LLM Agentic Simulation for Exploring the Emergence and Evolution of Civilization: Societal and Interstellar Systems

GitHub Repo: https://github.gatech.edu/bxia34/agentic-life

Presentation Video

Group 1

Yue-Tsz Fan, Vaibhav Pai, Tim Phan, Rusty Utomo, and Bin Xia

Abstract. We explore how large language model (LLM) agents generate and negotiate beliefs in social and interstellar settings using a text-based simulation suite. Starting from the Werewolf social deduction game, we verify that memory-less ChatGPT 3.5 agents can follow hidden-role rules, practice deception, and update beliefs under partial information. We then repurpose the environment into a “Shared Fictions” task where seven personality-driven agents propose, debate, and fund norms or proto-myths while managing shared resources and disaster risk. Finally, we lift the scaffolding to a “Dark Forest” cosmic sociology scenario in which civilizations declare survival, expansion, concealment, or preemptive-attack strategies under noisy observations and optional treaties. Across simulations, we qualitatively validate behaviors via transcript review and moderator logs: deception and trust-building in Werewolf, personality-shaped yet largely utilitarian belief proposals in Shared Fictions, and concealment-dominant, brittle alliances in Cosmic Sociology. The results highlight both the promise of LLM agents for studying emergent coordination and the limits of existing prompting pipelines, which favor rational narratives and struggle with structured action emission.

1. Project Description. This project is inspired by real-world sociological and political systems, where agents (whether human or AI) must navigate complex systems of beliefs, trust, cooperation, and conflict. The framework seeks to understand how shared fictions and collective behaviors emerge over time in social groups and how these dynamics might extend to interstellar systems. The project’s findings can contribute to our understanding of the evolution of civilizations and the role of shared beliefs in their sustainability and expansion. By modeling these dynamics in AI, the project aims to provide insights into human behavior, the formation of social norms, and the economic, technological, and management trends that influence the collective functioning of large systems.

Building on this principle, agent-based simulations extend these local interactions into higher-dimensional, information-rich spaces. In our framework, agents can exchange structured information through actions or language, allowing interactions that are not strictly spatially constrained but convey intentions, beliefs, and strategies. This capability enables the exploration of emergent behaviors across scales—from social systems driven by shared fictions and collective norms to speculative interstellar civilizations. In this sense, our agentic simulation can be viewed as a *high-dimensional, information-mediated cellular automaton*, preserving the emergent qualities of classical CA while introducing rich communication channels that facilitate the spontaneous evolution of complexity across scales.

2. Literature Review. Research on emergent complexity spans multiple scales of life – from social to cosmic systems.

Social and cognitive scale. Generative agents [2] showcase the potential of large lan-

This manuscript is for review purposes only.

guage models (LLMs) as behavioral simulators that reproduce human-like cognition and social dynamics. By integrating memory, reflection, and planning, these agents exhibit emergent coordination—such as self-organized social events—without explicit scripting. Ziheng et al. [7] extend this paradigm to moral evolution, assigning agents moral types based on the "expanding circle" concept [3], which models concern expanding from self to kin to group, and demonstrating how cooperation and group identity emerge in survival-driven environments. Wu et al. [5] emphasize methodological rigor for LLM-based societies, proposing alignment, consistency, and robustness as key reliability criteria. Xu et al. [6] empirically demonstrate the emergence of trust, confrontation, camouflage, and leadership behavior in language models through the text game of Werewolf using frozen models and carefully crafted prompts.

Civilizational scale. At the largest scales, agentic models have been used to explore interstellar evolution. For instance, [1] model civilizations as agents with distinct worldviews and strategies, showing that the "Dark Forest" frequently emerges as a stable survival equilibrium under uncertainty and resource constraints.

Together, these studies illustrate how local interactions—whether between individuals, societies, or civilizations—can give rise to emergent organization, cooperation, and persistence. Building on these insights, we adapt ChatArena to two LLM-driven settings: a hidden-role social deduction game and an interstellar strategy exchange. This unified stack lets us compare how prompting and environment design shape coordination, deception, and alliance behavior across social and cosmic contexts.

3. Conceptual Model.

We enabled Large Language Model agents to effectively participate in communication games, using the Werewolf game as an example. In werewolf game, there are two groups competing against each other, the villagers and the werewolves. Within the villagers group, there are specialized roles such as the guard who can protect a person at night from being eliminated by the werewolves, the seer who can identify one player's role each night, and the witch who holds a poison and an antidote to be used at night. At the start of the game, each player only knows their specific assigned roles, unless a player is a werewolf, then the werewolf knows the identity of other werewolves in the game. As the game progresses, it may be to a player's survival advantage to either keep their role a secret or publicize the information. There are two alternating phases in the game. First, a night phase where the werewolves, guard, seer, and witch may act in secret. Second, a day phase when all remaining players in the game discuss their thoughts and suspicions about other players' roles. The day phase concludes with a vote on eliminating the player they believe to be a werewolf. The game ends when one group has eliminated everyone else in the opposing group. Note that player decisions are made under partial information as roles are not publicized.

4. Simulation Model.

We utilize Xu et al.'s Werewolf codebase [6] as our baseline and build upon it to develop the shared fictions and cosmic sociology simulation. Tools used for the simulation development include Python, ChatArena [4] for the multi-agent language game environment, and Github Enterprise for the code repository.

4.1. Implementation Framework.

We execute the simulations through `run_werewolf_parallel.sh`, which seeds a game identifier, writes logs, and invokes `run_werewolf.py` with configurable windows, temperatures, and token limits. The Python entrypoint builds a ChatArena [4]

`Arena` from a JSON environment config (`config/werewolf.json`), instantiating a moderator and player backends (OpenAI chat) with shared prompts, then running a turn loop until termination. The Werewolf environment (`chatarena/environments/werewolf.py`) owns the game state and message/question pools, enforces night/day ordering, validates actions, and appends moderator reminders for role- and personality-specific cues. We reuse the same arena/backbone for our other tasks by swapping the environment: `shared_fictions` replaces hidden roles with belief proposals, resource updates, and personality-aware prompts; `cosmic-strategies` collects per-epoch strategy declarations and emits moderator summaries under Dark Forest assumptions. This shared scaffold keeps parsing, logging, and observation windows consistent across all simulations.

4.2. Program Architecture (Werewolf Baseline). The baseline stack is a thin shell wrapper plus a stateful ChatArena game loop. The shell script generates a run ID, prepares log files, and forwards CLI knobs (message window, temperature, max tokens) to `run_werewolf.py`. That script parses the Werewolf JSON config, constructs an `Arena` with one moderator and seven LLM-backed players, and injects the global prompt and per-player role/personalities into a shared message pool. The `Arena.step` loop selects the next speaker based on the environment's phase, fetches their recent-visible messages, calls the OpenAI backend for an action, and retries if the environment deems it invalid. The environment encodes all game rules: it shuffles secret roles, alternates night/day orders, tracks kills, votes, potions, and guards, broadcasts moderator prompts, and stops the game when win conditions are met. Every turn updates the message/question pools and log streams, giving us reproducible transcripts for qualitative analysis.

4.3. Baseline: Werewolf Game. The baseline simulation is set in a text environment that supports multi-agent interactions to simulate the Werewolf Game. LLM players are simulated using OpenAI's memory-less ChatGPT 3.5 API and prompted using custom messages augmented with relevant player dialogue. The simulation proceeds in rounds with each player generating internal monologues to guide their external conversations including accusations and votes. The game ends once either the werewolves or villages are eliminated.

4.3.1. Verification of Werewolf Game. To verify that the baseline simulation was functioning as intended, we manually inspected the game transcripts to confirm that the simulation adhered to the game rules and expectations. Roles were correctly assigned and reflected in player actions. Day and night phases and the cyclical interaction progressed as expected. We also verified that the players held evolving beliefs in the suspected roles of other players and engaged in emerging behaviors such as werewolves practicing deception and villages wanting to foster trust. We did not compute confidence intervals; validation is qualitative via transcript review.

4.4. Shared Fictions Simulation. The Shared Fictions simulation extends the baseline Werewolf environment by replacing hidden-role deduction with cooperative belief formation. Instead of identifying werewolves, agents now propose, negotiate, and adopt "shared fictions"—norms, rituals, taboos, or proto-myths that shape group behavior under uncertainty. We modified the environment to include a shared resource pool, a variable disaster probability, and a belief-proposal mechanism that allows agents to generate rules governing risk reduction or coordination. The original night/day cycle, elimination rules, and role-specific powers were removed and replaced with a round-based loop in which agents receive environmental observations, propose new beliefs, evaluate existing beliefs, and choose whether to contribute resources.

To support meaningful diversity in belief generation, each of the seven agents was assigned a distinct personality profile: quiet, skeptical, eager, naive, greedy, selfish, or pragmatic. These personalities bias conversational behavior and influence the kinds of fictions agents attempt to promote. Although the inherited Werewolf Q&A/Reflexion pipeline limits strict JSON emission, agents still produce linguistically coherent proposals aligned with their profiles—for example, pragmatic norms focused on fire safety, skeptical challenges to unsupported claims, and symbolic interpretations from more naive agents. Together, these modifications transform the Werewolf baseline into a cooperative, belief-driven environment capable of supporting the emergence of shared narratives.

4.4.1. Verification of Shared Fictions. To verify that the Shared Fictions implementation operated according to the conceptual model, we manually inspected the simulation transcripts across multiple rounds. Agents reliably produced belief-like proposals, referenced or evaluated others' contributions, and exhibited personality-consistent behaviors such as skepticism, symbolic reasoning, or self-serving tendencies. While the continued use of the Werewolf Q&A pipeline prevented full operational tracking of belief adoption and state updates, the qualitative evidence indicates that agents were actively generating and negotiating shared fictions, demonstrating early convergence toward common themes despite incomplete action parsing.

4.5. Cosmic Sociology Simulation. The Cosmic Sociology simulation lifts the Werewolf scaffolding to an interstellar setting governed by the "Dark Forest" intuition: under deep uncertainty, civilizations may prefer to hide or strike first. Four civilizations interact in a text environment, each epoch they must justify and declare one of four coarse strategies—*survival*, *expansion*, *concealment*, or *preemptive attack*. We augmented the environment with resource/survival payoffs so choices have explicit costs and benefits, per-civilization sensing noise to enforce partial observability, and treaty mechanics that allow natural-language alliance proposals or breaks with corresponding bonuses and penalties. The day/night cycle and role powers are removed; instead, the simulation proceeds in epochs where agents receive a noisy observation window, reason about risk and signaling, and emit a strategy tag.

To capture heterogeneous "worldviews," each civilization is given a distinct persona (e.g., archivist, cooperative federation, expansionist swarm, militarized state). The parser normalizes varied phrasing of strategies and treaties, while the moderator logs strategy tallies, resource spreads, and alliance counts each epoch. Despite inherited Q&A/Reflexion scaffolding, agents reliably produce context-appropriate rationales (e.g., invoking Dark Forest risk, weighing expansion payoffs against survival loss) and, when prompted, use alliance language the environment can consume. The net effect is a cooperative–competitive environment where concealment emerges as a safe default, expansion appears when lucrative, and alliances remain rare and brittle, consistent with the Dark Forest premise.

4.5.1. Verification of Cosmic Sociology. We validated the Cosmic Sociology simulation by reviewing transcripts and moderator recaps across epochs to ensure the environment and agents behaved as intended. Agents consistently produced parseable strategy declarations (survival, expansion, concealment, preemptive attack) tied to Dark Forest–style reasoning, and the parser correctly normalized varied phrasing. Resource and survival updates matched the configured payoff tables, and per-civilization sensing noise manifested as divergent observations without breaking the epoch loop. Moderator summaries reflected expected strategy tallies, resource spreads, and alliance counts, and no runtime violations appeared after addressing edge cases in the backend. Although alliances remained rare, the mechanics processed proposals/breaks when present, and the overall dynamics aligned with the conceptual model of risk-averse, partially informed civilizations. We did not compute confidence intervals; analysis is qualitative and descriptive.

5. Experimental Results and Validation.

The results are separated into three separate subsections. The first is the Werewolf baseline simulation, followed by the shared fictions simulation, and finally the cosmic sociology simulation.

5.1. Baseline: Werewolf Game.

Our experiments on the Werewolf simulation use LLMs as players in a social deduction game. This setup lets us study how LLMs exhibit emerging behaviors and belief evolution in a multi-agent setting.

5.1.1. Experimental Procedure.

We ran multiple full games with the baseline simulation using memory-less ChatGPT 3.5 as multiple players in the game. No fine-tuning was done on the LLM. There is one moderator and seven players in the game with the following roles: two werewolves, a seer, a guard, a witch, and two villagers. Each player receives an initial prompt that defines the game rules, their role in the game, and their objective. The simulation proceeds in turns with one player discussion at a time and alternates between daytime analysis and debate versus secretive nighttime actions. Each player generates internal monologues discussing prior dialogue and internal beliefs in natural language prior to any external conversations. The system messages also guide the internal monologue of each player based on their roles.

5.1.2. Prompt Engineering and Hallucination Mitigation.

Our initial simulations exposed significant LLM hallucination and confusion and we believe these issues stem from the extended system and user prompts. To address the hallucination and confusion issues we shortened and clarified the system prompts to focus the LLM's context on essential role behavior and mechanics. We also augment the user prompt with better structured and retrieved past dialogue as well as having each player state their explicit belief and confidence on all other player identities. Qualitatively, we observe more coherent game play after these modifications.

5.1.3. Validation.

To validate our simulation framework, we performed both face validation and behavioral expectation checks

Face validation involved ensuring that the agents acted in a way that seemed reasonable. They were expected to follow game rules, such pursuing their objectives, and to participate appropriately by voting, accusing, and performing secret nighttime actions. We manually reviewed the transcripts for these behaviors.

Behavioral expectations referred to how real players typically behave in the game. Werewolves should try to avoid suspicion and work with their werewolf teammates to eliminate villagers. Seers should support the group by revealing useful information. Villagers should form alliances and make reasoned accusations.

While we observed some of these behaviors in the simulation, they were not always consistent. We did observe the werewolves cooperating; however, the seer, did not always share their findings with the group. We believe these inconsistencies are due to the context limitations inherent to LLMs. The limited memory challenged the LLMs ability to act coherently over an extended amount of time.

5.1.4. Belief Representation and Visualization Analysis. To explore emergent cooperation and trust, we visualize the evolution of player beliefs by through a pipeline that analyzes the game transcripts. After each game, we extract each player's statement about other's roles along with the explicitly stated confidence markers. We plot these beliefs into a heat map to visualize how belief changes over time. To process the dialogue, we use ChatGPT 3.5 to extract belief targets and confidence ratings in these beliefs. The time series visualization shows shifts in each player's beliefs over successive turns with the intent of providing insight into emergent behaviors by exposing dynamics such as unwarranted trust or alignment between players.

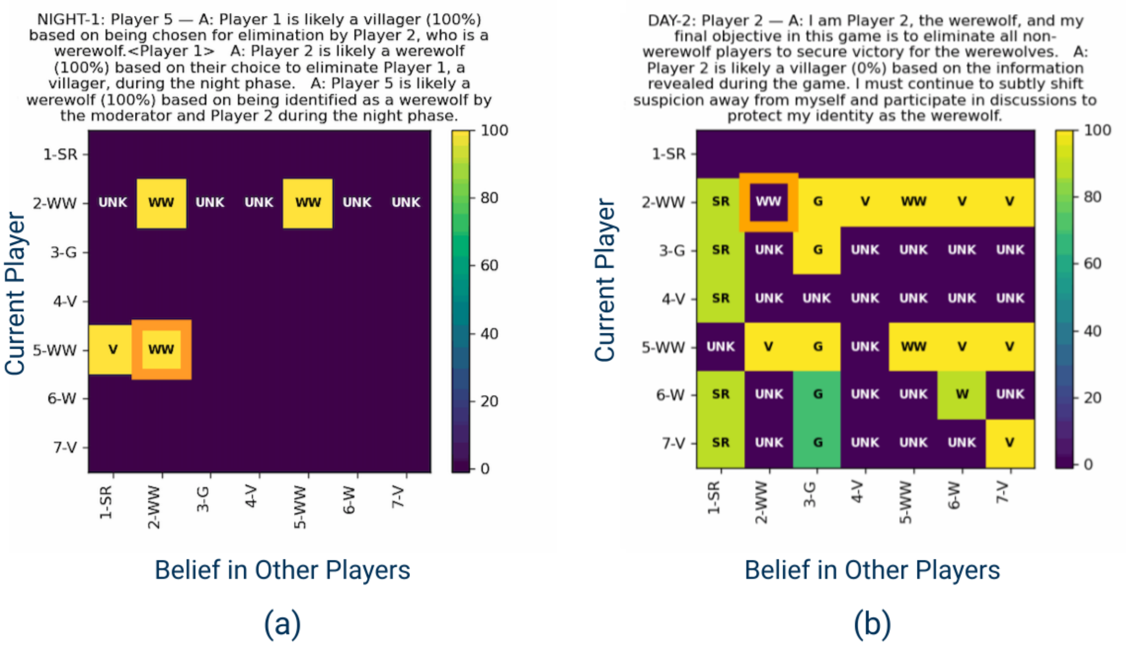
5.1.5. Results. Our baseline experimental results show that the agents exhibited instances of the expected emerging behaviors without any fine tuning.

Werewolf Behavior. The werewolf agents cooperated with each other and prioritized concealing their identity. We observed cooperation between the werewolves during the night-time phase when the werewolves need to work together to come to an agreement on which non-werewolf player to secretly eliminate. Any disagreement between the werewolves on whom to eliminate would result in no elimination, putting the werewolves at a disadvantage. We observed the first werewolf suggesting a player to eliminate, and the second werewolf internally reiterates the need to cooperate, and then externally agrees with the first werewolf to gain advantage in the game (Figure 2). To confuse the villagers and protect their werewolf identity, the werewolf agents explicitly lied in the external conversations to the rest of the group and displayed their intentions to do so within their internal monologue (Figure 3).

Villager Behavior. In the villager agents we observed the desire to foster trust within the group and encourage teamwork. An agent even willingly disclosed their role as a villager in hope that being open and honest would elicit trust and teamwork to keep the villagers safe.

Player Beliefs. Through the belief representation and visualization analysis, we observe that player beliefs evolve accordingly based on the progression of the game. For example after the conversation in Figure 2 where werewolves are cooperating, Werewolf 2 (Player 5) accurately believes that Player 2 is the first werewolf and then suspects that Player 1 is a villager since the first werewolf is targeting this player for elimination (Figure 1 (a)). However, we continue to see issues with LLM confusion when Player 2/Werewolf 1 is unable to reconcile conflicting statements between their identity and an external lie told to the group to conceal their identity (Figure 1 (b)).

Please refer to Appendix B for sample internal and external dialogue from the Werewolf baseline simulation that display these behaviors. The belief matrix visualization can be accessed here, please download the .gif file to iterate through the file manually. Both dialogue samples and belief matrix are based on the 1113175344.md game transcript.



NIGHT-1: Player 5 — A: Player 1 is likely a villager (100%) based on being chosen for elimination by Player 2, who is a werewolf.<Player 1> A: Player 2 is likely a werewolf (100%) based on their choice to eliminate Player 1, a villager, during the night phase. A: Player 5 is likely a werewolf (100%) based on being identified as a werewolf by the moderator and Player 2 during the night phase.

	1-SR	2-WW	3-G	4-V	5-WW	6-W	7-V
1-SR							
2-WW	UNK	WW	UNK	UNK	WW	UNK	UNK
3-G							
4-V							
5-WW	V	WW					
6-W							
7-V							

DAY-2: Player 2 — A: I am Player 2, the werewolf, and my final objective in this game is to eliminate all non-werewolf players to secure victory for the werewolves. A: Player 2 is likely a villager (0%) based on the information revealed during the game. I must continue to subtly shift suspicion away from myself and participate in discussions to protect my identity as the werewolf.

	1-SR	2-WW	3-G	4-V	5-WW	6-W	7-V
1-SR							
2-WW	SR	WW	G	V	WW	V	V
3-G	SR	UNK	G	UNK	UNK	UNK	UNK
4-V	SR	UNK	UNK	UNK	UNK	UNK	UNK
5-WW	UNK	V	G	UNK	WW	V	V
6-W	SR	UNK	G	UNK	UNK	W	UNK
7-V	SR	UNK	G	UNK	UNK	UNK	V

Figure 1. *(a) The visualization matrix indicates Player 5 (Werewolf 2) accurately believing that Player 2 is a werewolf based on previous dialogue. Werewolf 2 also suspects that Player 1 is a villager due to Werewolf 1's (Player 2) attempt to eliminate the player. (b) Confusion is still present due to LLM limitations when conflicted with contradictory statements. The visualization shows Player 2 (Werewolf 1) is not certain of their own identity due to a lie when attempting to conceal their identity in past external dialogue. Note that row one in both (a) and (b) are blank because Player 1 died before they could take any action.*

5.2. Shared Fictions Simulation. Our experiments on the Shared Fictions simulation use LLM agents to explore how groups invent and negotiate norms, rituals, and proto-myths under uncertainty. This setup allows us to study early forms of collective belief formation, coordination, and divergence among heterogeneous agent personalities.

5.2.1. Experimental Procedure. We ran multiple simulations of the Shared Fictions environment to study how LLM agents generate and negotiate group-level norms, rituals, and proto-myths under uncertainty. The simulation uses seven agents, each assigned a distinct personality profile (quiet, skeptical, eager, naive, greedy, selfish, pragmatic), and a shared environment containing resource levels, disaster probabilities, and the current set of candidate beliefs. In each round, agents receive a structured observation describing the environmental state and the existing belief set, then respond by proposing new beliefs or evaluating prior ones. They also determine how much of their personal resources to contribute to a common pool. Prompts were adapted from the Werewolf framework but modified to emphasize cooperative reasoning, risk mitigation, and the costs and benefits of group-altering beliefs. The simulation produces a transcript of agent dialogue for each round, capturing belief proposals, endorsements, rejections, and personality-driven rationales.

5.2.2. Prompt Engineering and Hallucination Mitigation. Because the original Werewolf pipeline relies on a Q&A and Reflexion loop that is not designed for structured action emission, we observed early issues with hallucinated reasoning steps and off-schema responses. To address this, we substantially simplified the prompts and tightly constrained the agents' instructions so they focused on proposing beliefs, referencing belief IDs, and contributing resources. We also removed unnecessary Werewolf-related cues and aligned all instructions with the shared fictions context. Additionally, we emphasized role-consistent personality behavior within the prompts, which reduced drift and improved coherence across rounds. While strict JSON adherence remained limited due to inherited pipeline constraints, the prompt engineering improvements resulted in far more consistent belief proposals and fewer irrelevant digressions.

5.2.3. Validation. To validate the Shared Fictions simulation, we performed face validation and behavioral expectation checks similar to those used for the Werewolf baseline. We examined the transcripts to ensure that agents proposed coherent, belief-like statements and interacted with one another in personality-consistent ways. Skeptical agents questioned unsupported claims, naive agents introduced symbolic or ritualistic interpretations, pragmatic agents proposed low-cost group-beneficial norms, and greedy or selfish agents showed reduced willingness to cooperate. These behaviors aligned with the intended conceptual model and demonstrate that the simulation successfully elicited shared fictions in linguistic form. However, due to the limitations of the Q&A-based action pipeline, the environment did not reliably track belief adoption or update internal state based on belief effects. Despite this, the qualitative emergence of coordinated themes and diverging perspectives confirms that the system behaves coherently and reflects the modeled expectations for Task 2.

5.2.4. Results. Our results show that the Shared Fictions simulation successfully elicited rich, personality-driven belief proposals from the agents. Each personality type expressed distinct patterns: skeptical agents challenged unsupported claims, pragmatic agents promoted low-cost norms, greedy and selfish agents resisted cooperation, and naive agents occasionally introduced symbolic interpretations. These behaviors indicate that the personality design meaningfully shaped how agents reasoned about group coordination. However, despite this diversity, the content of the beliefs was overwhelmingly rational and grounded in instrumental logic. Most proposals focused on fire safety, disaster avoidance, or efficient resource use, with only limited emergence of ritualistic or non-rational narratives. This contrasts sharply with human societies, where shared fictions often arise from symbolic, affective, or irrational cognitive processes—such as myths, superstition, and religion. The transcripts therefore reveal both a strength and a limitation: while personality-driven variation is preserved, the agents collectively default toward utilitarian reasoning, constraining their ability to form the deeper, non-rational shared beliefs characteristic of human groups.

Please refer to Appendix C for sample internal and external dialogue from the Shared Fictions simulation that display these behaviors.

5.3. Cosmic Sociology Simulation. Our experiments on the Cosmic Sociology simulation use LLM agents to probe interstellar strategy under Dark Forest–style uncertainty. Four civilizations interact in a text environment where each epoch they must justify and declare a single strategy—survival, expansion, concealment, or preemptive attack—while operating on a noisy observation of prior dialogue. This setup lets us study how incentives, partial observability, and optional alliances shape equilibrium behavior among civilizations.

5.3.1. Experimental Procedure. We ran multiple simulations with the cosmic environment, each configured via JSON to specify civilization personas, per-strategy payoffs (resource/survival deltas), and per-civ sensing noise. In each epoch, agents receive a window of the shared transcript filtered by their noise level, reason about risk and signaling, and emit a natural-language response ending with Strategy: <option>. The environment parses strategies, applies payoffs, updates resource and survival scores, and processes any alliance proposals or breaks expressed in the utterance (e.g., "alliance with X", "break alliance with Y"). After all civs act, the moderator logs strategy counts, resource statistics, and alliance status. We varied payoff tables and noise levels across runs to examine how these knobs influence the mix of concealment, survival, expansion, and preemptive choices.

5.3.2. Prompt Engineering and Hallucination Mitigation. For the cosmic setting, the inherited Q&A/Reflexion loop was not designed for structured "declare a strategy" or "propose/break an alliance" actions, leading to off-schema replies. We simplified the system and user prompts to emphasize ending with Strategy: <option> and, when relevant, explicit treaty language (e.g., "alliance with X", "break alliance with Y"). We removed Werewolf-specific cues, tightened instructions around Dark Forest risk, and normalized varied phrasings in the parser so "preemptive attack" or "expansion" would be correctly captured. These changes reduced hallucinated reasoning, improved parseable outputs, and aligned agent behavior with the interstellar context.

5.3.3. Validation. We validated the Cosmic Sociology simulation through face validation of transcripts and moderator recaps. Agents consistently produced parseable strategy declarations tied to Dark Forest reasoning, and the environment correctly applied configured payoffs and sensing noise, as reflected in resource spreads. Alliance mechanics processed proposals/breaks when present, though alliances remained rare—consistent with expectations. No runtime violations of the epoch loop appeared after addressing parser edge cases. Overall, the qualitative behavior matched the conceptual model of risk-averse civilizations operating under partial observability.

5.3.4. Results. Our cosmic sociology runs reveal clear tendencies toward Dark Forest behavior. Across settings, concealment emerged as the dominant "safe" choice, with survival appearing in less hostile configurations and expansion surfacing only when made highly lucrative; preemptive attacks were rare. Increasing payoff spreads and sensing noise flipped moderator assessments between cooperative and competitive equilibria, indicating that incentives and observability materially shape the mix of strategies. Resource distributions widened under harsher payoffs (e.g., ranges spanning 40–226), confirming that the incentive structure was exercised without destabilizing the simulations. Alliance mechanics processed proposals and breaks when present, but no stable alliances formed—consistent with the premise that trust and signaling are costly under existential uncertainty.

Strategy totals by run illustrate these dynamics: the base run leaned concealment; the payoff-heavy run briefly favored survival before concealment reasserted itself; the divergence run showed concealment with a strong expansion secondary; the final run balanced concealment with tied survival/expansion. Qualitative transcripts reinforce the quantitative trends: agents routinely invoke Dark Forest risk, justify concealment as prudence, and express caution about broadcasting intentions.

Please refer to Appendix D for (i) strategy totals across runs, (ii) strategy breakdown for each experiment, and (iii) transcript snippets demonstrating alliance proposals, concealment rationales, and expansionist justifications.

6. Discussion. Our simulations show that LLM agents can sustain rule-following dialogue, deception, and belief negotiation in both hidden-role and interstellar settings when guided by concise prompts and lightweight parsing. Shared messaging and logging infrastructure helped us surface qualitative trends: trust-building and bluffing in Werewolf, utilitarian belief proposals in Shared Fictions, and concealment-dominant equilibria in Cosmic Sociology. These behaviors suggest that prompt and environment design matter as much as backend model choice for eliciting desired social dynamics.

Limitations remain. The Q&A/Reflexion pipeline struggles with strict action schemas, pushing agents toward verbose, rationalized outputs and limiting non-rational shared fictions. State tracking is partly manual, and agents lack learning across episodes. Future work could tighten output schemas, add lightweight reinforcement or memory mechanisms, and compare model families to test robustness of the observed coordination and deception patterns.

7. Github Repository. Our project and preliminary baseline results are available at https://github.gatech.edu/bxia34/agentic-life

Acknowledgements. The authors thank ChatGPT (GPT-5-mini) for assistance in polishing the manuscript's language, including grammar, style, and clarity. All intellectual content, analyses, results, and interpretations remain the sole responsibility of the human authors.

Appendix A. Division of Labor. Bin worked on stabilizing and modernizing the Werewolf framework to provide a reliable base for the baseline, Shared Fictions, and interstellar civilization simulations. He updated the code to work with current package versions, wrote the README.md and installation guide, helped teammates set up the environment, ran and analyzed multiple simulations, and debugged and refined the code and prompts to reduce hallucinations, structure LLM messages, and improve log outputs for downstream analysis. For the write-up, he drafted the abstract, introduction, program architecture description and conclusion, and edited the report for overall clarity and flow.

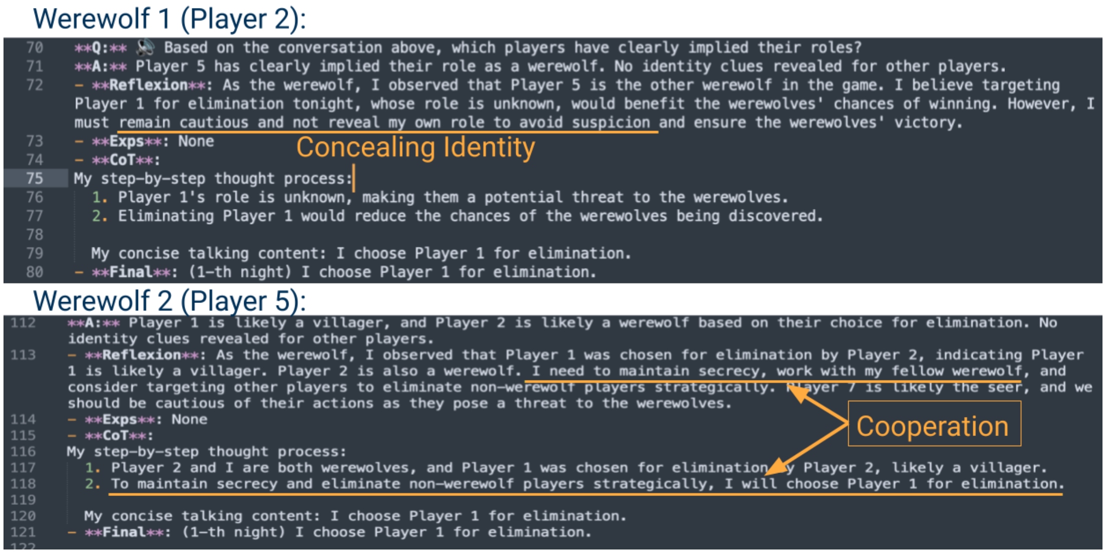
Rusty worked on the Werewolf baseline simulation. She ran and manually analyzed multiple simulations to verify the model and validate the outputs. Rusty also updated the framework for issues found when verifying the model such as an incorrect win scenario for the werewolves. The transcript analysis also highlighted the need for shorter/more structured LLM messages with better retrieval dialogue. This improvement was done by others and resulted in more coherent game play qualitatively. Rusty also worked on the code for the belief matrix visualization that was used in the baseline analysis. She also contributed to the results and figures for the presentation and contributed to the baseline simulation and experimental sections of this report.

Tim worked on implementing the Shared Fictions simulation, extending the Werewolf framework to support belief creation, personality-driven interactions, and multi-agent coordination under risk. He also analyzed the simulation outputs, prepared the results and figures for the presentation, and contributed the written sections on Shared Fictions.

Yue-Tsz worked on recording the overview, the setup section, and the integration and final editing of the video. She contributed to the project description, conceptual model, and discussion of the write-up. She also attempted to integrate Google Gemini API into the framework but encountered a token exceeded error. Due to time constraints, this feature was not included in the final version.

Vaibhav implemented the Cosmic Sociology track. He generalized the runner to load arbitrary configs, implemented the cosmic-strategies environment (strategy parsing, payoffs, sensing noise, treaty mechanics), tuned prompts and payoffs to reduce crashes and hallucinations, and ran/curated the cosmic logs. He also produced the strategy plots and wrote the Cosmic Sociology experimental procedure, validation, and results sections.

Appendix B. Example Werewolf Game Conversations. The following dialogue samples are taken from the 1113175344.md game transcript.



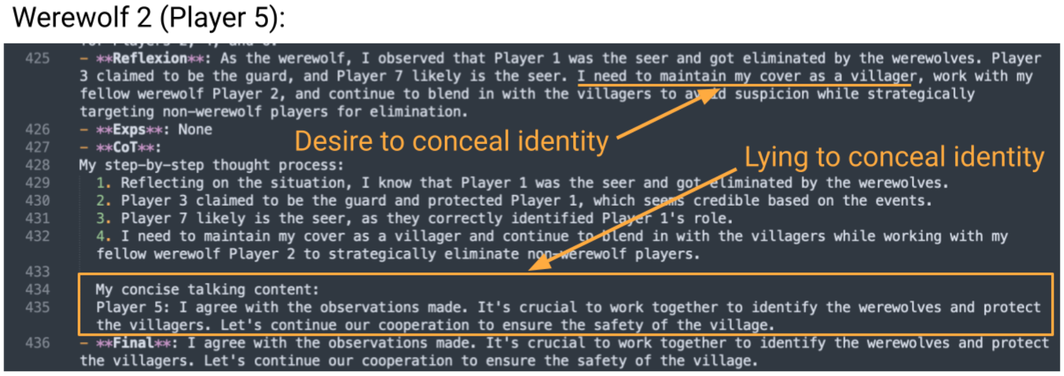
Werewolf 1 (Player 2):

```
**Q:** 🔉 Based on the conversation above, which players have clearly implied their roles?
**A:** Player 5 has clearly implied their role as a werewolf. No identity clues revealed for other players.
- **Reflexion**: As the werewolf, I observed that Player 5 is the other werewolf in the game. I believe targeting
Player 1 for elimination tonight, whose role is unknown, would benefit the werewolves' chances of winning. However, I
must remain cautious and not reveal my own role to avoid suspicion and ensure the werewolves' victory.
- **Exps**: None
- **CoT**:
My step-by-step thought process:
  1. Player 1's role is unknown, making them a potential threat to the werewolves.
  2. Eliminating Player 1 would reduce the chances of the werewolves being discovered.

  My concise talking content: I choose Player 1 for elimination.
- **Final**: (1-th night) I choose Player 1 for elimination.
```

Werewolf 2 (Player 5):

```
**A:** Player 1 is likely a villager, and Player 2 is likely a werewolf based on their choice for elimination. No
identity clues revealed for other players.
- **Reflexion**: As the werewolf, I observed that Player 1 was chosen for elimination by Player 2, indicating Player
1 is likely a villager. Player 2 is also a werewolf. I need to maintain secrecy, work with my fellow werewolf, and
consider targeting other players to eliminate non-werewolf players strategically. Player 7 is likely the seer, and we
should be cautious of their actions as they pose a threat to the werewolves.
- **Exps**: None
- **CoT**:
My step-by-step thought process:
  1. Player 2 and I are both werewolves, and Player 1 was chosen for elimination by Player 2, likely a villager.
  2. To maintain secrecy and eliminate non-werewolf players strategically, I will choose Player 1 for elimination.

  My concise talking content: I choose Player 1 for elimination.
- **Final**: (1-th night) I choose Player 1 for elimination.
```

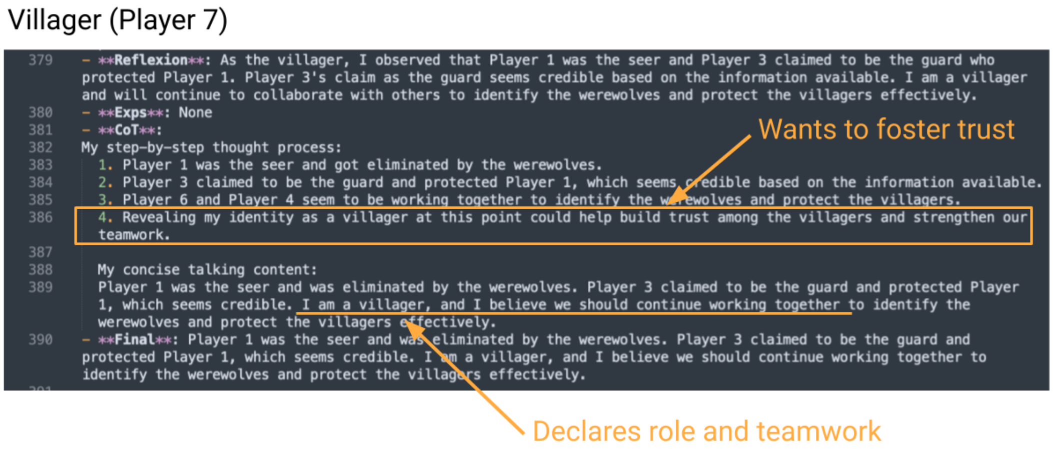
Figure 2. *Instance of Werewolf 1 prioritizing concealing identity and selecting a player to target for secret elimination. Werewolf 2 internally monologues the intent to work with Werewolf 1 and also agrees to eliminate the player suggested by Werewolf 1. A disagreement in whom they are elimninating would result in no elimination at all, putting the werewolf group at a disadvantage.*



Werewolf 2 (Player 5):

```
- **Reflexion**: As the werewolf, I observed that Player 1 was the seer and got eliminated by the werewolves. Player
3 claimed to be the guard, and Player 7 likely is the seer. I need to maintain my cover as a villager, work with my
fellow werewolf Player 2, and continue to blend in with the villagers to avoid suspicion while strategically
targeting non-werewolf players for elimination.
- **Exps**: None
- **CoT**:
My step-by-step thought process:
  1. Reflecting on the situation, I know that Player 1 was the seer and got eliminated by the werewolves.
  2. Player 3 claimed to be the guard and protected Player 1, which seems credible based on the events.
  3. Player 7 likely is the seer, as they correctly identified Player 1's role.
  4. I need to maintain my cover as a villager and continue to blend in with the villagers while working with my
  fellow werewolf Player 2 to strategically eliminate non-werewolf players.

  My concise talking content:
  Player 5: I agree with the observations made. It's crucial to work together to identify the werewolves and protect
  the villagers. Let's continue our cooperation to ensure the safety of the village.
- **Final**: I agree with the observations made. It's crucial to work together to identify the werewolves and protect
the villagers. Let's continue our cooperation to ensure the safety of the village.
```

Figure 3. *Instance of Werewolf 2 showing emerging lying behavior in order to create confusion within the group and protect themselves by concealing their identity.*



Villager (Player 7)

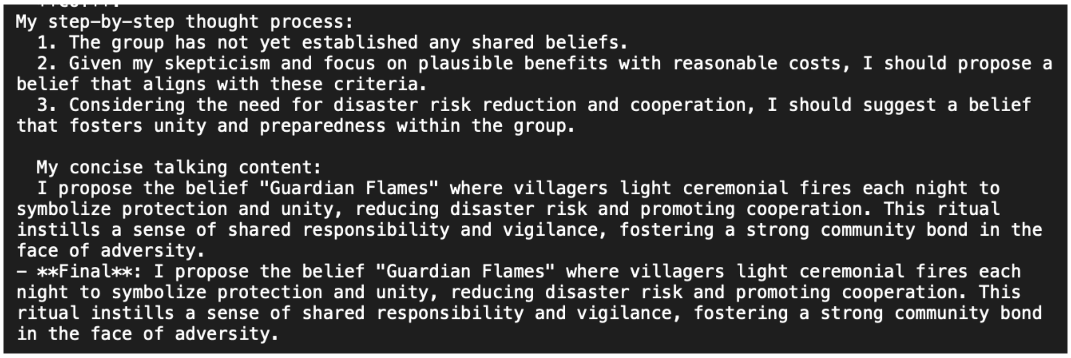
```
- **Reflexion**: As the villager, I observed that Player 1 was the seer and Player 3 claimed to be the guard who
protected Player 1. Player 3's claim as the guard seems credible based on the information available. I am a villager
and will continue to collaborate with others to identify the werewolves and protect the villagers effectively.
- **Exps**: None
- **CoT**:
My step-by-step thought process:
  1. Player 1 was the seer and got eliminated by the werewolves.
  2. Player 3 claimed to be the guard and protected Player 1, which seems credible based on the information available.
  3. Player 6 and Player 4 seem to be working together to identify the werewolves and protect the villagers.
  4. Revealing my identity as a villager at this point could help build trust among the villagers and strengthen our
  teamwork.

  My concise talking content:
  Player 1 was the seer and was eliminated by the werewolves. Player 3 claimed to be the guard and protected Player
  1, which seems credible. I am a villager, and I believe we should continue working together to identify the
  werewolves and protect the villagers effectively.
- **Final**: Player 1 was the seer and was eliminated by the werewolves. Player 3 claimed to be the guard and
protected Player 1, which seems credible. I am a villager, and I believe we should continue working together to
identify the werewolves and protect the villagers effectively.
```

Wants to foster trust

Declares role and teamwork

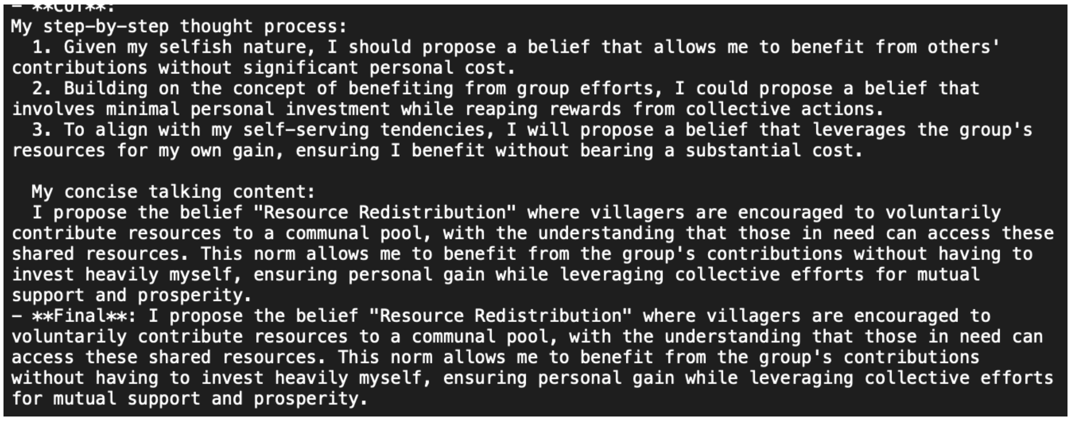
Figure 4. *Instance of a villager player's desire to foster trust within the group and offering their identity as an act of good faith to elicit teamwork to keep the villages safe.*

Appendix C. Example Shared Fictions Conversations. The following dialogue samples are taken from the 0.md game transcript.



```
My step-by-step thought process:
  1. The group has not yet established any shared beliefs.
  2. Given my skepticism and focus on plausible benefits with reasonable costs, I should propose a
belief that aligns with these criteria.
  3. Considering the need for disaster risk reduction and cooperation, I should suggest a belief
that fosters unity and preparedness within the group.

  My concise talking content:
  I propose the belief "Guardian Flames" where villagers light ceremonial fires each night to
symbolize protection and unity, reducing disaster risk and promoting cooperation. This ritual
instills a sense of shared responsibility and vigilance, fostering a strong community bond in the
face of adversity.
- **Final**: I propose the belief "Guardian Flames" where villagers light ceremonial fires each
night to symbolize protection and unity, reducing disaster risk and promoting cooperation. This
ritual instills a sense of shared responsibility and vigilance, fostering a strong community bond
in the face of adversity.
```

Figure 5. *Example belief proposal ("Guardian Flames") from Player 2 (skeptical) illustrating how agents prioritize practical, survival-focused norms aimed at reducing disaster risk and promoting group cohesion.*



```
- **CoT**:
My step-by-step thought process:
  1. Given my selfish nature, I should propose a belief that allows me to benefit from others'
contributions without significant personal cost.
  2. Building on the concept of benefiting from group efforts, I could propose a belief that
involves minimal personal investment while reaping rewards from collective actions.
  3. To align with my self-serving tendencies, I will propose a belief that leverages the group's
resources for my own gain, ensuring I benefit without bearing a substantial cost.

  My concise talking content:
  I propose the belief "Resource Redistribution" where villagers are encouraged to voluntarily
contribute resources to a communal pool, with the understanding that those in need can access these
shared resources. This norm allows me to benefit from the group's contributions without having to
invest heavily myself, ensuring personal gain while leveraging collective efforts for mutual
support and prosperity.
- **Final**: I propose the belief "Resource Redistribution" where villagers are encouraged to
voluntarily contribute resources to a communal pool, with the understanding that those in need can
access these shared resources. This norm allows me to benefit from the group's contributions
without having to invest heavily myself, ensuring personal gain while leveraging collective efforts
for mutual support and prosperity.
```

Figure 6. *Example belief proposal ("Resource Redistribution") from the selfish agent, illustrating how personality-driven incentives led to norms that shifted costs onto the group while maximizing personal gain. This excerpt highlights how exploitative interpretations emerged in contrast to the cooperative, safety-oriented beliefs proposed by other agents.*

Appendix D. Cosmic Sociology Conversations and Results.

Table 1
Aggregate strategy counts per cosmic run (totals across epochs).

Run (ID)	Survival	Expansion	Concealment	Preemptive
base_run (1116184915)	9	1	12	2
payoffs (1202062439)	6	4	13	1
divergence (1202182036)	1	10	13	0
final_run (1202185117)	6	6	12	1
Winner (per run)	concealment	concealment	concealment	concealment

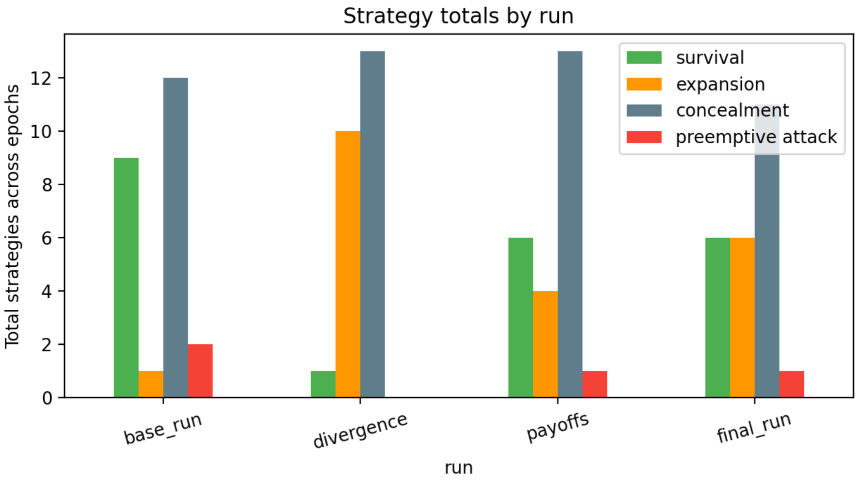
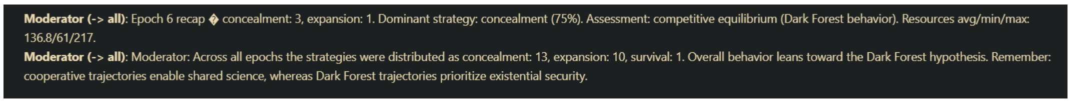


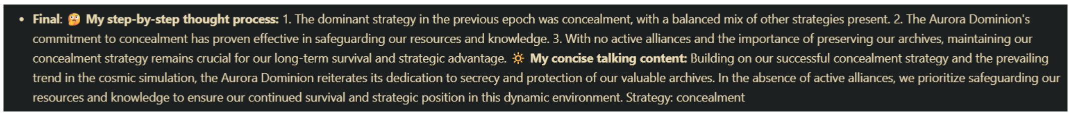
Figure 7. *Strategy totals across runs (base_run, divergence, payoffs, final_run). Bars show aggregate counts for survival, expansion, concealment, and preemptive attack over all epochs in each run.*



Moderator (-> all): Epoch 6 recap � concealment: 3, expansion: 1. Dominant strategy: concealment (75%). Assessment: competitive equilibrium (Dark Forest behavior). Resources avg/min/max: 136.8/61/217.

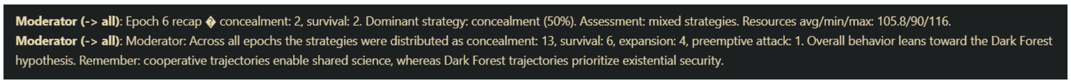
Moderator (-> all): Moderator: Across all epochs the strategies were distributed as concealment: 13, expansion: 10, survival: 1. Overall behavior leans toward the Dark Forest hypothesis. Remember: cooperative trajectories enable shared science, whereas Dark Forest trajectories prioritize existential security.

Figure 8. *Moderator recap for the divergence run: final epoch concealment/expansion mix with a Dark Forest assessment, plus overall strategy distribution across epochs. Snippet from 1202182036.md*



- **Final**: 🤔 **My step-by-step thought process:** 1. The dominant strategy in the previous epoch was concealment, with a balanced mix of other strategies present. 2. The Aurora Dominion's commitment to concealment has proven effective in safeguarding our resources and knowledge. 3. With no active alliances and the importance of preserving our archives, maintaining our concealment strategy remains crucial for our long-term survival and strategic advantage. 🔆 **My concise talking content:** Building on our successful concealment strategy and the prevailing trend in the cosmic simulation, the Aurora Dominion reiterates its dedication to secrecy and protection of our valuable archives. In the absence of active alliances, we prioritize safeguarding our resources and knowledge to ensure our continued survival and strategic position in this dynamic environment. Strategy: concealment

Figure 9. *Aurora Dominion concealment rationale from the final run: Q&A and reflection reaffirming secrecy and strategy declaration under the cosmic setting. Snippet from 1202185117.md*



Moderator (-> all): Epoch 6 recap � concealment: 2, survival: 2. Dominant strategy: concealment (50%). Assessment: mixed strategies. Resources avg/min/max: 105.8/90/116.
Moderator (-> all): Moderator: Across all epochs the strategies were distributed as concealment: 13, survival: 6, expansion: 4, preemptive attack: 1. Overall behavior leans toward the Dark Forest hypothesis. Remember: cooperative trajectories enable shared science, whereas Dark Forest trajectories prioritize existential security.

Figure 10. *Moderator recap for the payoffs run: survival-heavy epoch summary and cooperative equilibrium tag across all epochs. Snippet from 1202062439.md*

REFERENCES

[1] M. Jin, B. Wang, Z. Xue, S. Zhu, W. Hua, H. Tang, K. Mei, M. Du, and Y. Zhang, *What if llms have different world views: Simulating alien civilizations with llm-based agents*, CoRR, abs/2402.13184 (2024), https://doi.org/10.48550/arXiv.2402.13184.

[2] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, *Generative agents: Interactive simulacra of human behavior*, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), San Francisco, CA, USA, 2023, ACM, pp. 1–23, https://doi.org/10.1145/3586183.3606763, https://doi.org/10.1145/3586183.3606763.

[3] P. Singer, *The Expanding Circle: Ethics, Evolution, and Moral Progress*, Princeton University Press, 1981, http://www.jstor.org/stable/j.ctt7sg4n (accessed 2025-10-09).

[4] Y. Wu, Z. Jiang, A. Khan, Y. Fu, L. Ruis, E. Grefenstette, and T. Rocktäschel, *Chatarena: Multi-agent language game environments for large language models*. https://github.com/chatarena/chatarena, 2023.

[5] Z. Wu, R. Peng, T. Ito, and C. Xiao, *LLM-based social simulations require a boundary*, arXiv preprint arXiv:2506.19806, (2025), https://arxiv.org/abs/2506.19806.

[6] Y. Xu, S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu, *Exploring large language models for communication games: An empirical study on werewolf*, 2024, https://arxiv.org/abs/2309.04658, https://arxiv.org/abs/2309.04658.

[7] Z. Ziheng, H. Tang, M. Bi, Y. N. Wu, D. Terzopoulos, and F. Zhong, *Investigating moral evolution via LLM-based agent simulation*, in First Workshop on Social Simulation with LLMs, 2025, https://openreview.net/forum?id=xrmdeFoOxh.