WHY IS OURS BETTER?

- Ours will be able to significantly cut down on costs by not relying on LLMs as much as previous solutions, especially when it comes to memory
- We will be able to more fully and thoroughly test the project due to new ways to evaluate LLMs

IMPACT

- We introduce a new way to store memories in a realistic way to avoid excessive costs.
- By cutting down on the cost of running such a simulation, we allow smaller game developers, scientists, etc. to run such simulations
- Although games initially came to mind with this project, it could also be used in fields of study such as psychology and behavioral science

MATERIALS

- For the LLM, we are using Llama 3.1 with the Ollama Python package
- We are running the project entirely locally on a computer
- We are using pygame, a Python game development library, to develop the environment/visuals



METHODS – PLANNING

- Much like Park et al., we generate a list of tasks at the beginning of each day
- Each task will have a timestamp assigned to it
- When the time comes, more specific steps are outlined
- This list of tasks change dynamically as unexpected events happen

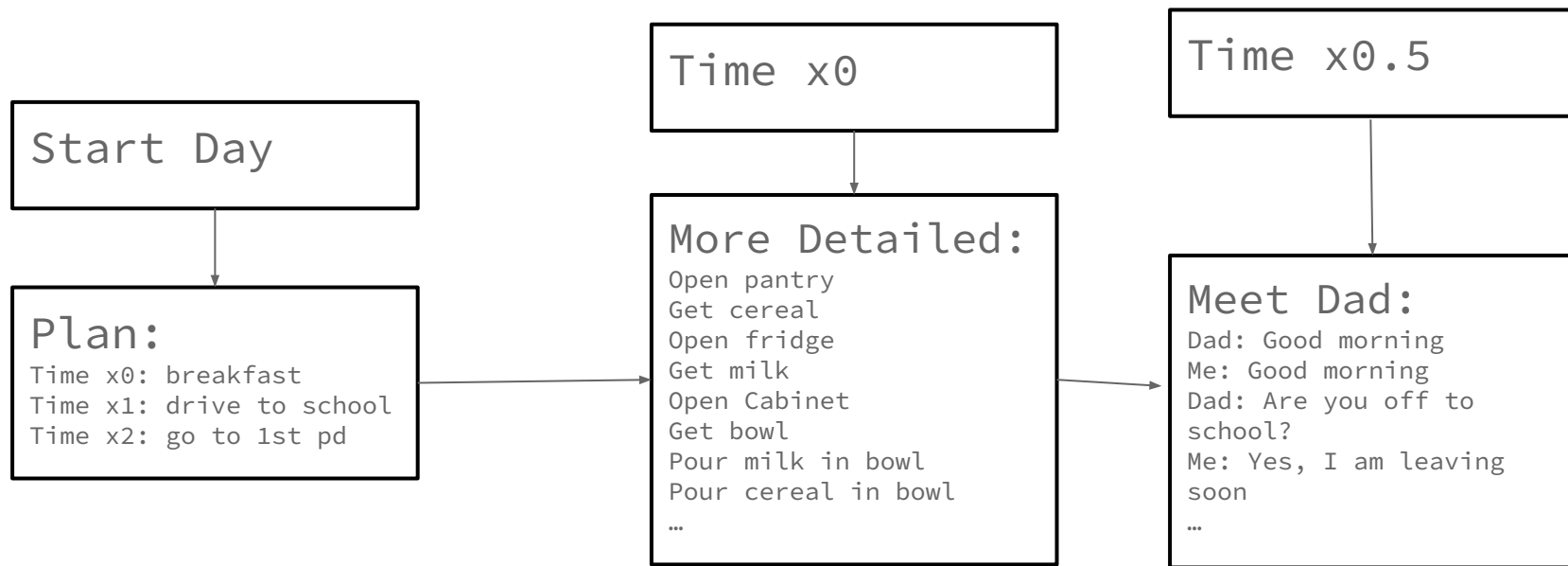
METHODS – MEMORY STORAGE

- Like Lin et al., we use a vector-based system to store
- For a given set of memories accomplishing one task, this task is stored, rather than the set of memories
- Each memory is tagged with a category (e.g. work, school, food, etc.)
- Each memory has an importance score

METHODS – MEMORY RETRIEVAL

- When a memory retrieval is necessary, past memories' relevance to the current situation and overall importance are weighed
- Relevancy is obtained via the categories
- An LLM then goes through the selected memories and obtains relevant information

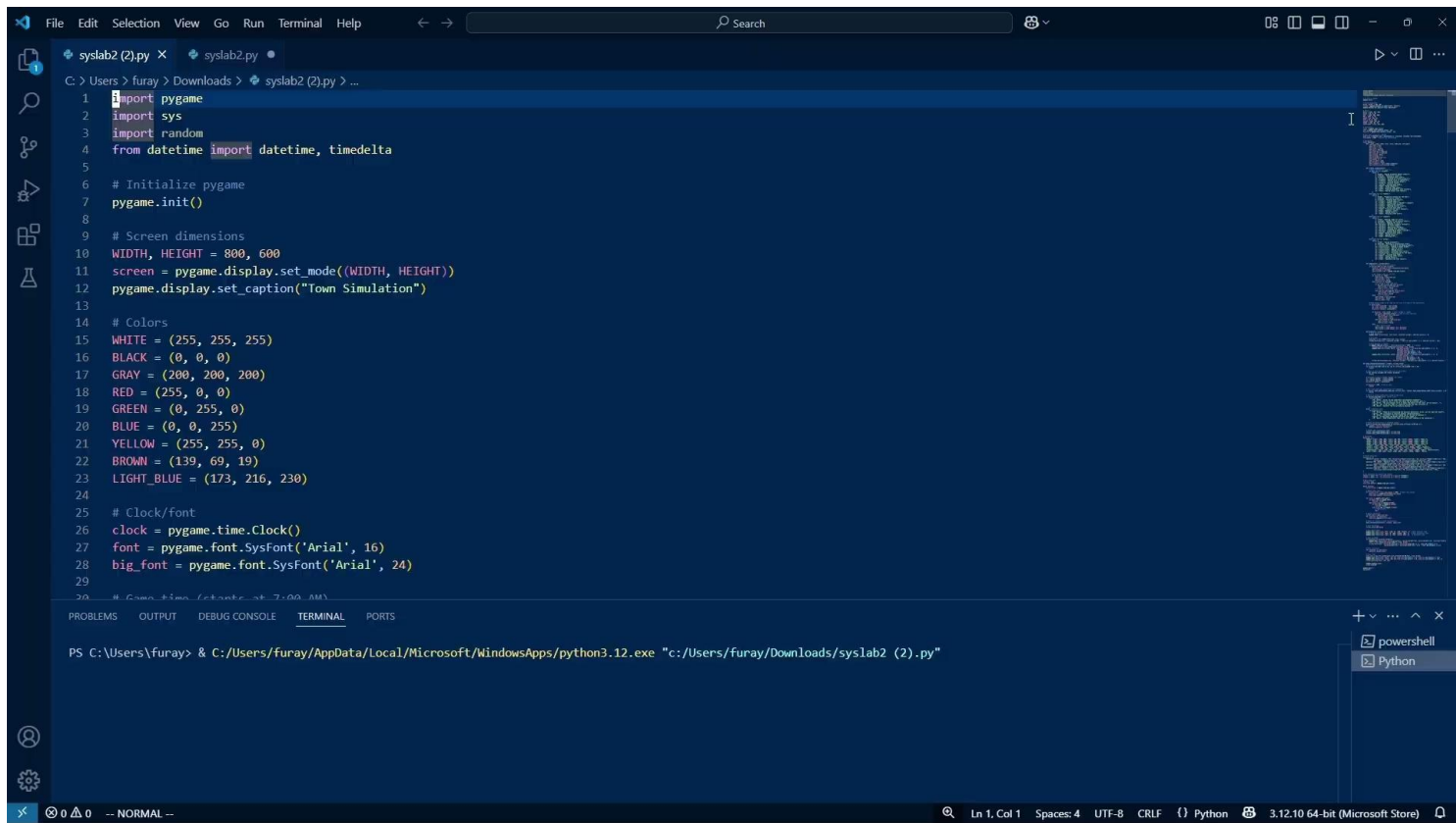
METHODS – DIAGRAM



RESULTS

- Turing Test: Have human evaluators separate between text of agents interacting with each other and two humans interacting with each other
- Task Success: How well the agents executed on the tasks they planned out

DEMO



```
File Edit Selection View Go Run Terminal Help
syslab2 (2).py x syslab2.py
C:\Users\furay> furay > Downloads > syslab2 (2).py > ...
1 import pygame
2 import sys
3 import random
4 from datetime import datetime, timedelta
5
6 # Initialize pygame
7 pygame.init()
8
9 # Screen dimensions
10 WIDTH, HEIGHT = 800, 600
11 screen = pygame.display.set_mode((WIDTH, HEIGHT))
12 pygame.display.set_caption("Town Simulation")
13
14 # Colors
15 WHITE = (255, 255, 255)
16 BLACK = (0, 0, 0)
17 GRAY = (200, 200, 200)
18 RED = (255, 0, 0)
19 GREEN = (0, 255, 0)
20 BLUE = (0, 0, 255)
21 YELLOW = (255, 255, 0)
22 BROWN = (139, 69, 19)
23 LIGHT_BLUE = (173, 216, 230)
24
25 # Clock/font
26 clock = pygame.time.Clock()
27 font = pygame.font.SysFont('Arial', 16)
28 big_font = pygame.font.SysFont('Arial', 24)
29
30 # Game time starts at 7:00 AM
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS C:\Users\Furay> & C:/Users/furay/AppData/Local/Microsoft/WindowsApps/python3.12.exe "c:/Users/Furay/Downloads/syslab2 (2).py"

powerhell
Python

Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Python 3.12.10 64-bit (Microsoft Store)

AGENT PROFILES

Alice: You are a student, and you enjoy playing basketball and video games

Mr. Smith: You are a teacher who teaches math and history, and enjoys reading

Bob: You are a cashier, and spends your free time watching TV

Charlie: You are a worker at a construction site and love spending quality time with your family



07:30 AM

Alice



Home 1

Having breakfast before school

Mr. Smith



Home 2

Preparing lessons for the day

Bob



Home 3

Getting ready for work

Charlie



Home 4

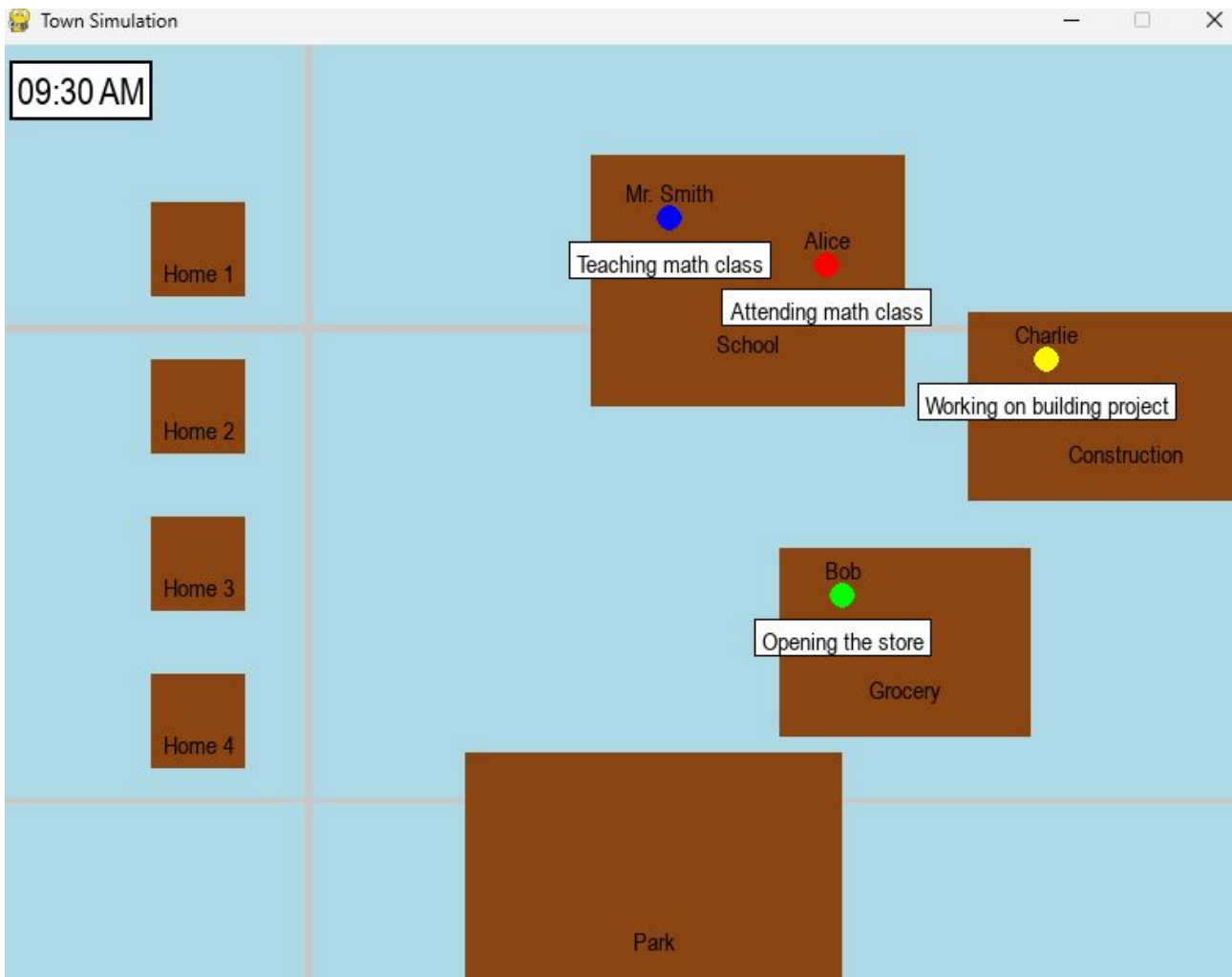
Eating breakfast

School

Construction

Grocery

Park



Classroom Conversation at 09:30 AM:

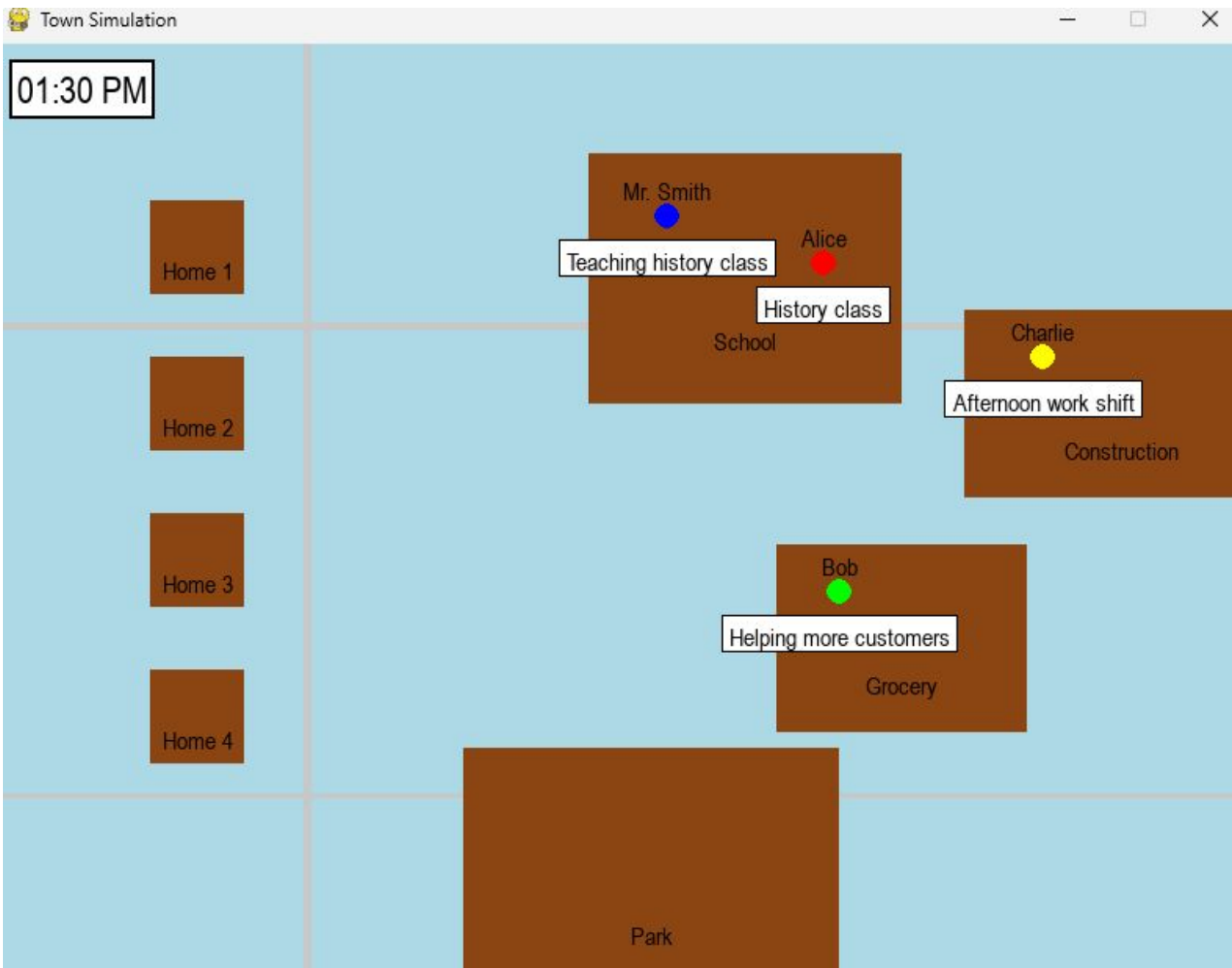
Mr. Smith: Alice, do you understand the quadratic formula?

Alice: I think so, but I'm not sure about the discriminant part.

Mr. Smith: The discriminant tells us about the nature of the roots. Let me explain...

Alice: Oh, I see now! When it's positive we get two real solutions.

Mr. Smith: Exactly!



Classroom Conversation at 01:30 PM:

Mr. Smith: Today we're discussing the American Revolution. Alice, can you name one cause?

Alice: Was it because of taxation without representation?

Mr. Smith: Excellent! That was indeed one of the main grievances.

Alice: I read about the Boston Tea Party last night.

Mr. Smith: Good preparation! That was a key event leading to the revolution.



05:30 PM

Home 1

Alice

Playing basketball with friends

School

Mr. Smith

Making dinner

Bob

Making dinner

Charlie

Having dinner

Construction

Grocery

Park

08:00 PM

Alice



Home 4

Having dinner with family

Mr. Smith



Home 3

Relaxing after work

Bob



Home 2

Watching TV

Charlie



Home 1

Spending time with family

School

Construction

Grocery

Park

LIMITATIONS

- Because we are running the project locally, we are limited by computer specifications. Our laptops aren't suited to run Ollama, but we have computers at home where Ollama works a lot better
- We are limited by time as we do not have the time to try this type of simulation in diverse environments
- When testing, we are limited by the people we are close enough with to test the simulation

FUTURE WORK

- Test the simulation in different, non-preset environments
- Test the simulation in an actual game



Q&A

REFERENCES

- <https://arxiv.org/pdf/2305.10250>
- <https://arxiv.org/pdf/2304.03442>
- <https://arxiv.org/pdf/2203.11171>
- <https://arxiv.org/pdf/2201.11903>
- <https://medium.com/autonomous-agents/mathematically-evaluating-hallucinations-in-llms-like-chatgpt-e9db339b39c2>

THANKS!