Mimic or Miss: An AI Pandemic Response Test

Veronika Wannack, Sharada Prasanna Mohanty, Marcel Salathé,
The Digital Epidemiology Lab, EPFL

Background

During a pandemic, epidemiological models help predict disease spread, shape policies & public sentiment.

However, these models cost a lot of time and resources and struggle with unpredictable parameters, like human behaviour.

Large language models (LLMs, such as ChatGPT, Claude, or Llama) may have the complexity to simulate how humans behave in a pandemic. Recent studies have shown that generate agents can reflect upon their memories, plan and make decisions.

In this research, we explore:

Can LLMs simulate human-like behaviour during global health emergencies?

What are the strengths, limitations, and biases of these models? What influences their ability to mimic human behaviour?

Methods

Initial Model

Spatial SIR (susceptible, infected, recovered), integrated with LLM API. ChatGPT-4o-mini selected due to cost and intelligence. Agent behaviour was refined with prompt engineering.

Agents were told the infection rates & their personality type. Actions determined by the agents were: more, less or no change in social interaction.

Refined Model

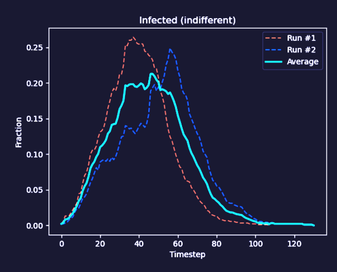
Spatial SEIR (SIR + "exposed"), integrating LangChain, a framework that allows agent-based modelling, and an implementation of the Park et al. (2023) paper, which introduced generative agents with planning capabilities.

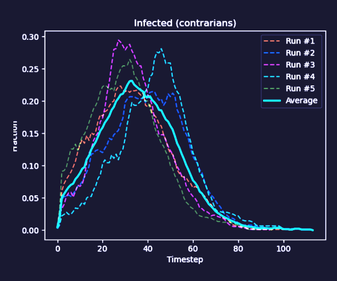
A large sandbox social environment was constructed to simulate real-world scenarios around this agentic implementation.

In this model, the agents watch the news, think and plan their day, make decisions to go to work and attend social hubs, have conversations with each other on current events and reflect on their symptoms.

The actions observed are the same as the initial model: do the agents increase or decrease social interaction? Do the dynamics follow the logic of real pandemics?

Results

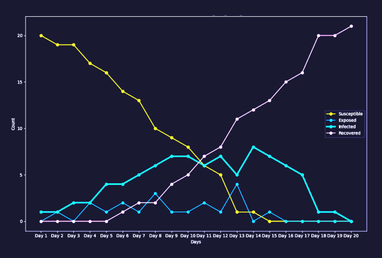




Initial Model

Behaviour at 0% infected: contrarian personalities increased social activities; indifferent personalities had no change.

Behaviour at >1%: all personality types, regardless of the actual infection rate, reacted by quarantining.
→ **The model fails!**



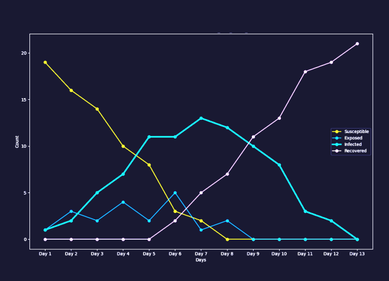
Refined Model

Agents tend to self-isolate as cases go up, but reasonably.

Can observe multiple peaks, like in real life.

Dynamic infection rates: outcome changes based on LLM decisions.

Personalities play an influence.
→ **Good approximation**



Demo

This demo shows the logic of the refined model. The agents go through their day in a virtual society, much like you and I do every day. They wake up, read the news, and plan their days.

The decisions to go anywhere or talk to others is entirely made by the LLM agent. The only influence is the memory & personality!

This simulation has 2 characters, **Alice** & **Bob.**

Alice (age: 25, personality: social, extroverted), describes herself as: “an energetic individual who finds joy in fostering connections”

Bob (age: 21, personality: risk taker, social), describes himself as: “enjoying thrilling experiences, and breaking the rules for fun.”

Home 1 - a cozy appartment	Home 2 - big house with backyard
Status: No conversation	Status: No conversation
Alice [Infected]	Bob [Susceptible]

Start of the simulation

Alice and Bob live at different homes and work at different locations.
Unknown to them, **Alice is infected. Bob is susceptible:** he has a chance of becoming infected if he comes in contact with Alice.

The agents make their daily decisions. Alice and Bob both decide to go to work, but since they do not work in the same place, Alice does not infect Bob.

Work 1 - research lab	Work 2 - tech company
Status: No conversation	Status: No conversation
Alice [Infected]	Bob [Susceptible]

After their hard day, Alice and Bob decide to catch up at the local park. **Alice infects Bob!**

Social 2 - community park	Social 2 - community park
Status: Conversing	Status: Conversing
Alice [Infected]	Bob [Exposed]
Bob [Susceptible]	Alice [Infected]

Initially, Bob is exposed: he cannot infect other people. Eventually, Bob becomes infected.

Home 1 - a cozy appartment	Home 2 - big house with backyard
Status: No conversation	Status: No conversation
Alice [Infected]	Bob [Infected]

The next day, Alice and Bob decide to stay home, feeling symptoms like fever and nausea.

Conclusions

RESEARCH RESULTS:

Initial model does not work: LLM agents will make the “safer” choice to quarantine each time. Very interesting behaviour: is it due to safety “hyperparameters? Biases in training data?

The more refined model improves drastically, but still suffers from similar biases. However, the result is much closer to reality.

Current state: environment has been built, but the models are computationally very expensive. I am continuing this research by optimising the processes. **Goal: paper by end of the year.**

AS A LEARNING EXPERIENCE:

This is the largest and most complex coding project I’ve ever done and I have dramatically increased my coding ability.

I have also learned so many invaluable things about LLMs, agent-based modelling (LangChain), memory management & parallelisation, and so much more from my (incredible) supervisor.

This is also my first long term experience in the lab. I loved spending time with colleagues and look forward to continuing our collaboration.

LAIDLAW SCHOLARS

EPFL

Sources:
Liang, X., Song, S., Zheng, Z., Wang, H., Yu, Q., Li, X., Li, R., Cheng, P., Wang, Z., Xiong, F., & Li, Z. (2024, July 19). Internal Consistency and Self-Feedback in Large Language Models: A survey. arXiv.org. https://arxiv.org/abs/2407.14507
Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, April 7). Generative Agents: interactive simulacra of Human Behavior. arXiv.org. https://arxiv.org/abs/2304.03442
Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., & Murahari, V. (2024, July 25). PersonaGym: Evaluating persona agents and LLMs. arXiv.org. https://arxiv.org/abs/2407.18416
Vatsal, S., & Dubey, H. (2024, July 17). A survey of prompt engineering methods in large language models for different NLP tasks. arXiv.org. https://arxiv.org/abs/2407.12994