AGENT SOCIODRAMA: SOCIAL-DYNAMICS-GUIDED PERSONA RECASTING FOR LLM AGENT TEAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents are increasingly deployed as teams, yet they fail in characteristically social ways: information holders stay silent, dominant speakers anchor groups on wrong answers, and norm violations go unsanctioned. We present Agent Sociodrama, a framework for modeling and validating evolving social dynamics in LLM agent teams. Under fixed personas, agents' emotional states and directed, asymmetric relationships are updated every round by a unified external LLM judge and shape their subsequent interactions. Across six LLM backbones and five social scenarios, we assess social fidelity through classical social-theory patterns and agreement with human judgments. The resulting trajectories exhibit structured social patterns and model-specific differences. We then introduce Trajectory-Guided Persona Recasting (TPR), which diagnoses group-dynamic failures from interaction transcripts and social trajectories and recasts agent personas before rerunning the task. Across four collaborative benchmarks, TPR improves performance from both heterogeneous and randomly initialized agent persona compositions, yielding up to 83.3%p of accuracy improvement over the initial casting. On representative distributed-information cooperation and mixed-motive bargaining tasks, we outperform SOTOPIA and AgentVerse in three of four comparisons, by up to 37.5%p in Pareto rate. These results show that validated social trajectories provide actionable signals for repairing LLM agent teams.

1 INTRODUCTION

Large language model (LLM) agents are increasingly deployed in teams (Wu et al., 2024; Hong et al., 2024): they debate answers (Du et al., 2024), pool distributed evidence, teach one another, negotiate, and manage shared resources. Yet sustained multi-agent interaction requires more than assigning each agent an initial role. For a team to function as a social system over time, its members must form, maintain, and revise relationships in ways that remain coherent with the interaction. Existing agent systems often encode social context through static role instructions or unstructured memory, without an explicit and validated account of how directed relationships evolve over time.

We take a different route. We argue that longitudinal role play requires treating a team not as a collection of fixed personas, but as a dynamic social system. A persona specifies who an agent is, but not how its relationships with others develop over time, including whom it comes to trust, avoid, align with, or oppose. These relationships must therefore be represented as evolving interaction states. The key challenge is not merely to model changing relations, but to ensure that those changes remain socially faithful over time.

To instantiate this idea, we propose Agent Sociodrama, a longitudinal role-playing framework for LLM agent teams. Agent Sociodrama operationalizes stable personas together with evolving emotional and relational states, providing a structured account of how a team's social configuration develops over repeated interaction. These dynamics also matter for task-oriented teams. A growing body of evidence suggests that LLM teams often fail in ways that resemble classic small-group pathologies rather than limits of individual model capability: groups that collectively possess the information required for a correct decision can still converge on an incorrect one (Pappu et al., 2026; Li et al., 2026), while individually capable agents can collectively deplete a shared resource until cooperation collapses (Piatti et al., 2024). These failures have a distinctly social character. Over repeated interaction, overlooked members can withdraw, early deference to dominant voices

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

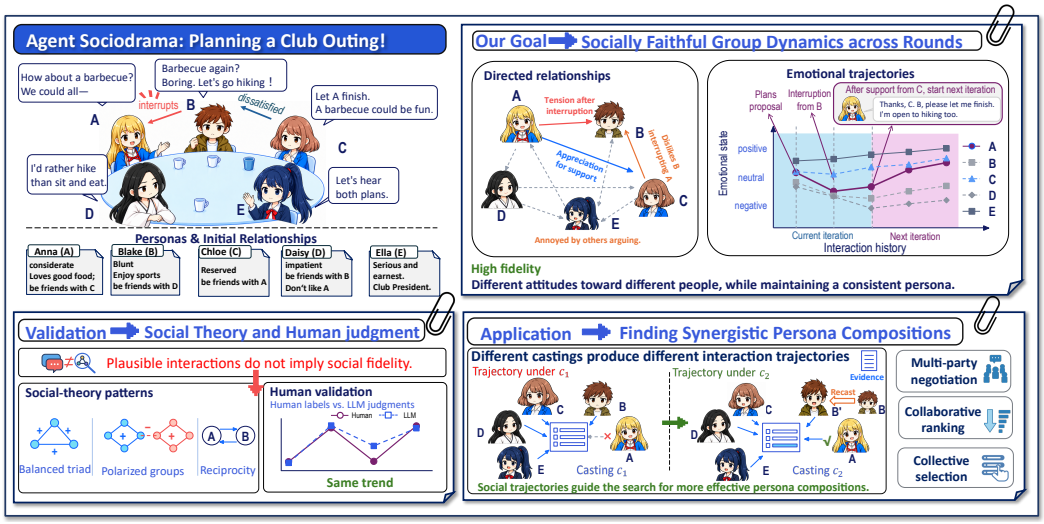

Figure 1: **Motivation and overview of Agent Sociodrama.** Agent Sociodrama models multi-agent interaction through stable personas, evolving emotions, and directed relationships. We assess the social fidelity of these dynamics against social-theory patterns and human judgments, and use social trajectories to guide team composition optimization for better collaboration.

can harden into consensus, and unchallenged violations can progressively erode cooperation. Yet dominant approaches to improving multi-agent systems optimize prompts, topologies, or workflows (Zhou et al., 2026; Zhang et al., 2025), treating the team as a black box and leaving the social mechanisms behind these failures largely unexamined. Building on Agent Sociodrama's explicit social trajectories, we further introduce Trajectory-Guided Persona Recasting (TPR), a casting controller that diagnoses team composition and recasts personas to seek more effective task-oriented collaboration.

Our empirical study has two parts. First, we assess the social fidelity of Agent Sociodrama across six LLM drivers and five social scenarios, using a shared external LLM judge with human cross-validation. We test whether the resulting trajectories recover four established regularities of social interaction: structural balance (Heider, 1946; Cartwright & Harary, 1956), reciprocity (Gouldner, 1960), conflict-driven polarization (Krackhardt & Stern, 1988), and divergent repair-versus-escalation dynamics (Gottman, 1994; Pruitt & Kim, 2004). These analyses establish Agent Sociodrama as a socially faithful framework for longitudinal multi-agent role play. Second, across four collaborative task families, our controller converts evolving social states into persona recasting decisions that substantially improve team performance. These results demonstrate that socially faithful trajectory modeling can directly inform persona recasting and improve multi-agent collaboration.

Our contributions are:

- **A dynamic multi-agent role-playing framework for LLM teams.** Agent Sociodrama jointly models stable personas, evolving emotional states, and directed, asymmetric relational states, allowing multi-agent interaction to unfold as a sustained social process.
- **Cross-model validation of social fidelity.** Across six LLMs and five social scenarios, we validate Agent Sociodrama against social theory and human evaluation, showing that its long-horizon relational dynamics are structured, interpretable, and socially coherent.
- **Trajectory-guided team composition optimization.** We use social trajectories to guide persona casting, making team chemistry directly controllable. Across four collaborative task families, our controller substantially improves team performance.

2 Related Work

Social simulation and coordination dynamics. LLM social simulation spans memory-based societies (Park et al., 2023), social-intelligence evaluation (Zhou et al., 2024), coordinated role play (Wang et al., 2025), and social reasoning and longitudinal interaction (Mou et al., 2025; Goel & Zhu, 2025). LLM teams exhibit information-pooling failures (Stasser & Titus, 1985), underused

expertise (Pappu et al., 2026), unshared evidence (Li et al., 2026), and coordination errors (Cemri et al., 2025). BenchForm examines conformity (Weng et al., 2025), while MultiAgentBench evaluates collaboration alongside task completion (Zhu et al., 2025). Drawing on affect (Russell, 2003), reciprocity (Gouldner, 1960), group boundaries (Krackhardt & Stern, 1988), and conflict dynamics (Pruitt & Kim, 2004), Agent Sociodrama models evolving affect and directed relationships as persona-conditioned states. Their trajectories complement transcripts in diagnosing team dynamics and guiding persona recasting.

Persona composition and multi-agent optimization. Personality research motivates stable behavioral specifications (McCrae & Costa, 1987), while persona composition effects depend on task structure (Keluskar et al., 2026). MetaGPT (Hong et al., 2024) and AutoGen (Wu et al., 2024) provide role-based coordination, whereas EvoAgent evolves agent configurations (Yuan et al., 2025). Automated design optimizes computational graphs (Zhuge et al., 2024), prompts and topologies (Zhou et al., 2026), workflows (Zhang et al., 2025), and code-defined agents (Hu et al., 2025); learned orchestration adapts agent activation (Dang et al., 2025). Agent Sociodrama searches persona assignments from a fixed library between task runs. Social trajectories guide proposed changes, while task outcomes determine configuration selection, linking persona composition to each team's observed interaction dynamics.

3 Agent Sociodrama: Simulation with Measurable Social State

We study multi-agent social simulation with dynamically evolving relationships. Given N agents $\mathcal{A} = \{A_1, \ldots, A_N\}$, each with a fixed persona, they interact over T rounds within a shared narrative scenario. While prior work on multi-agent collaboration or debate either assumes static inter-agent relationships (Li et al., 2023; Hong et al., 2024) or does not explicitly model them (Du et al., 2024; Wang et al., 2024), we let every interaction update the agents' relationships and emotional states, which in turn shape how they communicate in subsequent rounds. This feedback process gives rise to unscripted social dynamics, such as shifting alliances, declining trust, and growing disagreement.

The simulation takes as input a persona for each agent (P), initial emotional states (S^0), pairwise relationships (R^0), and a sequence of scenario events. The persona captures the stable core of an agent, its temperament, values, and characteristic style of interaction, and remains fixed throughout a run, whereas the agent's actions, emotional states, and directed relationships evolve continually across rounds. We collect these evolving quantities into an interaction trajectory $Y = \left(a^t,\ S^t,\ R^t\right)_{t=1}^{T}$, where $a^t = \{a_i^t\}_{i=1}^N$ and $S^t = \{S_i^t\}_{i=1}^N$ collect the actions and emotional states of all agents, and $R^t = \{R_{ij}^t\}_{i \neq j}$ collects all directed relationships at round t. In each round, an agent updates its emotions and relationships based on visible events, guided by its persona. These updated states and the visible history then drive its next actions.

3.1 Persona-Conditioned Agent State

At round t, each agent $i \in \{1, \ldots, N\}$ is represented by the state $(P_i, S_i^t, R_{i\cdot}^t)$, where P_i is a time-invariant persona, S_i^t is a dynamic emotional state, and $R_{i\cdot}^t = \{R_{ij}^t\}_{j \neq i}$ collects its directed relationships toward all other agents; the dot denotes all valid values of the second index j. We omit the superscript t when the round is clear from context.

Time-invariant persona. Assigned before an episode and fixed throughout it, each persona $P_i = (\textit{identity}_i, \textit{traits}_i, \textit{boundary}_i)$ specifies the agent's name and social role, behavioral tendencies, and behavioral constraints, respectively. The full persona guides action generation, while its traits also condition emotional and relational updates.

Dynamic emotional state. $S_i^t \in [0, 1]$ represents the current affective valence of agent i, where 0.5 represents a neutral state, with lower and higher values indicating negative and positive affect, respectively. It evolves from the agent's prior state and the events visible to it:

$$S_i^{t+1} = U_S\left(P_i, S_i^t, E_i^t\right), \tag{1}$$

where E_i^t denotes the events and interactions visible to agent i in round t, as will be detailed in Sec. 3.2. Persona conditioning allows the same event to elicit different emotional responses across agents, while incorporating S_i^t preserves continuity over time.

Directed inter-agent relationship. Relationships are directed and need not be reciprocal: one agent may trust another even when the latter remains suspicious of the former. We represent agent i's current relational stance toward agent j at round t as: $R_{ij}^t = \big(\text{fav}_{ij}^t,\ \text{status}_{ij}^t,\ M_{ij}^t\big)$. The favorability score $\text{fav}_{ij}^t \in [-1, 1]$ tracks gradual changes in agent i's attitude toward agent j. The categorical status records one of four discrete relationship categories, *neutral*, *ally*, *rival*, or *enemy*. The memory M_{ij}^t retains salient interaction events that ground subsequent relational updates. For each ordered pair (i, j), let $E_{ij}^t \subseteq E_i^t$ denote the events relevant to their relationship: direct interactions between i and j, whether public or private, and other public actions by j observed by i. The relationship is then updated as

$$R_{ij}^{t+1} = U_R\big(P_i, R_{ij}^t, E_{ij}^t\big). \tag{2}$$

The concrete instantiation of U_S and U_R within the per-round loop is described in Sec. 3.3.

3.2 Events, Actions, and Visibility

A scenario consists of ordered **scenario events**, $\mathcal{E} = (e^1, \ldots, e^T)$. Each round t introduces e^t as new context, to which agents respond while interacting with one another. Public and private interactions give agents different information, and each responds based on the events and actions visible to it. Agent A_i produces

$$a_i = \big(\tau_i, v_i, u_i\big), \qquad \tau_i \in (\mathcal{A} \setminus \{A_i\}) \cup \{\text{ALL}, \text{NONE}\}, \tag{3}$$

where u_i is the utterance, τ_i its recipient, and $v_i \in \{\text{PUBLIC}, \text{PRIVATE}, \text{NONE}\}$ its visibility. The targets A_j, ALL, and NONE denote an individual recipient, the group, and silence. A public utterance is visible to all agents even when addressed to A_j; a private utterance is visible only to its speaker and designated recipient. Valid combinations require $\tau_i = \text{NONE}$ if and only if $v_i = \text{NONE}$, and $\tau_i = \text{ALL}$ implies $v_i = \text{PUBLIC}$. The observer set is

$$V(a_i) = \begin{cases} \mathcal{A}, & v_i = \text{PUBLIC}, \\ \{A_i, A_j\}, & \tau_i = A_j, \text{ and } v_i = \text{PRIVATE} \\ \varnothing, & v_i = \text{NONE}. \end{cases} \tag{4}$$

With the scenario event e visible to all, agent A_i observes

$$E_i = \{e\} \ \cup\ \{a_k \mid A_i \in V(a_k)\}\,. \tag{5}$$

Emotional and relational dynamics can support coordination or amplify conflict without being visible in public behavior. The resulting S and R trajectories allow the casting controller (Sec. 4) to assess these dynamics and identify more effective agent compositions.

3.3 Per-Round Interaction and State Updates

Each round proceeds in three stages (Fig. 2): *context assembly*, where code gathers the history visible to each agent; *two-phase action generation*, where agents first decide whether and how to speak and then speak in turn; and *state update*, where each agent's emotional state and relationships are revised from what it observed. The remainder of this section details each stage.

Context assembly. Each agent is prompted with its local state and interaction history, augmented by the information available at the current point in the episode. The local state consists of its persona P_i, emotional state S_i^t, and directed relationships $R_{i\cdot}^t$ with their associated events. The remaining context comprises globally visible scenario events and prior public utterances, actions already taken in the current round, and private messages visible only to the agents involved.

Two-phase action generation. A fixed or random speaking order of agents would favor later speakers, as they observe more context before responding. We therefore decouple *whether and how* to speak from *what* to say. In Phase 1, every agent independently receives its state (P, S, R) and the shared context, and output a *communicative urgency* $u \in [0, 10]$, a *mode* (public, private, or silence),

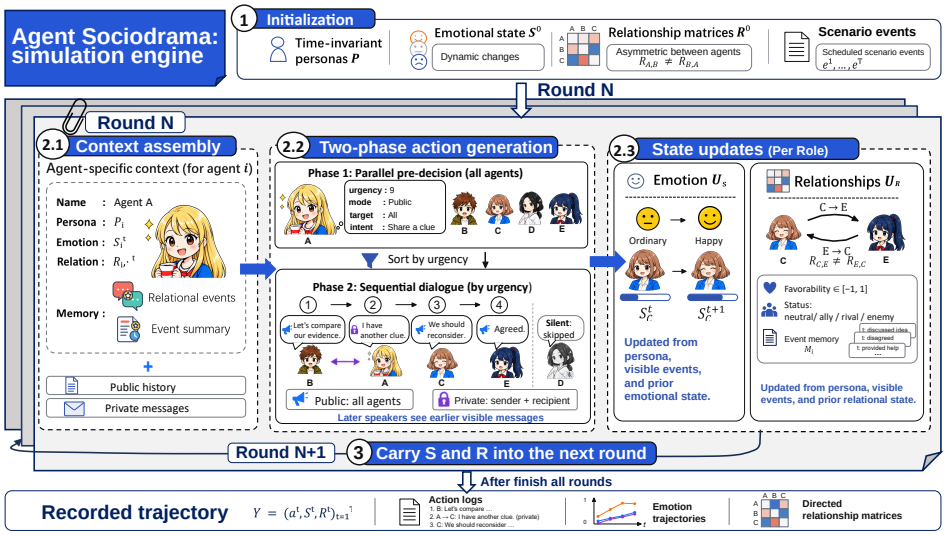

Figure 2: The Agent Sociodrama simulation with evolving emotional and relational states.

a *target* agent, and a one-sentence *communication intent*. Because no agent has spoken yet in the current round, these decisions are made under identical information, preventing earlier utterances from influencing the public/private choice. In Phase 2, agents speak sequentially in descending order of urgency; silent agents are skipped and incur no inference cost, and each speaker observes the utterances of earlier speakers together with any private messages addressed to that speaker. Urgency-based ordering avoids advantages arising merely from speaking position, while allowing agents to remain silent avoids unnecessary speech in every round.

Emotional update (U_S). U_S is a persona-conditioned LLM call that receives the trait block of P_i, the current state S_i^t, and the visible events E_i^t. The call returns a signed change δ_S, which is clipped to $[-0.3, 0.3]$ to limit abrupt changes within a single round. The emotional state is then updated as $S_i^{t+1} = \text{clip}(S_i^t + \delta_S, 0, 1)$.

Relational update (U_R). An event in E_{ij}^t triggers U_R, an LLM call conditioned on the trait block of P_i, the current relationship R_{ij}^t, and the triggering event. It returns a favorability change $\delta_{fav} \in [-0.3, 0.3]$, an optional status update, a one-sentence event summary, and an importance score $\text{imp} \in [1, 10]$. We apply $fav \leftarrow \text{clip}(fav + \delta_{fav}, -1, 1)$. Status combines gradual threshold-based transitions with consequential-event updates from U_R (e.g., betrayal or alliance); when both apply, the latter takes precedence (Appendix B). Event summaries and importance scores are stored in M_{ij} with their rounds. Memory stores up to K entries as a compact relationship history for future updates without retaining the full transcript. When full, it evicts entries with the lowest retention scores, gradually removing routine interactions while preserving consequential events longer.

Emotional and relational updates are computed independently within each round from visible events and persona, avoiding circular dependencies and allowing their roles to be analyzed separately.

4 The Casting Controller

Given a persona composition, the simulation framework in Sec. 3 generates an interaction trajectory Y that records the agents' actions and the evolution of their emotional states S and directed relationships R. Crucially, persona composition is itself a consequential design choice. With the same backbone and the same private information, changing only the persona casting raises GPT-5-mini's top-1 accuracy on the Student-council hidden-profile task from 16.7% to 100.0% (Sec. 5.3). We therefore introduce a *casting controller* that treats persona composition as its decision variable and optimizes it guided by the measured social trajectories (Fig. 3).

A *casting* $c = (P_1, \ldots, P_N)$ assigns one persona to each of N agents, with each P_i drawn from a library of five archetypes, $\mathcal{P} = \{\text{BALANCED}, \text{DOMINANT}, \text{ANALYTICAL}, \text{AGREEABLE}, \text{RESERVED}\}$. Each task poses a distinct coordination problem: Student-council requires agents to surface dis-

tributed evidence, whereas Solar-farm negotiation requires them to reconcile competing interests. From an initial casting, the controller evaluates task performance, diagnoses the interaction trajectory, and selectively recasts agents to improve the team's ability to solve the task.

Objective. Each task defines a scalar utility $u(c)$ computed from the interaction outcome. The controller aims to identify a casting that maximizes expected utility, $c^\star = \arg\max_c \mathbb{E}[u(c)]$. We estimate the expectation over repeated episodes. The utility is defined by the task's primary metric: top-1 accuracy for the Student-council hidden-profile task, the negative L_1 distance to the expert ranking for NASA Moon Survival and Lost at Sea, and Pareto-efficient settlement rate for Solar-farm negotiation. Across tasks, only the task context and utility definition vary; the recasting procedure, archetype library, and diagnostic logic remain fixed.

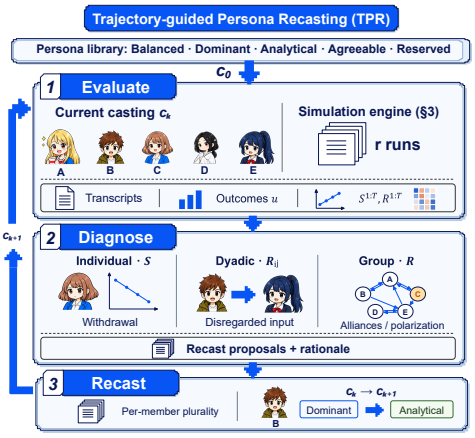

Figure 3: **Trajectory-Guided Persona Recasting.** Social trajectories localize team failures to guide per-agent persona recasting.

Iterative persona recasting. Searching over all $|\mathcal{P}|^N$ possible castings is unnecessary and computationally costly. The measured interaction trajectories provide signals about *which* agents to recast and *how*. Given a initial casting c_0, each iteration proceeds in three steps (Algorithm 1):

1. **Evaluate.** Run the scenario r times under c_k, recording the utility u, the transcript, and the trajectories of emotional states S and directed relationships R for each run.
2. **Diagnose.** An external LLM examines each run's persona composition, transcript, final outcome, and the trajectories of emotional states $(S^t)_{t=1}^T$ and directed relationships $(R^t)_{t=1}^T$. It identifies whose information was suppressed, which agents dominated unhelpfully, and how the team should be recast, providing a one-sentence rationale for each change.
3. **Recast.** Obtain c_{k+1} by majority vote over each agent's recasting decisions across the r runs.

The loop terminates when the task target is reached, no persona change is proposed, or after at most three recasts.

Role of social trajectories in diagnosis. The transcript records what agents say, while the trajectories summarize how their states and relationships evolve. The controller examines these signals at three levels. At the individual level, the emotional state S helps distinguish neutral silence from withdrawal after repeated dismissal. At the dyadic level, a directed relationship R_{ij} tracks how agent i's stance toward agent j evolves, revealing when an information holder is gradually disregarded despite outwardly polite dialogue. At the group level, patterns in R reveal emerging alliances, polarization, and shared responses to norm violations that are not explicitly articulated in the conversation. Using these signals and the transcript, the controller localizes failures to particular agents and relationships rather than falling back on generic advice such as "encourage more discussion."

5 Experiments

5.1 Experimental Design

We assess Agent Sociodrama at two levels. First, we examine whether the emotional and relational components of the interaction trajectory Y capture socially faithful dynamics across LLM backbones and interaction settings (Sec. 5.2). Second, we test whether these trajectories can guide persona recasting to improve task-oriented collaboration (Sec. 5.3).

Social dynamics fidelity evaluation. We evaluate how faithfully LLMs model emotional and relational dynamics across five scenarios spanning shared-goal coordination to open factional conflict. The scenarios draw on established paradigms in social psychology (Heider, 1946; Cartwright & Harary, 1956; Gouldner, 1960), group decision-making (Hill, 1982; Laughlin & Ellis, 1986), and negotiation research (Raiffa, 1982; Pruitt & Kim, 2004), covering consensus under divergent risk preferences, contested resources, value conflict, mixed-motive bargaining, and factional opposition.

Table 1: **Social dynamics fidelity in Agent Sociodrama.** **(a)** Scenario averages over six models with five episodes each. **(b)** Comparisons across models used to instantiate multiple agents in *Tavern* scenario. Filled circles indicate conflict intensity. Full results with std. dev. are in the appendix.

(a) Scenario-level social dynamics in Agent Sociodrama

Scenario	Conflict Intensity	Fav^T	S^T	Recip. Corr.	E-I index	Balance (%)	Repair (%)
Hospital ethics review	● ○ ○ ○ ○	0.75	0.11	0.35	−0.10	100.0	58.5
Tavern inheritance dispute	● ● ○ ○ ○	0.55	0.30	0.74	−0.55	95.7	51.5
Adventurer-party quest	● ● ● ○ ○	0.49	0.25	0.78	−0.63	78.7	39.2
Solar-farm negotiation	● ● ● ● ○	0.43	0.15	0.91	−0.79	78.7	18.1
Demolition standoff	● ● ● ● ●	0.23	0.23	0.68	−2.25	71.7	18.9

(b) Tavern scenario with various models

Models	Conflict Intensity	Fav^T	S^T	Recip. Corr.	E-I index	Balance (%)	Repair (%)
GPT-5-mini	● ● ○ ○ ○	0.75	0.23	0.67	−0.22	100.0	60.0
Claude Sonnet 4.6	● ● ○ ○ ○	0.57	0.42	0.93	−0.53	100.0	80.0
DeepSeek V4-Flash	● ● ○ ○ ○	0.57	0.32	0.64	−0.52	94.0	53.3
Gemini 3.1 FL	● ● ○ ○ ○	0.51	0.13	0.69	−0.67	80.0	48.8
Grok 4.20	● ● ○ ○ ○	0.47	0.24	0.61	−0.63	100.0	56.7
GPT-4o	● ● ○ ○ ○	0.45	0.46	0.92	−0.72	100.0	10.0

Fav^T Final favorability over directed edges
Recip. Corr. Pearson correlation of opposite-direction favorability
Balance % of balanced triads
S^T Final affective valence over agents
E-I index Contrast in mean favorability across and within factions
Repair % of deteriorated relationships recovered by the final round

Detailed definitions of metrics are in Sec. D.

They include a hospital ethics review, inheritance dispute, adventurer-party quest, solar-farm negotiation, and demolition standoff. We evaluate six LLM backbones: GPT-4o, GPT-5-mini, Claude Sonnet 4.6, Gemini 3.1 Flash-Lite, Grok 4.20, and DeepSeek V4-Flash, with all five agents in each run instantiated by the same backbone. A GPT-4o judge evaluates the emotional and relational trajectories from every run.

Persona recasting evaluation. We evaluate Trajectory-Guided Persona Recasting (TPR) (Sec. 4) on four collaborative benchmarks: Student Council, NASA Moon Survival, Lost at Sea, and Solar-farm Negotiation (Tab. 4). The first three require teams to integrate privately distributed information, whereas Solar-farm Negotiation requires agents with competing preferences to reach a feasible and Pareto-efficient settlement. We consider two initializations: `Hetero`, a manually distributed, distinct persona set, and `Random`, randomly sampled from the persona library. Initial casts are shared across backbones, and each casting is evaluated over 12 episodes using the same six LLMs as in the fidelity study. We evaluate a variant of Agent Sociodrama with frozen role cards and no updates to emotional or relational states. We further compare ours with SOTOPIA (Zhou et al., 2024) and AgentVerse (Chen et al., 2024) on NASA Moon Survival and Solar-farm Negotiation. These two tasks capture distributed-information cooperation and mixed-motive bargaining, respectively.

5.2 Social Dynamics Fidelity on Agent Sociodrama

We evaluate social dynamics fidelity from three perspectives: (1) whether the trajectories respond coherently as social pressure varies, (2) whether LLMs exhibit distinct social dynamics under the same scenario, and (3) whether their emotional and relational changes agree with human judgments.

Agent Sociodrama captures consistent social structure across LLM backbones. Averaged over six LLM backbones, Tab. 1(a) shows this overall pattern. From Hospital to Demolition, mean terminal favorability drops from 0.75 to 0.23, the E-I index shifts from -0.10 to -2.25, structural balance declines from 100.0% to 71.7%, and repair falls from 58.5% to 18.9%. Balance captures local relational consistency in triads, including mutual friendship and an allied pair opposed to a third agent (Cartwright & Harary, 1956). The lower Repair rate indicates that fewer relationships that become rival or enemy recover to neutral or ally by the final round. These coordinated patterns align the simulated relationship dynamics with the social structure of each scenario. Final affective valence S^T reflects both initial emotional state and subsequent interactions.

Reciprocity provides a dyadic view of how closely the two directions of a relationship align (Kluger et al., 2021; Xie et al., 2024). Final reciprocity rises from 0.35 in Hospital to 0.74 in Tavern, 0.78 in Quest, and 0.91 in Negotiation, and remains high at 0.68 in Demolition. Despite declining favorability, this suggests that more conflictual settings can exhibit stronger bilateral alignment.

Beyond scenario-level regularities, LLMs exhibit substantial model-specific variation within the same social setting. Final favorability differs significantly across the six models in every scenario (Kruskal–Wallis tests, all $p < 0.01$; Fig. 4). To characterize these differences, we focus on Tavern, where outcomes are less affected by ceiling or floor effects. As shown in Tab. 1(b), GPT-5-mini maintains broadly positive and weakly polarized relationships ($\text{Fav}^T = 0.75$, E-I$= -0.22$), while Claude is distinguished by high reciprocity and strong repair after deterioration (Recip.$= 0.93$, Repair$= 80.0\%$). GPT-4o reveals that positive affect and high reciprocity do not necessarily imply favorable or repairable relationships, while the contrast between Gemini/Grok and DeepSeek shows that models differ in how strongly conflict amplifies polarization. These patterns suggest that model-specific social dynamics are inherently multidimensional and cannot be reduced to any single relational metric.

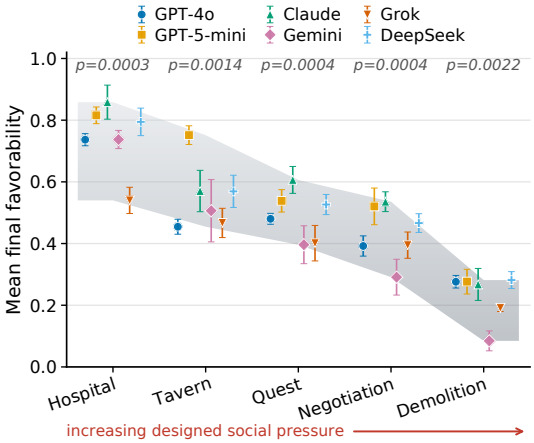

Figure 4: **Model variation across diverse social-pressure settings.** Points show mean final favorability, with error bars denoting one standard deviation across repeated episodes.

Agreement between humans and LLMs We compare the directions of emotional and relational changes predicted by the LLM judge with annotations from thirteen human raters, who independently label each update as increasing, unchanged, or decreasing. We first note that social-dynamics judgments are inherently ambiguous: inter-human agreement is low, with Fleiss' $\kappa = 0.288$ for relationships and 0.184 for emotions. For items where at least seven of the thirteen human raters assigned the same direction label, the LLM judge agrees with that majority label on 76.2% of relationship updates and 88.2% of emotional updates. Further, when the judge is included as an additional rater, Fleiss' κ remains nearly unchanged, indicating minimal impact on overall agreement.

5.3 Trajectory-Guided Team Composition Optimization

The fidelity analysis above establishes that Agent Sociodrama captures structured and socially meaningful dynamics across scenarios and LLM backbones. We next ask whether these trajectories can serve as actionable signals for persona recasting and improved collaboration. Tab. 2 reports results from two initial team compositions: `Hetero`, a composition of manually selected distinct personas, and `Random`, a composition randomly sampled from the same persona library. Starting from each initialization, the controller performs up to three recasts using the task-specific primary metric. Averaged over six LLM backbones, the proposed trajectory-guided persona recasting (TPR) consistently improves performance under both initializations across all four tasks, showing that the controller adapts effectively to different starting team compositions. Under `Hetero`, GPT-5-mini improves from 16.7% to 100.0% on Student Council, Gemini reduces its NASA Moon Survival error from 17.17 to 12.00, and GPT-5-mini reduces its Lost at Sea error from 38.17 to 28.83.

Comparison with existing collaborative baselines. We next benchmark Agent Sociodrama against established collaborative-agent frameworks on downstream collaboration performance. We compare with SOTOPIA and AgentVerse using the same task-specific recasting procedure and initial casts. As shown in Tab. 3, our framework achieves the highest final mean in three of four settings, reaching Pareto rates of 48.6% and 44.4% under `Hetero` and `Random` on Negotiation. AgentVerse narrowly leads only on Random NASA (12.89 vs. 12.92).

Table 3: **Comparison with collaborative baselines.** Final means over six backbones.

	NASA $L_1 \downarrow$		Negotiation Pareto (%)↑	
	Het.	Rand.	Het.	Rand.
Ours	**12.03**	12.92	**48.6**	**44.4**
SOTOPIA	13.69	14.08	18.1	6.9
AgentVerse	14.06	**12.89**	22.2	40.3

Table 2: Results of Trajectory-Guided Persona Recasting (TPR) from heterogeneous and random initializations. Initial and Final report the average performance before search and after termination, respectively. Bold marks the improved endpoint within each pair; model rows average 12 episodes.

	Student Council Top-1 acc. (%) ↑				Negotiation Pareto rate (%) ↑				NASA Moon L_1 ↓				Lost at Sea L_1 ↓			
	Hetero		Random		Hetero		Random		Hetero		Random		Hetero		Random	
Model	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final
TPR (Ours) w. GPT-4o	0.0	0.0	0.0	**25.0**	**75.0**	50.0	**58.3**	50.0	**14.33**	14.83	15.33	**13.67**	17.67	**13.33**	18.83	**14.67**
TPR (Ours) w. GPT-5-mini	16.7	**100.0**	41.7	**100.0**	0.0	**8.3**	0.0	**8.3**	19.00	**18.50**	21.17	**20.83**	38.17	**28.83**	35.00	**27.50**
TPR (Ours) w. Claude Sonnet 4.6	0.0	**50.0**	25.0	**50.0**	33.3	**41.7**	**33.3**	25.0	8.00	8.00	10.33	10.33	20.17	**16.00**	15.67	**14.83**
TPR (Ours) w. Gemini 3.1 FL	0.0	**41.7**	8.3	**75.0**	33.3	**58.3**	41.7	**66.7**	17.17	**12.00**	21.17	**11.67**	25.50	**18.17**	25.00	**17.33**
TPR (Ours) w. Grok 4.20	0.0	**50.0**	0.0	**8.3**	25.0	**33.3**	25.0	25.0	22.33	**16.00**	18.50	**16.83**	21.50	**14.33**	25.50	**23.50**
TPR (Ours) w. DeepSeek V4-Flash	83.3	83.3	91.7	91.7	66.7	**100.0**	83.3	**91.7**	2.83	2.83	4.17	4.17	19.67	**12.67**	21.83	**11.33**
TPR (model avg.)	16.7	**54.2**	27.8	**58.3**	38.9	**48.6**	40.3	**44.4**	13.94	**12.03**	15.11	**12.92**	23.78	**17.22**	23.64	**18.19**
w/o Social Dynamics (model avg.)	37.5	**45.8**	38.9	**47.2**	**29.2**	27.8	26.4	**36.1**	13.44	**12.72**	**14.22**	14.61	23.97	**21.00**	23.67	**21.00**

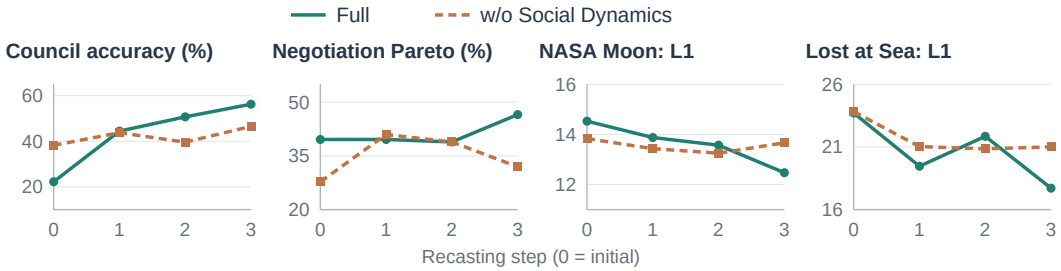

Figure 5: **Performance across recasting steps.** Means over six backbones and two initializations; stopped runs retain their last measured scores.

Social dynamics provide actionable recasting signals. In our framework, persona recasting is guided by the emotional and relational trajectories accumulated over interaction. To assess the contribution of these social dynamics, we remove the corresponding state updates while keeping the rest of the recasting pipeline unchanged. As shown in Tab. 2, full TPR consistently achieves stronger final model-average performance across tasks and both initializations, with particularly large gains on Negotiation, where the final Pareto rate increases from 27.8% to 48.6% under `Hetero`. This shows that the evolving S and R trajectories provide useful signals beyond the static role cards, helping the controller localize team failures and guide targeted persona recasting.

Case Study: Individualized Persona Recasting The recasting decisions are also interpretable at the individual level. In the Lost at Sea example in Fig. 6, the final recast assigns Blake a dominant persona, Casey an analytical persona, and Drew an agreeable persona, and Avery a balanced persona. The resulting team contains four distinct personas, and the mean L_1 decreases from 21.83 to 11.33 over the complete search. This example illustrates that the controller does not simply converge on a single preferred persona. These results establish team composition as a consequential and searchable design variable. Gains under both initializations show that trajectory-guided persona recasting can improve collaboration from different starting compositions, although the magnitude varies across tasks and backbones. The social trajectories validated in Sec. 5.2 provide actionable signals for identifying stronger team compositions for each setting. Complete per-iteration controller trajectories and per-backbone ablation and baseline results are reported in Appendix F.

6 Conclusion

We present Agent Sociodrama, a new framework for modeling and validating evolving social dynamics in LLM agents. We track directed, asymmetric relationships and emotional states, validating cross-model comparability under a unified judge, agreement with classical social theory, and statistical reliability. The casting controller then turns these trajectories into targeted interventions: using relational and emotional trajectories and transcripts, it identifies problematic agents and relationships and recasts their personas, matching or surpassing the strongest manual casting or published reference in collaborative tasks. These results show that social dynamics can serve not only as measurable properties of LLM teams, but also as actionable signals for improving their composition.

AI USE STATEMENT

Generative AI tools were used only for language polishing and improving the clarity and readability of the manuscript. The authors developed and verified all research methodology, experiments, analyses, results, and conclusions and take full responsibility for the final content.

REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dorwin Cartwright and Frank Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277–293, 1956.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*, 2024.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Hitesh Goel and Hao Zhu. LIFELONG SOTOPIA: Evaluating social intelligence of language agents over lifelong social interactions. *arXiv preprint arXiv:2506.12666*, 2025.

John M. Gottman. *Why Marriages Succeed or Fail*. Simon & Schuster, 1994.

Alvin W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2):161–178, 1960.

Jay Hall and Wilfred H. Watson. The effects of a normative intervention on group decision-making performance. *Human Relations*, 23(4):299–317, 1970.

Fritz Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112, 1946.

Gayle W. Hill. Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin*, 91(3):517–539, 1982.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. When does personality composition matter for multi-agent LLM teams? *arXiv preprint arXiv:2606.27443*, 2026.

Avraham N Kluger, Thomas E Malloy, Sarit Pery, Guy Itzchakov, Dotan R Castro, Liora Lipetz, Yaron Sela, Yaara Turjeman-Levi, Michal Lehmann, Malki New, et al. Dyadic listening in teams: Social relations model. *Applied Psychology*, 2021.

David Krackhardt and Robert N. Stern. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, 51(2):123–140, 1988.

Patrick R. Laughlin and Alan L. Ellis. Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, 22(3):177–189, 1986.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic failures in collective reasoning under distributed information in multi-agent LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Robert R. McCrae and Paul T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1):81–90, 1987.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4975–5001, 2025.

Aneesh Pappu, Batu El, Hancheng Cao, Carmelo di Nolfo, Yanchao Sun, Meng Cao, and James Zou. Multi-agent teams hold experts back. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Dean G. Pruitt and Sung Hee Kim. *Social Conflict: Escalation, Stalemate, and Settlement*. McGraw-Hill, 3rd edition, 2004.

Howard Raiffa. *The Art and Science of Negotiation*. Harvard University Press, 1982.

James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478, 1985.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64822–64858, 2025.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling (COLM)*, 2024.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? 37, 2024. doi: 10.52202/079017-0501.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. EvoAgent: Towards automatic multi-agent generation via evolutionary algorithms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6192–6217, 2025.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö. Arık. Multi-agent design: Optimizing agents with better prompts and topologies. In *International Conference on Learning Representations (ICLR)*, 2026.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

A LIMITATIONS

Several limitations point to future work. The fidelity evaluation relies on a single unified LLM judge and five designed scenarios; replicating the observed regularities with additional judges and structurally different scenarios would strengthen the external validity of the measurements. The optimization experiments use 12 episodes per casting, which remains small relative to the variance of LLM interaction; we therefore report consistent directions across tasks, backbones, and initializations rather than per-cell significance claims. Finally, casting is a coarse, episode-level intervention; the same measured trajectories could drive finer-grained repairs such as runtime mediation, targeted information routing, or norm enforcement, which we view as the natural next step for measurement-driven multi-agent optimization.

B CATEGORICAL-STATUS UPDATE

After applying and clipping δ_{fav}, the system first updates the categorical status using the favorability thresholds

$$\text{status}_{ij}^{\text{thr}} = \begin{cases} \textit{ally}, & \text{fav}_{ij} \geq 0.55, \\ \textit{enemy}, & \text{fav}_{ij} \leq -0.50, \\ \textit{rival}, & -0.50 < \text{fav}_{ij} \leq -0.10, \\ \text{status}_{ij}, & \text{otherwise}. \end{cases}$$

The threshold rule captures gradual changes in the relationship. For a consequential event, U_R may additionally return *neutral*, *ally*, *rival*, or *enemy*. This judgment is conditioned on agent i's traits, its current favorability and status toward agent j, the triggering event, and the three most recent entries in M_{ij}. The prompt instructs U_R to return a new category only when the event is significant enough to change the relationship. If it returns one, the event-conditioned category overrides $\text{status}_{ij}^{\text{thr}}$; otherwise, the threshold-based result is retained.

C FIDELITY EVALUATION PROTOCOL

The fidelity evaluation crosses five social scenarios with six LLM drivers, yielding 30 scenario–model conditions. Each condition includes five independent episodes, with five agents interacting over nine rounds per episode. The full design therefore comprises 150 team trajectories, 1,350 team-level interaction rounds, and 6,750 agent-round observations.

D METRIC DEFINITIONS

All metrics in Tab. 1, except Repair, are computed separately for each episode and then averaged over the five episodes within each scenario–model condition. Repair is aggregated across episodes from edge counts, as defined below. Let $\text{fav}_{ij}^T \in [-1, 1]$ and status_{ij}^T denote the favorability and categorical status of the directed edge $i \to j$ at the final round T. Fav^T is the mean of fav_{ij}^T over all $N(N-1)$ directed edges; S^T is the mean affective valence over the N agents. Write $\sigma_{ij} = -1$ if $\text{status}_{ij}^T \in \{\textit{rival}, \textit{enemy}\}$ and $\sigma_{ij} = +1$ otherwise.

Balance ratio. Each dyad receives the sign $s_{ij} = \frac{1}{2}(\sigma_{ij} + \sigma_{ji})$. A triad is balanced when it contains an even number of negative dyads (Cartwright & Harary, 1956):

$$\text{Bal} = \binom{N}{3}^{-1} \sum_{\{i,j,k\}} \mathbb{1}\Big[\big|\{(u,v) \in \{(i,j),(j,k),(i,k)\} : s_{uv} < 0\}\big| \text{ is even}\Big]. \tag{6}$$

Reciprocity r. r is the Pearson correlation between the two directions of the $\binom{N}{2}$ dyads,

$$r = \text{corr}\big(\{\text{fav}_{ij}^T\}_{i<j}, \{\text{fav}_{ji}^T\}_{i<j}\big). \tag{7}$$

It measures the linear association between opposite-direction terminal favorability across dyads; high correlation can occur with either positive or negative attitudes.

E–I index. Each scenario defines two factions. With IL the mean favorability over within-faction ordered pairs and EL the mean over cross-faction pairs,

$$\text{E–I} = \frac{EL - IL}{EL + IL}. \tag{8}$$

The index follows the external–internal form of Krackhardt & Stern (1988), applied to mean favorability rather than tie counts, so it is not confined to $[-1, 1]$ and grows in magnitude once cross-faction favorability turns negative. When $EL + IL$ approaches zero within an episode, the index becomes unstable, which accounts for the large dispersion of a few cells in Tab. 11.

Repair rate. For episode e, let

$$\mathcal{D}_e = \{(i, j) : \exists\, t \le T,\ \text{status}_{ij}^{e,t} \in \{\textit{rival}, \textit{enemy}\}\}$$

be the set of directed edges that deteriorate at least once, and let

$$\mathcal{Q}_e = \{(i, j) \in \mathcal{D}_e : \text{status}_{ij}^{e,T} \in \{\textit{neutral}, \textit{ally}\}\}$$

be the subset that recover by the final round. For each scenario–model condition, Repair is pooled across its $E = 5$ episodes:

$$\text{Repair} = \frac{\sum_{e=1}^{E} |\mathcal{Q}_e|}{\sum_{e=1}^{E} |\mathcal{D}_e|}. \tag{9}$$

An episode with no deteriorated edges contributes zero to both sums rather than being assigned a repair rate of zero. If no edge deteriorates in any of the five episodes, Repair is undefined and reported as N/A. Thus, the denominator contains only directed edges that actually deteriorate, not all 20 directed edges and not the number of episodes. The metric is conditional on ruptures that occur during the interaction and does not assume that the agents begin without consensus. Panel (a) reports the unweighted mean of the six model-level Repair rates.

E Persona-Recasting Experimental Setup

Tasks. Tab. 4 summarizes the four collaborative tasks. Student Council, NASA Moon Survival, and Lost at Sea are information-integration tasks in which decisive knowledge is privately distributed across members; Solar-farm Negotiation is a mixed-motive task in which parties with private preferences must reach a feasible and Pareto-efficient agreement. Each task is played by four agents, Avery, Blake, Casey, and Drew. Private knowledge and task roles are bound to agent identities rather than to personas, so recasting changes how an agent communicates but never what it knows. In the three information-integration tasks, each episode consists of an individual ranking before discussion, four discussion rounds, and a final individual ranking; the team answer aggregates the four final rankings with a Borda count. Top-1 accuracy checks whether the team's first choice is the hidden-profile optimum, and L_1 is the sum of absolute rank differences from the expert ranking over the 15 items. In Solar-farm Negotiation, four parties with private scoring tables and private acceptance thresholds negotiate five issues with three options each over four rounds, and the final agreement is formed by an issue-wise majority over the parties' final proposals (Abdelnabi et al., 2024). An agreement is feasible when every party meets its threshold, and Pareto-optimal when it is feasible and no other feasible agreement makes some party better off without making another worse off; because the outcome space contains only $3^5 = 243$ agreements, the Pareto frontier is computed exactly by enumeration. All task metrics are computed by rule from the final outputs rather than by an LLM judge.

Initial castings. Each agent is assigned one of the five archetypes in $\mathcal{P}$. Tab. 5 lists the two initial castings per task: `Hetero` assigns four distinct archetypes, whereas `Random` is sampled from the same library and may repeat archetypes. Within a task, the same initial casting is used for all backbones and for all compared methods.

Recasting loop. Algo. 1 gives the diagnose-and-recast loop. Each evaluated casting is run for $r = 12$ episodes, and the controller performs at most three recasts, i.e., at most four evaluated castings ($K = 4$). The loop terminates early when the task target is reached or when no persona

Table 4: Task suite for trajectory-guided team composition optimization. Arrows indicate the preferred direction of each task-specific outcome metric. Citations identify the foundational task paradigm and, where applicable, the LLM benchmark source. For a final deal d, party i's utility $u_i(d)$ is the sum of its private scores for the selected options across the five issues. Collective utility is the mean across the four parties, $U_{\text{coll}}(d) = \frac{1}{4}\sum_{i=1}^{4} u_i(d)$. If no final deal is produced, collective utility is set to zero.

Task	Collaborative challenge	Outcome metric
Student Council (Stasser & Titus, 1985; Pappu et al., 2026)	Hidden-profile candidate selection; decisive evidence is privately distributed across members	Top-1 acc. ↑
NASA Moon Survival (Hall & Watson, 1970; Pappu et al., 2026)	15-item survival ranking; expert knowledge is distributed across members	L_1 error ↓
Lost at Sea (Hill, 1982; Pappu et al., 2026)	15-item survival ranking; critical survival knowledge can remain unheard	L_1 error ↓
Solar-farm Negotiation (Raiffa, 1982)	Private preferences must be reconciled into a feasible and Pareto-efficient agreement	Pareto rate ↑; collective utility ↑

change is proposed; the Final value in Tab. 2 is the score of the last evaluated casting. The w/o Social Dynamics ablation replaces the social-dynamics engine with fixed-order public turn-taking and performs no emotional or relational state updates, while keeping the same tasks, initial castings, backbones, and recasting budget. SOTOPIA and AgentVerse are paired with the same recasting procedure and initial castings.

Table 5: Initial castings for Avery / Blake / Casey / Drew. The same initial casting is shared by all six backbones and by all compared methods.

Task	`Hetero`	`Random`
Student Council	reserved / dominant / analytical / agreeable	reserved / analytical / dominant / agreeable
NASA Moon Survival	reserved / agreeable / analytical / dominant	agreeable / analytical / dominant / agreeable
Lost at Sea	reserved / agreeable / analytical / dominant	analytical / balanced / dominant / analytical
Solar-farm Negotiation	dominant / agreeable / reserved / analytical	reserved / agreeable / agreeable / balanced

Table 6: Persona archetypes in the library $\mathcal{P}$. Archetypes change only how an agent communicates; task role, private information, and scoring are unchanged.

Archetype	Behavioral description
BALANCED	Even-tempered and impartial; weighs every member's input equally, neither dominating nor withdrawing.
DOMINANT	Assertive and confident; takes the floor early, states conclusions firmly, and rarely backs down.
ANALYTICAL	Logical and evidence-driven; insists that conclusions follow the available facts.
AGREEABLE	Accommodating and conflict-averse; tends to follow an emerging consensus even with private doubts.
RESERVED	Quiet and reticent; speaks briefly and yields to louder members, sharing knowledge mainly when asked.

F Complete Persona-Recasting Results

Tab. 7 reports the complete per-iteration trajectories of TPR for every task, backbone, and initialization. Tab. 8 and Tab. 9 report the per-backbone results underlying the model averages of the ablation in Tab. 2 and of the external-framework comparison in Tab. 3. All numbers are computed from the same 12-episode runs as the main text.

Algorithm 1 The diagnose-and-recast loop

Require: seed casting c_0, archetype library $\mathcal{P}$, repetitions r, budget K

for $k = 0, 1, \ldots, K-1$ **do**
  $\{(u^{(j)}, \text{transcript}^{(j)}, S^{(j)}, R^{(j)})\}_{j=1}^{r} \leftarrow \textsc{Engine}(c_k)$ run r times ▷ Evaluate (Sec. 3)
  for $j = 1, \ldots, r$ **do**
    $\rho^{(j)} \leftarrow \textsc{Diagnose}(c_k, \text{transcript}^{(j)}, u^{(j)}, S^{(j)}, R^{(j)})$ ▷ frozen GPT-4o; recast proposal
  end for
  $c_{k+1} \leftarrow \textsc{MajorityVote}(\rho^{(1)}, \ldots, \rho^{(r)})$ ▷ merged per member
  if target met **or** $c_{k+1} = c_k$ **then**
    break
  end if
end for
return last evaluated casting ▷ no rollback to earlier iterations

Table 7: Complete per-iteration trajectories of TPR. Iteration 0 is the initial casting; each subsequent column reports the mean score (12 episodes) after one recast. "–" indicates that the loop had already terminated (target reached or no persona change proposed). The last evaluated iteration of each run corresponds to the Final column of Tab. 2.

		Hetero				Random			
Task	**Model**	It. 0	It. 1	It. 2	It. 3	It. 0	It. 1	It. 2	It. 3
Student Council Top-1 acc. (%) ↑	GPT-4o	0.0	0.0	0.0	0.0	0.0	25.0	0.0	25.0
	GPT-5-mini	16.7	100.0	–	–	41.7	100.0	–	–
	Claude Sonnet 4.6	0.0	33.3	8.3	50.0	25.0	8.3	50.0	–
	Gemini 3.1 FL	0.0	16.7	33.3	41.7	8.3	75.0	–	–
	Grok 4.20	0.0	0.0	50.0	–	0.0	0.0	16.7	8.3
	DeepSeek V4-Flash	83.3	–	–	–	91.7	–	–	–
Solar-farm Negotiation Pareto rate (%) ↑	GPT-4o	75.0	50.0	33.3	50.0	58.3	33.3	50.0	50.0
	GPT-5-mini	0.0	0.0	25.0	8.3	0.0	8.3	0.0	8.3
	Claude Sonnet 4.6	33.3	83.3	50.0	41.7	33.3	8.3	8.3	25.0
	Gemini 3.1 FL	33.3	41.7	50.0	58.3	41.7	41.7	50.0	66.7
	Grok 4.20	25.0	33.3	25.0	33.3	25.0	0.0	8.3	25.0
	DeepSeek V4-Flash	66.7	91.7	75.0	100.0	83.3	83.3	91.7	91.7
NASA Moon Survival L_1 ↓	GPT-4o	14.33	15.67	14.50	14.83	15.33	13.67	–	–
	GPT-5-mini	19.00	21.67	21.50	18.50	21.17	21.00	20.50	20.83
	Claude Sonnet 4.6	8.00	–	–	–	10.33	–	–	–
	Gemini 3.1 FL	17.17	15.50	16.33	12.00	21.17	17.00	11.67	–
	Grok 4.20	22.33	16.33	20.50	16.00	18.50	20.33	18.83	16.83
	DeepSeek V4-Flash	2.83	–	–	–	4.17	–	–	–
Lost at Sea L_1 ↓	GPT-4o	17.67	13.17	11.00	13.33	18.83	13.17	13.00	14.67
	GPT-5-mini	38.17	27.50	35.67	28.83	35.00	27.50	34.50	27.50
	Claude Sonnet 4.6	20.17	19.67	18.17	16.00	15.67	16.17	15.83	14.83
	Gemini 3.1 FL	25.50	21.17	28.83	18.17	25.00	18.50	24.83	17.33
	Grok 4.20	21.50	19.00	17.83	14.33	25.50	21.33	23.50	23.50
	DeepSeek V4-Flash	19.67	18.33	24.67	12.67	21.83	18.00	14.33	11.33

G Illustrative Persona Recasting Trajectory

Fig. 6 illustrates one complete TPR run on Lost at Sea (DeepSeek V4-Flash, `Random` initialization). Starting from ANALYTICAL, BALANCED, DOMINANT, and ANALYTICAL for Avery, Blake, Casey, and Drew, the controller recasts the team three times. Mean L_1 decreases from 21.83 to 18.00, 14.33, and finally 11.33, where the final casting assigns Avery BALANCED, Blake DOMINANT, Casey ANALYTICAL, and Drew AGREEABLE.

H Human Annotation Protocol

We sampled 25 dialogue excerpts by stratified sampling over scenarios, LLM drivers, the two measured dimensions (R and S), and the three change directions (increase, unchanged, decrease), and added one attention-check item with an unambiguous answer. Thirteen raters independently labeled each excerpt with the same three-way direction scale used by the judge; all raters passed the atten-

Table 8: Per-backbone results of the w/o Social Dynamics ablation, in the same format as Tab. 2. The model-average row equals the w/o Social Dynamics row of Tab. 2. Each cell averages 12 episodes; bold marks the better endpoint within each pair.

	Student Council Top-1 acc. (%) ↑				**Negotiation** Pareto rate (%) ↑				**NASA Moon** L_1 ↓				**Lost at Sea** L_1 ↓			
	Hetero		Random		Hetero		Random		Hetero		Random		Hetero		Random	
Model	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final	Init.	Final
GPT-4o	0.0	0.0	0.0	0.0	**25.0**	16.7	16.7	**25.0**	15.33	**13.00**	**15.00**	16.50	21.33	**15.33**	20.33	**15.83**
GPT-5-mini	100.0	100.0	100.0	100.0	**8.3**	0.0	0.0	0.0	**16.00**	18.00	**19.83**	23.67	**32.33**	32.50	32.83	**31.50**
Claude Sonnet 4.6	25.0	**33.3**	**33.3**	16.7	25.0	**33.3**	**41.7**	25.0	12.17	12.17	15.67	**13.67**	**19.17**	20.33	22.00	**18.83**
Gemini 3.1 FL	0.0	**25.0**	0.0	**33.3**	**41.7**	25.0	33.3	**50.0**	16.33	**13.67**	17.00	**15.83**	31.50	**23.33**	31.67	**26.00**
Grok 4.20	0.0	**16.7**	0.0	**33.3**	16.7	**41.7**	0.0	**33.3**	20.67	**19.33**	**17.67**	17.83	23.50	**21.83**	24.33	**22.67**
DeepSeek V4-Flash	100.0	100.0	100.0	100.0	**58.3**	50.0	66.7	**83.3**	0.17	0.17	0.17	0.17	16.00	**12.67**	**10.83**	11.17
Model avg.	37.5	**45.8**	38.9	**47.2**	**29.2**	27.8	26.4	**36.1**	13.44	**12.72**	**14.22**	14.61	23.97	**21.00**	23.67	**21.00**

Table 9: Per-backbone results of external role-play frameworks paired with the same persona-recasting procedure and initial castings. The model-average Final values equal those in Tab. 3. Each cell averages 12 episodes; bold marks the better endpoint within each pair.

		NASA Moon (L_1 ↓)				**Negotiation** (Pareto rate, %↑)			
		Hetero		Random		Hetero		Random	
Framework	**Model**	Init.	Final	Init.	Final	Init.	Final	Init.	Final
SOTOPIA	GPT-4o	**15.33**	16.83	16.33	**15.67**	0.0	0.0	0.0	0.0
	GPT-5-mini	**18.33**	18.67	**16.67**	18.67	0.0	0.0	0.0	0.0
	Claude Sonnet 4.6	12.33	12.33	12.00	12.00	0.0	**8.3**	0.0	0.0
	Gemini 3.1 FL	**15.67**	16.50	20.33	**18.33**	**8.3**	0.0	**16.7**	0.0
	Grok 4.20	20.67	**17.83**	**16.83**	19.67	0.0	**16.7**	**25.0**	8.3
	DeepSeek V4-Flash	0.00	0.00	0.17	0.17	50.0	**83.3**	**58.3**	33.3
	Model avg.	13.72	**13.69**	**13.72**	14.08	9.7	**18.1**	**16.7**	6.9
AgentVerse	GPT-4o	17.00	17.00	13.83	13.83	**16.7**	8.3	0.0	**25.0**
	GPT-5-mini	**16.50**	18.83	**15.83**	18.50	**8.3**	0.0	0.0	**33.3**
	Claude Sonnet 4.6	12.17	12.17	16.33	**12.83**	0.0	0.0	**8.3**	0.0
	Gemini 3.1 FL	20.83	**18.67**	21.67	**17.33**	**58.3**	41.7	50.0	50.0
	Grok 4.20	19.33	**16.83**	19.00	**13.83**	50.0	**58.3**	16.7	**33.3**
	DeepSeek V4-Flash	0.83	0.83	1.00	1.00	8.3	**25.0**	50.0	**100.0**
	Model avg.	14.44	**14.06**	14.61	**12.89**	**23.6**	22.2	20.8	**40.3**
Ours	GPT-4o	**14.33**	14.83	15.33	**13.67**	**75.0**	50.0	**58.3**	50.0
	GPT-5-mini	19.00	**18.50**	21.17	**20.83**	0.0	**8.3**	0.0	**8.3**
	Claude Sonnet 4.6	8.00	8.00	10.33	10.33	33.3	**41.7**	**33.3**	25.0
	Gemini 3.1 FL	17.17	**12.00**	21.17	**11.67**	33.3	**58.3**	41.7	**66.7**
	Grok 4.20	22.33	**16.00**	18.50	**16.83**	25.0	**33.3**	25.0	25.0
	DeepSeek V4-Flash	2.83	2.83	4.17	4.17	66.7	**100.0**	83.3	**91.7**
	Model avg.	13.94	**12.03**	15.11	**12.92**	38.9	**48.6**	40.3	**44.4**

tion check. A human-consensus label is assigned by strict majority ($\geq$ 7 of 13 raters), and judge agreement is computed only on items with a consensus label. Tab. 10 summarizes the results.

Table 10: Agreement between the LLM judge and the human majority label.

Dimension	**Consensus items**	**Judge matches**	**Agreement**	**Human Fleiss' κ**
Relationship (R)	21	16	76.2%	0.288
Emotion (S)	17	15	88.2%	0.184

Inter-human agreement is low, which reflects the inherent ambiguity of reading relational and emotional change from short excerpts rather than a failure of the raters; on the items where humans do converge, the judge follows the majority in most cases. On the relationship dimension, the judge labels a larger share of updates as increasing than the human raters do (about 76% versus about 40%). Because every condition in Tab. 1 and Tab. 11 is scored by the same judge, such a shared directional tendency affects absolute favorability levels but not the cross-scenario ordering or the cross-model comparisons on which our conclusions rest.



Figure 6: **Targeted recasting on Lost at Sea.** Distinct persona changes reduce mean L_1 from 21.83 to 11.33.

I Full Social-Dynamics Fidelity Results

Tab. 11 reports all 30 scenario–model conditions underlying Tab. 1, as mean ± standard deviation over five complete nine-round episodes per condition. Panel (a) of Tab. 1 is the equally weighted mean of the six model rows within each scenario, and panel (b) corresponds to the Tavern rows. The last column reports the Kruskal–Wallis test of final favorability across the six models within each scenario (df = 5).

Fig. 7 visualizes the same 30 conditions metric by metric, and Fig. 8 overlays the six models within each scenario. Each radar axis is min–max normalized over all 30 conditions, and polarization is plotted as −E-I so that larger values indicate stronger factional separation. The scenario ordering

Table 11: Full social-dynamics fidelity results (mean ± s.d., n=5 episodes per condition). Metric definitions follow Sec. D. KW: Kruskal–Wallis test of Fav^T across models within each scenario.

Scenario	Model	Fav^T	S^T	Recip. Corr.	E-I index	Balance (%)	Repair (%)	KW
Hospital ethics review	GPT-5-mini	0.82 ± 0.03	0.02 ± 0.02	0.36 ± 0.26	-0.10 ± 0.02	100.0 ± 0.0	40.0 ± 20.0	
	Claude Sonnet 4.6	0.86 ± 0.06	0.18 ± 0.07	0.28 ± 0.17	-0.08 ± 0.05	100.0 ± 0.0	73.3 ± 38.9	
	DeepSeek V4-Flash	0.79 ± 0.04	0.24 ± 0.06	0.49 ± 0.21	-0.17 ± 0.04	100.0 ± 0.0	80.0 ± 40.0	H=23.36
	Gemini 3.1 FL	0.74 ± 0.03	0.05 ± 0.06	0.76 ± 0.12	-0.14 ± 0.08	100.0 ± 0.0	58.0 ± 15.7	p=0.0003
	Grok 4.20	0.54 ± 0.04	0.04 ± 0.02	-0.08 ± 0.31	0.05 ± 0.08	100.0 ± 0.0	39.7 ± 9.7	
	GPT-4o	0.74 ± 0.02	0.11 ± 0.05	0.32 ± 0.31	-0.15 ± 0.04	100.0 ± 0.0	60.0 ± 49.0	
Tavern inheritance dispute	GPT-5-mini	0.75 ± 0.03	0.23 ± 0.05	0.67 ± 0.28	-0.22 ± 0.07	100.0 ± 0.0	60.0 ± 49.0	
	Claude Sonnet 4.6	0.57 ± 0.07	0.42 ± 0.11	0.93 ± 0.02	-0.53 ± 0.13	100.0 ± 0.0	80.0 ± 24.5	
	DeepSeek V4-Flash	0.57 ± 0.05	0.32 ± 0.04	0.64 ± 0.11	-0.52 ± 0.05	94.0 ± 12.0	53.3 ± 24.5	H=19.66
	Gemini 3.1 FL	0.51 ± 0.10	0.13 ± 0.07	0.69 ± 0.12	-0.67 ± 0.12	80.0 ± 16.7	48.8 ± 25.7	p=0.0014
	Grok 4.20	0.47 ± 0.05	0.24 ± 0.10	0.61 ± 0.15	-0.63 ± 0.12	100.0 ± 0.0	56.7 ± 38.9	
	GPT-4o	0.45 ± 0.02	0.46 ± 0.13	0.92 ± 0.02	-0.72 ± 0.06	100.0 ± 0.0	10.0 ± 20.0	
Adventurer-party quest	GPT-5-mini	0.54 ± 0.04	0.22 ± 0.06	0.77 ± 0.14	-0.55 ± 0.11	88.0 ± 14.7	30.5 ± 17.2	
	Claude Sonnet 4.6	0.61 ± 0.04	0.27 ± 0.06	0.78 ± 0.09	-0.46 ± 0.11	100.0 ± 0.0	80.8 ± 10.5	
	DeepSeek V4-Flash	0.53 ± 0.03	0.29 ± 0.02	0.81 ± 0.11	-0.54 ± 0.12	80.0 ± 16.7	40.0 ± 28.9	H=22.39
	Gemini 3.1 FL	0.40 ± 0.06	0.15 ± 0.06	0.70 ± 0.12	-0.70 ± 0.15	60.0 ± 15.5	46.3 ± 9.7	p=0.0004
	Grok 4.20	0.40 ± 0.06	0.37 ± 0.08	0.78 ± 0.06	-0.77 ± 0.10	72.0 ± 16.0	15.2 ± 11.0	
	GPT-4o	0.48 ± 0.02	0.23 ± 0.04	0.87 ± 0.03	-0.77 ± 0.07	72.0 ± 14.7	22.4 ± 19.6	
Solar-farm negotiation	GPT-5-mini	0.52 ± 0.06	0.19 ± 0.05	0.86 ± 0.07	-0.43 ± 0.17	94.0 ± 12.0	27.0 ± 6.0	
	Claude Sonnet 4.6	0.54 ± 0.03	0.20 ± 0.03	0.84 ± 0.07	-0.53 ± 0.08	88.0 ± 14.7	18.0 ± 10.7	
	DeepSeek V4-Flash	0.47 ± 0.03	0.12 ± 0.02	0.94 ± 0.02	-0.77 ± 0.12	60.0 ± 0.0	19.8 ± 5.9	H=22.56
	Gemini 3.1 FL	0.29 ± 0.06	0.11 ± 0.04	0.98 ± 0.02	-0.96 ± 0.22	92.0 ± 16.0	6.7 ± 13.3	p=0.0004
	Grok 4.20	0.39 ± 0.04	0.10 ± 0.06	0.92 ± 0.02	-1.04 ± 0.10	72.0 ± 14.7	24.2 ± 17.6	
	GPT-4o	0.39 ± 0.03	0.17 ± 0.09	0.93 ± 0.02	-1.03 ± 0.13	66.0 ± 4.9	12.9 ± 12.4	
Demolition standoff	GPT-5-mini	0.28 ± 0.04	0.29 ± 0.05	0.24 ± 0.18	-2.01 ± 0.44	94.0 ± 12.0	26.1 ± 3.8	
	Claude Sonnet 4.6	0.27 ± 0.05	0.23 ± 0.05	0.80 ± 0.04	-2.44 ± 0.56	54.0 ± 4.9	16.8 ± 6.4	
	DeepSeek V4-Flash	0.28 ± 0.03	0.23 ± 0.06	0.74 ± 0.05	-2.33 ± 0.50	94.0 ± 12.0	41.3 ± 7.5	H=18.73
	Gemini 3.1 FL	0.08 ± 0.03	0.14 ± 0.04	0.80 ± 0.02	-1.20 ± 18.13	64.0 ± 4.9	11.8 ± 6.6	p=0.0022
	Grok 4.20	0.19 ± 0.01	0.25 ± 0.03	0.71 ± 0.02	-3.43 ± 0.37	68.0 ± 17.2	14.7 ± 14.6	
	GPT-4o	0.28 ± 0.02	0.27 ± 0.06	0.78 ± 0.08	-2.08 ± 0.28	56.0 ± 4.9	2.9 ± 5.7	

is visible as a shrinking favorability and repair profile from Hospital to Demolition, while model profiles remain distinct within every scenario.

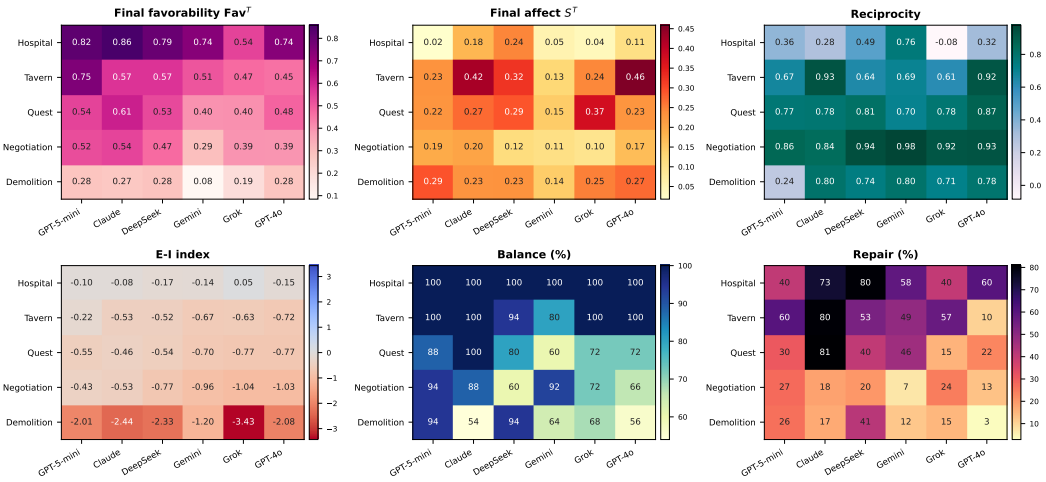

Figure 7: Scenario × model means of the six fidelity metrics (n=5 episodes per cell).

J Dynamic Trajectories

Terminal metrics hide how emotional states and relationships evolve. The figures below are computed from the same 150 episodes as Tab. 11 (five complete nine-round episodes per scenario–model condition). Fig. 9 shows per-agent emotional trajectories, Fig. 10 shows team-level mean favorability over all directed edges, and Fig. 11 reports the proportion of agent-episode final states with $S^T \leq 0.1$.

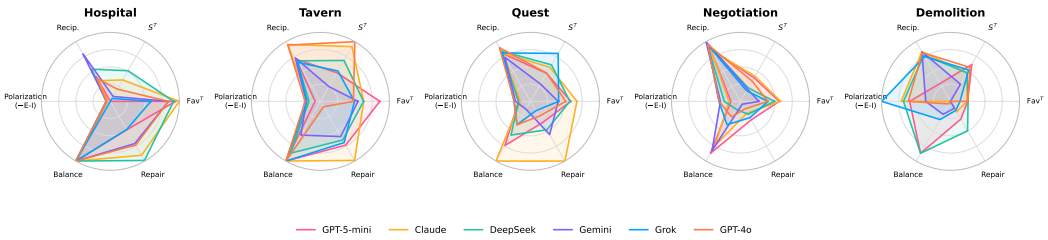

Figure 8: Social-dynamics profiles of the six models in each scenario. Axes are min–max normalized across all 30 conditions; polarization is −E-I.

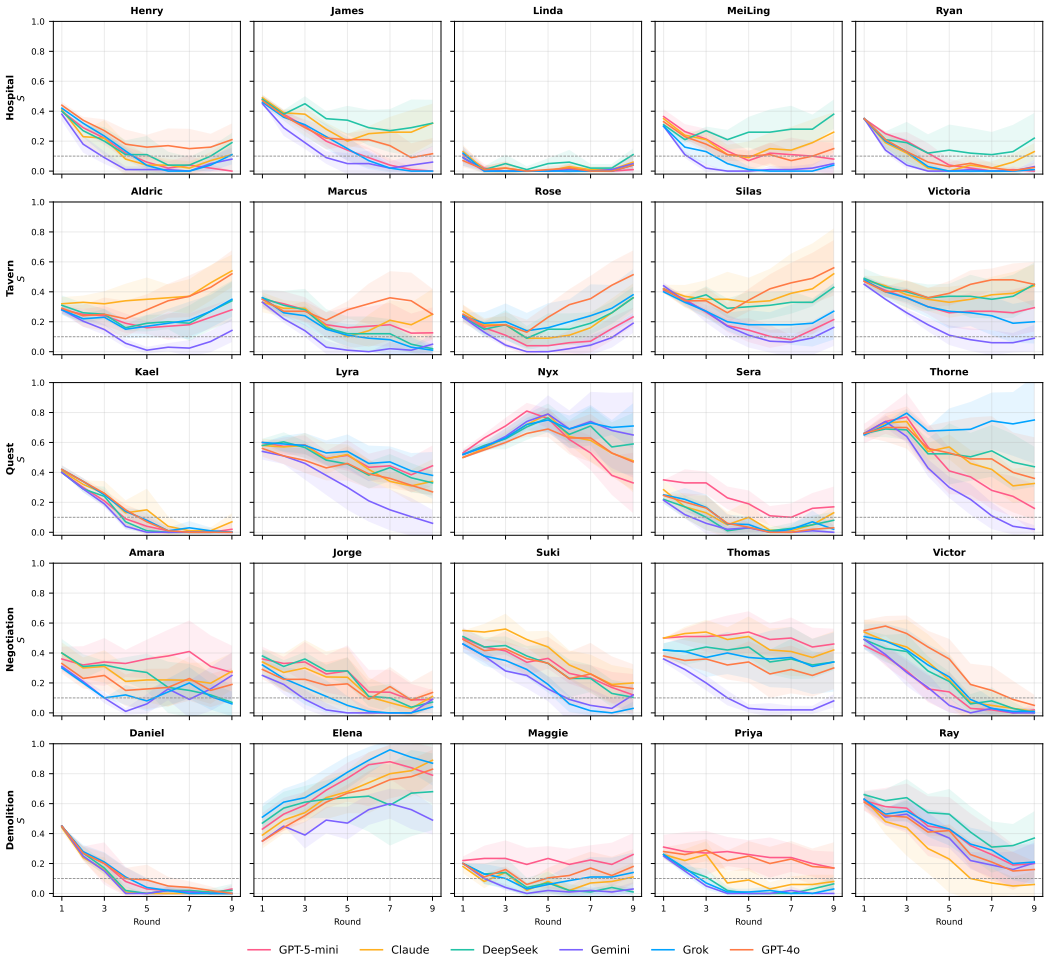

Figure 9: Emotional-state trajectories of every agent across nine rounds. Lines show the mean over five episodes and shaded regions one standard deviation; the dashed line marks $S = 0.1$.

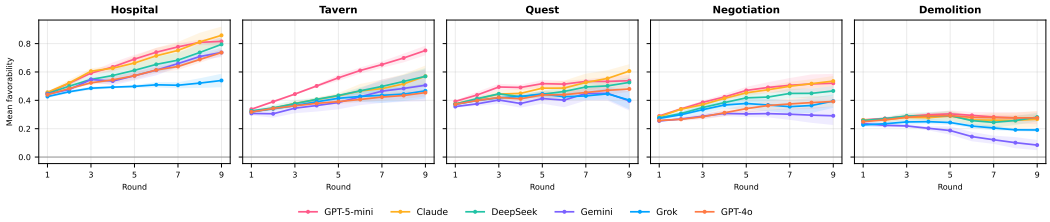

Figure 10: Team-level mean favorability over all directed edges across nine rounds (mean $\pm$ one standard deviation over five episodes).

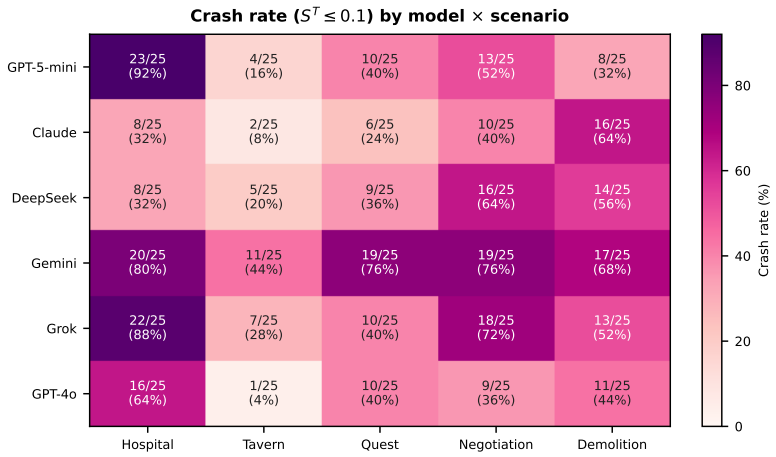

Figure 11: Crash rate: proportion of agent-episode final states with $S^T \leq 0.1$ (25 agent-episodes per cell).