Constructing Agent Brains Using Large Language Modeling and Applications in Natural Language Processing

Soheil Saneei, Shreya Banerjee

Abstract

Large-Language Models (LLMs) have re-popularized research into general autonomous agents with human-like cognitive abilities, particularly in creating simulacra for human behavior. Our research contributes to this area by exploring methods for simulating believable multi-agent communication through LLMs imbued with 'brains' that we constructed inspired by components from foundational papers. While our broader research initiative is focused on developing a comprehensive full 'brain' architecture for these agents, this paper hones in on a critical subset of its functionalities that shape dialogue. We connected this LLM 'brain' to ChatGPT-4 using an API and embedded our agents in a turn-based simulation. We compared both individual components and also evaluated their combined performance. We also documented token usage to further analyze prompt optimization. This work aims to advance the development of more nuanced, psychologically grounded AI agents capable of complex and realistic social interactions within simulated environments.

Keywords

Large Language Models, Generative Agents, Agent-Based Simulation, Personality Modeling, Multi-Agent Communication, Prompt Engineering, Cognitive Architecture

1. Introduction

Simulating human behavior has long been a central goal in artificial intelligence, bridging fields such as cognitive science, computational psychology, and human-computer interaction. With the advent of Large Language Models (LLMs), this ambition has taken on new momentum. Unlike traditional symbolic systems or handcrafted agent frameworks, LLMs offer a powerful means of generating flexible, context-aware behavior, raising the possibility of autonomous agents that not only act but also reason, reflect, and socialize like humans. Recent studies have begun exploring this potential. Park et al.'s Generative Agents: An Interactive Simulacra for Human Behavior [1] demonstrated that LLMs could underpin agents capable of believable daily routines, memory-driven behavior, and social interactions within a small virtual town. These agents were embedded into a Sims-like 2D environment and exhibited socially-emergent behavior. For example, in the simulation, one of the agents, Isabella, decides to host a Valentine's Day party and begins inviting other agents during her daily interactions. Through memory and planning, agents remember the event, spread the invitation organically, and adjust their own schedules to attend, creating believable emergent social dynamics. The main mechanisms within the work include:

- Identity & Backstory: Agents are initialized with a prompt containing a backstory that produces an identity for them. A summary of this backstory is stored within the memory stream and is subject to changing depending on significant events.
- Memory Stream: Agents have an evolving memory of past experiences, which is selectively retrieved based on relevance and recency to inform future actions and conversations. Memory relevance is primarily determined using embedding-based retrieval, and cosine similarity is used to measure how relevant a stored memory is to the agent's current situation or query.

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

soheilsaneei@gmail.com (S. Saneei); shreyabbanerjee@gmail.com (S. Banerjee)

- Reflection: Agents periodically review their memories to infer higher-level insights (e.g., realizing someone likes them) and update their beliefs, influencing future plans and behaviors.
- Planning: Agents autonomously create daily schedules and adjust their actions based on internal goals, needs, and social information, allowing them to coordinate activities like attending events.
- Reactive Behavior: Agents respond to immediate stimuli from the environment and other agents in real time, balancing between following plans and adapting to new interactions.
- Movement: Agents calculate paths and use animations to get to locations that are developed through their plans and scheduling.

This foundational work sparked a wave of interest in using LLMs to simulate interpersonal dynamics. Piao et al.'s *AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society* [2] further scaled these ideas, using 10,000 generative agents to investigate large-scale societal phenomena. They were inspired by the mechanisms within the Park et al. project and expanded:

- Agent Design (LLM-driven Social Generative Agents) Each agent is built with: Profile & Status: Demographics (name, age, etc.), economic status, mental state, social relationships.
- Three Mental Layers:
  - Emotion: Based on six basic emotions (e.g., joy, sadness, fear) influencing decision-making.
  - Needs: Modeled using Maslow's hierarchy (e.g., physiological, safety, belonging).
  - Cognition: Attitudes, thoughts, memory, and perception influenced by emotion and experience. Agents dynamically plan and act based on their emotions, needs, and cognition.
- Mind-Behavior Coupling
  - Psychological states (emotion, needs, cognition) directly drive behavior selection. Uses theories like Theory of Planned Behavior to explicitly map mental states to actions.
- Behavioral Modules
  - Mobility Behaviors: Movement is needs-driven (e.g., moving to a café to satisfy social needs). Places are chosen using a Gravity Model (attractiveness vs. distance).
  - Social Behaviors: Online/offline interactions, relationship strength (family, friends, colleagues). Conversations influence emotions, attitudes, and can lead to events like meetings.
  - Economic Behaviors: Employment and consumption behaviors. Dynamic adjustment based on income, savings, taxes, and interest rates.
- Memory and Adaptive Behavior
  - Stream Memory: Event Flow (records actions/events) and Perception Flow (records thoughts and emotional reactions). Enables agents to learn and adapt based on past experiences.

While each project focuses on social behavior at their respective scale (small vs. large), each project's model of the individual lacks in-depth personality modeling. Klinkert et al. explored the accuracy in LLMs representing personality profiles based on the Big 5 personality framework in their work, *Driving Generative Agents With Their Personality* [3]. The Big 5 personality framework consists of openness, conscientiousness, extroversion, agreeableness, and neuroticism (OCEAN). It is widely considered the most robust personal assessment framework. This paper explores how Large Language Models (LLMs), particularly GPT-4, can be used to create more human-like Non-Playable Characters (NPCs) in video games by integrating psychometric personality values. Using the Big Five personality framework, the researchers prompt LLMs to generate behavior and dialogue that reflect a given personality profile. Their evaluation, based on repurposing the International Personality Item Pool (IPIP) questionnaire, shows that GPT-4 can accurately and consistently embody assigned personalities, outperforming earlier models like text-davinci-003 and gpt-3.5-turbo. The results suggest that combining Affective Computing systems with advanced LLMs offers a promising path toward developing emotionally rich and believable NPCs, with future potential to incorporate evolving emotions, attitudes, and social dynamics for even deeper realism.

2. A Comprehensive and Efficient LLM Brain

Each of these papers contributes new insights to the literature and highlights the gaps that remain for future exploration. In our paper, we aim to explore which combination of these concepts yields the best performance for simulating dialogue within a conversation between agents. We use prompting that incorporates OCEAN for personality modeling, Maslow's hierarchy for needs, and a reflection mechanism for updating emotions and analyzing the other agent's beliefs, desires, and intentions.

3. Methodology

This research employed an agent-based simulation framework built in Python to investigate the impact of prompt engineering on generating believable, personality-driven agent interactions. The core components were autonomous agents powered by a Large Language Model. Each agent was initialized with a unique psychological profile defined in a JSON file, specifying scores for the Big Five (OCEAN) personality traits and initial urgencies for Maslow's hierarchy of needs. Beliefs, desires, and intentions for each agent were also developed. The simulation proceeded in discrete turns, where each agent generated a dialogue response based on the conversation history, task context, and its internal state. A reflection mechanism, inspired by the *Generative Agents* paper, was developed in accordance with Bryant et al.'s *Theory of Mind experience sampling in typical adults*[4]. The empirical study found that participants thought about mental states 30 percent of the time during conversations. When they thought about mental states, they thought about their own mental state 67 percent of the time and the mental state of the person they were talking to 37 percent of the time. To simulate this reflection process, we generated a random number between 0 and 1 every time an agent took a turn to speak in the conversation. If the number was below .30, the agent would be prompted to reflect. We then generated another random number which determined whether the agent would self reflect or reflect about the other agent. When the agent self-reflected, it would assess its own six basic emotions: joy, sadness, fear, anger, surprise, and disgust. It would also think about its own beliefs, desires, and intentions at that point of the conversation, updating them accordingly. When the agent reflected about the other agent, it would be prompted to develop three highly-salient questions about the conversation so far and answer them. The agent would also guess the other agent's beliefs and desires. The primary methodological intervention involved systematically modifying the prompt template provided to the LLM for response generation.

Simulations were executed under different prompt configurations, and the resulting dialogue logs were qualitatively analyzed to assess the effectiveness of each prompt modification in eliciting human-like behavior. First we tested the full system, every component was prompted and fed into the LLM. Then we tested dialogue with just OCEAN modeling. Next we tested the dialogue with strictly Maslow's hierarchy of needs and self-reflection. Finally, we tested with just the self-reflection and other-reflection mechanism. We then conducted additional tests with the full model, completely without belief, desire, and intention. Finally, we conducted contrastive testing with the inputs in the OCEAN model, with the full model active. We simulated each conversation for 20 turns and each conversation 3 times for each configuration.

4. Results

4.1. Comparing Configurations

In the full configuration, where the prompt includes an O C E A N 5-tuple, Maslow needs, and both self- and other-reflection, the dialogue is colorful, emotionally expressive, and purpose driven. Alex and Milo spontaneously voice feelings ("I feel a sense of joy and excitement about the potential friendship...") and raise meta-level questions such as "What specific interactive elements can we incorporate into the Garden project?", which in turn drive concrete planning of workshops, murals, and sensor-based sculptures. The combination of high Openness/Agreeableness from OCEAN, Maslow-driven need shifts, and the theory-of-mind prompts injected by other-reflection yields the richest, most collaborative conversation in the study.

When the BDI bookkeeping is disabled but the rest of the prompt is unchanged, the live dialogue is virtually indistinguishable from the full run: the same exuberant wording, the same reflective questions, and the same cadence of idea generation appear, confirming that post-simulation BDI analysis affects logs, and token cost, but not the conversation itself.

In contrast, the OCEAN-only condition (traits injected, but no Maslow and no reflection) retains the imaginative vocabulary—"It's such a creative way to blend functionality with artistry..."—yet feels emotionally flat and drifts into polite repetition because no reflective mechanisms push it forward. Without Maslow there are no affect swings, and without reflection the partners seldom probe one another's intentions, so plans emerge slowly and remain vague.

The Maslow + Self-reflection mode removes the trait line but keeps need updates. Here the agents' dominant needs flip (Milo shifts from Belonging to Esteem after turn 2) and joy steadily rises, giving the talk an upbeat emotional arc. However, diction becomes generic enthusiastic ("Let's make this smart organiser... a true piece of art!") because the model no longer receives personality cues, and—lacking other-reflection—the speakers mostly echo each other's enthusiasm until late in the exchange.

Finally, the Reflection-only run (self + other reflection, no OCEAN, no Maslow) regains perspective-taking depth—"How can we structure the mural so visitors feel ownership?"—but the language is plain and the emotional register steady because neither traits nor needs are in play. The dialogue is purposeful, thanks to other-reflection's "salient question" injections, yet stylistically flat and interchangeable across agents.

Across these modes, personality text chiefly colors wording and increases collectivist phrasing; Maslow and self-reflection supply affect and motivational shifts; other-reflection is the engine of theory-of-mind questions and concrete next actions—though it is also the primary token-cost driver. Disabling post-run BDI analysis trims tokens without altering dialogue, while omitting OCEAN or Maslow removes distinctive voice or emotional dynamism respectively, and omitting other-reflection pares back strategic depth.

4.2. Contrasting OCEAN Traits

After conducting our initial tests, we generated contrastive values for the OCEAN model. We modeled Alex to be highly open, have low conscientiousness, moderate extroversion, low agreeableness, and low neuroticism. We modeled Milo to have low openness, high conscientiousness, moderate extroversion, high agreeableness, and low neuroticism.

The simulation results demonstrated varied consistency in reflecting agents' OCEAN profiles through LLM-generated dialogue. Milo Bennett's profile (high Conscientiousness C=0.9, high Agreeableness A=0.8) was consistently reflected. His high Conscientiousness clearly manifested through methodical planning, such as suggesting specific materials like ”smooth stones or soft moss” for a sensory wall and proposing testing phases like ”creating a prototype in the Academy first” (Log 5, Turn 10), alongside logistical focus like setting ”a specific time for that visit to keep ourselves accountable” (Log 6, Turn 4). His high Agreeableness was also consistently evident in supportive, cooperative language, often validating Alex's points with phrases like, ”I see where you're coming from, Alex, and I think your idea... makes a lot of sense” (Log 6, Turn 6). Conversely, simulating Alex Carter's low Agreeableness (A=0.1) proved inconsistent across runs. While initial simulations showed highly agreeable dialogue, potentially due to LLM bias or overriding BDI needs, later simulations (Logs 5 and 6) successfully demonstrated low agreeableness via critical questioning and skepticism, such as Alex stating, ”While I get what you're saying... I wonder if the distractions... might actually hinder our focus” (Log 6, Turn 3) or expressing doubt like, ”I'm a bit skeptical about how effective that transition would be” (Log 5, Turn 5). We first tried prompting descriptions of OCEAN, however, this proved ineffective. We then introduced conditional behavioral guidance directly into the prompt, instructing the agent to exhibit behaviors consistent with specific high or low agreeableness scores. This method yielded sufficient results, indicating that LLMs are naturally programmed to be agreeable.

We then adjusted Alex's neuroticism to be very high (N=1.0). It became clearly evident in the later simulations (especially Log 6), layering onto the low agreeableness. This manifested as frequent expressions of worry and a focus on potential negative outcomes, evident in statements about the "risk losing the depth of engagement" (Log 6, Turn 13) or the feeling that parameters were needed to "save us from potential chaos" lest they "end up with a disjointed project" (Log 6, Turn 17).

5. Discussion

Our results demonstrate that the full prompt configuration—including OCEAN traits, Maslow needs, and both self- and other-reflection mechanisms, yields the most realistic and collaborative dialogue, even without explicit BDI bookkeeping during live generation. Disabling BDI post-processing reduced token cost without altering dialogue quality, suggesting that lightweight after-the-fact analysis is sufficient for logging purposes and unnecessary for live conversational realism. In contrast, removing any major component (OCEAN traits, Maslow needs, or other-reflection) impaired distinct aspects of the dialogue: OCEAN traits colored the emotional tone and phrasing, Maslow dynamics drove affective shifts and motivation, and other-reflection propelled theory-of-mind depth and action planning.

These findings carry important implications for video game development, the simulation-to-reality (sim-to-real) gap, and AGI research. For video games, especially narrative-driven and social simulation titles, our results suggest that embedding lightweight psychological models, especially affective need shifts and theory-of-mind questioning, can greatly enhance the emotional believability and depth of NPC interactions without incurring prohibitive computational costs. For sim-to-reality research, maintaining motivational structures like Maslow and simulating reflective social cognition can make AI agents more behaviorally rich and goal-adaptive, critical for transferring trained behaviors into open, real-world environments. In AGI development, these experiments point toward a minimal viable architecture where layered affective, personality, and social-cognitive drivers produce realistic and adaptive communication without needing costly symbolic or BDI-heavy processes during runtime. Particularly, they suggest that scalable AGI dialogue will rely more on integrating emergent personality, motivation, and perspective-taking models into prompt-engineering and less on explicit symbolic representations.

Additionally, our contrastive tests of OCEAN trait modeling highlighted current LLM tendencies, particularly a natural bias toward agreeableness. Achieving low agreeableness or high neuroticism required direct behavioral conditioning rather than simple trait descriptions, pointing to current limitations in the direct expression of personality dimensions through prompting alone. Future research will be needed to develop more nuanced conditioning methods to fully unlock trait-based variation in large language models.

6. Future Work

For future work, the prompts can be further optimized for token usage. Different ideas for components can be tested to create better LLM 'brains'. We plan to implement this brain into a 3D gaming environment, to test agent behavior for simulation and video game applications. In addition, different combinations of values for OCEAN can be tested. Further testing of Maslow's hierarchy of needs within a dialogue setting. We expect the needs-based modeling to be highly impactful in the gaming environment.

Declaration on Generative AI

The core methodology of this research involved the use of GPT-4 (via API access) for agent brain construction and dialogue generation, as described in the paper. No generative AI tools were employed for the writing or editing of this manuscript beyond the research itself. The author takes full responsibility for the publication's content.

References

[1] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, arXiv preprint arXiv:2304.03442 (2023).

[2] J. Piao, Y. Yan, J. Zhang, N. Li, J. Yan, X. Lan, Z. Lu, Z. Zheng, J. Y. Wang, D. Zhou, C. Gao, F. Xu, F. Zhang, K. Rong, J. Su, Y. Li, Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society, arXiv preprint arXiv:2502.08691 (2025).

[3] L. J. Klinkert, S. Buongiorno, C. Clark, Driving generative agents with their personality, in: AIIDE'23: Experimental AI in Games, CEUR Workshop Proceedings, 2023. URL: https://gitlab.com/humin-game-lab/artificial-psychosocial-framework/-/tree/master/LLM%20Personality.

[4] L. Bryant, A. Coffey, D. J. Povinelli, J. R. Pruett, Theory of mind experience sampling in typical adults, Consciousness and Cognition 22 (2013) 697–707. doi:10.1016/j.concog.2013.04.005.