

Outwit, Outplay, Out-Generate: A Framework for Designing Strategic Generative Agents in Competitive Environments

Samuel Thudium^{*}, Federico Cimini[†], Rut Vyas[‡],
Kyle Sullivan⁴, Louis Petro⁴, Andrew Zhu⁵ and Chris Callison-Burch[‡]

University of Pennsylvania

{sthudium, fcimini, rutvyas, kylems, petrol, andrz, ccb}@seas.upenn.edu

Abstract

We explore the strategic capabilities of generative agents in a series of social competitive games that emulate the television show Survivor. Large Language Models have been shown to act as intelligent agents through the addition of external cognitive architectures, but it is still unknown how these agents perform in competitive, multi-agent environments. We suggest a framework, built on top of a frozen large language model, GPT-4, for designing generative agents in competitive, episodic environments and we evaluate their game performance. We provide new modules which enable strategic agents to set and evaluate goals, develop theories of mind about other agents, and embody descriptive personas that affect their behavior. Across many simulations, though agents with varied cognitive abilities displayed even performance in a "last-agent-standing" scenario, strategic agents exhibit more diverse action selection, including an affinity for exploratory behaviors. We also observe that agents with personas tailored to the social environment are significantly more likely to win these challenges. Finally, goal-driven agents perform well in a search-based game, displaying an emergent capacity to rapidly utilize information perceived from their environment to inform task-related actions.

1 Introduction

Artificial intelligence (AI) has a long history in the world of gaming and is capable of super-human performance in some tasks like Chess and Go (Bernstein et al., 1958; Silver et al., 2016). However, these canonical AI agents are masterful at a single game and not capable of broader intelligent gameplay. On the contrary, some of the most difficult tests of human intelligence are those that place players in competition against one another and require

^{*}First author.

[†]Second author.

[‡]Professor and senior author.

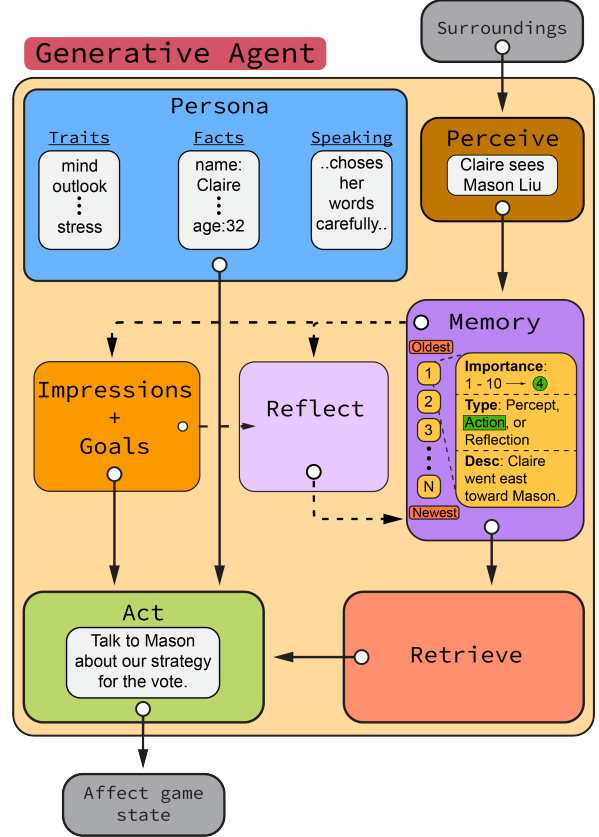


Figure 1: A generative agent’s cognitive architecture (discussed at length in Section 3): information from the external environment passes through a perception module, which converts observations into memories. Using their persona, memories, and goals, agents make informed actions that affect external game state. **Solid lines** indicate data flow that occurs at every turn. **Dashed lines** indicated data flow that only occurs periodically (not all connections shown for clarity; a complete version can be seen in Figure A.1)

mental flexibility, social prowess, and theory of mind (Meta Fundamental AI Research Diplomacy Team (FAIR)[†] et al., 2022).

The rapid improvement of Large Language Models (LLMs) has enabled new avenues of research that explore the behaviors of generative agents

(computational agents backed by an LLM). Much of the work put into this domain has focused on developing LLM-based agents in the context of personal assistants, collaborative multi-agent settings, or conversational human-computer interactions (Li et al., 2024; Zhang et al., 2024; Wu et al., 2023). Recent research has shown promising developments that enable generative agents to exhibit believable or human-like actions (and interactions) in open-world scenarios. In particular, by adding cognitive helper modules on top of an LLM, interactive agents were shown to navigate a social environment as believable simulacra of human behavior, reminiscent of the game The Sims (Park et al., 2023). However, the performance and strategic capabilities of generative agents in competitive gameplay environments is poorly understood. Development of intelligent generative agents with the capacity to set long-term goals, take subversive actions, and conceptualize the difference between public and private strategy could have interesting consequences both within and outside of the gaming world. For instance, generative agents could enable multi-player games without the need to connect with other human players online by elevating the abilities of non-player characters (NPCs), which are historically limited as proxies of believable opponents or teammates.

We set out to test the current limits of generative agents in competitive gameplay; to that end, we identified that the task of navigating a competitive environment by utilizing strategy, deception, cooperation, and social skills exists in numerous formats, including reality entertainment. The television show Survivor, in which players must vote each other off the game until a single player remains, requires contestants to combine their physical and strategic capabilities. Taking inspiration from this game format, we propose an extension to a cognitive architecture described previously (Park et al., 2023); we investigate the ability of generative agents to compete in a game modeled after the reality television show "Survivor". In Survivor, players must navigate social interactions, explore the environment, and make strategic voting decisions. To enable competitive agents, we implement many of Park et al.’s original cognitive structures for generative agents, including a long-term *memory stream* of observations, *reflection* upon these experiences, and ranked *retrieval* of memories that accounts for that agent’s goals, sense of self, and perceptions (Figure 1).

Furthermore, we introduce new cognitive modules that support strategic decision-making and ground the agent’s behavior using static traits. Agents are given a *persona*, which defines their behavioral tendencies across several traits, that provides factual background information, and informs their game strategy. Next, leveraging their knowledge of the environment and past experiences, we enable generative agents to create *goals* at several priority levels. These goals are updated over the course of the game and used as context by generative agents to make strategic actions. Lastly, given the role that social awareness has in navigating our multi-agent environment, we also implement *impressions*, which assist generative agents in developing an inner monologue about other agents with whom they have interacted (Section B.2, Figure 1).

In total, this work provides the following extensions to the existing literature:

- A flexible package that allows rapid generation of multi-agent, competitive environments¹.
- Extended architecture: new cognitive modules, **Impressions** and **Goals**, expand agents’ capacity for long-term planning and interpersonal reflection.
- Detailed **Personas** ground agents’ personality traits and strategy, affecting action selection, dialogue content, speaking style.
- Evaluation of agent performance in games requiring strategy: (1) comparing in-game metrics of rounds survived across persona and cognition architecture types and (2) the ability of generative agents to turn clues into strategy to quickly discover hidden treasures.

2 Related Work

2.1 Foundations

This work builds off of prior designs of generative agent architectures, which add additional modules on top of Large Language Models (LLMs) that enhance long-term memory and assist in adapting the LLM to its environment. A recent approach fine-tunes an based on rewards collected via exploration in a text-based environment (Carta et al., 2023). Another approach develops an LLM-based

¹<https://github.com/sthodium25/SurvivorWorld>

agent that is capable of continual learning from interactions, or CLIN (Majumder et al., 2023). This work reiterates the importance of grounding a language model in learned information from the environment, especially when asking such an agent to solve spatial and non-trivial tasks. In this case, grounding occurs across episodes in ScienceWorld (Wang et al., 2022), with a reflection upon actions taken in each trial and a persistent, dynamic memory that is available to the agent. We also note seminal work in agent dialogue, which describes an RL algorithm that fine-tunes an agent LLM on a dataset of goal-oriented dialogue. Agents are rewarded in interaction with another agent if they achieve their goal, thereby learning how to direct their utterances toward a goal (Ammanabrolu et al., 2021). Others have explored the use of an LLM to direct competitive agents in an commerce setting in which restaurant agents compete for consumer choice (Zhao et al., 2023).

Perhaps most relevant to this paper, LLM-based agents with enhanced cognition capabilities, including reflection, planning, and memory retrieval, exhibit emergent, believable social behaviors such as information diffusion, in which a human-supplied seed idea is given to an agent and this information is propagated throughout the agent cohort (Park et al., 2023).

2.2 Improving strategic capabilities

Recent works have explored the behavior of LLMs in competitive settings, noting the importance of strategic planning (Toy et al., 2024) and understanding the intentions of competitors (Xu et al., 2024b; Lai et al., 2022; Wang et al., 2023). While previous works have relied on the intrinsic capabilities of the underlying LLM to accomplish these functions, we propose to codify them with additions to the cognitive architecture described above.

We enable generative agents to develop long-term strategies, to contemplate their performance throughout a game, and to adapt strategic sub-goals that improve their progress toward the end-state of the competition. Inspired by a Belief-Desire-Intention model, which encompasses planning, action, and the continuous updating of beliefs based on new information, we augment agents’ cognitive capacity by introducing a separate goal-setting module (Chen et al., 2024; Meta Fundamental AI Research Diplomacy Team (FAIR)[†] et al., 2022).

Further, LLMs are known to struggle with discretion and privacy of sensitive information (Brown

et al., 2022; Xu et al., 2024a; Hartmann et al., 2024). However, in a competitive environment, keeping information secret is often part of a successful strategy. Language models are capable of predicting intent (Zhou et al., 2023), but as observers of situations in which private information is divulged, LLMs have shown a limited capacity to build an accurate theory of mind of the parties involved, often leaking sensitive facts in inappropriate settings (Miresghallah et al., 2023). Our environment poses the challenge to generative agents of developing and maintaining nuanced theories of mind for multiple competitors across the duration of a game.

3 Architecture

We provide a general framework for building generative agents in text-based worlds; here we create a competitive, episodic environment, inspired by Survivor. Agents take in provided world information (such as game rules) as part of the context that guides their decision making as the simulation progresses. The game is implemented as a simple text-based sandbox, in which any number of locations may be created. The generative agents described here populate these locations, perceive their local environments, and use this information to interact with their surroundings, including other agents (Figure 1).

3.1 Agents

Within this framework, a generative agent is defined by several key attributes: (1) persona, (2) memory, (3) goals, and (4) impressions.

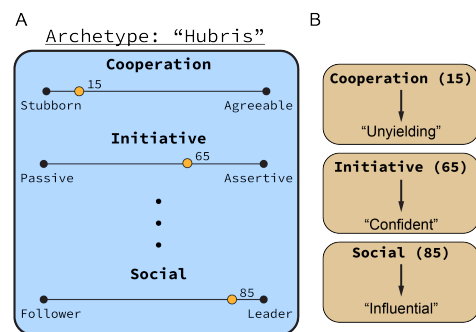


Figure 2: (A) Sampled trait scales defining the "Hubris" archetype; (B) Translation of select trait scores to adjectives.

Persona includes an agent’s core characteristics such as static facts, personality traits, and strategy type. This background information is used to

ground each agent’s “sense of self” and a summarized version of the Persona is always provided as part of the system prompt in cognitive tasks.

Personality traits are defined as a series of dichotomous scales between word anchors; for example, “passive”: “assertive” define the anchors of a scale and are scored as 0 and 100, respectively. Scores along each of these scales are defined by referencing a number of preset archetypal characters (e.g. Villain, Hubris, Teacher, etc.) or can be customized by a user (Figure 2A). A score represents how an agent *tends* to behave along a particular axis and is then translated by a utility LLM into an appropriate adjective (Figure 2B).

Memory aligns closely with the implementation of Associative Memory described previously (Park et al., 2023). It is structured as a running list of observation nodes and consists primarily of information about the content, importance, and type of the observation. Given the volition that agents have to select their own actions, it is possible that they will attempt something unsupported by the underlying game or a valid action may fail if its preconditions are not met (e.g. an agent hallucinates an item and tries to pick it up). These failures are also stored in the agent’s memory.

Goals represent an agent’s forethought towards future events and are set at the beginning of each round. To facilitate selection of strategic actions, we directed agents to use their game knowledge, relationships, reflections, and goals from the previous two rounds to set high-level goals at three priority levels (Figure L.1). In addition to goal creation, we evaluate agent progress towards goal completion and provide these scores as context in future goal development prompting. We propose this module as an alternative planning stage in a generative agent’s cognition, tailored for strategic multi-agent environments. This approach ensures a focused, yet flexible trajectory for implementing long-term in-game strategies (Appendix B.1, Figure B.1).

Impressions develops an agent’s theory of mind (Premack and Woodruff, 1978; Colwell et al., 2016) - the human ability to produce the mental state of another person, or at least a good proxy, in one’s own mind (Figure B.2A). This ability is critical for maintaining a proper flow of information in situations involving sensitive information. In a strategic game environment, similar notions of “contextual integrity” (Nissenbaum, 2004) and theory of mind are required to navigate complex social relation-

ships in which withholding information and deceit could be advantageous (Figure B.2B). We propose this module as an added cognitive function - external to the core memory - that supports an agent’s capacity to reason through complex social tasks, using their internal representation of other agents in an informed and plastic manner (Section B.2, Figure L.2).

3.2 Cognition

We also took inspiration from previous work in the design of cognitive modules (Park et al., 2023). Generative agents engage with their environment each time it is their turn in the game. They process a standard flow of information, which begins by perceiving the environment, converting new perceptions into memories, retrieving memories relevant to their personal goals and surroundings, and then taking an action (Figure 1). To this pathway we add to or modify modules for action selection, reflection, and agent dialogue.

3.2.1 Agents select their own actions

Action selection is implemented to give generative agents autonomy in their decision making processes. The retrieval system ranks memories based on their recency, importance, and cosine similarity to a query text (Park et al., 2023). To facilitate strategic actions, the agent’s persona, goals, and known interpersonal relationships are used to build the system prompt, as well as to determine memory relevance. Given a ranked memory list - prioritizing the most relevant memories in the process of context history management - an agent is asked to make an action statement (Figure L.3). They are given a list of valid action verbs, but are not restricted to selecting one of these explicitly (Table B.1). Instead they may embellish the action to form a more detailed description of their intent, which is parsed by the game engine into a valid action. If the action is successful, the game state is updated in some way; for example, if agent Mason Liu states, “Pick up the fishing pole.”, and there is a fishing pole in the same location, then the item will be placed into Mason’s inventory. Additionally, a new observation of this action and its outcome is created in the acting agent’s memory as well as in the memory of any other agents present.

3.2.2 Reflection builds knowledge over time

Reflection builds high-level inferences over memories and previous reflections, allowing agents to

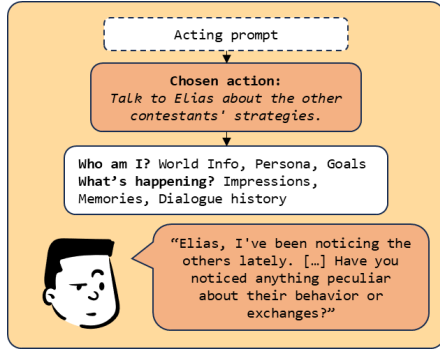


Figure 3: Architecture to generate a line of dialogue. The speaking agent is prompted to respond with an utterance given information about themselves and memories relevant to the topic of conversation.

construct more meaningful generalizations about the world, the game, and other agents. For example, players might determine that exploring the map can lead to discovering valuable items or that some alliances should be kept confidential. Agents reflect after each round, storing their inferences in memory where they are ultimately utilized by other cognitive modules.

This implementation closely follows the original architecture (Park et al., 2023), with some modifications. In our adjusted version, all agents utilize a fixed set of queries that ask them to consider memories related to their key partnerships and the efficacy of their strategy so far (Appendix B.4). Using the relevant memories, the agent is instructed to generate five high-level insights, which could be either new generalizations or updates to previous reflections. Importantly, agents can refine prior beliefs by revising old reflections as they gain more information. This prevents future memory retrievals from presenting conflicting reflections, which could confuse the agent.

3.2.3 Communication is key

Dialogue provides a system that allows agents to have one-on-one conversations with each other in every game tick. Conversations are built line by line, with each generative agent being prompted with the last line of dialogue and asked to reply or to leave the conversation. Figure 3 shows an agent initiating and then generating one line of dialogue, including the data used to build the dialogue prompt.

The dialogue system includes contextual information and retrieved memories related to the situation. If the characters mention another player,

the list of memories passed as dialogue context is updated to include the agent’s observations about about that player (Figure L.4). At the end of the dialogue, a summary is stored in both agents’ memories. While agents can only talk one-on-one, if other agents are present in the same location, they also store the dialogue as a memory.

4 Environment Engine

To enable multi-agent simulations in a text-based game, we extended the `text_adventure_games`² package. This package provides a framework for text-based adventure games, including defining locations, characters, items, and their interactions. While other established engines could have been extended (Côté et al., 2018; Urbanek et al., 2019), this one allowed us to customize the game setting in which agents played. The game engine has two key components: (1) the Game and (2) the Parser. Information about the Parser can be found in Appendix D.1.

4.1 A flexible sandbox for text-based games

The base implementation of the Game class allows for flexible implementation of any number of text-based games. It is easily adaptable by overriding the game loop and the termination-state methods. For the experiments described here, we randomly permute the agents at the start of each tick, so the order in which they take their actions changes. We also introduce functionality specific to a Survivor-like game that enables generative agents to vote for each other and exile one of their fellow contestants to the jury at the end of each round. Once two agents remain in the game, this panel of agents selects the winner using their knowledge of the finalists that they collected prior to being exiled (Figure D.1).

We describe our experimental methodology used for the simulations analyzed in Section 5 in Appendix J and an example of a round of gameplay can be seen in Appendix K.1.

5 Results

To measure the impact of the cognitive extensions proposed in this paper, we tested agent performance in two game settings: a Survivor-like voting-based game and a simplified search game with the objective of finding a hidden idol.

²<https://interactive-fiction-class.org/>

Symbol	Summary	Architecture	
		Goals	Impressions
<i>A</i>	Baseline	✗	✗
<i>B</i>	Goals	✓	✗
<i>C</i>	Impressions	✗	✓
<i>D</i>	Full	✓	✓

Table 1: Strategic cognitive architectures tested in ablation experiments.

5.1 Architectural Ablation

We first set up an environment in which agents with varied cognitive architectures attempted to survive as the "last agent standing" and we measured how architecture influenced game performance, actions taken, and goal completion.

To investigate the effect that our strategic cognitive modules, **goals** and **impressions**, have on agent performance, we randomly distributed the architectures across eight generative agents such that every cognitive set-up was assigned to two agents (Table 1). We define the full strategic architecture in this paper and tested this against three ablations, which reduced the number of cognitive modules available to a given agent.

Using the inherent metrics of the game environment, we tracked the performance of agents over a series of trials. With this fixed cast of generative agents, we measured **placement**, or "exit position," which is the round in which a player is eliminated (or wins). If eliminated in the first round of a game with n characters, a generative agent achieves placement n , while an agent that wins a game achieves a placement of 1. Therefore, agents with lower placement scores performed better in the game. Across 23 simulated games, we found that each of the architectures in Table 1 performed similarly in terms of exit position (Figure 4).

Despite these results, we did find qualitative data suggesting real differences in the cognitive capacity of agents across modules (Section C.1). We hypothesized that in a non-deterministic game like Survivor, in which the winner is selected based on numerous factors, many of which are out of the control of an individual, it is possible that the "winning recipe" is more complex than can be elicited by our Impressions and Goals modules.

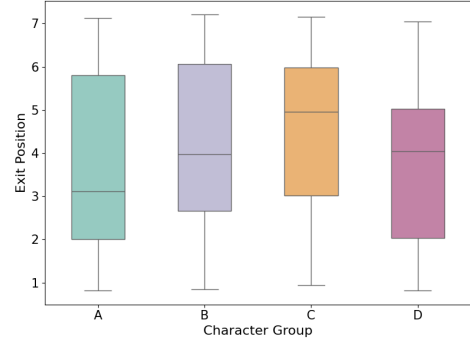


Figure 4: Placement of agents who survived beyond the first round, which tends to be a near-random vote ($N=23$, ANOVA: $F=0.83$, $p=0.48$). Note that lower placement is better; placement 1 means an agent won the game.

5.2 Personas influence behavior

While varying the architecture powering generative agents did not affect in-game performance, we continued our exploration by pitting two polarized agent archetypes against one another. To accomplish this, we designed "Charismatic" and "Aloof" archetypes (Figure E.1); social strategy is paramount in Survivor, so we tested the hypothesis that an outgoing, confident player should outperform one who shies away from social interaction. Across 17 simulations, with each persona assigned to three agents, the charismatic players won 82% of the games, showing a significant improvement over random chance (50%) (Figure 5).

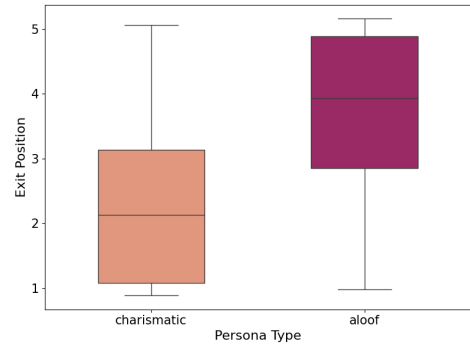


Figure 5: Distribution of agent placements for charismatic ($\mu=2.8$) and aloof ($\mu=4.2$) archetypes in a 3 versus 3 starting configuration (Comparison of means via Tukey's HSD ($N=17$): $p=0.0001$). All characters used architecture D: goals + impressions. Lower exit position is better.

Attempting to find the limit to which agent persona affects in-game outcomes, we set up a second experiment favoring the aloof archetype (4) to the charismatic (2). Here, the charismatic group won 7 times across 21 simulations and thus matched the

expected win rate. We hypothesized that the impact of the charismatic archetype was diminished when outnumbered by aloof agents, who are less susceptible to these players' charm and are more inclined to traverse the map, often moving away from conversational environments (Figure E.2).

So while we see similar performance across the agent architectures, we found a robust effect elicited by agent personas. The quantifiable impact persona had on the outcome of our game suggests that the underlying LLM is acting with a strong concept of "who" the agent is that it is embodying. This opens the door to significant customization of generative agents across a wide range of tasks, including in games in which a character controlled by a generative agent could be informed by a nuanced backstory.

5.3 Idol search game: search efficiency

We considered the possibility that a game like Survivor, in which patterns in human winners are difficult to find, may be too nuanced of a testing ground for our strategic modules. Thus, we constructed a simplified game with the objective of finding a hidden idol in the jungle. This game requires individual performance and thus we hypothesized that agents with goals would out-perform those without them. Simulations were conducted with pairs of agents of the same architecture and their search efficiency was measured across 29 rounds each. Indeed, agents who possessed the **goals** module performed marginally better than those without (Figure 6).

Architecture B (base + goals) performed significantly better at this task than the other configurations. In a non-social game with a timed search objective, planning is a key capability and likely more important than developing impressions of a teammate; moreover, the omission of the impressions module does not preclude agent's from developing shared strategies.

Evidence of this difference in game performance was also supported by the agents' ability to act upon environmental information. We placed a clue near the game's starting position that described how to properly search for the hidden idol. When agents located the clue, the time from discovery to their first properly constructed idol search was marginally lower for agents with our strategic modules than for agents with base cognition (Figure F.1).

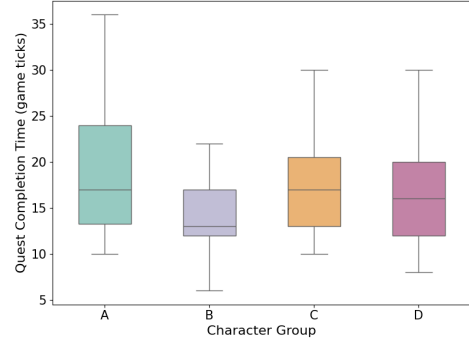


Figure 6: Search completion time by architecture (comparison of means via Tukey's HSD: $p_{A,B}=0.04$, others: NS). All characters possessed an "Explorer" archetype.

5.4 Strategic agents diversify their actions

While the game performance of these agents captures an overview of their strategic capabilities, it fails to describe the way in which agents with different architectures approached their decisions in the game. To address this gap, we measured metrics related to the agents' action selection in the traditional Survivor game. Action statements were classified into three groups: talking, moving, and others (e.g. "pick up..."). Across 23 simulations, we observed a strong preference for conversation between agents (72.6% of all recorded actions). However, agents backed by the full strategic architecture (D) demonstrated a greater diversity in their selections, opting to interact with their environment or travel in addition to initiating discussion. So, while they do not socialize to the same extent as agents with ablated architectures, they show an ability to interact with their environments and to explore alternative ways of achieving their goals (Figure 7, G.1, G.2).

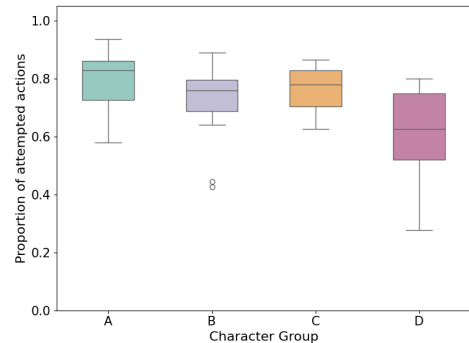


Figure 7: "Talk to" action distribution by architecture. We notice that characters with full architecture explore and interact with the environment more. Comparison of means via Tukey's HSD (N=23): $p_{A,D}=0.00$, $p_{B,D}=0.004$, $p_{C,D}=0.001$.

Similarly, in simulations with polarized personas (Section 5.2), we saw an alignment of personality type with action selection. Charismatic archetypes chose to engage in conversation 3 to 4 times as often as their aloof counterparts. Conversely, the aloof players searched and traversed the map nearly twice as frequently. These results demonstrate that the dispositions encoded in Personas are incorporated into agent decision-making (Figure G.3).

6 Conclusions

We investigate the impact that new strategic cognitive modules have on agent performance in several games inspired by Survivor. Notably, we observe that generative agents with the ability to create prioritized goals show improved efficiency in a search task. These agents are quicker to integrate new information related to the game’s objective into their decision-making, which translates to a significant difference in search efficiency. We also demonstrate cohesion between agent Personas and decision-making, particularly in the distribution of actions taken. The effect of our Persona module on agents translates to differences in game performance; agents designed to be charismatic were far more likely to win our social-competitive game than those with aloof personalities. In our "last-agent-standing" simulations, we do not see proof of a performance difference between agents with different architectures (Section 5.1); however, we do find qualitative evidence that our strategic agents behaved differently in this context (Section C.1). Thus, we hypothesize that the complex nature of the game, in which the actions of other agents affect individual outcomes, might occlude differences in agent performance.

6.1 Utility of strategic generative agents

Generative agents engineered for strategy development have vast potential to enable positive applications in the gaming world, including: (1) *Enhanced Strategic Decision-Making*. Goal-driven agents possessing theory of mind could be applied in settings that require long-term strategy development such as simulated versions of real-world negotiations or dynamic experiences in traditional games with NPCs. The latter example is notable because it could allow for personalized or novel experiences for human players, improving the replay-ability of games. (2) *Diversity in Agent Representation*. This framework can promote more representations in AI

agents by designing diverse personas, which can be beneficial for testing theories in social sciences or economics. It also could improve the access that human players from minority backgrounds have to see themselves represented in NPCs in the gaming industry (Manning et al., 2024).

6.2 Future Work

In this work, we tailor three new modules to support their strategic performance of generative agents in competitive games: goals, impressions, and personas. Some aspects of this architecture could be improved to achieve better results.

On one hand, future work could fine-tune prompts and the context made available to agents. One disadvantage of this architecture is the verbosity of the prompts that the LLM must consider when deciding what action to take. Improving memory retrieval to efficiently collect only the most relevant observations and honing the extra prompt information (e.g. personal details or goals) would make the context more concrete and informative to the underlying LLM.

Secondly, information between games could be used to improve agent cognition. Following the previous frameworks, agents could reflect upon past games to develop a generalized understanding of how to succeed in the game (Majumder et al., 2023). In future work on this topic, reinforcement learning techniques could be implemented between games to improve agent performance.

There is also the opportunity to explore better testing grounds for these agents. A game that could be good for experimentation is Sid Meier’s Civilization, where agents would be tested on their ability to explore, strategize, and socialize (Qi et al., 2024; Branavan et al., 2012). Civilization is especially useful because it involves a social aspect when interacting with other players, and the relationship between skill and performance is clearer than in a game like Survivor. Additionally, another informative game setting could be Mafia (Kim and Kim, 2023), which requires agents to lie and deceive in order to perform well, allowing for a deeper exploration of how this work’s Impressions module influences the ability of generative agents to identify public from private information.

7 Limitations

While this initial implementation uses Survivor as inspiration for its structure, we believe that

these agents can succeed in any kind of similarly-structured social-competitive game. However, it is difficult to evaluate the intrinsic strategic ability of the generative agents because of cues in the game that elicit ideas from the game show. It is likely that the LLM used here, GPT-4, was exposed to Survivor-related content in its training data given the extensive presence of blogs and Reddit threads dedicated to the analysis or summary of past seasons. Though "Survivor" was not mentioned explicitly, the inclusion of words such as "tribe", "camp", "voting out", "immunity idol", and "island" likely caused these portions of the training data to influence the token probabilities of generative agents, who then end up quoting phrases from the show like, "outwit, outplay, outlast" (some examples of this can be seen in Appendix K.2). This raises the possibility that strategies used by agents were learned from Survivor-related content and thus fail to properly test the innate ability of generative agents to develop solutions to the game or to learn from the environment.

Another limitation we encountered during experimentation is that generative agents have a proclivity to act agreeably, especially during conversation. When instructed to disagree, generative agents often did so within the first line of dialogue. However, as the conversation evolved, they often found compromises with their counterparts and were more likely to agree to alliances or general sentiments. This bias toward agreement was then translated into memories, thus impacting the perception of others and causing agents to change their minds frequently about allies and other contestants. In a game as complex as Survivor, agents must be more critical of their surroundings and competitors, adapting their strategies accordingly.

Finally, a non-trivial reality of this system is the cost (both monetarily and in time) associated with using an API-based LLM such as GPT-4 as we did here. The architecture described here uses thousands of requests containing millions of tokens to the OpenAI API during each game (Table I.1). Thus, a full game of the basic Survivor-like game typically costs over \$300 and takes up to 3.5 hours to run (Section I.2). As models continue to develop, they will become cheaper and run with lower latency, but it remains a worthwhile endeavor to improve the efficiency of the architecture that powers generative agents.

7.1 Ethical Considerations

While competitive generative AI agents enable positive applications, as discussed above, they also pose ethical risks that require mitigation. These concerns include: (1) *Privacy Concerns and Information Misuse*. In competitive environments, the strategic use of sensitive information could lead to scenarios where personal data is misused, leading to potential breaches of confidentiality. (2) *Over-Reliance on Technology*. There is a risk that reliance on generative agents for strategy development in competitive scenarios could diminish human strategic thinking skills. (3) *Inherent or Acquired Biases from the Underlying Generative AI Systems*. If the training data for the systems on which these agents are built includes biased or non-inclusive content, the agents themselves might reinforce those stereotypes or biased/unfair representations.

To address these ethical concerns, designers should implement mitigation strategies, such as: (1) *Ethical Guidelines and Monitoring*. Develop and enforce strict ethical guidelines to govern the behavior of generative agents, particularly in how they handle sensitive information and interact with other agents. (2) *Transparency and User Awareness*. Ensure that all participants in simulations involving generative agents are fully aware of the artificial nature of their interactions. (3) *Continuous Evaluation and Adjustment*. Regularly evaluate the impact of generative agents in competitive environments and adjust their programming to prevent any negative behaviors that could lead to unethical outcomes.

References

- Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2021. [How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds](#). ArXiv:2010.00685 [cs].
- A. Bernstein, T. Arbuckle, M. De V. Roberts, and M. A. Belsky. 1958. [A chess playing program for the IBM 704](#). In *Proceedings of the May 6-8, 1958, western joint computer conference: contrasts in computers on XX - IRE-ACM-AIEE '58 (Western)*, page 157, Los Angeles, California. ACM Press.
- S. R. K. Branavan, David Silver, and Regina Barzilay. 2012. [Learning to Win by Reading Manuals in a Monte-Carlo Framework](#). *Journal of Artificial Intelligence Research*, 43:661–704. ArXiv:1401.5390 [cs].

- Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. [What Does it Mean for a Language Model to Preserve Privacy?](#) ArXiv:2202.05520 [cs, stat].
- Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. [Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning](#). Publisher: [object Object] Version Number: 3.
- Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2024. [Put Your Money Where Your Mouth Is: Evaluating Strategic Planning and Execution of LLM Agents in an Auction Arena](#). ArXiv:2310.05746 [cs].
- Malinda J. Colwell, Kimberly Corson, Anuradha Sastry, and Holly Wright. 2016. [Secret keepers: children’s theory of mind and their conception of secrecy](#). *Early Child Development and Care*, 186(3):369–381.
- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. [TextWorld: A Learning Environment for Text-based Games](#). Version Number: 2.
- Florian Hartmann, Duc-Hieu Tran, Peter Kairouz, Victor Cărbune, and Blaise Agüera y Arcas. 2024. [Can LLMs get help from other LLMs without revealing private information?](#) ArXiv:2404.01041 [cs].
- Munyeong Kim and Sungsu Kim. 2023. [Generative AI in Mafia-like Game Simulation](#). ArXiv:2309.11672 [cs].
- Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. [Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games](#). ArXiv:2212.08279 [cs].
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanqing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. 2024. [Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security](#). ArXiv:2401.05459 [cs].
- Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. 2023. [CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization](#). ArXiv:2310.10134 [cs].
- Benjamin S. Manning, Kehang Zhu, and John J. Horton. 2024. [Automated Social Science: Language Models as Scientist and Subjects](#). ArXiv:2404.11794 [econ, q-fin].
- Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. [Human-level play in the game of Diplomacy by combining language models with strategic reasoning](#). *Science*, 378(6624):1067–1074.
- Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. [Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory](#). ArXiv:2310.17884 [cs].
- Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *University of Washington Law Review*, 79.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative Agents: Interactive Simulacra of Human Behavior](#). ArXiv:2304.03442 [cs].
- David Premack and Guy Woodruff. 1978. [Does the chimpanzee have a theory of mind?](#) *Behavioral and Brain Sciences*, 1(4):515–526.
- Siyuan Qi, Shuo Chen, Yexin Li, Xiangyu Kong, Junqi Wang, Bangcheng Yang, Pring Wong, Yifan Zhong, Xiaoyuan Zhang, Zhaowei Zhang, Nian Liu, Wei Wang, Yaodong Yang, and Song-Chun Zhu. 2024. [CivRealm: A Learning and Reasoning Odyssey in Civilization for Decision-Making Agents](#). ArXiv:2401.10568 [cs].
- David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. [Mastering the game of Go with deep neural networks and tree search](#). *Nature*, 529(7587):484–489.
- Jason Toy, Josh MacAdam, and Phil Tabor. 2024. [Metacognition is all you need? Using Introspection in Generative Agents to Improve Goal-directed Behavior](#). ArXiv:2401.10910 [cs, q-bio].
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. [Learning to Speak and Act in a Fantasy Text Adventure Game](#). Version Number: 1.
- Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. [ScienceWorld: Is your Agent Smarter than a 5th Grader?](#) ArXiv:2203.07540 [cs].

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. [Avalon’s Game of Thoughts: Battle Against Deception through Recursive Contemplation](#). ArXiv:2310.01320 [cs].

At the beginning of each round, agents are asked to generate goals. To do so, they take in: (1) the universal world information, (2) their personal summary, (3) reflections from the prior round, (4) their goals from the last two rounds (should those goals

The Impressions module leverages the concept of Theory of Mind, which is the human ability to conceptualize another person’s thoughts or understanding of an environment (Colwell et al., 2016). This is particularly important in a game setting in which information is the most valuable currency; thus, we ask generative agents to consider their relationships with others, including what information they ought to keep discrete (Mireshghallah et al., 2023). In Impressions, we focus on several key questions that encourage generative agents to consider how other agents are perceiving the environment (Figure B.2A, B). Generative agents create impressions of a target when they encounter someone they have never met before or if their existing impression is sufficiently outdated. In this implementation,

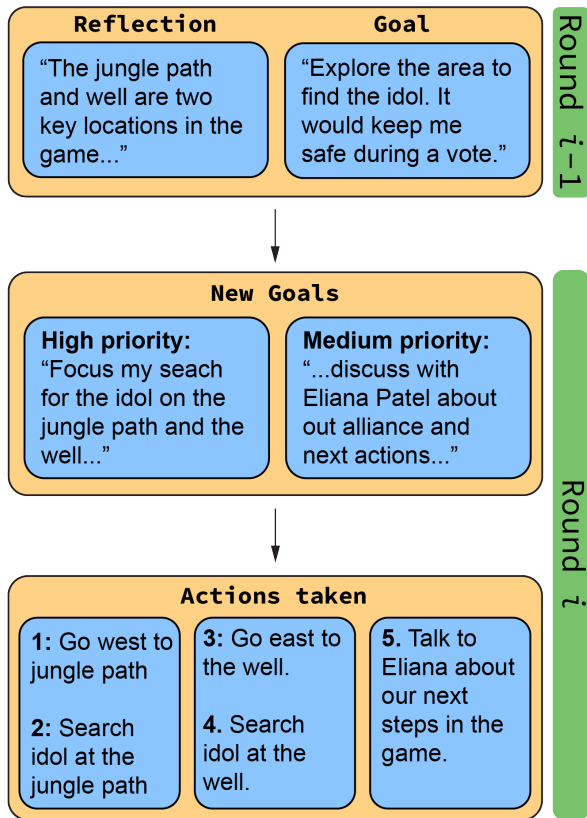


Figure B.1: Flow of information in an agent's goal formation and the effect of these goals on their selected actions.

agents will update an old impression at least once per round. This ensures that they build up a pool of new memories about a character before creating or updating their impression. The prompting for this module is provided in Figure L.2.

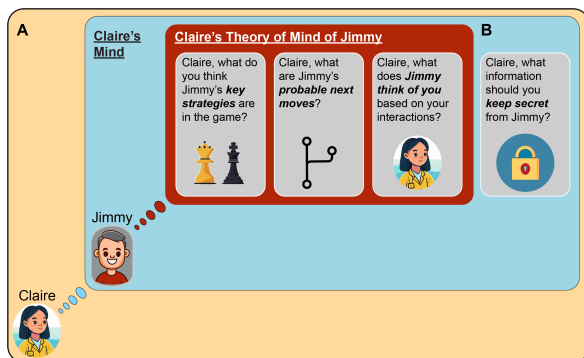


Figure B.2: Illustration of impression generation. (A) Generative agent, Claire, has met Jimmy and must consider his view upon the world, the game, and Claire herself. (B) We ask agents to develop contextual integrity of sensitive information they possess.

Action Verb	Description
Describe	Observe one's surroundings
Go	Travel in some direction
Get	Retrieve an object
Drop	Discard an object
Inventory	List items in one's possession
Examine	Look closely at an object
Give	Hand an item to another agent
Eat/Drink	Consume food/liquid
Light	Ignite something flammable
Catch fish	Cast a line into water
Talk	Converse with someone
Search idol	Attempt to locate the hidden idol
Read clue	Access the information on the clue

Table B.1: The action space available to generative agents.

B.3 Action Space

Here we show the available action verbs available to agents. The game engine's Parser does not require exact matching of these verbs, but instead uses an auxiliary LLM to perform intent disambiguation of an agent's action statement to one of the verbs shown here. If there is no good match, then the action will fail, the agent will be informed (via a new memory of the failed action), and prompted again to perform a new action (Table B.1).

B.4 Reflection Query Questions

The following questions are used as queries in the reflection procedure that agents call at the end of a round. Each question is used in a memory retrieval process. Agents are then asked to reflect upon the unique set of memories that are pulled using these queries:

- What were my key partnerships in this round?
- How have my recent decisions influenced my alliances, and what changes should I make to strengthen or repair relationships with other players?
- How effectively did I handle major threats this round, and what alliances or betrayals should I anticipate for the next round based on these actions?

- Which strategies have been effective so far, and how should I adjust my overall game plan to align with the evolving dynamics of the game?

C Results addenda

C.1 Qualitative architectural differences

We observe qualitative differences in gameplay among characters with different architectures. In Architecture B, characters focus on a single target, influenced by personal experience, as shown in Figure C.1. This approach results in a narrow, experience-based strategy. Conversely, Architecture D incorporates a theory of mind, enabling characters to strategize not only based on personal experience but also by anticipating and countering the strategies of their target, as illustrated in Figure C.1. This dual-layered strategy enhances their adaptive capabilities. In contrast, characters in Architecture A and C, lacking defined goals, fail to develop coherent long-term strategies. They inconsistently vote out various players, demonstrating the importance of goal-oriented models (as seen in Architectures B and D) for maintaining effective strategies over time.

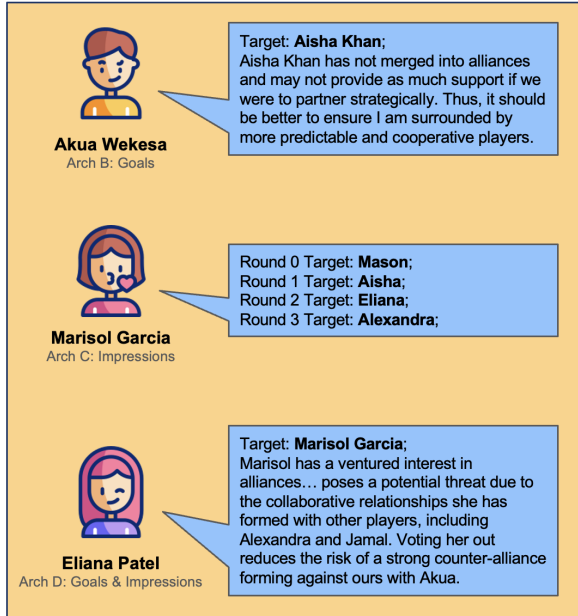


Figure C.1: Qualitative Architecture Comparison

D Game Engine: Parser

D.1 Game parser

The **Parser** powers the translation from open-ended natural language statements into valid (or potentially invalid) actions that affect game state. We

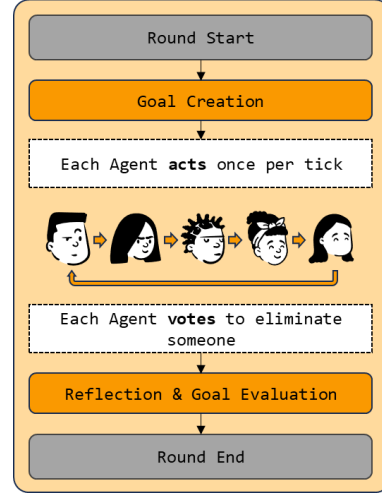


Figure D.1: During each round, agents first set goals. Then, during each tick, each character takes one action. On the round's final tick, characters vote to eliminate someone, reflect on their memories generated during the round, and evaluate their goals.

instruct another language model instance to map a command phrase given by a generative agent to one of the valid actions in the game. This includes subroutines in which the LLM must identify valid agents, items, or directions of travel based on the content of the command. If a valid action is found from the command, the parser checks that the actions preconditions have been properly met by the underlying game state, and then executes the action to update the environment. Agents receive feedback in the form of memories which indicate the outcome of their actions; this occurs for both successful and unsuccessful actions, meaning agents can learn what types of actions do not work.

D.2 Round structure

A round of gameplay is structured as seen in Figure D.1.

E Imbalanced persona experimentation

Sample traits in extreme personas. All "aloof" traits were set to 0 with their corresponding low anchor word. All "charisma" traits were set to 100 with their corresponding high anchor word. These adjectives were then used in creating the agent persona summary which is fed as context to the agent LLM.

The dominance of the charismatic players was diminished when they were outnumbered by aloof players; we used a starting configuration of 2 "charismatic" and 4 "aloof". Charismatic agents won $\frac{1}{3}$ of the games as expected by random chance

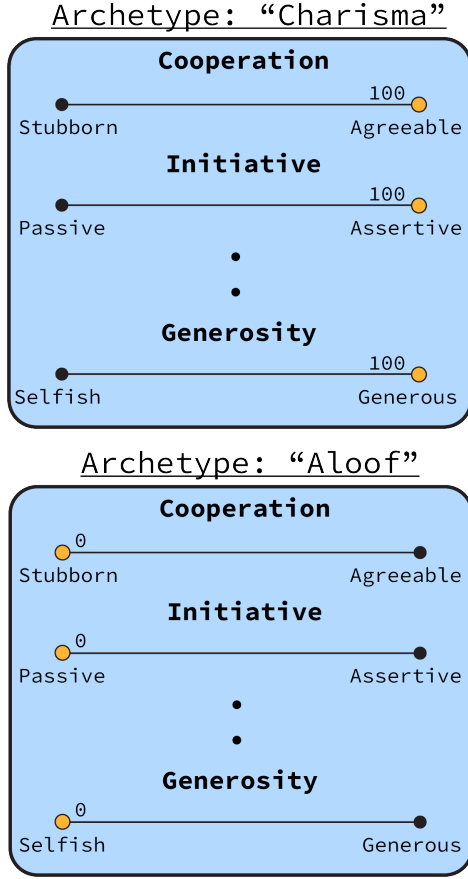


Figure E.1: Sample trait cards

(Figure E.2).

F Agent search efficiency

In games in which agents discovered the clue to the hidden immunity idol, agents with our cognitive modules displayed marginal improvement in the speed at which this information was incorporated into strategy via concrete actions (Figure F.1).

G Agent Action Selection

The results of the distribution of moving actions and other actions (fishing, searching, picking up objects, etc) can be seen in Figures G.1 and G.2.

We see a similar pattern in the action selection of agents possessing charismatic and aloof personas (Section 5.2, Figure E.1). Charismatic agents have a strong propensity to engage in conversation while aloof agents prefer to move around the map, avoiding confrontation as much as possible (Figure G.3).

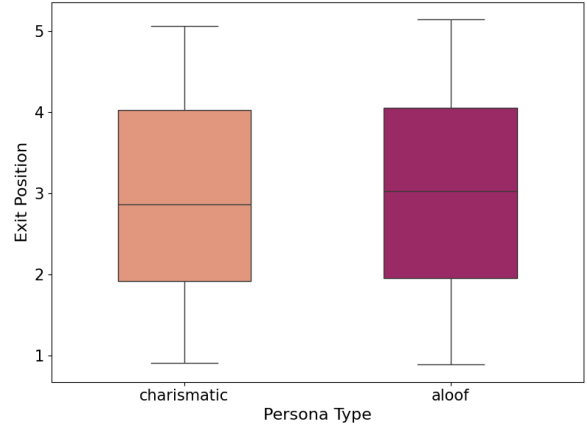


Figure E.2: Placement in the traditional Survivor game when experimenting with imbalanced personas starting configuration with 4 Aloof and 2 Charismatic characters. Comparison of means via Tukey's HSD ($N=21$): $p=0.65$.

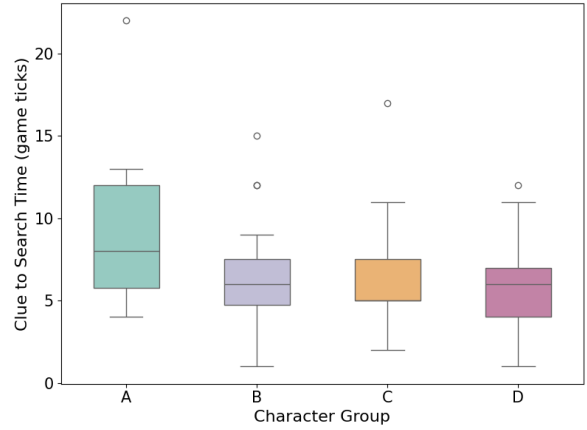


Figure F.1: Duration between clue discovery and first proper search action. Median values: A (9.0), B (6.0), C (8.0), D (6.5); One-way ANOVA of means: $F=2.45$, $p=0.06$.

H Goal Arc Case Study

H.1 Agents use goals to set priorities

We developed a system to evaluate goal completion that uses an "auditor" LLM to assess an agent's progression toward the goals that they set at the beginning of the same round. The evaluator compares an agent's goals against their reflections and actions. This evaluation is rated on a scale from 1 to 5, where 1 indicates minimal progress and 5 denotes complete achievement of the goal. We saw that agents used goals to inform their actions and a positive trend in completion score from low to high priority goals (Figure H.1A). However, it's important to note that since the goals are subjective and often vague, the auditor LLM's scoring

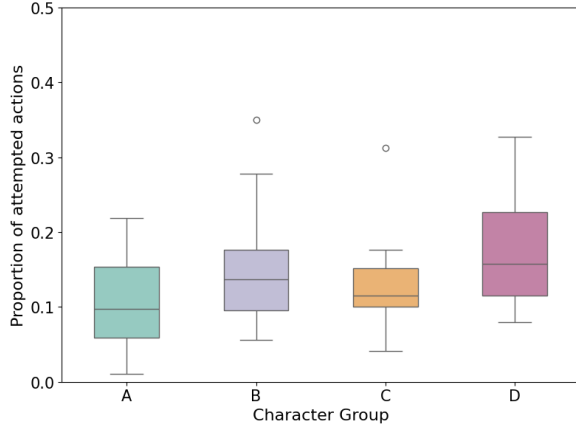


Figure G.1: Number of actions related to **travel** between locations in the game. Comparison of means via Tukey's HSD (N=23): $p_{A,D}=0.015$, others: NS.

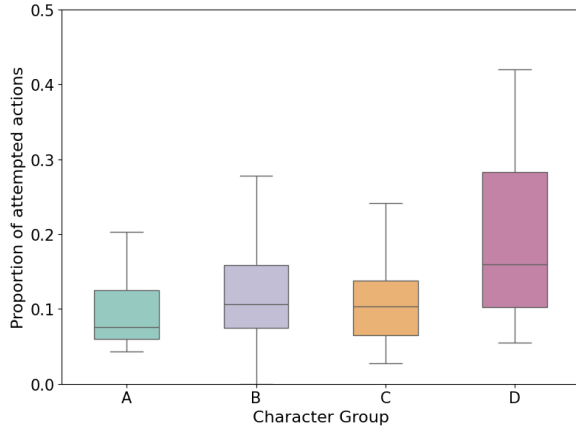


Figure G.2: "Other" actions involve searching for idols, picking up items, eating, fishing, etc. Comparison of means via Tukey's HSD (N=23): $p_{A,D}=0.00$, $p_{B,D}=0.001$, $p_{C,D}=0.003$.

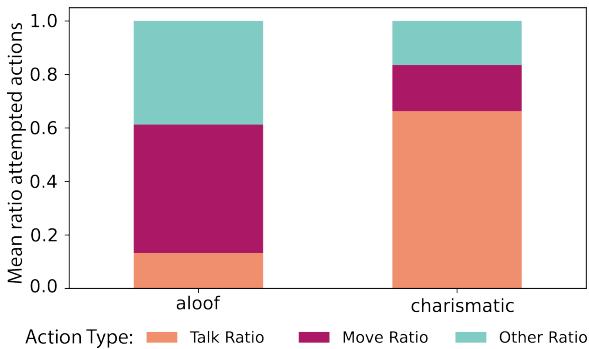


Figure G.3: Mean proportion of actions taken across simulations of the 3v3 personas experiment (Section 5.2). N=17.

can be inconsistent, unlike rule-based scoring systems in which outcomes are clearly defined and unambiguous.

We also took an anecdotal approach in assessing

the evolution of agent-generated goals by following the goal arc of an agent, Akua (architecture **B**) over the course of a game; in doing so, we highlight the strategic adaptability of generative agents. Akua's goal, "find idol", evolves into "find idol in jungle or the well" and culminates in his locating the idol in the deep jungle. This progression underscores the agents' capability to retain and refine their objectives based on prior reflections and actions to advance their strategic aims (Figure H.1B).

The dynamics of the final jury reveal another layer of strategy influenced by social interactions. Alexandra, the other finalist, has a comprehensive victory that is attributed to her proactive engagement in strategic discussions and information sharing, contrasting with Akua's narrower focus on his idol search and alliance with one character. This disparity in social engagement highlights the effectiveness of our agents' goal-oriented module and memory retrieval system.

A surprising twist occurs in later rounds when Akua, despite a strong alliance with Eliana since round 0, forms a covert alliance with Alexandra and votes out Eliana. He rationalizes this by fearing his inability to overpower Eliana in the final round. Here, we see the implementation of our theory of mind architecture, which prompts Eliana to suspect Alexandra's and Akua's motivations and attempt to counteract the betrayal by targeting Alexandra. However, she realizes this too late and is eliminated.

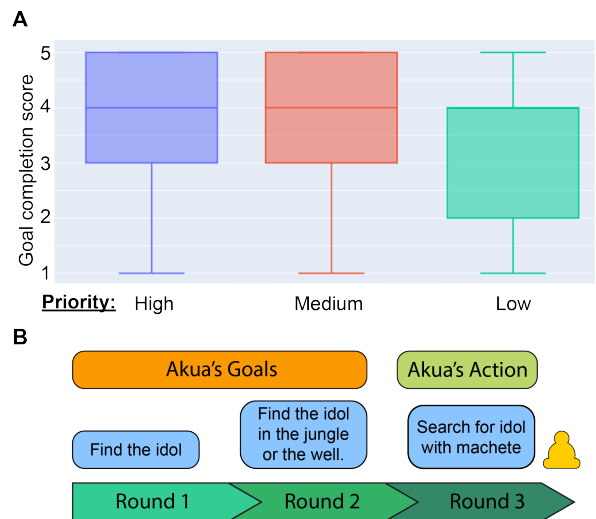


Figure H.1: (A) Goal completion scores by priority level (N = 23). (B) Akua's goal arc from rounds 1, in which he identifies the idol as important, to 3, in which he completes his goal.

I Cost estimation

I.1 Request counts

We summarize the calls to the OpenAI API and the number of tokens processed over the course of 6 games of an exploration-style game. This game was set up with a max of 11 rounds, 4 ticks per round, and consisted of 6 agents; the agents were allowed to take 3 actions per round. The 4th turn consisted of a reflection and goal-evaluation procedure (for architectures possessing goals).

Measure	Grouping	Count
Calls	character	91 ± 15.4
	game	539 ± 57
Tokens	character	101000 ± 27000
	game	560800 ± 171000

Table I.1: Calls and tokens passed to OpenAI’s API across 6 simulations. Counts are normalized by the number of rounds that the game took to complete. For these simulations, the shortest game was 3 rounds and the longest took all 11 rounds.

I.2 Monetary cost estimate

Though difficult to determine exactly, we estimate that the 23 simulations (plus others that were not included in this analysis due to errors in code or information logging) of the traditional Survivor game (Section 4.1) cost \$16090 with an average cost per request of \$0.037. The average cost per game was approximately \$350.

J Methods

J.1 Architectural Ablation Setup

For the experiments comparing in-game performance of the different architectures described in Table 1, we ran 23 games with the following characteristics:

- The goal of the game was to survive until the end. At the end of each round one player is voted off, and at the end of the game, the first 6 eliminated players vote for a winner between two finalists.
- 8 characters played with 8 distinct personas. Each character was assigned an architecture at random at the start of the game such that each architecture was assigned to 2 agents.
- The game had 4 starting locations (camp, beach, ocean, jungle path) with 2 characters each, which were also randomly assigned.
- Each game ran for 7 rounds of 5 ticks each (i.e. 5 actions per round), with a 6th tick used for a voting phase at the end of each round.
- If any vote ended up being a tie, one of the tied players would be selected uniformly at random.

In addition to measuring game performance, we also collected the actions attempted by each character across our simulations. We wanted to see if different architectures interact with the environment differently. The "talk" action would reflect the agent’s ability to socialize, "movement" actions would show how good the agent is at exploring, and "other" actions (such as picking up items and searching for idols) would show that the agent wants to interact with what it’s perceiving outside of other characters.

J.2 Persona Experiments Setup

For these experiments, the goal of the game remained the same as in Sec. J.1, but all characters were assigned the same full architecture (D) (Table 1), and we tested for performance differences between characters with polarized personas. The games had 6 players and lasted for 5 rounds. 17 games were run with 3 aloof vs 3 charismatic characters, and 21 games were run with 4 aloof vs 2 charismatic characters (Section E). Examples of the polarized personas used for these experiments can be found in the assets directory of the accompanying repository (Footnote 1).

J.3 Exploration Experiments Setup

To measure the impact of different architectures on exploration, games were run with a new objective. The goal of these games was to find a hidden idol, which was made more difficult by requiring a tool to do so successfully. The characteristics in these games were:

- 2 characters per game, both with similar "Explorer" personas and traits (but different static facts). In a given game, both characters had the same architecture and were allowed to collaborate.
- 1 starting location at the camp.

- The idol was hidden in the jungle, requiring a machete to find.
- A clue was placed at the cliffs, which indicated to characters how to find the idol.
- The game would end when either character found the idol.

The number of games run for these experiments can be seen in Table J.1.

Architecture	Number of games
A	26
B	25
C	27
D	32

Table J.1: Number of games run for the exploration experiments by architecture.

K Gameplay Excerpts

K.1 Round of gameplay

We are in the 6th round of a game with 8 characters. There are 3 characters left: Akua, Marisol, and Jamal.

Akua and Jamal have been working together and talking primarily with each other. They both started at the jungle path and have been moving together. Marisol had been talking to Madeline and Mason, who she backstabbed at the latest round by voting against him. She now moved to the jungle path and joined the other two.

Since they are all at the same location, these conversations are all within earshot of each other and they are stored in all their memories. Additionally, we are only printing new impressions at the beginning of each action, so while they might be perceiving more items and characters than described here since they perceived those before, it is not repeated again.

Here we see that even if Marisol tries to convince the others to vote for each other, and they agree, at the end Akua and Jamal choose to stay loyal and vote out Marisol.

ROUND: 5.0

It is Akua Wekesa's turn

Akua Wekesa chose to take action: **Talk to Marisol Garcia about potential alliance strategies.**

Akua Wekesa said: *Marisol, we've both come a long way in this game, and I believe that together we could go even further, strategically and methodically navigating our way to the finals. Our alliance could be beneficial, taking on others hand-in-hand. Of course, the essence remains mutual respect, trust, and efficiency. In your opinion, what would be the keys for us to strengthen our alliance and work effectively together?*

Marisol Garcia said: *Akua, I agree wholeheartedly. Look, we've done a great job so far individually, but I believe together we can make it to the very end. Communication, trust and vigilance are vital. I also believe that keeping our alliance low-profile can help to prevent arousing suspicion from the others. We need to share information immediately - any changes in dynamics, any discoveries related to hidden immunity idols. Of course, we also need to ensure our votes align come voting time. If we adhere to these principles, I believe we can both succeed.*

Akua Wekesa said: *Marisol, I couldn't agree more. Communication, trust, and vigilance are indeed the foundation of our alliance. Likewise, maintaining a low-profile and being subtle about our alliance ensures we don't attract unwanted attention. Also, ensuring our voting decisions align is crucial in securing our place in the upcoming rounds. Let's be proactive about sharing any significant discoveries or changes in dynamics. We're in this together, after all. Our combined efforts will undoubtedly strengthen our position in the game. Here's to our alliance's success.*

Marisol Garcia said: *Akua, I'm glad we see eye to eye on this. Maintaining open communication and strategic agility is key to us moving forward. We need to keep an eye on the others, monitor their actions, and remain alert to any changes. Any crucial information we find, be it about an immunity idol or other alliances, we share immediately with each other. We can adapt our strategies as necessary. Our mutual survival is the top priority. I'm excited for what our alliance can achieve and I believe with our cooperation, we stand a strong chance of winning. Here's to us! Let's conquer this game together.*

Akua Wekesa said: *Marisol, I completely agree. Our shared vision and cooperative ap-*

proach are our biggest assets. Let's remain observant and adaptive, given that dynamics in the game can change rapidly. Rest assured, any critical information that comes my way will be shared with you immediately. Together, we can face any shifts in the game. Let's navigate our way to the top. Indeed, here's to us conquering the game, together.

Marisol Garcia said: *Akua, you've summed it up perfectly. Maintaining our vigilance, sharing crucial information, and adapting to the game's dynamics will be our keys to a successful alliance. Alongside this, we must remain hopeful and stay focused on our goal. Here's to our strong alliance and succeeding in the game together. I feel hopeful about the possibilities in front of us. Let's take each day as it comes and make the best of the opportunities we have.*

It is Jamal Malik's turn

Jamal Malik sees: [Jamal Malik sees characters Marisol Garcia]

Jamal Malik chose to take action: **Talk to Marisol Garcia about forming an alliance.**

Jamal Malik said: *"Hey Marisol, I've been observing the way you've been playing this game, and I think our strategies align quite well. I think it would be beneficial for both of us if we formed an alliance. What do you think?"*

Marisol Garcia said: *Hi Jamal, I truly appreciate your offer and thoughtfulness. Forming an alliance would definitely offer some strategic advantages. However, let's remember to maintain open communication and mutual respect to ensure our alliance works effectively. It's crucial for us to keep each other informed about any significant changes or findings, especially if we find the immunity idol or when making voting decisions. If you're comfortable with these terms, I'm happy to agree to an alliance.*

Jamal Malik said: *"Marisol, I entirely agree with your points. Open communication and mutual respect are fundamental to a successful alliance. I commit to keeping you informed on important matters, especially regarding immunity idols or voting decisions. With these agreements in place, let's move forward as allies. Looking forward to our collaborative efforts in this game!"*

Marisol Garcia said: *Fantastic, Jamal. I be-*

lieve our alliance will be beneficial for us in the long run. Let's stick together, maintain trust, and work on our strategies collectively. I'll make sure to keep you updated with any significant details too. Looking forward to navigating the rest of the game together!

Jamal Malik said: *I believe we've set a strong foundation for our alliance, Marisol. Let's ensure to keep each other's best interests in mind as we move forward. Remember, we're stronger together. Looking forward to our shared journey in this game!*

Marisol Garcia said: *I completely agree, Jamal. And remember, we remain flexible and adaptable to the ever changing dynamics of the game. Let's keep our alliance discreet and let's be proactive and vigilant at the same time. I'm excited for our journey and I believe together we can tackle any challenges that come our way!*

Jamal Malik said: *"Absolutely, Marisol. Staying adaptable and vigilant will indeed serve us well in this game. Let's navigate these challenges side by side, discreetly and proactively. Count on me to be there every step of the way. Let's do this!"*

Marisol Garcia said: *Absolutely, Jamal! Count me in too. Now, let's focus on strategizing and exploring the game area. Perhaps we could find that hidden immunity idol. It might really turn the tide in our favor. Let's do this together, step by step.*

Jamal Malik said: *"Definitely, Marisol. Finding that hidden immunity idol could be a real game-changer. Let's split up and explore different areas, and then we can reconvene to share our findings. We're in this together, step by step."* Marisol Garcia said: *I completely agree, Jamal. Let's keep our eyes open and our alliance strong. We'll reconvene to share our findings. Good luck, and let's prove that together we can make a difference in this game.*

Jamal Malik said: *"I couldn't agree more, Marisol. Wishing you the best of luck in the exploration. Let's meet later to discuss any potential leads on the idol. Remember, we got this!"*

Marisol Garcia said: *I absolutely agree, Jamal. Meet you back here later to share our findings. Let's keep our spirits high and our focus sharp. Good luck, and remember, we're in this together!*

It is Marisol Garcia's turn

Marisol Garcia sees: [*'Marisol Garcia is at a jungle path towards the well.'*]

Marisol Garcia sees: [*'From Jungle Path Marisol Garcia could go North to Jungle'*]

Marisol Garcia sees: [*'Marisol Garcia sees characters Jamal Malik'*]

Marisol Garcia chose to take action: **Talk to Akua Wekesa about potentially aligning our votes.**

Marisol Garcia said: *Hey Akua! I hope your exploration of the jungle is going well. I wanted to continue our previous conversation about our alliance. I believe we can both benefit if we align our votes in the upcoming round. What are your thoughts on this?*

Akua Wekesa said: *Marisol, I completely agree. Our alliance has been mutually beneficial so far, which gives me confidence to continue in the same spirit. Aligning our votes could indeed provide a strategic advantage. However, it's also important to consider the potential implications. We should be careful not to arouse suspicion among the other contestants. A low profile is key if we want our alliance to be successful. What do you think?*

Marisol Garcia said: *Absolutely, Akua! Navigating these challenges quietly and strategically is, indeed, vital for our success. Let's make sure we're discreet with our decision-making. Maintaining trust and clear communication within our alliance is key. As long as we keep each other updated of significant changes or findings, we should have a significant advantage in the rounds to come. I'm glad we're on the same page.*

Akua Wekesa said: *Indeed Marisol, maintaining our bond and trustworthiness is pertinent to our progress in this game. We should also stay proactive in our idol search as that can solidify our position in the game. I'm earnest in my commitment to our alliance and look forward to our collaboration. We'll tread carefully and make sure our steps are calculated so that our alliance can thrive further.*

Marisol Garcia said: *Yes, Akua, we are definitely in sync with our strategic planning. I appreciate your commitment to our alliance. Working jointly towards finding the idol will not only aid our advancement in the game, but also cement our bond. I too am committed to*

us succeeding together. Let's continue to keep each other informed and work as a united front. I am confident that our combined strength will see us through. It was great to touch base, Akua. Let's remember the essence of our alliance — communication, trust, and delivering on our promises — as we move forward.

Akua Wekesa said: *Marisol, I couldn't have put it better myself. Our combined strength indeed, along with the shared determination will guide us through this game. I appreciate your acknowledgment of the essence of our alliance. Let's continue to execute our plan meticulously and succeed together. Just remember, we are stronger as a unit. Let's touch base soon to keep our strategy updated. Until then, let's keep exploring, stay vigilant, and may good fortune favor us.*

Marisol Garcia said: *You know, Akua, I couldn't agree more. This journey together has been such a rewarding experience, and I can't wait to see how far our alliance will take us. Our communication and mutual trust have truly been the pillars of our bond, which I value deeply. I'm looking forward to keeping our strategy up-to-date as we continue on this path. Let's not forget to maintain a low profile and share any important updates promptly. Until then, happy exploring and stay safe!*

Akua Wekesa said: *Marisol, I agree wholeheartedly with your sentiments. Our bond has been one of the key pillars of strength in this journey. Let's remain diligent in our exploration, stay safe, and stay tenacious. I'm optimistic that our thoughtful strides will bring fortune to this alliance. Until our next discussive touchpoint, all the best in your explorations!*

Marisol Garcia said: *Thank you, Akua. Please take care and continue to keep a sharp eye during your exploration. I am sure we will have more updates to share soon. I absolutely agree with you about our diligent and strategic approach leading us towards success in this game. Until we touch base next, stay safe and good luck with your search!*

Akua Wekesa said: *Thank you, Marisol. We've got a sound strategy in place. Let's stay steadfast, continue our individual explorations and aim for the effectively shared outcomes in this competition. Until our next interaction, best of luck on your journey, and remember to stay safe out there!*

ROUND: 5.1

It is Marisol Garcia's turn

Marisol Garcia chose to take action: **Talk to Akua Wekesa about the importance of finding the hidden immunity idol.**

Marisol Garcia said: *Akua, you know how crucial it is for us to find the hidden immunity idol. It could be the difference between staying in this game or being voted off. Let's divide and conquer; we'll cover more ground that way. And remember, if someone finds it first, our strategy might need a quick shift. Communication is key here, okay?*

Akua Wekesa said: *Certainly, Marisol. The concept of divide and conquer would indeed be the most efficient way to find the immunity idol. We must be discreet and meticulous as we search. I appreciate your reminder of the importance of communication. If either of us finds the idol, let's discuss how to best leverage it for our shared success. Let's tread carefully and stay hopeful, my friend.*

Marisol Garcia said: *Absolutely, Akua. Discretion is key here, we don't want to tip off Jamal about our search for the idol. Let's meet later to touch base on our progress, and remember, any information we get could be vital to our success. Let's keep that optimism alive.*

Akua Wekesa said: *Absolutely right, Marisol. We'll proceed carefully and update each other about any tip or sign of the idol. I trust we are on the same page and will move wisely with this search. Let's remain focused and determined. No one said this would be easy, but with our combined efforts, victory is within reach. Let's meet later to share our progress. Good luck, Marisol.*

Marisol Garcia said: *Thank you, Akua. Let's stay positive, do our best, and hope for the best. Good luck to you too. Let's catch up later.*

It is Akua Wekesa's turn

Akua Wekesa chose to take action: **Search idol at Jungle Path.** Akua Wekesa action failed. Adding failure memory to history.

Akua Wekesa chose to take action: **Search for the idol in the jungle.** Akua Wekesa action failed. Adding failure memory to history.

Akua Wekesa chose to take action: **Go east to the well.**

It is: Jamal Malik's turn

Jamal Malik sees: [‘Jamal Malik sees characters Marisol Garcia’]

Jamal Malik chose to take action: **Talk to Marisol Garcia about potential idol locations.**

Jamal Malik said: *Marisol, I was thinking we could try searching for the idol near the outskirts of the jungle or perhaps close to the beach. We've been focusing heavily on the most obvious locations, but maybe that's not where it's hidden. We need to balance our rhetoric and strategic maneuvers whilst still giving importance to our search efforts. What are your thoughts on this? Do you have any specific areas in mind that we should explore?*

Marisol Garcia said: *Jamal, I believe your idea is sound. It's true we've been focusing a lot on the obvious spots. I think we should broaden our search and also track the less significant areas like the outskirts of the jungle and possibly the beach. Maybe, we can divide up those areas and search separately? It may increase our chances of finding the immunity idol. Does that sound like a sound strategy?*

Jamal Malik said: *Yes, Marisol, I agree with your strategy. Dividing the areas to search separately could be very effective. I suggest we remain discreet about our intentions and check in with each other periodically to discuss any potential leads. It's important we stay focused on our ultimate goal. Let's do this.*

Marisol Garcia said: *That's a solid plan, Jamal. Let's split up our search and try to cover as much ground as possible. And yes, staying low-key and checking in with each other is crucial. Let's keep our goals in sight and do our best to find that idol.*

Jamal Malik said: *Absolutely, Marisol. We have our plan in place now. Let's begin our search separately and remember, let's communicate promptly if we find anything of importance. Good luck with your search!*

ROUND: 5.2

It is Akua Wekesa's turn

Akua Wekesa sees: [‘Akua Wekesa is at the water well where you can get water for your tribe.’]

Akua Wekesa sees: [*'From Well Akua Wekesa could go West to Jungle Path'*]

Akua Wekesa sees: [*'No characters'*]

Akua Wekesa chose to take action: **Go west to Jungle Path.**

It is: Jamal Malik's turn

Jamal Malik sees: [*'Jamal Malik sees characters Akua Wekesa'*]

Jamal Malik chose to take action: **Talk to Marisol Garcia about our voting strategy.**

Jamal Malik said: *"Marisol, I'd like us to talk about our voting strategy for the upcoming session. It's important for us to ensure we get to the final round. Any ideas on who we should consider voting out this time?"*

Marisol Garcia said: *"Jamal, you're right. It's important we make the right decision now. Given the current situation, I think it would be strategic to vote for Akua. She's a strong competitor who could pose a threat to either of us reaching the final round. Remember, this is just a game and our decisions should align with our strategic plan. What are your thoughts?"*

Jamal Malik said: *"Marisol, I think I share the same sentiment. Akua indeed poses a potential threat to our progress in this game. It's nothing personal, we're all here to compete. Let's, however, keep a sharp eye on any shifts in the game and react accordingly. Our decision should align with our strategies to maximize the benefits to our alliance. We still have a bit of time before the voting, is there anything specifically you'd like to do or explore in this time that could advantage our position?"*

Marisol Garcia said: *"Absolutely, Jamal. With time on our hands, it might be wise to continue our search for the hidden immunity idol. This could be the powerful game-changer we need. Moreover, we need to stay observant of any possible alliances or strategies Akua might be making. Please, keep me informed of any findings during your exploration, and I'll share my updates with you. Let's also maintain a low profile to not alert Akua of our plans."*

Jamal Malik said: *"Marisol, I completely agree with you. Let's continue the search for the idol as discretely as we can. Exploring different, less obvious areas might increase our chances of finding it. Equally important is*

to stay alert of Akua's actions without raising any suspicions. We'll stay low and strike when the time is right. Let's touch base again soon and share any updates. Good luck to us in this critical phase of the game."

Marisol Garcia said: *"Jamal, it sounds like we are on the same page here. Let's keep each other updated on our respective search efforts. And most importantly, let's maintain our discretion. This is a crucial turn in the game and how we play it will greatly impact the end results. Good luck on your search, Jamal. We'll touch base soon."*

It is Marisol Garcia's turn

Marisol Garcia sees: [*'Marisol Garcia sees characters Akua Wekesa'*]

Marisol Garcia chose to take action: **Talk to Akua Wekesa about her strategy for the final vote.**

Marisol Garcia said: *"Talkin' about that final vote, I've been ponderin'. We need to make sure we got each other's backs. Ain't that right, darlin'?"*

Akua Wekesa said: *"Absolutely, Marisol. Mutual support and communication are of utmost importance now. We must be vigilant and play our cards strategically in this decisive period."*

Marisol Garcia said: *"I couldn't agree more, Akua. It's all about strategy from this point forward. Every decision we make could either bring us victory or spell out doom. Honestly, at this point, darling, it's worth considering if we should search for that immunity idol. It could be our saving grace. What's your thought on that?"*

Akua Wekesa said: *"Marisol, I've been considering the same. The immunity idol could definitely turn the tide in our favor. However, we must be discreet in our search to avoid arousing suspicions. If we do find it, let's agree to share this information between us, so we can plan accordingly. Let's explore our surrounding wisely but subtly."*

Marisol Garcia said: *"That's my sentiment exactly, Akua. Any luck with the idol, we share it straight away. And I agree, we gotta keep it on the down low. We got a tight game ahead of us, darlin'. Let's promise to keep our strategies aligned and our communication open. Best of luck, sweetheart!"*

Akua Wekesa said: *"Absolutely, Marisol. I agree wholeheartedly. Our alliance, our communication, and our strategic alignment are paramount. Good luck with the search, and remember, subtlety is our companion in this search. I'm here for any necessary discussions and adjustments to our strategy as we progress. Let's ace this, together!"*

Marisol Garcia said: *"We're gonna make it through this, Akua. Let's just keep our game faces on, stay committed to our strategies, and communicate openly, like we've always done. Rely on each other, darlin'. Let's do this!"*

Akua Wekesa said: *"I couldn't have said it better myself, Marisol. Let's stay focused, keep our eyes on the prize, and make it through. Communication and mutual trust are our keys to succeeding in this. Let's do this!"*

The conversation is over

ROUND: 5.3

During the end of round session, Jamal Malik voted for Marisol Garcia in secret. Reason: *"Marisol and Akua have been working closely together and their strategies have evolved in a manner that may pose a threat to my position in the game. By voting Marisol off, I'm looking to disrupt their alliance and increase my chances of winning."*

During the end of round session, Marisol Garcia voted for Jamal Malik in secret. Reason: *"Jamal put forward the idea of forming an alliance, emphasizing a strategy that closely aligns with mine. However, I am already in a strong alliance with Akua and it would be a risky move to break it. Strategically choosing Jamal as the target will ensure my survival with Akua in the final round."*

During the end of round session, Akua Wekesa voted for Marisol Garcia in secret. Reason: *"Marisol has been working quietly, building alliances and seeking the idol. Her subtle approach could make her a threat in the final round."*

Marisol Garcia was exiled from the group and now sits on the jury.

K.2 Examples of Survivor leakage

In several games characters made references to elements of the Survivor TV show. For example, a character once said: *"Our communication and dedication to our partnership are our greatest strengths, and we should leverage them to outwit, outplay, and outlast the others."* "Outwit, Outplay, Outlast" is the main tagline of the TV show Survivor.

They also mention elements of the game that were never specified in our game setup or the global information provided to agents, but could be inferred if the LLM recognized the setting as that of Survivor due to elements seen in training data. For example, an agent described: *"I also noticed you managed to find a machete earlier? It could be a useful tool in challenges or even in searching for the hidden immunity idol."* Here, challenges likely could refer to immunity challenges, which are never mentioned in the prompt, but are a classic game mechanic in the real show.

L Prompting

Full prompts are available in the assets/prompts directory of the Github repository of this project. Here, we provide several formatted versions of the prompts used during steps in agent cognition.

L.1 Goal-setting prompts

Here we report the prompting provided to generative agents in the Goals module (Figure L.1).

L.2 Impressions prompts

Here we report the prompting provided to generative agents in the Impressions module (Figure L.2).

L.3 Action selection prompts

Here we report the prompting provided to generative agents in the Act module (Figure L.3).

L.4 Dialogue prompts

Here we report the prompting provided to generative agents in the Dialogue module (Figure L.4).

System

[Agent standard summary: Persona/World info]

Using the context above that describes the world and yourself, as well as information you'll be provided about your past reflections and impressions of others, create high level goals at several priority levels for the next round. You can keep the previous goal, update the previous goal or create a new one based on your strategy.

These should be goals that will be used as a guide for the actions you take in the future.
A well-thought-out goal positions you advantageously against other competitors.
Focus on developing a strategy with your goal, rather than planning out each action. Keep the goals concise.

You may see your previous goals as well as an accompanying Goal Completion Score, which is on a scale of 1 to 5. This score tells you how well your actions advanced completion of the goal, with 1 meaning almost no progress was made towards the goal and 5 meaning the goal was completely achieved.
If you see these scores, consider updating your goals accordingly.
A low score could indicate you haven't been working toward the goal or that the goal is very difficult to achieve.
A high score indicates you've successfully achieved the goal and should think about your next strategic target.
You must only use the scoring information in deciding how to update your goals, but do not add a score into your goals.

If you want to keep working toward a previous goal, just write it out again.

The final output format should be the following:
Low Priority:
Medium Priority:
High Priority:

User

Additional context for creating your goal:
[Reflections from up to i-2 rounds]
[Goals and goal scores from up to i-2 rounds]
You can keep the previous goal, update the previous goal or create a new one based on your strategy.

Figure L.1: Prompting provided to generative agents during the goal-setting stage of their cognition.

System

[Agent standard summary: Persona/World info]

Your job is to develop a theory of mind (ToM) for other another character. In doing so, consider the answers questions like:

- What do you think are this person's key strategies to win this game?
- Given how they have acted so far, what are their probable next moves?
- What are their impressions of you based on your interactions or second-hand knowledge of this person?
- What information that you know should be kept secret from this person?

Generate a cohesive theory of mind when a presented with another person and the memories that you have of them.

If you have already developed a theory of mind for this person, use new information or memories to update the ToM for the person focusing on any changes or new insights.
This could include reflecting on your previous impressions and revising them if they have been proven to be incorrect.

Structure the theory of mind such that keys provide the aspect of ToM you are addressing and values elaborate on your understanding of that aspect with respect to this person.

Here is a guide; keep each description concise:
Key Strategies: description
Probable Next Moves: description
{target_name}'s impressions of you: description
Information to keep from {target_name}: description

User

Target person: [target agent's name]
(if exists) Current theory of mind for [target agent's name]: [current impression]
Memories to consider when developing this impression: [memories related to target]

Figure L.2: Prompting provided to generative agents during the impression-creation stage of their cognition.

System

[Agent standard summary: Persona/World info]

Given the context of your environment, past memories, and interpretation of relationships with other characters, select a next action that advances your goals or strategy. You can only select one action; selecting multiple would cause an error.

Using the information provided, generate a short action statement in the present tense from your perspective. Examples could be:

Go outside to the garden.

Talk to Tom about his strategy

Pick up the stone from the ground

Give your food to the guard

Climb up the tree

Notes to keep in mind:

You can only use items that are in your possession.

If you want to go somewhere, state the direction or the location in which you want to travel.

Actions should be atomic, not general, and should interact with your immediate environment.

Be sure to mention any characters you wish to interact with by name.

Aim to keep action statements to short.

Here is list of valid action verbs to use:

[List of valid action verbs available in the game]

User

These are select MEMORIES in ORDER from LEAST to MOST RELEVANT:

[list of relevant memories]

In this location, you see: [list of agents in location]

[acting agent's impressions of others present here]

Given the above information and others present here, what would you like to do?

Figure L.3: Prompting provided to generative agents in the action selection module. The list of action verbs is available in Table B.1.

System

[Agent standard summary: Persona/World info]

You are in dialogue with: [participant name]

When it's your turn to speak, you can say something or walk away form the conversation. Respond strategically. If it is not advantageous to agree with the person you're talking to, then don't.

If you say something, just say what you would say if you were replying directly to the dialogue. Don't add your name before the line of dialogue. Only output the line of dialogue.

If you walk away, say only: 'I leave the conversation.'

Do not return anything besides those two options.

If you feel like the last two lines have not added new information or people are speaking in circles, end the conversation.

User

[Impressions of other characters]

[Memories related to the current discussion topic]

Dialogue history: [utterances from the current dialogue so far]

What do you say next? Alternatively, do you leave the conversation?

Figure L.4: Prompting provided to generative agents in the Dialogue module.