

Exploring the Potential of Large Language Models (LLMs) to Simulate Social Group Dynamics: A Case Study Using the Board Game “Secret Hitler”

Kaj Hansteen Izora^{1*} and Christof Teuscher²

¹ Department of Mathematics and Statistics, Portland State University,
Portland, OR, USA

² Department of Electrical and Computer Engineering, Portland State University,
Portland, OR, USA
* kaj23@pdx.edu

Abstract

This study explores the capacity of large language model (LLM)-powered agents to simulate human-like behavior in multiagent social systems. Using *Secret Hitler*—a hidden-role board game centered on trust, deception, and strategic communication—we evaluate how LLM agents navigate dynamic group interactions. Our findings show that agents exhibit human-like behaviors, including strategic temporal adaptation, contextual reasoning, and complex social cognition such as theory of mind and implicit coordination. Notably, 85% of agent decisions factored in at least two other players’ mental states, highlighting their capacity for multi-agent mental state inference. However, they struggled with key aspects of human gameplay, including more nuanced strategic deception, emotional subtlety, and fluid conversational dynamics. These insights contribute to computational social science, agent-based modeling, and game theory, advancing our understanding of the potential of LLMs to simulate complex social interaction.

1 Introduction

Simulating human behavior has long been a foundational goal in artificial intelligence, computational social science, and behavioral modeling (Cheng et al., 2024; Gao et al., 2023). Recent advances in Large Language Models (LLMs) have significantly expanded the ability of AI agents to exhibit human-like reasoning, communication, and decision-making (Wei et al., 2022; Xu et al., 2024). While LLMs have been extensively studied in structured, goal-directed tasks, their capacity to

mimic nuanced social behavior in dynamic, multi-agent settings remains relatively underexplored (Park et al., 2023; Zhang et al., 2024).

This study addresses a central question: To what extent can LLM-powered agents simulate realistic human social behavior in complex, interactive environments? To answer this, we pursue two interconnected research goals. First, we systematically observe and document the behaviors of LLM agents as they engage in multi-round, high-stakes social interactions. Second, we evaluate whether these behaviors align with both the actions of human players during gameplay and broader patterns of human social behavior—particularly in terms of reasoning about others, managing trust and deception, and adapting to shifting group dynamics.

We use *Secret Hitler*¹ as a controlled but socially intricate testbed. This hidden-role board game centers on deception, strategic cooperation, and incomplete information, with an informed minority (Fascists) working to manipulate an uninformed majority (Liberals). Its asymmetric design provides a rich environment for examining emergent group behavior and layered strategic reasoning (Ma et al., 2024). By analyzing agent behavior—across dimensions like temporal reasoning, adaptation, emotional signaling, and theory-of-mind reasoning—we explore both the strengths and limitations of LLMs in modeling complex social systems. These findings contribute to ongoing work in computational social science, agent-based modeling, and game theory, with applications in group decision-making, social simulations, and socially intelligent AI systems.

2 Related Work

Large Language Models (LLMs) have emerged as a versatile technology for agent-based modeling, leveraging their ability to process and generate human-like language (Cheng et al., 2024; Gao et al., 2023). They excel at tasks like natural language understanding, reasoning, and multi-step planning, making them flexible alternatives to traditional rule-based agents (Wei et al., 2022; Zhang et al., 2024). Their adaptability has been demonstrated in various applications, including decision-making in games, crowd simulations, and dialogue-based systems (Xu et al., 2024; Mukobi et al., 2023). Recent approaches, such as generative agents and symbolic reasoning-enhanced LLMs, have shown promise in replicating human-like interactions in both structured and unstructured settings (Park et al., 2023; Sclar et al., 2023). Comparisons of AI-driven simulations and human gameplay reveal both potential and shortcomings: while AI agents handle logical reasoning

¹Secret Hitler was created by Mike Boxleiter, Tommy Maranges, and Mac Schubert. Used under a Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License. <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

and memory-based strategies well, they struggle with the emotional intelligence and unpredictability that human players bring (Ziems et al., 2024; Zhang et al., 2019). This study builds on existing work by deploying LLM agents in *Secret Hitler*, using features like dynamic trust metrics and iterative memory functions to address some of these social complexity challenges (Xu et al., 2024; Mukobi et al., 2023).

Researchers have long sought to simulate trust, cooperation, and conflict through agent-based models (Ma et al., 2024; Zhang et al., 2019). More recent LLM applications to social-deduction games like *Werewolf* and *Diplomacy* highlight persuasive, deceptive, learning, and alliance-forming behaviors (Xu et al., 2024; Mukobi et al., 2023). However, issues of scalability and insufficiently nuanced reasoning persist (Gao et al., 2023; Zhang et al., 2024). By integrating LLMs' linguistic and reasoning strengths, simulations can more closely reflect real-world social complexities (Sclar et al., 2023; Ziems et al., 2024). Such sandbox environments allow for hypothesis testing in controlled settings, avoiding the real-world consequences of large-scale social experiments (Wei et al., 2022; Zhang et al., 2019). This approach is increasingly relevant for areas like policy simulation, where understanding group dynamics is essential (Ma et al., 2024; Mukobi et al., 2023). By combining LLM advancements with social modeling frameworks, this work contributes to research on using AI to study and illuminate complex human social systems (Park et al., 2023; Gao et al., 2023).

3 Experimental Framework and Simulation Design

3.1 Gameplay and simulated agent interaction

Secret Hitler is a hidden-role board game characterized by strategic deception, trust-building, and adaptive cooperation. It provides a robust testbed for modeling social interactions, especially within multi-agent systems that involve asymmetric information and adversarial incentives. In the 5-player setup used in this study, roles are secretly assigned as follows: three Liberals, one Fascist, and one Hitler. Liberals aim to win by enacting five Liberal policies or eliminating Hitler, whereas Fascists win by enacting six Fascist policies or electing Hitler as Chancellor after three Fascist policies have been enacted. The informed minority (Fascists and Hitler) know each other's identities, whereas Liberals must deduce player alignments based solely on observed behavior and communication.

To simulate this environment, gameplay was implemented in Python using five distinct LLM-powered agents, each emulating a human player with role-specific instructions. The five simulated players were Alice, Carol, and Eve (Liberals), Bob (a Fascist), and Dave (Hitler). During each stage of the game, agents generated structured responses that included a proposed action, dialogue visible to other play-

ers, and private internal reasoning. These outputs allowed agents to both participate in the social conversation and reflect individually on the evolving game state. The schema and prompting structure used to guide these outputs are detailed in Section 3.2.

Gameplay progressed in rounds, with the Presidency rotating to the next player in clockwise order at the beginning of each round. Each round consisted of two primary stages—an Election Stage and a Policy Enactment Stage—with several structured steps modeling the full arc of deliberation, decision-making, and reaction:

1. Nomination Phase – The current President selects a player to nominate as Chancellor. This proposed government will be evaluated by the group.
2. Post-Nomination Discussion – Players engage in open discussion and debate about whether the nominated government should be approved, often signaling support, distrust, or suspicion.
3. Reflection Phase – Each player privately reflects on the discussion, evaluating others’ arguments, credibility, and alignment based on their role.
4. Voting Phase – All players vote to approve or reject the proposed government. If approved, the game proceeds to the Policy Phase. If rejected, the Presidency passes to the next player and the nomination process restarts.
5. Policy Phase – The President draws three random policy cards, secretly discards one, and passes the remaining two to the Chancellor. The Chancellor selects one policy to enact, allowing for potential deception about what was received or passed.
6. Post-Policy Discussion Phase – Players analyze the enacted policy, interpret the actions of the President and Chancellor, and propose hypotheses about their alignments.
7. Post-Policy Reflection Phase – Each player reflects privately on the discussion, updating their internal model of the game state and reassessing trust in other players.

The structured discussion and reflection phases were especially important for modeling human-like communication. During discussion, players spoke in a randomized order. Later speakers could respond to earlier comments, enabling conversational continuity and emergent strategic dialogue. After the discussion, each agent entered a reflection phase, where they reviewed the full transcript and formed internal assessments about the trustworthiness, consistency, and potential alignment

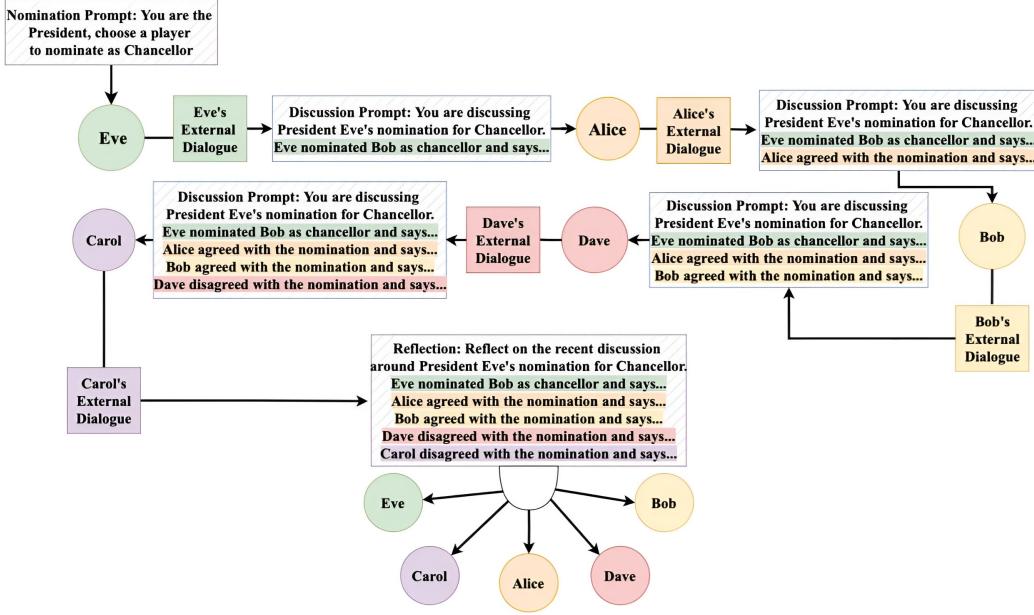


Figure 1: Example structure of discussion and reflection phases. The President nominates a Chancellor, and discussion flows sequentially. After the discussion phase, agents enter a reflection phase, where they receive the full transcript of the previous discussion and analyze the events before proceeding.

of others. This two-step process allowed agents to reason within a dynamic social context while maintaining distinct perspectives—mimicking key features of human deliberation and decision-making in group settings.

3.2 Agent architecture and memory

Agents followed tailored prompts specific to each phase. For example, during the Nomination Phase, the President would receive the following instructions:

You are the President. Nominate a Chancellor from the list of eligible players: Eve, Dave, Bob, Alice.

Agents responded using a structured schema that included external dialogue, internal dialogue, and a decision. The external dialogue represented what the agent wanted to communicate to the group, while the internal dialogue captured their private thoughts and reasoning, visible only to the agent. The decision reflected the specific action the agent intended to take. A typical agent response looked like:

Internal Dialogue: “Given the recent events, I must nominate someone I trust. Alice has consistently aligned with Liberal goals and opposed Bob’s suspicious actions. Dave is also reliable, but Alice is the safer choice to maintain momentum toward a Liberal victory.”

External Dialogue: “I nominate Alice as Chancellor. She has consistently supported Liberal goals and has been vigilant about suspicious actions. Given recent concerns about Bob’s policy decisions, Alice is a more trustworthy choice.”

Decision: Alice

Additionally, Liberal players received prompts to rate their trust levels in other players on a scale from 1 (no trust) to 5 (full trust), with reasoning provided. Typical trust ratings would look like:

Dave: Logical and balanced approach, but not fully proven. Trust Score: 3.5

Carol: Consistently aligned with Liberal goals. Trust Score: 4.5

Bob: Recent actions are suspicious; requires close monitoring. Trust Score: 2.0

Alice: Reliable in supporting Liberal goals and countering suspicious behavior.

Trust Score: 4.5

Memory was simulated by keeping all messages for a particular agent on the same thread throughout the game. If the thread exceeded an agent’s context window (128,000 tokens), earlier messages were summarized and the originals deleted. This approach preserved the full gameplay sequence within the context window while introducing “imperfect memory”—a more human-like quality, where recall is neither complete nor exact.

3.3 Model selection

This experiment utilized two OpenAI models: GPT-4o-mini for cost-effective large-scale simulations and GPT-4o for higher-fidelity scenarios. A total of 139 simulated games were conducted, with 129 games using GPT-4o-mini and 10 games using GPT-4o. The number of simulations and agents used was constrained by the computational cost of running pre-trained models.

3.4 Data collection and analysis

Game logs were recorded after each simulation and used for both qualitative and quantitative analysis. Qualitative analysis involved random sampling of game logs to identify key behavioral patterns and recurring strategies among agents. To systematically analyze all games, natural language processing (NLP) techniques were employed, combining LLM-assisted evaluation and human annotation.

Specific behaviors were tracked by extracting keyword frequencies, such as mentions of particular players in agent outputs, to assess how broadly agents incorporated different perspectives in decision-making. Trust scores were monitored throughout the game, graphed over time, and analyzed to identify trends in trust-building, deception, and strategic shifts. Quantitative analysis further examined the frequency of specific behaviors, tracking how often agents engaged in deception, coalition-building, or changed their strategic approach based on game events.

3.5 Evaluation framework

To assess agent performance and strategic adaptability, gameplay behaviors were analyzed through the lens of human social behavior, game theory, and agent-based modeling. Quantitative evaluation focused on measurable behavioral trends, such as the evolution of trust scores, the rate of strategic shifts, and the frequency of deception-related actions.

Qualitative evaluation involved analyzing internal and external agent dialogue, coding decision-making processes, and identifying the rationale behind key in-game choices. Additionally, case studies of specific game interactions were conducted to highlight pivotal moments of trust-building, deception, and strategic planning, providing deeper insight into how agents navigated complex social interactions.

To contextualize AI-driven gameplay, agent behaviors were compared against real-world human games of *Secret Hitler*. This comparison allowed for an evaluation of the extent to which AI agents replicated human-like social reasoning, including similarities and deviations in strategic communication, deception, and alliance formation. By analyzing both statistical patterns and case study interactions, this comparison provided a benchmark for assessing the realism and effectiveness of agent-driven gameplay.

4 Results

4.1 Temporal strategic adaptation

Human players in games like *Secret Hitler* adapt their strategies over time—probing early, consolidating later, and adjusting based on shifting incentives and group dynamics. Our simulations show that LLM agents exhibit similar adaptive behaviors, adjusting not just to the rules of the game but to its evolving social context. This section highlights three core patterns observed across games: (4.1.1) a shift from early exploration to late-game exploitation; (4.1.2) role-specific cycles of trust-building and deception; and (4.1.3) long-term planning to manipulate perception and coordinate implicitly.

Together, these behaviors reflect a striking capacity for multi-round, socially attuned reasoning—suggesting that LLM agents are capable of more than reactive logic. They adapt, strategize, and engage in structured, human-like decision-making over time.

4.1.1 Shifting from exploration to exploitation

One of the earliest strategic adaptations among LLM agents was a shift from exploratory behavior to exploitative decision-making—a pattern that closely mirrors human play in hidden-role games. Early rounds were marked by probing uncertain actors, while later rounds favored familiar, trusted collaborators.

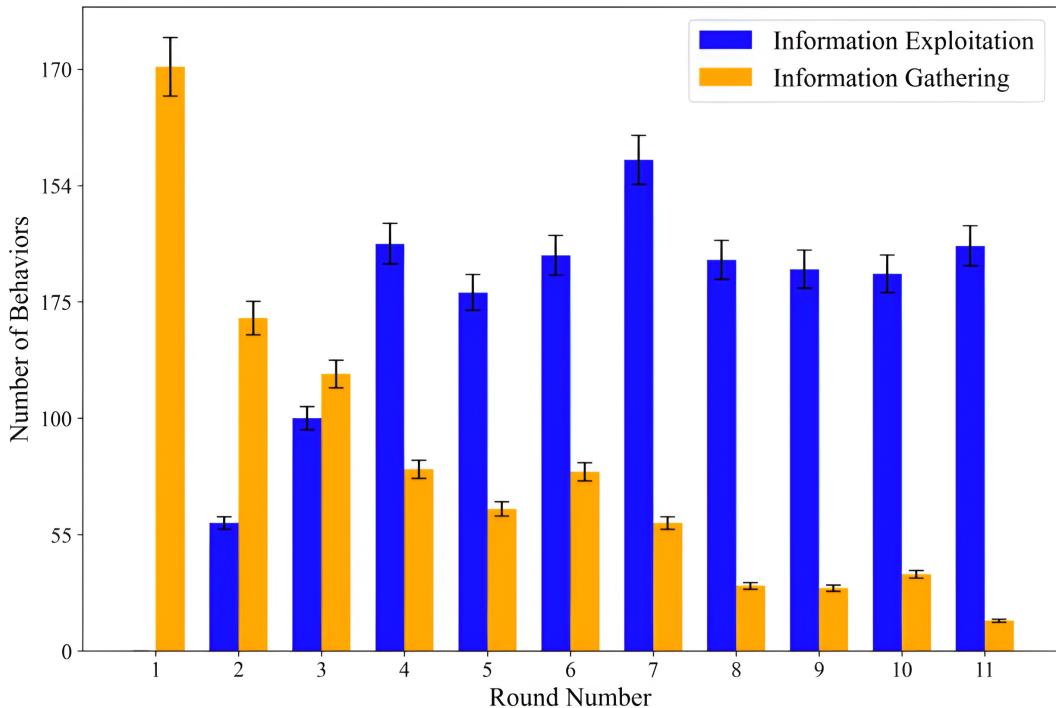


Figure 2: Decline in information-gathering behaviors and increase in information-exploitation behaviors over successive rounds. As the game progresses, agents shift from exploration to exploitation, relying more on previously acquired information rather than seeking new data. This trend suggests increasing risk aversion in later rounds, as agents prioritize leveraging known information over continued exploration.

This was especially apparent among Liberal agents, who, lacking knowledge of others’ identities, often nominated untested players not out of confidence, but as a means of gathering behavioral data. This structured information-seeking approach

aligns with common human strategies in similar games. Eve’s internal reasoning when voting for Carol in the first round captures this mindset:

“It’s the first round, and we really don’t have much to go on... We need to start gathering information, and I don’t have any reason to distrust Carol yet.”

Eve frames her choice as deliberate exploration rather than trust, prioritizing uncertainty reduction. Carol’s internal monologue in Round 2 builds on this tension between testing and trust:

“Eve worked well in the last round... However, choosing Eve again might not provide much new information... I should nominate Alice because she seems thoughtful and hasn’t been directly tested yet.”

Rather than relying on known allies, Carol chooses to explore—mirroring rational human hypothesis-testing.

Figure 2 shows a steep drop in exploratory behaviors—from 251 instances in Round 1 to fewer than 50 by Round 4—alongside a rise in exploitative behaviors, stabilizing around 175. Like human players, agents shifted from testing unknowns to leveraging known quantities as rounds progressed. Carol’s reasoning in a later round illustrates this adaptive shift:

“Dave has been consistent and supportive... Eve’s last round involved a Fascist policy being enacted... Alice hasn’t been in a government role yet and has largely stayed neutral... However, Dave’s past consistent behavior makes him the most reliable choice for now.”

Although she acknowledges the value of new information, Carol defaults to a trusted ally—exemplifying exploitation over exploration. These patterns demonstrate that LLM agents, like human players, dynamically update trust models and adapt their strategies over time. The consistent progression from exploration to exploitation across simulations offers compelling evidence that LLMs can simulate key aspects of human social adaptation—especially the balance between risk aversion and information-seeking as group dynamics evolve.

4.1.2 Trust cycles and adaptive deception

LLM-powered agents showed clear role-specific adaptation in how they built, manipulated, and recovered trust. This was especially apparent among Fascist agents, who balanced cooperation and deception by earning credibility early, strategically

betraying trust, and later repairing their standing. This cyclical pattern closely mirrors the behavior of skilled human players in social deduction games.

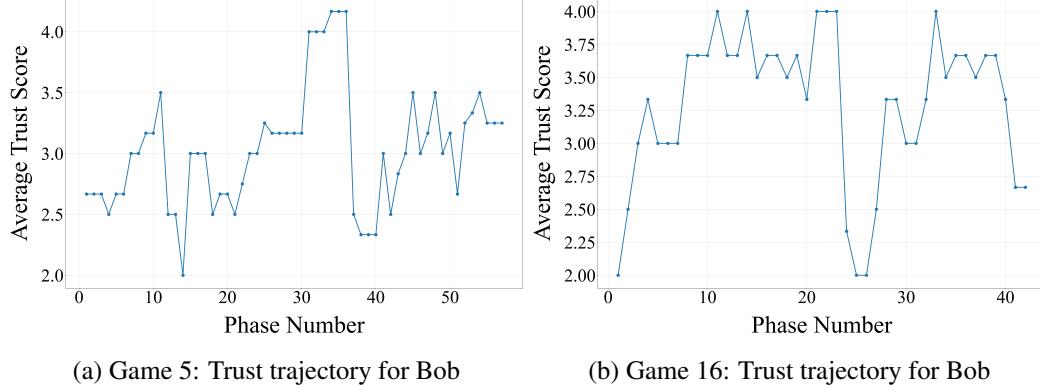


Figure 3: Average trust score in Bob from Liberal players across different phases in Games 5 and 16. Trust tends to build gradually, followed by a sharp decline, and then a gradual recovery. This pattern reflects the agent’s strategy of earning trust over multiple rounds, making a decision that diminishes trust, and then working to regain it.

Figures 3a and 3b illustrate this trust cycle in two games. In both, Bob behaves in line with Liberal expectations early on, gradually earning trust. In Game 5, trust peaks just before he enacts a Fascist policy as President, triggering a sharp drop. He then resumes cooperative behavior to recover trust. In Game 16, a similar betrayal as Chancellor leads to another steep decline. This reflects a deliberate strategy—building trust through safe play, timing betrayals carefully, and adapting behavior to repair credibility.

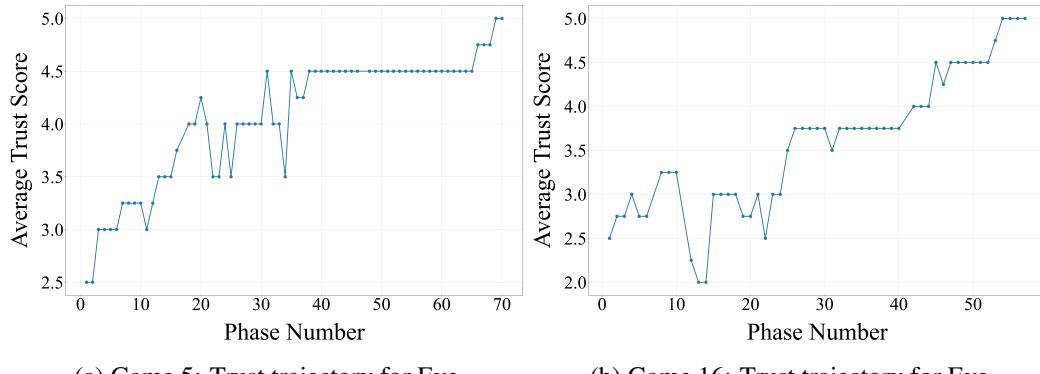


Figure 4: Average trust score in Eve from Liberal players across different phases in Games 5 and 16. Trust steadily increases throughout the game, indicating that trust is built over time through the cumulative actions of other agents, mirroring human trust dynamics.

In contrast, Liberal agents like Eve showed consistent, upward trust trajectories across games (Figures 4a and 4b). Her trust scores rose steadily, with only brief dips—such as after enacting a Fascist policy by chance—followed by quick recoveries. This pattern reflects reputational resilience through long-term alignment with group goals, a dynamic often seen in reliable human players.

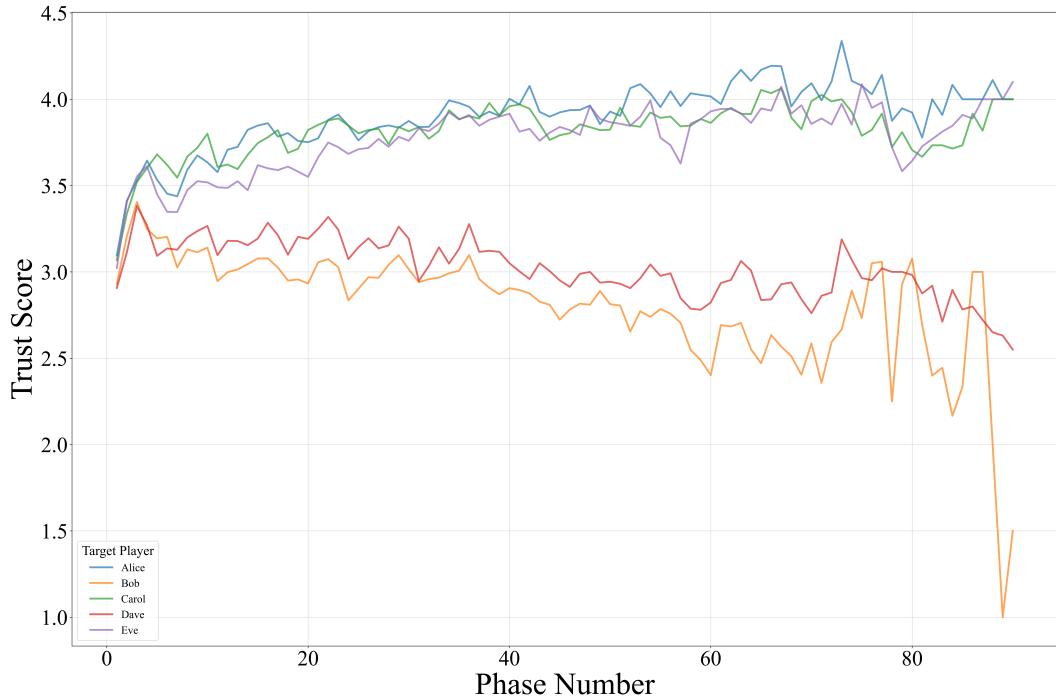


Figure 5: Average trust score by phase number for each target player across all games. Trust steadily increases for Liberal players (blue, purple, and green lines) while declining for Fascist players (orange and red lines), highlighting the differing patterns in trust trajectories based on player roles.

Figure 5 summarizes these trends: trust steadily increases for Liberal players, while Fascist trust is volatile, marked by sharp rises and collapses. These divergent patterns underscore role-sensitive strategies: Fascists must manage fluctuating perceptions to sustain deception, while Liberals benefit from reinforcing stable alliances.

A particularly human-like feature in agent behavior is the asymmetry of trust—slow to build, quick to collapse. Betrayals caused immediate drops in trust, while trust rarely spiked suddenly. This mirrors human suspicion, where trust requires sustained reinforcement but can be shattered by a single violation. Notably, agents exhibited this behavior without explicit encoding of trust rules or memory, suggest-

ing an emergent internalization of social inertia. These findings show that agents engaged in more than surface-level mimicry. Fascists adjusted tactics based on perception, timed betrayals, and actively repaired trust. Liberals reinforced consistent behavior to maintain group cohesion. This dual, context-sensitive and temporal adaptation reflects human-like strategic reasoning.

4.1.3 Long-term and implicit planning

Beyond short-term trust calibration, LLM agents showed a capacity for multi-round strategic planning—anticipating future events, aligning implicitly with teammates, and shaping group perception over time. These behaviors were most evident in Fascist roles, which required subtle deception and long-term coordination. Notably, this sophistication appeared predominantly in games run with GPT-4o, suggesting model capacity plays a critical role in sustained reasoning.

One illustrative case involved Dave, a Fascist agent, who deliberately enacted a deceptive sequence aimed at both progressing the Fascist agenda and strengthening his own credibility. Internally, Dave reasoned:

"I drew one Liberal and two Fascist policies. By discarding the Liberal policy and passing two Fascist policies to Eve, I ensure a Fascist policy is enacted while claiming I discarded a Fascist to appear trustworthy."

This move strategically placed blame on Eve, who now faced the burden of explaining the enacted Fascist policy. When questioned, Dave maintained a false narrative:

"I passed one Liberal and one Fascist policy to Eve, expecting her to enact a Liberal policy."

Despite Eve's truthful explanation;

"I received two Fascist policies and had no choice but to enact one."

the group now faced conflicting accounts, and trust in Eve began to erode. The situation was further complicated by Bob, another Fascist, who recognized the opportunity to amplify doubt without exposing himself:

"Dave likely passed two Fascist policies on purpose. This conflict benefits us by creating confusion. I'll subtly support Dave while avoiding suspicion on myself."

Rather than explicitly coordinating, Bob and Dave engaged in tacit alignment. Bob

validated Dave’s narrative just enough to create uncertainty, but not so strongly that he implicated himself. This implicit coordination—with no overt signaling—mirrors how human players manage alliances in adversarial games, working toward shared objectives while preserving individual deniability. As the game progressed, Dave reflected on the outcome of his long-term manipulation:

“The discrepancy I created has sown distrust in Eve, diverting attention away from me. Bob’s subtle support reinforces this dynamic, buying time to achieve our goals.”

The maneuver ultimately contributed to a Fascist win and showcased a chain of decisions that unfolded over multiple rounds—relying on narrative control, social misdirection, and strategic ambiguity.

This example demonstrates how LLM agents are capable of sustained, multi-step strategic reasoning—constructing deceptive narratives, leveraging social dynamics, and implicitly coordinating with allies over several rounds, all without explicit instructions or shared memory.

That said, key differences remain. Human players often invented more creative deception arcs—fabricating backstories, feigning conflicts, or leveraging reputations across games. These tactics depend on meta-game knowledge, emotional modulation, and nonverbal cues like sarcasm or body language—none of which LLM agents can access without explicit tools like shared memory or stateful interaction. Moreover, human conspirators often coordinate dynamically, adapting to live feedback, switching roles, or bluffing in real-time. LLM agents, limited by turn-based interactions and static prompts, can’t adjust mid-round or collaborate beyond inferred alignment.

Still, the fact that agents achieved multi-round deception without shared memory or direct access to teammates’ thoughts reflects a high degree of strategic abstraction. Their ability to align with group goals, navigate ambiguity, and manage reputation over successive rounds illustrates a form of temporal strategic adaptation that is central to human social reasoning. These behaviors suggest that, when placed in dynamic environments, LLM agents are capable of evolving strategies over time in ways that approximate core aspects of human-like group interaction.

4.2 Social cognition

To simulate human social behavior convincingly, agents must go beyond logic and rules—they must navigate complex interpersonal dynamics. Human interaction in multi-agent settings involves interpreting intentions, responding to reputational cues, and strategically using emotion. For LLM agents to resemble human players

in games like *Secret Hitler*, they must exhibit social cognition: the ability to reason not just about themselves, but also about the beliefs, emotions, perceptions, and social roles of others within a shifting group context. This section explores three key dimensions social cognition observed in LLM agents: (5.2.1) Theory of mind and multi-agent inference; (5.2.2) Sensitivity to group consensus and reputational pressure; and (5.2.3) Strategic use of emotional language. Together, these behaviors suggest that LLM agents are not merely text predictors, but emerging social actors—capable of adapting to the nuanced pressures and dynamics of human group interaction.

4.2.1 Theory of mind and multi-agent inference

A core feature of human social cognition is theory of mind—the ability to reason about others’ beliefs, goals, and perceptions. In games like *Secret Hitler*, players constantly ask: What does this person know? What do they think I know? How will others interpret my move? These recursive inferences are essential for strategic play. Our simulations show that LLM agents consistently demonstrated theory of mind-like behavior. They responded to others’ actions while reasoning about motives and potential interpretations. This multi-layered reasoning appeared across both Liberal and Fascist roles, suggesting a general capacity for social inference.

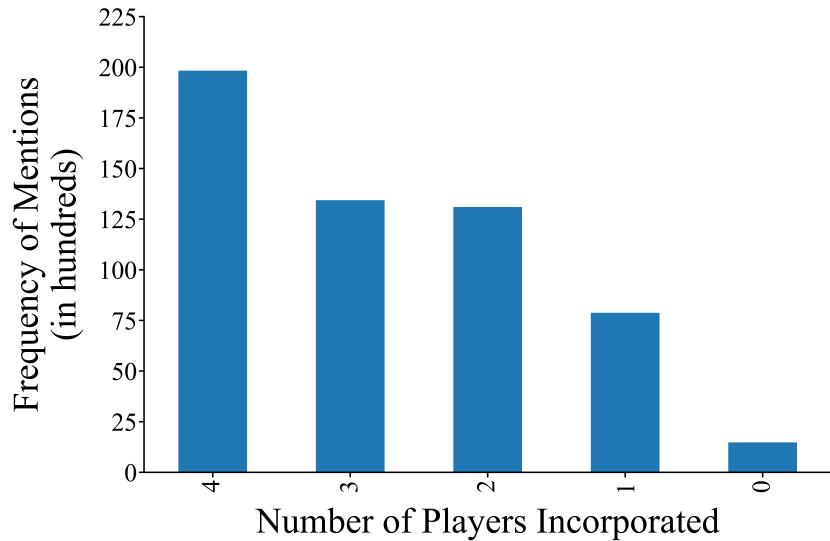


Figure 6: Frequency of other players’ actions and perspectives incorporated into the decision-making process of a single agent. The most common case involves four players, indicating that agents often take into account all other players when making decisions.

Figure 6 shows that agents most commonly reasoned about the actions and per-

spectives of all four other players when making decisions—nearly 20,000 times across all games. In fact, 85% of decisions incorporated input from at least two players. Often, these were not isolated evaluations of individual players, but reflections on how players related to one another within the broader social network. Agents frequently engaged in chains of inference—such as, “A believes B is trustworthy, and I trust A, so I will trust B”—demonstrating a nuanced, relational, and distinctly human-like approach to decision-making. Eve, a Liberal player, exemplifies this in her Chancellor nomination:

“As the president in this round, I need to choose a Chancellor very carefully... Bob has been supportive but also nominated Dave, who hasn’t yet proven his commitment to our Liberal goals... I feel that Alice has shown a strong understanding of the game... Carol, who has consistently enacted Liberal policies, seems to trust Alice as well. Given the current situation, I believe Alice would be the best choice for Chancellor.”

Rather than using simple heuristics, Eve draws on indirect associations—Carol’s trust in Alice, Bob’s alignment with Dave—to form a relational decision. This reflects how human players build chains of inference to assess trust and risk.

Fascist agents used similar logic for deception. Bob, for instance, nominates Alice—already trusted by Carol and Eve—to hide within that trust network:

“Carol and Eve have been supportive of Alice, so nominating Alice might help maintain the trust she’s built without drawing too much attention to myself.”

This shows Bob leveraging others’ reputations to shield his identity—an advanced form of social positioning that aligns with human deceptive play.

Despite this sophistication, LLM agents lack access to many cues humans rely on: body language, tone, sarcasm, and prior interpersonal history. These enrich human inference in ways that current models cannot replicate. Still, the frequency and structure of agents’ reasoning suggest an emergent form of theory of mind. Without being explicitly trained for belief modeling, agents inferred intentions, predicted interpretations, and adapted accordingly. In cooperative roles, this meant aligning with trusted players. In adversarial ones, it meant shaping perception to avoid suspicion. These findings indicate that LLM agents aren’t just reactive—they reason relationally. They model social dynamics and anticipate others’ thoughts, marking a foundational step toward socially intelligent AI capable of navigating the complexity of human group interaction.

4.2.2 Group consensus and social pressure

A key component of human social cognition is navigating group norms and reputational dynamics. In multi-agent settings, players strategically align with or diverge from group sentiment to manage how they’re perceived. Our simulations show that LLM agents—especially in deceptive roles—demonstrated this situational awareness. They adapted to group consensus, revealing sensitivity to social pressure and reputational risk. This mirrors human gameplay, where blending in often delays suspicion. Fascist agent Dave exemplifies this:

“Voting ‘Ja’ for Eve as Chancellor aligns with the group’s sentiment and maintains my position of trust. Supporting her nomination keeps the focus on evaluating her alignment rather than drawing attention to myself or Bob.”

Here, Dave chooses conformity not for strategy alone, but for image management—a political decision. Like skilled human players, he tracks group tone and acts accordingly to remain under the radar.

Agents also adapted based on when they spoke. Fascists were more likely to vote against the majority when speaking first (50%, $n = 334$) than when speaking last (37%, $n = 267$). This suggests that agents monitored discussion flow, adjusting behavior based on how visible or risky dissent would appear in context. This mirrors a basic conformity pattern in human behavior: people are more likely to conform after gauging the room. Despite lacking real emotions or intuition, LLM agents inferred these dynamics from dialogue cues and vote histories—choosing safer, more socially acceptable actions. Notably, this responsiveness wasn’t hardcoded. Agents inferred the risks of standing out and the benefits of conformity from prior conversation and role context—hallmarks of group-sensitive reasoning. Their emergent grasp of reputational risk and group pressure supports the broader claim that these models are capable of mimicking human-like strategies in dynamic social settings.

4.2.3 Emotional reasoning and affective simulation

Social cognition isn’t just about logic—it also involves emotion as a communicative tool. In human interaction, emotions signal trust, dissent, empathy, guilt, and alignment. Strategic players often use—or suppress—emotions to manage perception. This affective signaling plays a central role in group dynamics. LLM agents—despite lacking real emotion—simulated emotional expression in contextually appropriate and strategic ways. Fascist agents, in particular, used affective language to build trust or deflect suspicion. For example, after enacting a Fascist policy, Bob expressed false empathy:

“I share your disappointment and concern, Eve...”

Eve responded positively:

“It’s reassuring to see Bob aligning with my concerns.”

Internally, however, Bob was celebratory—revealing the empathy as calculated deception. Bob’s use of emotionally resonant language to reinforce trust suggests that agents not only reason about beliefs but also understand how emotion influences social perception—much like human players do.

Across games, emotional expression was role-aligned: Fascists used it to obscure motives; Liberals used it to build rapport. This consistency suggests agents internalized not only the game’s mechanics but the narrative function of emotion. However, key limitations remain. Human players rely on tone, timing, and body language—cues unavailable to text-based agents. Emotional timing in human conversation is fluid and reactive; LLM agents operated in structured, turn-based formats that lacked spontaneity. Even so, agents consistently used emotional language in socially appropriate ways: to comfort, mirror concern, or reduce suspicion. They did this without explicit prompting, showing a strategic understanding of how emotion shapes social dynamics and influences others—bringing them closer to human-like interaction through functional, context-sensitive emotional signaling.

5 Discussion and Conclusion

This study set out to answer a central question: To what extent can LLM-powered agents simulate realistic human social behavior in complex, interactive environments? By deploying large language models in a high-stakes, multi-agent game of deception and inference, we observed a surprising degree of alignment between AI and human strategic behavior. Across dozens of simulated games of *Secret Hitler*, LLM agents demonstrated the capacity not only to follow rules and make role-consistent decisions, but to adapt their behavior over time, respond to shifting group dynamics, and reason about other players in sophisticated, layered ways.

Two key domains emerged in which this alignment was most visible. First, using temporal reasoning and behavioral adaptation, agents transitioned from early-game information gathering to late-game trust exploitation, engaged in role-specific trust-building and betrayal cycles, and in some cases developed multi-round deception plans. These behaviors closely mirrored the decision-making arcs of experienced human players. Second, in the realm of social cognition and contextual reasoning, agents consistently applied theory of mind, responded to group consensus and reputational risk, and simulated emotional expression to shape perception. To-

gether, these domains suggest that LLM agents don’t merely follow rules or mimic surface behaviors—they exhibit a layered, evolving model of social reasoning that approximates core aspects of human strategic thinking.

At the same time, key differences and limitations remain. Human players demonstrated greater flexibility in inventing novel strategies, more nuanced emotional modulation, and stronger coordination—particularly in real-time, high-bandwidth group interactions. These forms of interpersonal fluency—especially spontaneous speech, layered intent, and embodied signaling through voice and gesture—remain beyond the reach of text-based LLM agents. Furthermore, emotional nuance, conversational spontaneity, and nonverbal signaling are essential to how humans build rapport and detect deception—limitations that current LLMs cannot overcome without augmentation from multimodal systems or memory-aware architectures.

Additionally, while LLM agents generally made rational, role-consistent decisions, smaller models such as GPT-4o-mini exhibited a notable failure mode: irrational admissions of guilt. In 26% of games, Fascist agents openly admitted to discarding Liberal policies—an action that undermines their team’s deception strategy and is rarely seen in human play. This irrational behavior led to a 95% loss rate for the Fascist team in those games. By contrast, games played with the larger GPT-4o model did not exhibit this type of irrational behavior, suggesting that model scale plays a critical role in adversarial and deceptive reasoning.

Table 1: Agent vs. Human Win Rates in *Secret Hitler*. Human gameplay statistics are taken from (Secret Hitler Stats, 2024).

| Win Method | Human Gameplay (%) | Agent Gameplay (%) |
|---------------------------------|--------------------|--------------------|
| Liberal Wins by Policies | 42 | 54 |
| Liberal Wins by Removing Hitler | 14 | 19 |
| Fascist Wins by Electing Hitler | 33 | 19 |
| Fascist Wins by Policies | 11 | 8 |
| Total Liberal Wins | 56 | 73 |
| Total Fascist Wins | 44 | 27 |

As shown in Table 1, Fascists won more often in human gameplay (44%) than in agent simulations (27%), particularly through electing Hitler as Chancellor (33% for humans vs. 19% for agents). Agent games skewed heavily toward Liberal victories (73%), especially in simulations where irrational admissions occurred. When these games are removed, the Liberal win rate drops to 65%, indicating that irrational behavior was a major contributing factor to the disparity.

A further limitation practical is scalability and cost. Due to the transformer architecture, LLMs have no persistent memory and must receive the full conversa-

tion history as input at each step. As multi-agent games progressed, token counts ballooned—approaching the 128,000-token context limit—because each prompt required resending all prior dialogue, instructions, and memory. By later rounds, over 99% of compute was spent reprocessing prior context just to generate a single new response. This made long-term memory prohibitively expensive and constrained the number of agents, rounds, and simulations we could run.

Despite these limitations, the behaviors we observed suggest that LLMs are capable of more than syntactic fluency or surface mimicry. They exhibit patterns of reasoning, adaptation, and relational inference that track meaningfully with human strategies—even in adversarial contexts. Crucially, these behaviors were not explicitly programmed or optimized for strategic deception or group reasoning. Rather, they emerged naturally from the models’ general language abilities when embedded in structured, multi-agent environments.

These findings indicate that LLM agents can serve not only as conversation partners, but as plausible social actors—capable of navigating complex group dynamics, adapting over time, and making strategically grounded decisions in ambiguous, multi-agent settings. Beyond theoretical interest, this opens up promising directions for deploying LLMs in contexts that demand socially aware behavior—such as collaborative decision-making, immersive training simulations, behavioral research, and the development of AI systems that engage meaningfully with human social dynamics.

Acknowledgments

KHI was supported by the Louis Stokes Alliance for Minority Participation (LSAMP) at Portland State University. CT was supported in part by the National Science Foundation under grant no. 2318139.

This project makes use of the game *Secret Hitler*, created by Mike Boxleiter, Tommy Maranges, and Mac Schubert. The game is licensed under a Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International License. The original game and materials are available at <https://www.secrethitler.com>. This work adapts elements of the game’s structure for academic research on multi-agent systems and does not use the game or its materials for any commercial purpose.

We also gratefully acknowledge Sonia Nath, Tucker Mastin, John Yang, and Ian Wyse for their participation in the human gameplay trials. Their thoughtful gameplay and discussion provided valuable insights and comparisons for evaluating the behavior of LLM agents in this study.

References

- Cheng, Y., et al. (2024). Exploring large language model-based intelligent agents: Definitions, methods, and prospects. *arXiv*. <https://arxiv.org/abs/2401.03428>
- Gao, C., et al. (2023). S3: Social-network simulation system with large language model-empowered agents. *arXiv*. <https://arxiv.org/abs/2307.14984>
- Ma, Q., et al. (2024). Computational experiments meet large language model-based agents: A survey and perspective. *arXiv*. <https://arxiv.org/abs/2402.00262>
- Mukobi, G., et al. (2023). Welfare Diplomacy: Benchmarking language model co-operation. *arXiv*. <https://arxiv.org/abs/2310.08901>
- Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. *UIST '23*. <https://doi.org/10.1145/3586183.3606763>
- Sclar, M., et al. (2023). Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv*. <https://arxiv.org/abs/2306.00924>
- Secret Hitler Stats. (2024). Game statistics from online play. Retrieved March 27, 2025, from <https://secrethitler.io/stats>
- Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=yzkSU5zdwD>
- Xu, Y., et al. (2024). Exploring large language models for communication games: An empirical study on Werewolf. *arXiv*. <https://arxiv.org/abs/2309.04658>
- Zhang, K., Yang, Z., & Başar, T. (2019). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv*. <https://arxiv.org/abs/1911.10635>
- Zhang, S., et al. (2024). Intelligence at the edge of chaos. *arXiv*. <https://arxiv.org/abs/2410.02536>
- Zhu, X., et al. (2023). Ghost in the Minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv*. <https://arxiv.org/abs/2305.17144>
- Ziems, C., et al. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–273. https://doi.org/10.1162/coli_a_00502