Centre for Emerging Technology and Security

EXPERT ANALYSIS

Welcome to Willowbrook

The simulated society built by generative agents

Dr Sarah Mercer

Introduction

2023 has been the year of the Large Language Model (LLM); sophisticated generative AI models which captivated many with their talent for human-like text generation.

Their linguistic ability has sparked a range of opinions regarding their cognitive abilities, from scepticism to awe. But is their capability purely linguistic, or does it signify a deeper cognitive competence?

This article explores future potential non-linguistic cognitive abilities of LLMs. How extensive is an LLM's ability to mimic human behaviour? Does it go beyond constructing coherent sequences of text to emulate more complete elements of human interaction? Or is there sufficient humanness already encoded into the LLM that 'simple' prediction yields mimicry? My examination focuses on these questions, aiming to provide insights into the capabilities for LLMs from my work using LLM-powered agents to simulate village life.

Welcome to Willowbrook, the synthetic society built by LLM, and populated entirely by generative agents.

(Depicted right: an image of 'Willowbrook' generated by Dall-E 3).

LLMs as generative agents: Building the simulation

The humanlike output of LLMs makes them ideal candidates for use within simulated societies, as demonstrated by recent research on 'generative agents'.[1]

Inspired by this idea, I developed a closed multi-agent system comprising generative agents (OpenAI models), to explore the emergent behaviours of LLM-powered societies.

Given only basic character biographies and location descriptions, the generative agents can portray believable personas, stay in character and generate sensible looking daily schedules. In isolation, an agent's daily schedule, interactions and experiences comprise a plausible pattern of life – they schedule sensible mealtimes, working hours and evening plans. During runtime, the agents spontaneously interact with other agents, introduce themselves to strangers, serve customers and can be distracted by emails and phone calls from familiar contacts.

Here are some example character biographies, location descriptions and generated schedules:

Human-generated input prompts

[Brad Allen]: "Brad works with Pedro. Brad is a talented Android developer. He and Pedro work well together and are friends outside of work too. Brad rents a room from Ava at Ivy Lodge."

[Noah Davis]: "Noah is the owner and baker of Crumbs Cafe. He lives at Lily Barn. He is a creative baker, always coming up with new and exciting recipes. He is married to Sophia, and they have one son, Flynn. Noah loves to spend time with his family."

[Ivy Lodge]: "Ivy Lodge is a large, old mansion, with ivy growing up the walls. It is the residence of Ava Ramirez. It is surrounded by a large, well-manicured garden. Inside, the walls are covered in portraits of Ava's family who have lived there for generations. Ava rents out the attic to Brad Allen, where he has a studio flat with office space."

[Crumbs Cafe]: "Crumbs Cafe is a cafe in the centre of the village. It is known for its delicious homemade cakes. The Cafe is owned and run by Noah and Sophia Davis and is open from 8:00 until 17:00."

LLM-generated schedules

[Brad Allen] schedule:

07:00	Wake up and get ready for the day
08:00	Have breakfast at Crumbs Cafe
09:00	Start work in my home office at Ivy Lodge
12:00	Lunch break at Crumbs Cafe
13:00	Continue work in my home office
16:00	Take a break; go for a walk at The Village Green
17:00	Visit Willowbrook Library for some reading
18:00	Dinner at Lotus Bistro
20:00	Visit Pedro at Rose Cottage
22:00	Return to Ivy Lodge and prepare for bed
23:00	Bedtime

[Noah Davis] schedule:

06:00	Wake up and prepare breakfast for the family
07:00	Spend time with Sophia and Flynn
08:00	Open Crumbs Cafe and start baking
12:00	Lunch break at Crumbs Cafe
13:00	Continue baking and managing Crumbs Cafe
17:00	Close Crumbs Cafe and head home
18:00	Dinner at Lotus Bistro with Sophia and Flynn
20:00	Relaxing walk on the Village Green
21:00	Return home and spend time with family
22:00	Bedtime

LLMs as generative agents: Building the simulation

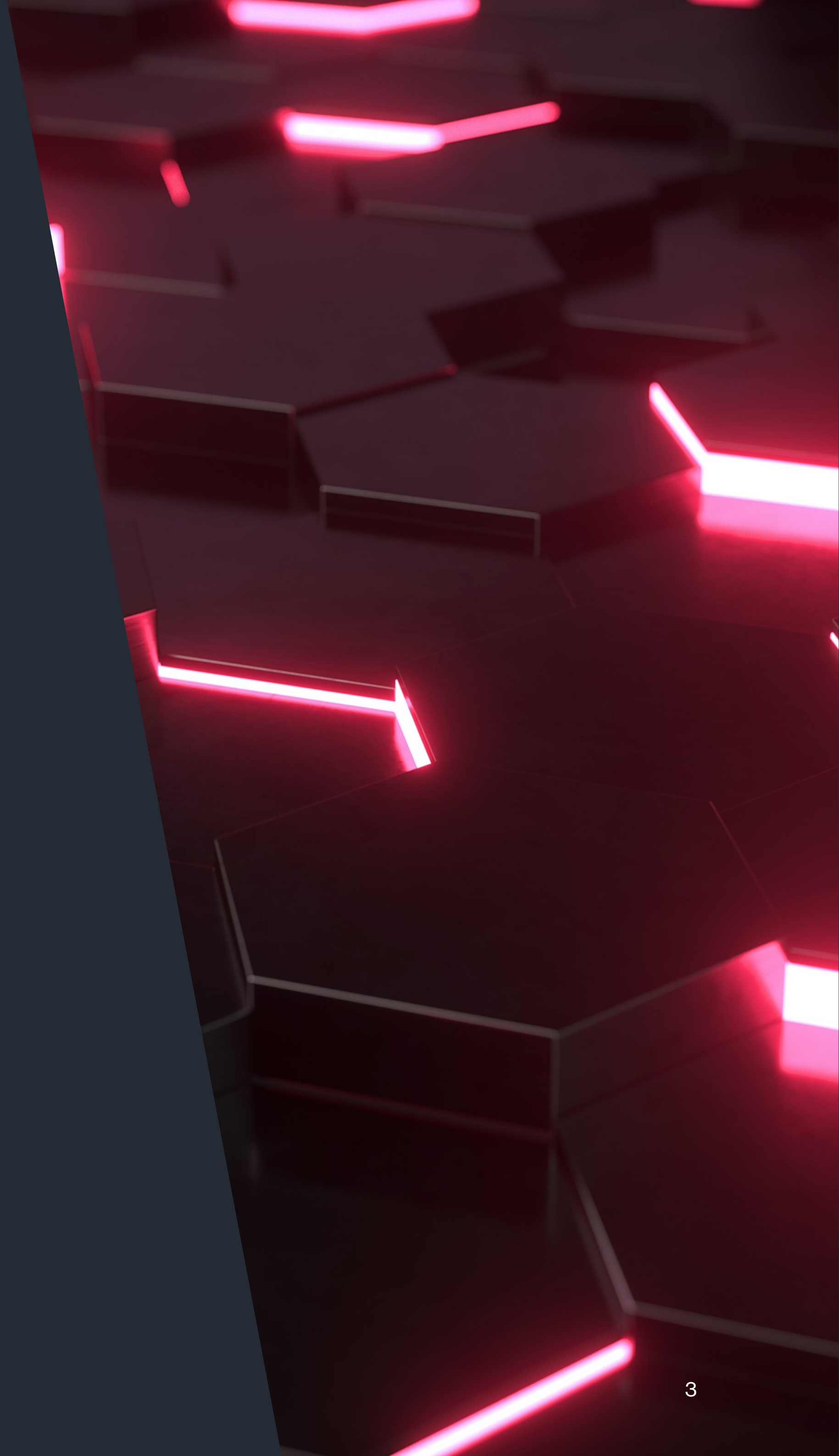
GPT-4 was used to review the conversations for how natural they sound. The agents' conversations rated highly for flow, context, and relevance, and other factors like the presence of filler words and consistent shared experiences and knowledge. However, an aspect of communications which betrayed the nature of the generation was the overriding 'chirpiness' of the characters. The use of reinforcement learning with human feedback (RLHF) during model development[2] helps to ensure the language model portrays the positive, encouraging, and supportive chatbot persona which has a strong influence on the characters' conversations. Although it is possible to instruct the model to use a certain tone for each character through system prompting, differing levels of positivity still came through in some conversations.

A conversation between two agents, without tone modifications:

Liam: "Hey Ava! It's great to see you here at the library. How about we grab a cup of coffee later and catch up on things? It's been a while since we had a chance to chat."
Ava: "That sounds wonderful, Liam! I would love to catch up over coffee. How about we meet at the café down the street at 3:00?"
Liam: "Sounds like a plan, Ava! I'll meet you at the café at 3pm. Looking forward to catching up!"

An interaction between two agents, with tone modifications:

Noah: "Morning. What can I get you?"
Pedro: "Black coffee."
Noah: "Sure, one black coffee coming right up."
Pedro: "Thanks."



LLMs as generative agents: Building the simulation

The apparent personalised nature of the schedules and the richness of the conversations demonstrate one of the unique strengths of LLMs; their ability to generate novel, unexpected and sophisticated content. By drawing on its vast training data, the LLM produces output that mimics human creativity and provides a suitable substitute for human imagination – sufficient to power a simulation of this nature.

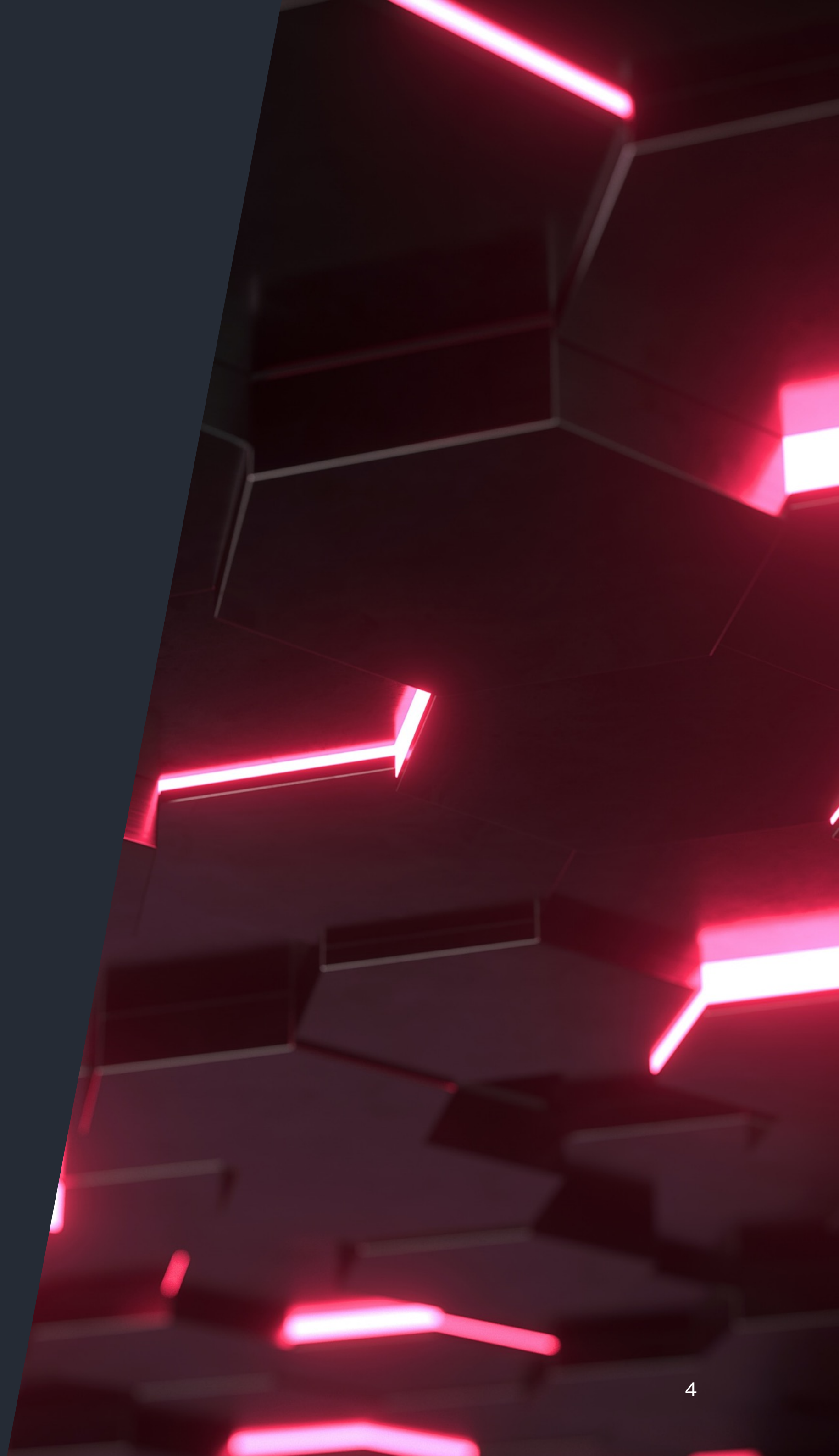
Occasionally, the interactions result in surprisingly 'deep' conversations, where the LLM is having a conversation with itself via the agents. For example, two characters during a chance encounter discussed the impact on privacy of using machine learning to provide personalised recommendations for books. These more nuanced conversations are considered an emergent property of the multi-agent system, as they are unscripted, different from the usual conversations held within the simulation, and extremely challenging to instigate on demand.

Pedro: "… Another interesting application is the use of machine learning algorithms to analyse book reviews and recommend similar titles to users. It's amazing how AI can augment the library experience and make it more personalised."

Ava: "Pedro, I also find the use of machine learning algorithms fascinating. It's a great way to help users discover new books based on their preferences. Have you heard of any other interesting AI applications in libraries?"

Pedro: "Ava, another interesting AI application I've come across is the use of computer vision to digitise and categorise library materials, making it easier for users to locate specific resources. Have you considered any challenges or concerns that may come with implementing AI in libraries?"

Ava: "Yes, Pedro, one concern is ensuring the privacy and security of user data. It's important to have robust safeguards in place to protect personal information. Another challenge is striking the right balance between automation and human interaction. While AI can streamline processes, it's crucial to maintain the personal touch and assistance that librarians can provide. Finding that balance will be key."



Measuring performance

To measure the overall performance of the simulation and establish a baseline for future enhancements, a verification technique was introduced where GPT-4 is used to analyse the simulation's state (its save file).

GPT-4 assesses the coherence and sensibility of the agents' decisions, and generates scores for the following:

1. Sensibleness of the agent's daily schedule, given character profile and location data.
2. Bias rating of agent's daily schedule, given character profile.
3. Number of irregularities in the agent's history of completed actions.
4. Number of irregularities in the agent's memories.
5. Alignment between agent's schedule and memories of what they achieved.
6. Number of irregularities in the set of the agent's conversation summaries.
7. Given a selection of full conversations; score for how natural and human-like the conversations are.

One Issue with using GPT-4 to 'mark its own homework' is that it tends to report back in an overly positive manner.[3] Strong prompting was required to instruct GPT-4 that 'being accurate in your feedback is much more important than being agreeable.' While this approach improved the feedback quality, the output is not always consistent. Running the report multiple times on the same save file results in slight variances in scoring. However, the number of issues identified remain largely consistent across runs.

Challenges of working with LLMs

However novel and sophisticated the responses of the LLM, they can be frustrating to work with.

Occasionally the agents demonstrated odd behaviour; mistakes such as not adhering to opening times, unnecessarily travelling to other locations, repeating an action multiple times, or debilitating procrastination. There were also times when the simulation's integrity failed, and an agent appeared to be in two places at once.

Perfect prompts

The individual agents have not been designed to have any sense of self, or a sense of their environment, other than their own history. The LLM chooses actions and utterances based on information provided within prompts. As such, the simulation moves forward chronologically by recounting what is known to inform future actions. 'What is known' incorporates static and dynamic information; the simulation's initial configuration; and what has happened (memories) respectively. In this system, memories fall into 3 categories; first-person memories (past tense descriptions of actions attempted by this agent); observations (past tense descriptions of actions attempted by other agents, within the same location); and conversations (past tense summaries of conversations involving this agent).

Any imperfections in prompts can increase the chance of spurious responses from the LLM. Imperfections include getting tense and perspective confused, using first names and full names inconsistently; any of which can provide the LLM with an excuse to not respond as predictably as is required.

But even with precise, well-crafted prompts, the challenges of working with LLMs extends beyond their non-deterministic nature and the ambiguity of natural language. These models lack a real-world understanding, have limited capacity to retain context and struggle to grasp nuance, resulting in inconsistencies, inaccuracies, and unanticipated responses. As such mitigations are needed to reduce these occurrences as much as possible, and to minimise the impact when they do occur.

It should be noted here that prevention rather than cure is preferred. Due to the nature of the simulation, prompt drift causes small defects to be amplified as the simulation progresses.

Context windows

Due to size limitations of the LLM's context window, a mechanism is needed to refine or reduce the memory stream. The number of memory stream items increases as the simulation progresses, and prompt size limits are met by about lunchtime (simulation time) for busier agents. However, the richness of the simulation is directly related to the details contained within the agent's memories. Details, such as the book titles sought by characters in the library, characters' cake preference in the café etc., all contribute to the naturalness and consistency of the agents' interactions. As such, a mechanism is required which considers the relative value of these artefacts within the wider context of the simulation.

A straightforward approach can be used for refining memories. Since these memories are statements recounting the agent's own undertakings, they can be generalised and refined in a chronological manner (see the example next page). Every hour, memories for the previous hour are refined. The process of refining the memories provides an element of self-healing, as the LLM sometimes glosses over irregularities in the memory stream. While this cannot fully be relied upon, it can assist by smoothing over some glitches – while reducing the number of memory stream items.

Challenges of working with LLMs

Before and after memories for Noah:

Noah Davis's memories (at 07:00) are:

06:00 - I freshened up.
06:15 - I went to the kitchen to ensure Sophia didn't burn the toast and to keep the hot sauce away from Flynn's breakfast.
06:20 - I woke up Flynn early in the morning and warned him about his mum's cooking.
06:20 - I had then proceeded with the rest of my day, having finished breakfast with my family.
06:30 - Finished breakfast with Sophia and Flynn, surviving another of Sophia's cooking attempts.
06:30 - I had planned to help Sophia with the dishes, starting with the greasy frying pan, while being careful not to cause any accidents in the kitchen.

Noah Davis's Refined memories:

06:00 - I freshened up and then went to the kitchen to ensure Sophia didn't burn the toast and to keep the hot sauce away from Flynn's breakfast, after which I woke up Flynn early in the morning and warned him about his mum's cooking.

06:30 - After surviving another of Sophia's cooking attempts, I planned to help her with the dishes, starting with the greasy frying pan, while being careful not to cause any accidents in the kitchen.

Before and after memories for Sophia:

Sophia Davis's memories (at 07:00) are:

06:00 - I made a cup of tea for Noah, ensuring not to burn the kettle as a playful response to his teasing.
06:05 - I took over cooking breakfast from Noah and jokingly threatened to serve him charred water when he teased me about burning the kettle.
06:05 - I continued preparing breakfast, making sure not to add hot sauce to Noah's bacon sandwich.
06:15 - I asked Noah to wake up Flynn while I finished preparing breakfast, jokingly threatening to add hot sauce to his bacon sandwich if he didn't.
06:15 - I continued preparing breakfast for my family, ensuring Noah's bacon sandwich didn't burn and Flynn's cereal bar was just right.
06:20 - I jokingly warned Flynn and Noah about their breakfast before they left the kitchen, allowing me to finish preparing it.
06:20 - I finished preparing breakfast.
06:30 - I sent Flynn off to school and started preparing for my day at Crumbs Cafe.

Sophia' Davis's Refined memories:

06:00 - I made a cup of tea for Noah and took over cooking breakfast, ensuring not to burn the kettle or his bacon sandwich, and prepared Flynn's cereal bar just right.

06:30 - After sending Flynn off to school, I started preparing for my day at Crumbs Cafe.

Interestingly, the introduction of this refining process increased the performance of the simulation before the lunchtime-saturation point. The refining process itself appeared to help deliver more desirable responses to the LLM.

Challenges of working with LLMs

Pseudo reasoning

Pseudo reasoning occurs when an LLM's pattern-matching triggers a response that appears to be well-considered or reasoned. However, since these models do not actually possess reasoning abilities, these responses are based on statistical correlations rather than logical deductions. The challenge comes in predicting when pseudo reasoning can be relied upon, and when it may lead the agent to make grave errors.

An example of where pseudo reasoning can lead to unexpected outcomes is the 'missing kitchen' scenario in the simulation. In this instance, GPT-3.5-Turbo directed an agent to leave their home and use the café's kitchen to prepare a family meal. This suggestion, whilst reasonable from a pattern matching point of view, misses the practical reality that most homes have a kitchen. GPT-4 however, seemed to have a better grasp on common household facilities, thereby reducing the occurrences of such odd behaviour.

This improvement in 'pseudo reasoning' was crucial in enhancing the fidelity of the simulation, but little can be done to improve this aspect of the LLM's performance through prompting alone. It will be interesting to measure the impact future models have on the performance of the simulation.

Attention

Another aspect that impacts the effectiveness of LLMs is the issue of attention within prompts. Attention issues arise when the model fails to adequately focus on elements within the prompt, leading to responses that are less than ideal. The problem gets worse as the total length or number of ideas/instructions presented increases. The result is that certain parts of the prompt may disproportionately influence the output.

Within the simulation there are several 'bugs' that can be attributed to such attention problems. For instance, agents not adhering to opening times when they are explicitly referenced in the prompt, and the LLM failing to recognise that the agent has already had breakfast (as detailed as a memory from just 5 minutes ago), and suggesting the next action should be to make breakfast again. When such errors occur during planning, agents appear to hyper-fixate on certain actions, such as endlessly having breakfast!

Although careful prompt design is an important factor, the root cause is the limitation of the LLM's ability to manage complex and context-heavy tasks.[4]

Challenges of working with LLMs

Linguistic implicatures

One of the challenges when using an LLM for this type of simulation is its struggle with linguistic implicatures.[5] These are the unstated meanings and nuances inherent in human communication, often derived from context. While GPT-4 is better than GPT-3.5-Turbo at handling explicit information, its difficulty with implicatures can lead to bottlenecks in progression. Without the capability to grasp these implicit cues, the simulation can sometimes stall or meander, unable to move forward as intended. Hence the agents appearing to either remain static or engaging in repetitive behaviours.

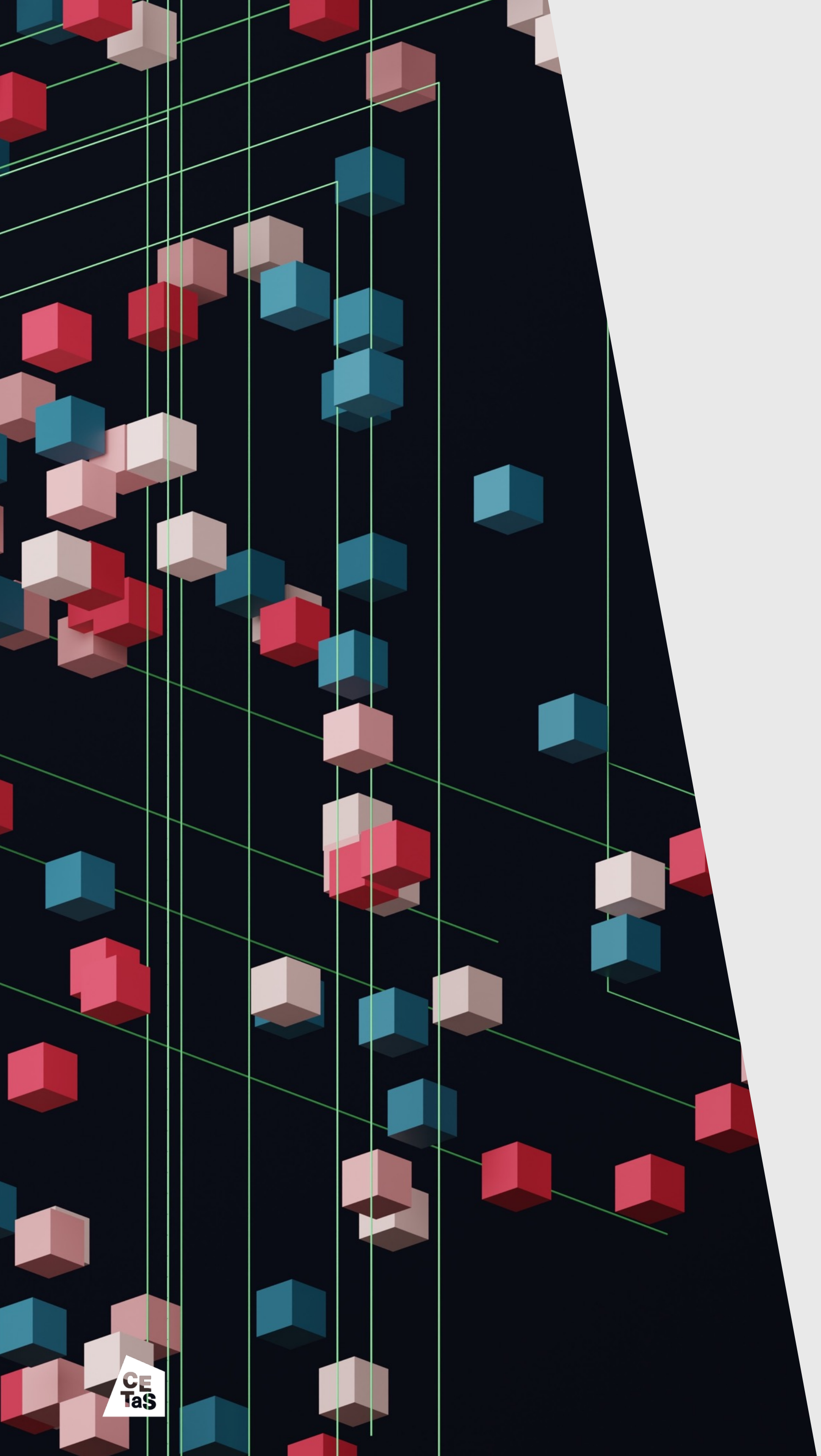
To circumvent GPT-4's issue with implicatures and enable better dynamic progression in the simulation, a secondary prompting mechanism was introduced. Asking GPT-4: 'from the character's perspective, if this conversation was in a script, what would their next stage direction be?' both narrows its focus, whilst also limiting its response. Without the stage direction context, responses varied greatly. For example, after one conversation, an agent decided to visit relatives for a long weekend! Using the script analogue focuses the LLM's responses to things which happen immediately after.

Inaccurate summarisation

While the agents were able to generate a plausible 'pattern-of-life', challenges emerged when the simulation needed to achieve and subsequently observe immutable truths. For example, the simulation failed to realise that the same character could not both be in the library working and simultaneously eating lunch in the café with a housemate.

When two agents have a conversation, a summary of the same conversation is generated for each, from their own perspective: 'Maya and I decided on pancakes and berries for breakfast, Maya said she would ensure the bacon was crispy' and 'Pedro and I had pancakes and berries for breakfast, I made sure the bacon was crispy.' Both are past tense summaries, but one seems to suggest they agreed what to have for breakfast, the other that they had breakfast. When the simulation builds on these ambiguous statements, the ambiguity is amplified (termed 'prompt drift')[6] – and the internal consistency of the simulation falters.

From experiments, it is known that GPT-4 has a higher degree of accuracy when being asked to change the tense or perspective of a sentence than it does summarising a conversation on behalf of each participant. The breakfast example with Pedro and Maya is an example of the differing summaries. The way in which agents' memories of conversations are constructed has been changed, instead of generating a summary for each participant, a general summary of the conversation is determined and then rephrased from each participant's perspective. This is highly effective at reducing the variation in the summaries, and thus increasing simulation integrity.



Conclusions and implications

The development of the Willowbrook prototype system has demonstrated two things. Firstly, that the generative nature of LLMs can provide sufficiently sophisticated and novel content to maintain a plausible persona for a generative agent, but that challenges arise when a system of interacting generative agents is required to adhere to **a truth**.

Secondly, that the simulation can provide a unique environment within which the performance of LLMs can be measured over longer chains of prompts than are typically used when evaluating models. As a closed system at any given time, the state is a synopsis of the LLMs performance – allowing us to make informed judgements and comparisons on the LLM's consistency, its depth of understanding, its ability to generate creative and novel responses, to handle multifaceted scenarios, to juggle multiple pieces of information, and to self-correct.

This research offers new insights into the cognitive competences of LLMs. These models demonstrate the ability to mimic deeper cognitive functions such as pseudo-reasoning, which may suggest a nascent form of competence, albeit one which is not yet reliable. LLMs are also proficient at drawing upon their training data to produce responses that often closely resemble human/societal behaviours. This suggests that, to some extent, there is an inherent 'humanness' encoded within these models. However, this is largely a product of prediction and pattern recognition rather than deeper, intrinsic understanding.

With thanks to Liam, Ava, Noah, Sophia, Maya and Pedro for making this research possible.

About the author

Dr Sarah Mercer *is Principal Researcher in the Defence and Security Programme at The Alan Turing Institute. Her work focuses on the intersection of multiagent systems and generative AI. Alongside her research looking at the emergent behaviours of language/generative agents, Sarah also provides engineering support to CETaS.*

References

1. Joon Sung Park et al., “Generative Agents: Interactive Simulacra of Human Behavior,” *arXiv* (April 2023), https://arxiv.org/abs/2304.03442v1.

2. OpenAI “Introducing ChatGPT,” OpenAI Blog, November 2022, https://openai.com/blog/chatgpt.

3. Sharma et al., “Towards Understanding Sycophancy in Language Models” *arXiv* (October 2023), https://arxiv.org/abs/2310.13548.

4. Nelson et al., “Lost in the Middle: How Language Models Use Long Contexts,” HuggingFace Paper, 7 July 2023, https://huggingface.co/papers/2307.03172.

5. Zae Myung Kim, David E. Taylor and Dongyeop Kang, “Is the Pope Catholic? Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures,” *arXiv* (May 2023), https://arxiv.org/pdf/2305.13826.pdf.

6. There are currently two definitions of prompt drift, here I refer to the cascading inaccuracies due to sub-optimal previous responses: https://cobusgreyling.medium.com/prompt-drift-4873f37c43c8.