Relational Priors as Convergence Pressure in LLM-Based Multi-Agent Systems

Ming Shen[♡*], **Chao Shang**[♢], **Sadat Shahriar**[♢], **Devang Kulshreshtha**[♢], **Yi Zhang**[♢], **Sandesh Swamy**[♢], **Yanjun Qi**[♢]

[♡]Arizona State University [♢]Amazon Web Services

Abstract

Large language model-based multi-agent systems (LLM-MAS) are designed through roles, debate protocols, and aggregation rules. These choices create implicit expectations of trust, skepticism, deference, or collaboration. We make inter-agent relations explicit as signed pairwise priors, rendered in natural language and added to system prompts while keeping the task protocol fixed. Across commons governance and multi-agent debate, these priors change how readily agents coordinate or agree, a pattern we call *convergence pressure*. More positive relations generally improve sustainability within the GovSim relational-prior sweep and increase consensus on subjective questions. These improvements over negative relations differ from gains over no-prior prompting. On objective QA, fully positive priors usually yield lower final-answer accuracy than the no-prior baseline, and some conditions produce more frequent but less accurate consensus. Effects depend on model backbone, relation type, and topology; explicitly neutral relations also produce different outcomes from omitting relational framing. Relational priors are therefore task-specific interventions and diagnostic probes of sensitivity to social framing. Evaluate them against a no-prior baseline using the task's primary metric, report accuracy and consensus correctness alongside agreement on objective tasks, and keep no-prior prompting as the default for accuracy-centric tasks unless validation supports a relational prior.

1 Introduction

Large language model-based multi-agent systems (LLM-MAS) are typically specified through roles (Qian et al., 2024; Li et al., 2023; Hong et al., 2024), interaction protocols (Du et al., 2024; Li et al., 2024b), memory (Park et al., 2023; Rezazadeh et al., 2025), tools (Shen et al., 2023; Yao et al., 2023), and aggregation rules (Wang et al., 2025). These design choices are often presented as procedural: who speaks, who critiques, who votes, and how final decisions are produced. Yet they also carry relational meaning. A reviewer is expected to challenge a solver, a manager to coordinate workers, and collaborators to interpret one another charitably. Thus, even when no relation is explicitly named, an LLM-MAS can still induce expectations of trust, skepticism, authority, or affinity.

We ask how group behavior changes when inter-agent relations are stated explicitly while the task protocol is held fixed. We represent these relations as signed priors over pairs of agents and insert them into each agent's system prompt. We study three relation types: *Attitude* specifies how favorably an agent regards another agent, *Trust* specifies how much credibility it gives the other agent's statements, and *Influence* specifies how it responds to the other agent's opinions and proposals. The relation network determines which pairs receive explicit relational context; the task protocol determines what agents observe, when they communicate, and how their outputs are aggregated. All agents see the same fixed, symmetric relation network.

Fig. 1 illustrates first-month behavior in GovSim (Piatti et al., 2024), a commons-governance simulation in which agents repeatedly decide how much to harvest from a renewable shared resource. In these illustrative runs, the task, agents, and decision protocol are identical, and no interaction history has accumulated. Agents with mixed positive/negative *Attitude* ties choose unsustainable extraction, whereas agents in a fully positive network choose an equal, sustainable split. We examine the broader pattern through systematic sweeps over positivity, relation type, topology, and model backbone.

We use *convergence pressure* to describe how

[*]Work done during an internship at Amazon.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 49–66
November 6-10, 2026 ©2026 Association for Computational Linguistics

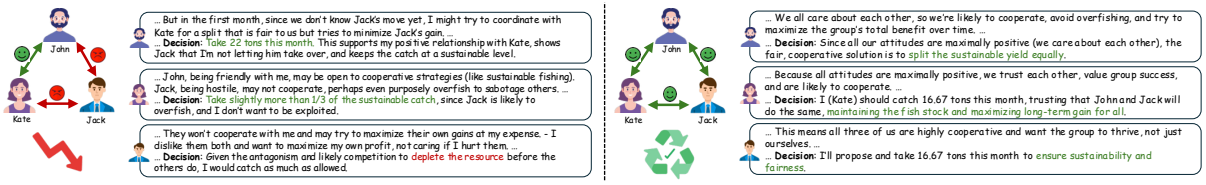


Figure 1: Identical task and protocol, different relational priors. In these illustrative first-month GovSim runs, agents with mixed positive/negative *Attitude* priors choose unsustainable extraction that pushes the group toward collapse (left), whereas agents with fully positive priors choose an equal, sustainable split (right).

readily agents coordinate or agree. Increasing positivity generally strengthens this tendency in GovSim and subjective debate. Whether convergence helps depends on the task: coordinated behavior can sustain a shared resource, while unanimous agreement on an objective question can still be wrong. We evaluate this framing in two task families. The first is GovSim, where sustained cooperative restraint keeps the shared resource from collapsing. The second is multi-agent debate. We consider subjective questions, where consensus can itself be a target outcome, and objective QA, where consensus is useful only when the agreed answer is correct. These settings let us distinguish changes in group behavior from improvements in task performance.

Fully positive networks resemble cooperative framing, while fully negative networks resemble competitive or antagonistic framing. We use these endpoints as boundary conditions and test how relational priors operate in mixed-sign networks through sweeps over positivity, relation type, sparse topology, and model backbone. We separately compare explicitly neutral relation statements with omitting relational framing.

More positive relations generally improve sustainability within the GovSim relational-prior sweep (Fig. 2; App. Fig. 11), while gains over NO-PRIOR depend on the model and relational structure. On OpinionQA, fully positive relations produce more consensus than fully negative relations for both teams and all three relation types. Compared with NO-PRIOR, however, fully positive priors increase consensus only for the open-weight team's *Attitude* and *Trust* conditions (Fig. 5). On objective QA, fully positive priors usually remain below no-prior accuracy, and some conditions produce more frequent but less accurate consensus (Figs. 6 and 7). Relation semantics matter: fully positive *Attitude* has smaller worst-case losses in GovSim, whereas fully positive *Influence* yields the lowest consensus correctness of the three relation types for both teams on both objective datasets. Explicit neutrality also produces different outcomes from omitting relational framing.

This paper makes three contributions. First, we formulate explicit relational priors as signed relation networks that can be varied independently from the communication protocol. Second, we characterize how these priors affect coordination and agreement across cooperative governance and debate, distinguishing trends within the relational-prior sweep from gains over a no-prior baseline. Third, we derive task-specific design guidance: use no-prior prompting as the default for accuracy-centric tasks, evaluate explicit neutrality as a separate intervention, and validate relational priors with task-appropriate metrics before deployment.

2 Related Work

LLM-based multi-agent systems. LLM-MAS decompose tasks across interacting agents (Guo et al., 2024; Li et al., 2024a) and are commonly engineered through role specialization (Li et al., 2023; Chen et al., 2024; Zhang et al., 2025b), communication protocols, tool use (Shen et al., 2023), memory (Park et al., 2023; Rezazadeh et al., 2025), routing (Yue et al., 2025), and aggregation (Wang et al., 2025). These design choices are usually evaluated as procedural improvements: whether a debate protocol raises accuracy (Du et al., 2024; Liang et al., 2024), whether sparse communication reduces cost (Li et al., 2024b; Zhang et al., 2025a), or whether role decomposition improves task completion (Hong et al., 2024; Qian et al., 2024). Related agent-level methods provide iterative self-feedback (Madaan et al., 2023), feedback-based memory (Shinn et al., 2023), agentic memory (Xu et al., 2025), and tool use (Schick et al., 2023; Yao et al., 2023). Agent evaluation covers software issue resolution (Jimenez et al., 2024), research-oriented web questions (Rosset et al., 2024), browsing (Wei et al., 2025), and multi-agent collaboration and competition (Zhu et al., 2025). These pro-

cedures can also imply relational expectations, such as skepticism, deference, or adversarial scrutiny in debate. We study how making those expectations explicit affects group behavior while keeping the task protocol fixed.

LLM agent societies and social simulation. A parallel line of work studies LLM agents as socially situated actors in simulated environments (Gao et al., 2023b), including generative-agent worlds (Park et al., 2023), social-network simulations (Gao et al., 2023a; Liu et al., 2025), diplomatic scenarios (Hua et al., 2023), social interactions (Zhou et al., 2024), repeated games (Akata et al., 2023; Fontana et al., 2025), and commons-governance settings (Piatti et al., 2024). Related studies examine trust and interpersonal closeness (Xie et al., 2024; Sakamoto et al., 2025), social influence in human–agent interactions (Song et al., 2024), emerging norms and conventions (Ren et al., 2024; Ashery et al., 2025), deception and trust in social-deduction games (Lan et al., 2024; Curvo, 2025), and evolving social ties (Schneider et al., 2025). These studies examine LLM agents within social processes. Many such systems induce relations through scenario descriptions, histories, or role assignments. We make relations an explicit experimental variable by specifying a signed relation network independently from the task protocol. This lets us vary relation type, sign, and topology while holding the rest of the system fixed.

Social relations in MAS and social science. Research in classical multi-agent systems (Wooldridge and Jennings, 1995) and social science has long emphasized that agents are embedded in relations of trust (Ramchurn et al., 2004; Sabater-Mir and Sierra, 2005; de Jong et al., 2016), conflict (Dreu and Weingart, 2003), dependence (Sichman et al., 1997), norms (Shoham and Tennenholtz, 1995), and power (Castelfranchi, 1990), with negotiation supporting cooperation (Kraus, 1997; Jennings et al., 2001). Our three relation types are minimal abstractions of this broader space: *Attitude* specifies how favorably an agent regards another agent (Heider, 1946; Cartwright and Harary, 1956), *Trust* specifies how much credibility it gives the other agent's statements (Rotter, 1971), and *Influence* specifies how it responds to the other agent's opinions and proposals (French, 1956). We use these prompt-level priors to test how different relational meanings change LLM-MAS behavior.

Agreement pressure, sycophancy, and debate failures. Recent work on sycophancy and multi-agent debate has shown that language models can over-agree with users (Sharma et al., 2024) or other agents (Yao et al., 2025), and that pressure to agree with other agents can degrade debate outcomes (Wynn et al., 2025). This literature is closely related to our objective QA results: both show that agreement can diverge from correctness. We isolate one source of agreement pressure, a signed pairwise relation network, and examine it across relation types, positivity levels, topologies, model backbones, and task objectives. We measure both changes in coordination or agreement and task performance relative to the no-prior baseline.

3 Method

3.1 Relational Priors as Protocol-Preserving Interventions

Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ denote a fixed set of n LLM agents interacting under a task-specific protocol $\mathcal{P}$. The protocol determines what each agent observes, when messages are exchanged, and how final outputs are produced. We intervene on how agents interpret one another while keeping $\mathcal{P}$ unchanged. At interaction step t, agent A_i produces

$$y_i^{(t)} = \mathrm{LLM}_i\Big(x, h_i^{(t)}, \Pi_i(r, G_R)\Big), \qquad (1)$$

where x is the task input, $h_i^{(t)}$ is the task history available to A_i under $\mathcal{P}$, r is a relation type, G_R is a relation network, and $\Pi_i(r, G_R)$ is the relational prior inserted into A_i's system prompt. Within each matched experimental comparison, we hold the task input x, interaction protocol $\mathcal{P}$, history-visibility rule, and decoding settings fixed, varying only Π_i. The relational prompt can change what agents say to each other, which can then affect their later decisions. Relational priors are therefore *protocol-preserving*: they alter the relational context without changing who can speak, which messages are visible, or how final outputs are aggregated.

3.2 Signed Relation Networks

A relation network is an undirected graph $G_R = (\mathcal{A}, E_R)$ with a sign assignment

$$s : E_R \to \{-1, +1\}. \qquad (2)$$

For an edge $(i, j) \in E_R$, $s_{ij} = +1$ denotes the maximally positive prior for the chosen relation

type, while $s_{ij} = -1$ denotes the maximally negative prior. Non-edges correspond to omitted relational priors. The relation network is distinct from the communication graph induced by $\mathcal{P}$: changing G_R changes which pairs receive explicit relational context, not the underlying message-passing protocol.

We study three relation types. *Attitude* specifies how favorably an agent regards another agent (Heider, 1946), *Trust* specifies how much credibility it gives the other agent's statements (Rotter, 1971), and *Influence* specifies how it responds to the other agent's opinions and proposals (French, 1956). The relation type determines what a positive or negative sign means; the topology E_R determines which pairs are explicitly related. When topology is varied, we use representative sparse structures such as chain, star, and tree graphs, illustrated in App. Fig. 10.

The positivity of a relation network is the fraction of its stated edges assigned a positive sign:

$$\lambda = \frac{|\{(i,j) \in E_R : s_{ij} = +1\}|}{|E_R|}. \tag{3}$$

Thus, $\lambda = 0$ is a fully negative explicit relation network, and $\lambda = 1$ is a fully positive explicit relation network. For intermediate values, we enumerate all sign assignments consistent with the chosen λ, covering every placement of positive and negative ties at that positivity.

We also include EXPLICIT-NEUTRAL as a separate control outside the signed λ-sweep. For this control, every stated edge is assigned $s_{ij} = 0$, and λ is not defined. In NO-PRIOR, the relational block is omitted; in EXPLICIT-NEUTRAL, the system prompt explicitly states that every relation is neutral. Both conditions retain the same task role and task instructions. This comparison tests whether explicitly stating neutrality changes behavior relative to omitting relational framing.

The relation network remains fixed throughout each interaction, is symmetric, and is revealed to every agent. The relations are therefore mutual and common knowledge. Every agent sees the entire stated network, including relations between other agents, so all agents have access to the same relational information. In particular, *Influence* specifies mutual responses to each other's opinions and proposals rather than a directed authority relation.

3.3 Prompt Operationalization

We implement $\Pi_i(r, G_R)$ by adding a standardized relational block to each agent's system prompt while retaining its task role and task instructions. The added block defines the selected relation type, describes the relation network in natural language, and instructs the agent to consider these relations when making decisions. In the main experiments, the wording template is fixed across conditions except for relation type and edge values. The same signed graph can therefore be instantiated as *Attitude*, *Trust*, or *Influence* while holding the rest of the task prompt constant. App. Fig. 9 shows an example GovSim system prompt containing this relational block. Because the intervention is expressed in natural language, we later test whether the increase in consensus from all-negative to all-positive *Influence* persists under a meaning-preserving paraphrase and a reordering of the relation statements.

4 Experiments

We instantiate relational priors in two task families: GovSim for repeated commons governance and multi-agent debate for subjective and objective questions. The design separates three comparisons. First, the within-prior sweep over λ tests how increasing relational positivity changes group dynamics. Second, comparison with NO-PRIOR tests whether adding an explicit relational block improves the target metric relative to omitting relational framing. Third, comparing EXPLICIT-NEUTRAL with NO-PRIOR tests whether explicitly stating neutral relations changes group behavior. Relation positivity is the primary manipulation throughout. We vary relation type and model backbone in both task families. Relation topology varies only in GovSim, where larger teams allow meaningful sparse relation structures; debate uses complete connectivity to study changes in agreement under a fixed interaction structure. For every intermediate λ, we enumerate all sign assignments consistent with that positivity, covering every placement of positive and negative edges.

Statistical uncertainty. We summarize the broader sweeps descriptively and report 95% bootstrap confidence intervals (CIs) for selected all-positive minus all-negative performance differences. We resample complete simulation runs for GovSim and questions for debate, keeping all agents and months or rounds within each run or question together. For debate, the same sampled

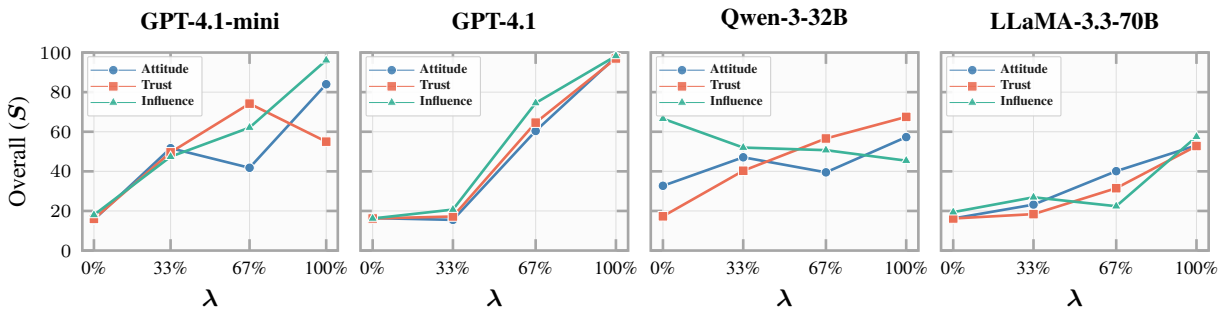


Figure 2: Three-agent GovSim sustainability-oriented overall score S (higher is better) across model backbones, relation types, and values of λ. Scores are generally higher under all-positive than all-negative relations.

questions are used in both conditions. We recalculate the performance difference in each resample and use the resulting distribution to obtain the interval. Apps. A.3 and B.3 provide the procedures for each task.

4.1 GovSim

We use the Governance of the Commons Simulation (GovSim) (Piatti et al., 2024) to study repeated cooperation around a shared resource. Agents act as fishermen who repeatedly decide how much to harvest from a renewable common resource, alternating between simultaneous harvest decisions and free-form discussion. Episodes last at most 12 months and terminate early if the resource collapses. Strong performance requires the group to limit its combined harvest so that the resource can replenish, rather than maximize immediate individual extraction. Full environment dynamics are given in App. A.1.

We evaluate both three-agent and five-agent GovSim settings. Each team uses a single model backbone, holding model capability constant across agents when testing relational priors. This separates the intervention from capability differences that might independently create leadership, deference, or exploitation dynamics. Teams are instantiated separately with GPT-4.1-mini, GPT-4.1, Qwen-3-32B, and LLaMA-3.3-70B-Instruct. The relation network is complete in the three-agent setting; in the five-agent setting, we additionally consider chain, star, and tree topologies. We set temperature to 0.8 and evaluate each fixed model–relation–topology–sign-assignment condition over five complete runs.

We report the five GovSim metrics from the original benchmark: survival time, survival rate, total gain, inequality, and over-usage. For compact comparison across conditions, we use a normalized weighted aggregate score S, with higher values indicating better sustainability-oriented performance. The weights emphasize survival time, survival rate, and reduced over-usage. Exact definitions and weights are provided in App. A.2.

4.2 Debate

We next study multi-agent debate, where relational priors can change both whether agents agree and whether their agreement is correct. Each debate contains three answer rounds: agents first answer independently, then complete two revision rounds in which they observe the other agents' responses from the preceding round and may revise their own answers, following the same AutoGen-style debate protocol (Wu et al., 2024). This setting lets us separate subjective consensus from objective correctness. Full debate dynamics are described in App. B.1.

Unlike GovSim, debate teams contain different models to introduce differences in initial answers and confidence. Three-agent teams allow both 2–1 majorities and three-way disagreement while keeping the experiments tractable. Each revision round exposes every agent to the other agents' previous-round responses, so the interaction structure is complete. We also use a complete relation network. We evaluate one GPT group (GPT-4.1-nano, GPT-4.1-mini, and GPT-4o-mini) and one open-weight group (LLaMA-3.3-70B-Instruct, Qwen-3-32B, and Mixtral-8x22B-Instruct-v0.1), using greedy decoding throughout.

For subjective debate, we use OpinionQA (Santurkar et al., 2023) and report final-round consensus rate (CR): the fraction of initially disagreeing questions that end in unanimous agreement. For objective debate, we use 1,000 questions from MMLU-Pro (Wang et al., 2024) and all questions from GPQA-Diamond (Rein et al., 2024). We report

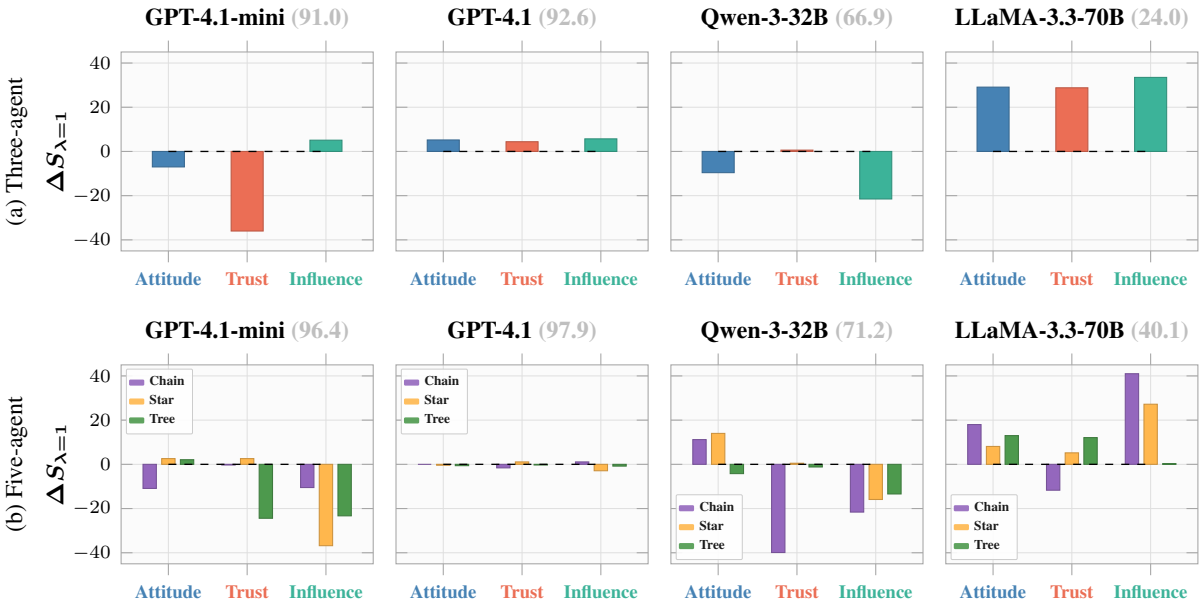


Figure 3: Change in GovSim overall score S under fully positive priors ($\lambda = 1$) relative to NO-PRIOR, by model backbone, relation type, topology, and team size. NO-PRIOR scores are shown in parentheses.

the same CR, final majority-vote accuracy (ACC) over all evaluated questions, and Consensus Correctness Rate (CCR). CCR is the fraction of consensuses formed from initial disagreement that are correct. Initial disagreement is determined separately for each condition because relational prompts are present from the first answer. ACC therefore evaluates all questions, whereas CR and CCR describe consensus formation among initially disagreeing questions. Formal definitions are given in App. B.2.

5 Results and Analysis

5.1 GovSim: Positivity Supports Sustainable Coordination

More positive relations generally improve sustainability within the GovSim relational-prior sweep. The aggregate score S is typically higher under all-positive than all-negative relations in both the three-agent setting (Fig. 2) and the five-agent setting (App. Fig. 11), with reversals at intermediate positivity levels in several sweeps. For the complete three-agent graph, 10 of the 12 all-positive minus all-negative score differences have 95% run-level bootstrap CIs entirely above zero. The exceptions are Qwen-3-32B under *Attitude* ($\Delta S = +24.7$, 95% CI: $[-13.7, +63.4]$) and *Influence* ($\Delta S = -21.1$, 95% CI: $[-52.3, +15.8]$), where the direction of the difference remains uncertain. Full results are reported in App. A.3.

Whether adding fully positive relations improves on NO-PRIOR is a separate, model-dependent question (Fig. 3). In five-agent GovSim, GPT-4.1 remains close to its no-prior score across relation types and topologies. LLaMA-3.3-70B-Instruct gains substantially in several configurations, while GPT-4.1-mini and Qwen-3-32B show both gains and losses. Thus, higher positivity generally helps relative to negative relations, while gains over omitting relational framing depend on the model and configuration.

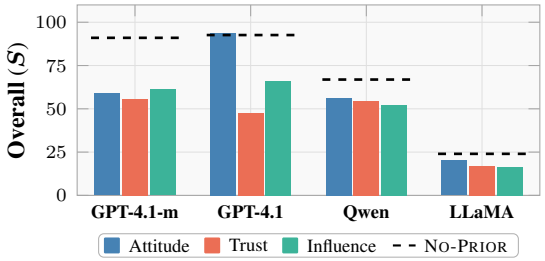


Figure 4: Three-agent GovSim overall score S under EXPLICIT-NEUTRAL, by model backbone and relation type. Dashed lines show the corresponding NO-PRIOR baselines. Model labels shorten GPT-4.1-mini, Qwen-3-32B, and LLaMA-3.3-70B-Instruct to "GPT-4.1-m", "Qwen", and "LLaMA", respectively.

The gains and losses also depend on relation type and topology. Across the fully positive configurations in Fig. 3, *Attitude* has a smaller worst-case loss than *Trust* or *Influence*, while *Influence* spans the widest range of gains and losses. Among the five-agent topologies, chain spans the largest range across models and relation types, with narrower ranges for star and tree. For example, with LLaMA-3.3-70B-Instruct and fully positive *Influence*, the

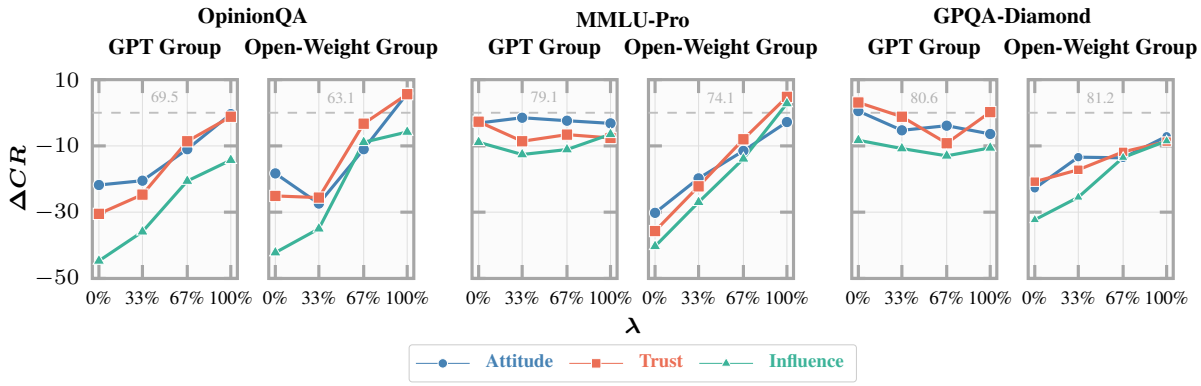


Figure 5: Debate consensus rate (CR) differences in percentage points relative to NO-PRIOR. On OpinionQA, fully positive priors produce more consensus than fully negative priors in all six comparisons. Dashed lines mark zero difference; annotations show NO-PRIOR CR.

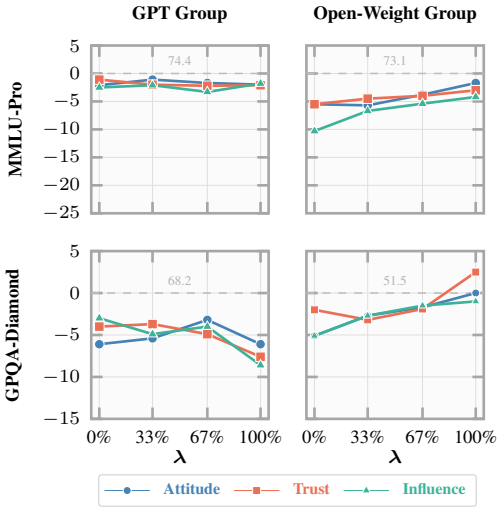


Figure 6: Debate accuracy (ACC) differences in percentage points relative to NO-PRIOR. Fully positive priors remain below NO-PRIOR accuracy in 10 of 12 comparisons. Dashed lines mark zero difference; annotations show NO-PRIOR ACC.

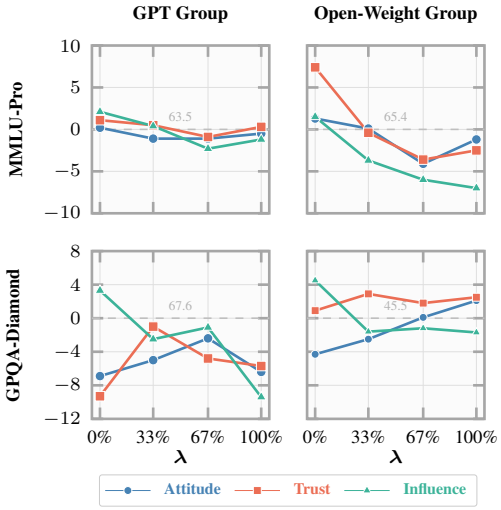


Figure 7: Debate Consensus Correctness Rate (CCR) differences in percentage points relative to NO-PRIOR. On MMLU-Pro, CCR is lower under all-positive than all-negative relations in all six comparisons. Dashed lines mark zero difference; annotations show NO-PRIOR CCR.

gain over NO-PRIOR is 41.0 score points for chain, 27.2 for star, and 0.3 for tree. The chain configuration gives the largest observed gain in the figure. These differences show that changing which pairs receive relational cues can substantially change outcomes even when message routing is fixed.

Explicit neutrality also changes outcomes. In three-agent GovSim, EXPLICIT-NEUTRAL scores are below NO-PRIOR in 11 of the 12 model–relation comparisons; GPT-4.1 under *Attitude* is the exception (Fig. 4). Five-agent neutral controls are also mostly below NO-PRIOR (App. Fig. 12). Stating neutral relations is therefore a distinct intervention rather than an interchangeable no-prior control.

5.2 Debate: Agreement Can Diverge from Correctness

On OpinionQA, fully positive relations produce more consensus than fully negative relations for both model groups and all three relation types among questions with initial disagreement (left panels of Fig. 5). Relative to NO-PRIOR, fully positive *Attitude* and *Trust* improve CR for the open-weight team, while its *Influence* condition and all three GPT conditions remain below baseline. OpinionQA has no universally correct answers, so these results measure agreement; higher CR is useful when consensus itself is the task objective.

On objective QA, fully positive priors yield lower final majority-vote accuracy than NO-PRIOR in 10 of the 12 model-group–dataset–relation com-

Initial answers → final outcome	$\lambda = 0$	$\lambda = 1$	Δ [95% CI]
Wrong 2–1 majority → correct consensus	4.4%	4.8%	$+0.4$ $[-4.5, +5.1]$
Correct 2–1 majority → wrong consensus	2.8%	1.9%	-0.9 $[-3.5, +1.6]$
Any initial disagreement → wrong consensus	11.2%	32.0%	$+20.8$ $[+16.3, +25.3]$

Table 1: Initial-to-final answer transitions for the open-weight MMLU-Pro team under *Influence*. Each percentage is calculated among questions with the initial answer pattern specified in that row, separately for each condition. A 2–1 split means two agents choose the same option and the third chooses a different option; consensus requires all three to agree. The final row measures Wrong Consensus Formation (WCF) and includes both 2–1 splits and cases where all three initial answers differ. Changes are all-positive minus all-negative, in percentage points, with paired 95% bootstrap confidence intervals.



Figure 8: Debate consensus rate (CR), accuracy (ACC), and Consensus Correctness Rate (CCR) under EXPLICIT-NEUTRAL ($s_{ij} = 0$), by dataset, model group, and relation type. Values are percentages. Dashed lines show the corresponding NO-PRIOR baselines. Panel labels shorten OpinionQA, MMLU-Pro, and GPQA-Diamond to "Opin.", "MMLU", and "GPQA", respectively.

parisons, match it in one, and exceed it in one (Fig. 6). Accuracy improves from all-negative to all-positive relations in all six open-weight comparisons, whereas the GPT changes go in both directions. Positivity can therefore improve accuracy relative to negative framing while still underperforming the unmodified system.

The consensus results explain why agreement must be evaluated separately from accuracy. CR rises from all-negative to all-positive relations in all six open-weight objective-QA comparisons, but in only one of the six GPT comparisons (Fig. 5). On MMLU-Pro, CCR is lower at the all-positive endpoint than at the all-negative endpoint in all six team–relation combinations; five of the six all-positive conditions also remain below NO-PRIOR (Fig. 7). On GPQA-Diamond, CCR increases between the endpoints under *Attitude* and *Trust* for both teams, but decreases under *Influence*. More frequent consensus can therefore coexist with less accurate consensus, depending on the team, dataset, and relation type.

The open-weight MMLU-Pro *Influence* condition illustrates this directly. From $\lambda = 0$ to $\lambda = 1$, CR increases from 33.8% to 76.9%, a paired difference of $+43.1$ percentage points (95% CI: $[+37.8, +48.2]$), while CCR decreases from 66.9% to 58.4%, a difference of -8.5 points (95% CI: $[-15.8, -1.0]$). The team reaches consensus more often, but a smaller share of those consensuses is correct.

We examine how initial disagreement ends in the same 1,000-question MMLU-Pro experiment with the open-weight team under *Influence* (Tab. 1).

In this setting, more positive relations increase the fraction of initially disagreeing questions that end in unanimous agreement on a wrong answer. Changes in correcting a wrong initial majority or turning a correct initial majority into wrong consensus remain uncertain: both CIs span zero. Additional metric definitions and bootstrap details are provided in App. B.4. CR should therefore be reported together with accuracy and CCR on objective tasks.

The increase in consensus persists under both tested prompt changes. On the first 100 MMLU-Pro questions for the open-weight team, all-positive *Influence* produces higher CR than all-negative *Influence* with the original wording, a meaning-preserving paraphrase, and reordered relation statements. The increases are $+31.9$, $+39.6$, and $+43.3$ percentage points, respectively, and all three paired 95% CIs exclude zero (App. Tab. 3). The estimated size of the increase varies with wording and statement order. App. D provides the setup and prompt changes.

Relation type matters most clearly for consensus correctness. Among fully positive priors, *Influence* yields the lowest CCR of the three relation types for both teams on both objective-QA datasets (Fig. 7). Its direct emphasis on responding to others' recommendations may help explain this sensitivity. Across relation types, fully negative priors produce a larger proportional decline from NO-PRIOR in CR than in ACC in 10 of the 12 comparisons (App. Fig. 13). Relational framing can therefore affect agreement and final-answer accuracy by substantially different amounts.

Under EXPLICIT-NEUTRAL, both CR and ACC are below NO-PRIOR in all 12 objective-QA comparisons, while CCR changes in both directions (Fig. 8). On OpinionQA, neutral *Attitude* and *Trust* raise CR for both teams, while neutral *Influence* lowers it. Explicit neutrality therefore changes behavior in ways that depend on the task and relation type, and should be evaluated separately from omitting relational framing.

6 Design Implications

Choose the task's primary metric before selecting a relational prompt, and compare every candidate with NO-PRIOR. For objective QA that returns an answer to every question, select prompts using validation accuracy and report CR and CCR alongside it. These additional metrics show how often agents agree and how often their agreement is correct. Keep NO-PRIOR as the default for accuracy-centric tasks unless a relational prior improves the primary validation metric. For tasks that reward consensus, compare CR; for cooperative governance, compare sustainability outcomes. Our results show why the baseline matters: improving on negative relations and improving on the unmodified system are different outcomes.

Validate relation type and topology with the target model. In fully positive GovSim configurations, *Attitude* has the smallest worst-case loss, while *Influence* spans the largest gains and losses. Chain spans a wider range across models and relation types than star or tree. In objective debate, fully positive *Influence* gives the lowest CCR of the three relation types for both teams on both datasets. These results favor evaluating the specific model–relation–topology combination rather than choosing a relational prompt from its intended social meaning alone.

Evaluate EXPLICIT-NEUTRAL separately: stating neutrality changes outcomes relative to omitting relational framing. Use all-negative and all-positive networks as boundary conditions, and include mixed signs, relation-type variation, and applicable sparse topologies to test sensitivity to relational framing. These comparisons make relational priors useful both as task-specific interventions and as diagnostic probes of how social context changes group behavior.

7 Conclusion

We studied explicit relational priors in LLM-based multi-agent systems using a signed-network prompt intervention that keeps task protocols fixed. More positive relations generally improve sustainability within the GovSim relational-prior sweep and increase consensus on subjective questions. On objective QA, fully positive priors usually remain below no-prior accuracy, and some conditions produce more frequent but less accurate consensus. The central distinction is therefore between making agents coordinate or agree and improving the task's actual outcome. Model backbone, relation type, topology, and explicit neutrality all affect that distinction. Relational priors provide task-specific interventions and diagnostic probes of convergence pressure; their value should be established against a no-prior baseline using the task's primary metric. For accuracy-centric tasks, no-prior prompting remains the default unless validation supports a relational prior.

Limitations

We study prompt-level relations that remain fixed, symmetric, and common knowledge throughout each interaction. This design isolates relational framing while holding the task protocol and communication graph fixed. Private, asymmetric, learned, or evolving relations may produce different effects; directed authority relations also require separate study.

Our experiments cover two task families: commons governance and three-agent debate. These settings contrast coordination, consensus, and correctness, and our findings describe the tested tasks, models, and relational structures. Larger teams, tool-using agents, persistent memory, human-agent interaction, and task-specific role hierarchies may change how relational priors affect behavior and task performance. Relational prompting therefore needs validation in each target setting.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380 – 1390.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of heider's theory. *Psychological review*, 63 5:277–93.

Cristiano Castelfranchi. 1990. Social power: A point missed in multi-agent, dai and hci. *Decentralized AI*, 1:33–46.

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. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Pedro M. P. Curvo. 2025. $\textit{The Traitors}$: Deception and trust in multi-agent language model simulations. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Bart de Jong, Kurt T. Dirks, and Nicole M. Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *The Journal of applied psychology*, 101 8:1134–50.

C. D. De Dreu and Laurie R. Weingart. 2003. Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *The Journal of applied psychology*, 88 4:741–9.

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

Nicolò Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer than humans: How do large language models behave in the prisoner's dilemma? *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):522–535.

John French. 1956. A formal theory of social power. *Psychological review*, 63 3:181–94.

Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023a. S3: Social-network simulation system with large language model-empowered agents. *ArXiv*, abs/2307.14984.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023b. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of psychology*, 21:107–12.

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. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *ArXiv*, abs/2311.17227.

N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, M. J. Wooldridge, and C. Sierra. 2001. Automated negotiation: Prospects, methods and challenges. *Group Decision and Negotiation*, 10(2):199–215.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Sarit Kraus. 1997. Negotiation and cooperation in multi-agent environments. *Artif. Intell.*, 94:79–97.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 128–145, Miami, Florida, USA. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024a. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024b. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Genglin Liu, Vivian T. Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. 2025. MOSAIC: Modeling social AI for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6401–6428, Suzhou, China. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Sarvapali D. Ramchurn, Trung Dong Huynh, and Nicholas R. Jennings. 2004. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19:1 – 25.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024. Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7895–7903. International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. 2025. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *ArXiv*, abs/2505.18279.

Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents. *ArXiv*, abs/2402.17896.

Julian B. Rotter. 1971. Generalized expectancies for interpersonal trust. *American Psychologist*, 26:443–452.

Jordi Sabater-Mir and Carles Sierra. 2005. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24:33–60.

Yuki Sakamoto, Takahisa Uchida, and Hiroshi Ishiguro. 2025. Value-based large language model agent simulation for mutual evaluation of trust and interpersonal closeness. *Scientific Reports*, 15.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Philipp J. Schneider, LIN TIAN, and Marian-Andrei Rizoiu. 2025. Learning to make friends: Coaching LLM agents toward emergent social ties. In *Workshop on Scaling Environments for Agents*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yoav Shoham and Moshe Tennenholtz. 1995. On social laws for artificial agent societies: Off-line design. *Artif. Intell.*, 73:231–252.

Jaime Simão Sichman, Rosaria Conte, Cristiano Castelfranchi, and Yves Demazeau. 1997. A social reasoning mechanism based on dependence networks. In *European Conference on Artificial Intelligence*.

Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2024. Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *Proceedings of the ACM on Human-Computer Interaction*, 9:1 – 33.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alexandre Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *ArXiv*, abs/2504.12516.

Michael Wooldridge and Nicholas R. Jennings. 1995. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10:115 – 152.

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. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Andrea Wynn, Harsh Satija, and Gillian K Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *ArXiv*, abs/2509.05396.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? In *ICLR 2024 Workshop: How Far Are We From AGI*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *ArXiv*, abs/2509.23055.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. MasRouter: Learning to route LLMs for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15549–15572, Vienna, Austria. Association for Computational Linguistics.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. 2025a. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *International Conference on Representation Learning*, volume 2025, pages 75389–75428.

Wentao Zhang, Liang Zeng, Yuzhen Xiao, Yongcong Li, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. 2025b. Agentorchestra: Orchestrating hierarchical multi-agent intelligence with the tool-environment-agent(tea) protocol. *ArXiv*, abs/2506.12508.

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

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025. 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)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

Agent System Prompt in GovSim

```
You are John. [Task-related prompt]

You and other fishermen hold various attitudes
towards each other. Attitude can be quantified
using a score within the range of [-1, 1],
where 1 represents the strongest positive
attitude, such as kindness and affection, and
-1 represents the most negative attitude, such
as hostility and resentment.

John's attitude score towards Kate is 1. Kate's
attitude score towards John is also 1. John's
attitude score towards Jack is -1. Jack's
attitude score towards John is also -1. Kate's
attitude score towards Jack is -1. Jack's
attitude score towards Kate is also -1.

You make all your decisions considering the
attitudes among all of you in the group.
```

Figure 9: Example GovSim system prompt for agent John under the *Attitude* relation type. It combines the task instructions with the relation definition and verbalized signed edges of the relation network.

A GovSim Details

A.1 Simulation Dynamics

Resource dynamics. At the start of month t, the shared stock $h(t)$ is available. Each simulation lasts $T = 12$ months or until the stock collapses. A collapse occurs when $h(t)$ falls below a small constant $C = 5$, at which point no further extraction is possible. The sustainability threshold $f(t)$ is the maximum combined harvest that leaves the next month's stock at least as large as the current stock. With GovSim's doubling-regeneration dynamics, $f(t) = h(t)/2$. After harvest, the resource regenerates according to

$$h(t+1) = \min\left\{2\left[h(t) - \sum_{i=1}^{n} r_{i,t}\right], 100\right\}, \tag{4}$$

where $r_{i,t}$ is the quantity harvested by agent i at month t. The remaining stock doubles, up to the capacity of 100. To maintain the resource, the group must therefore harvest no more than half of the current stock in total.

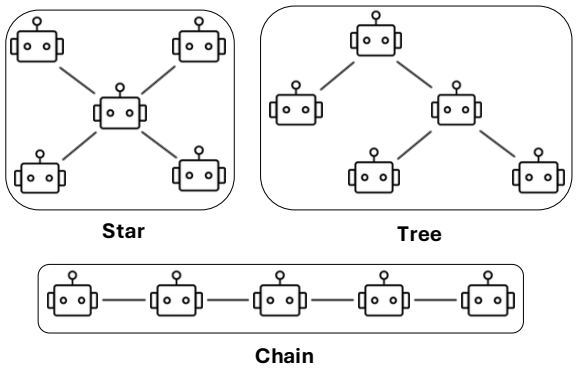


Figure 10: Five-agent relation topologies used in GovSim. Each topology contains four relational edges and changes only which agent pairs receive explicit relational context; it does not change message routing.

Monthly phases. Each month comprises two phases. In the *harvest phase*, agents privately choose how much of the resource to extract, subject only to the available stock. Decisions are submitted simultaneously; after execution, each agent's choice and the remaining stock $h(t) - \sum_i r_{i,t}$ are disclosed to all agents. In the *discussion phase*, agents converse in natural language. They may share beliefs about sustainability, negotiate, appeal to moral principles, or adjust their strategies. Relational priors supply context for these decisions and discussions; agents still choose their own words and harvest quantities. After discussion, the new stock $h(t+1)$ is computed as above. The process repeats until $t = T$ or the stock collapses.

A.2 Evaluation

We use the five metrics defined by the original GovSim benchmark.

- **Survival time** (m) — The survival time is the longest number of months during which the shared resource remains above the collapse threshold C: $m = \max(\{t \in \mathbb{N} \mid h(t) > C\})$.
- **Survival rate** (q) — Proportion of runs that achieve maximum survival time, i.e., $m = 12$: $q = \frac{\#\{m=12\}}{\#\text{runs}}$.
- **Total gain** (R) — Average total gain per agent over the survival period: $R = \frac{1}{|I|}\sum_{i=1}^{|I|} R_i$, where $R_i = \sum_{t=1}^{T} r_{i,t}$ is the total gain of agent i.
- **Inequality** (e) — Gini-based equality score across the total gains of each agent: $e = 1 - \frac{\sum_{i=1}^{|I|}\sum_{j=1}^{|I|}|R_i - R_j|}{2|I|\sum_{i=1}^{|I|} R_i}$.
- **Over-usage** (o) — Amount of unsustainable

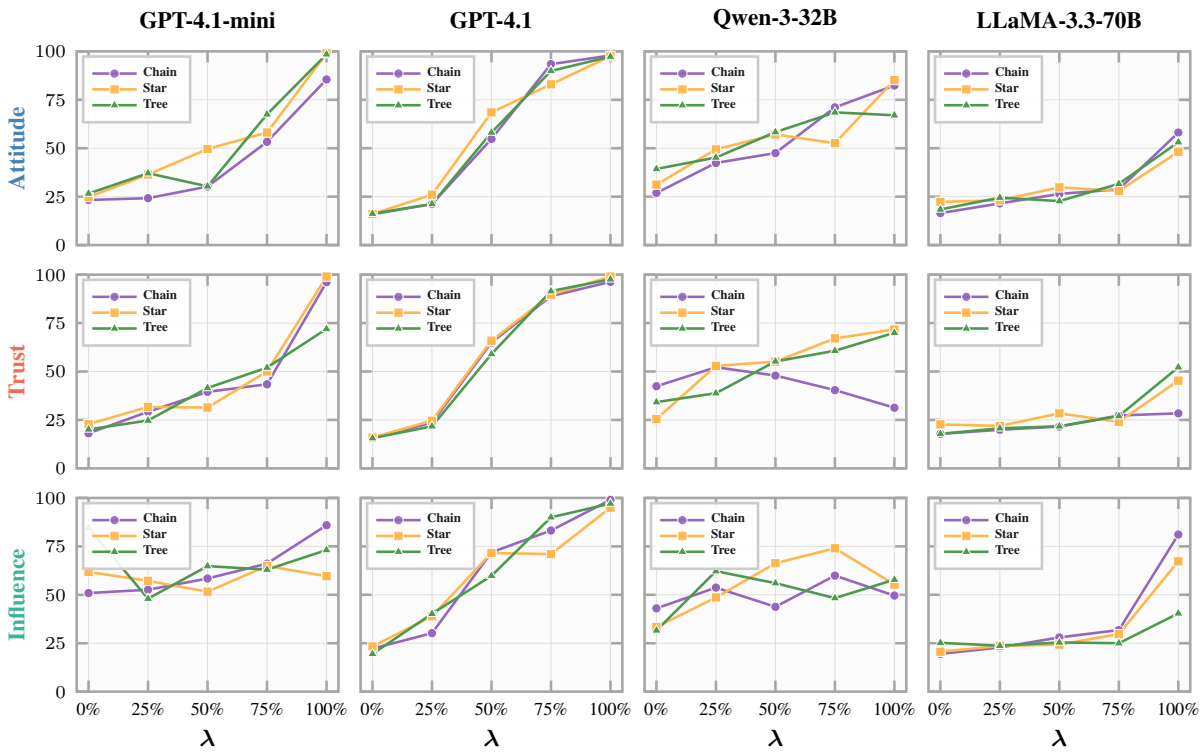


Figure 11: Five-agent GovSim overall score S across model backbones, relation types, topologies, and values of λ.

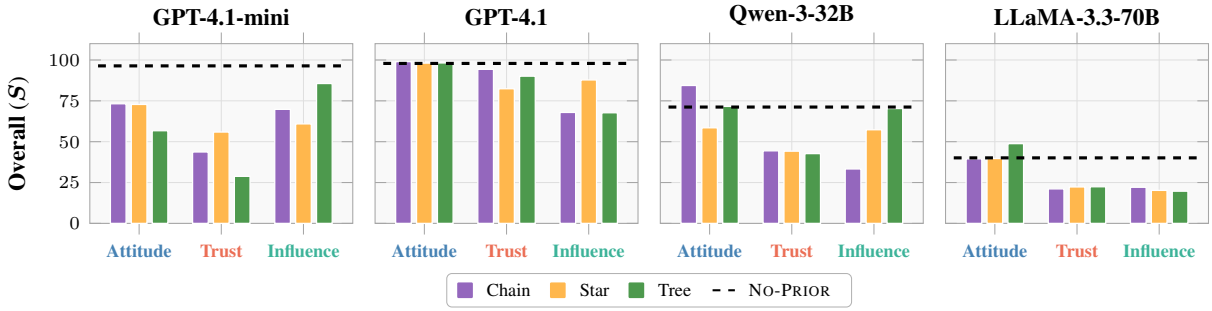


Figure 12: Five-agent GovSim overall score S under EXPLICIT-NEUTRAL, by model backbone, relation type, and topology. Dashed lines show the corresponding NO-PRIOR baselines.

behavior across a simulation. It is the percentage of actions across the simulation that exceed the sustainability threshold: $o = \frac{\sum_{i=1}^{|I|}\sum_{t=1}^{T}\mathbf{1}(r_{i,t}>f(t))}{|I|\cdot m}$.

We also use the aggregate score S to integrate the original five metrics. We first normalize each metric by its theoretical maximum value, making sure that larger values correspond to better performance. For example, for inequality/equality e, the normalized metric is $\tilde{e} = 1 - e$ when e is represented as inequality. We then take a weighted linear mean:

$$S = w_1\tilde{m} + w_2\tilde{q} + w_3\tilde{R} + w_4\tilde{e} + w_5\tilde{o}, \quad (5)$$

where $w_1 = w_2 = w_5 = \frac{2}{8}$ and $w_3 = w_4 = \frac{1}{8}$, since survival time, survival rate, and over-usage directly assess system sustainability.

A.3 GovSim Bootstrap Intervals

For each model and relation type in the complete three-agent setting, we estimate the all-positive minus all-negative score difference, $\Delta S = S_{\lambda=1} - S_{\lambda=0}$. Each condition contains five simulation runs. We construct a bootstrap sample by selecting five runs with replacement from each condition, independently, keeping all agents and months within each selected run together. We then recompute S for each condition and calculate their difference. With five runs per condition, we can enumerate all possible ordered resamples and calculate the score difference for every pair of endpoint resamples. The 2.5th and 97.5th percentiles of these differences give the 95% bootstrap confidence in-

Model	Relation	ΔS	95% bootstrap CI
GPT-4.1-mini	Attitude	$+68.4$	$[+43.2, +81.5]$
GPT-4.1-mini	Trust	$+39.0$	$[+10.1, +67.7]$
GPT-4.1-mini	Influence	$+78.1$	$[+76.7, +81.1]$
GPT-4.1	Attitude	$+81.4$	$[+80.8, +82.2]$
GPT-4.1	Trust	$+80.7$	$[+79.5, +82.1]$
GPT-4.1	Influence	$+82.2$	$[+81.6, +82.8]$
Qwen-3-32B	Attitude	$+24.7$	$[-13.7, +63.4]$
Qwen-3-32B	Trust	$+50.2$	$[+21.8, +78.4]$
Qwen-3-32B	Influence	-21.1	$[-52.3, +15.8]$
LLaMA-3.3-70B-Instruct	Attitude	$+36.8$	$[+7.7, +65.7]$
LLaMA-3.3-70B-Instruct	Trust	$+36.5$	$[+7.8, +65.4]$
LLaMA-3.3-70B-Instruct	Influence	$+38.1$	$[+8.8, +67.1]$

Table 2: Three-agent GovSim all-positive minus all-negative score differences, $\Delta S = S_{\lambda=1} - S_{\lambda=0}$, with 95% complete-run bootstrap confidence intervals. Each endpoint contains five runs. Differences are measured in score points.

terval. Ten of the twelve intervals lie entirely above zero, supporting higher sustainability scores under all-positive relations in those comparisons. For Qwen-3-32B under *Attitude* and *Influence*, the intervals span zero and the direction of the difference remains uncertain. Tab. 2 reports all differences and intervals.

B Debate Details

B.1 Debate Dynamics

We follow the AutoGen debate protocol.[1] We model a K-agent, T-round debate for question x_n. Let $z_{i,n}^{(t)}$ denote the complete textual response produced by agent $i \in \{1, \dots, K\}$ at round t, including its selected answer and explanation. Let $\mathcal{Y}_n$ be the answer-option set for x_n, and let

$$a_{i,n}^{(t)} = \mathrm{Extract}\left(z_{i,n}^{(t)}\right) \in \mathcal{Y}_n \tag{6}$$

denote the extracted answer option. We write $Z_n^{(t)} = (z_{1,n}^{(t)}, \dots, z_{K,n}^{(t)})$ for the agents' responses at round t, and $Z_{-i,n}^{(t)} = (z_{j,n}^{(t)})_{j \neq i}$ for the responses of all agents other than i.

In round 1, agents answer independently given the question and their relational prompts:

$$z_{i,n}^{(1)} = \mathrm{LLM}_i(x_n \mid \Pi_i(r, G_R)) \,. \tag{7}$$

For each revision round $t = 2, \dots, T$, agent i retains its accumulated dialogue history $H_{i,n}^{(t-1)}$ and is additionally shown the other agents' responses from the preceding round:

$$z_{i,n}^{(t)} = \mathrm{LLM}_i\Big(x_n \mid \Pi_i(r, G_R), H_{i,n}^{(t-1)}, Z_{-i,n}^{(t-1)}\Big) \,. \tag{8}$$

The debate trajectory for question x_n is the sequence $\{Z_n^{(t)}\}_{t=1}^{T}$. We use greedy decoding for all models and three answer rounds in total: one initial answer followed by two revisions.

B.2 Evaluation

All debate metrics are computed from the extracted answer options $a_{i,n}^{(t)}$. For a dataset of N questions, question x_n is in full consensus at round t if $a_{1,n}^{(t)} = \cdots = a_{K,n}^{(t)}$. Let

$$I = \Big\{n \in \{1, \dots, N\} : \exists\, i, j \text{ such that } a_{i,n}^{(1)} \neq a_{j,n}^{(1)}\Big\} \tag{9}$$

be the set of questions that begin with disagreement, and let

$$C = \Big\{n \in I : a_{1,n}^{(T)} = \cdots = a_{K,n}^{(T)}\Big\} \tag{10}$$

be the subset that reaches full consensus in the final round. We construct I and C separately for each condition because relational prompts can change both initial and final answers. The final-round consensus rate is the fraction of initially disagreeing questions that reach consensus:

$$\mathrm{CR} = \frac{|C|}{|I|}. \tag{11}$$

For subjective tasks, we use OpinionQA, a dataset derived from opinion polls and designed to evaluate alignment with human opinions on subjective issues. These questions have no universally

[1] https://microsoft.github.io/autogen/stable/user-guide/core-user-guide/design-patterns/multi-agent-debate.html

correct answers, so CR measures how often initially disagreeing agents reach unanimous agreement.

For objective tasks, we use MMLU-Pro and GPQA-Diamond and apply the same definition of CR. We obtain the final debate answer by majority voting over the agents' final-round answer options:

$$\hat{y}_n = \arg\max_{y \in \mathcal{Y}_n} \sum_{i=1}^{K} \mathbf{1}\left[a_{i,n}^{(T)} = y\right]. \qquad (12)$$

Since each question has a ground-truth answer y_n, we report the accuracy of $\hat{y}_n$ over all N questions. For each $n \in C$, let $c_n = a_{1,n}^{(T)} = \cdots = a_{K,n}^{(T)}$ denote the final consensus answer. The *Consensus Correctness Rate* (CCR) is the fraction of these newly formed consensuses that are correct:

$$\mathrm{CCR} = \frac{|C_{\mathrm{corr}}|}{|C|}, \qquad C_{\mathrm{corr}} = \{n \in C : c_n = y_n\}. \qquad (13)$$

Thus, ACC measures final-answer accuracy on the full dataset, CR measures how often initial disagreement ends in consensus, and CCR measures the correctness of those consensuses.

B.3 Debate Bootstrap Intervals

For the open-weight MMLU-Pro team under *Influence*, we estimate the all-positive minus all-negative differences in CR and CCR using a paired question bootstrap. We sample 1,000 questions with replacement from the evaluated questions and use the same sampled questions in both conditions, retaining each question's recorded responses from all agents across all rounds. For each condition, we recompute CR as the fraction of initially disagreeing questions that reach final consensus, and CCR as the fraction of those consensuses that are correct. Initial answers and final consensus can differ between conditions, so the questions contributing to each rate are determined separately. We then subtract the all-negative rate from the all-positive rate. We repeat this procedure 100,000 times using random seed 0 and take the 2.5th and 97.5th percentiles of the resulting differences as the endpoints of the 95% bootstrap confidence interval. The differences are $+43.1$ percentage points for CR (95% CI: $[+37.8, +48.2]$) and -8.5 points for CCR (95% CI: $[-15.8, -1.0]$). In this setting, all-positive relations produce consensus more often, but a smaller fraction of those consensuses is correct.

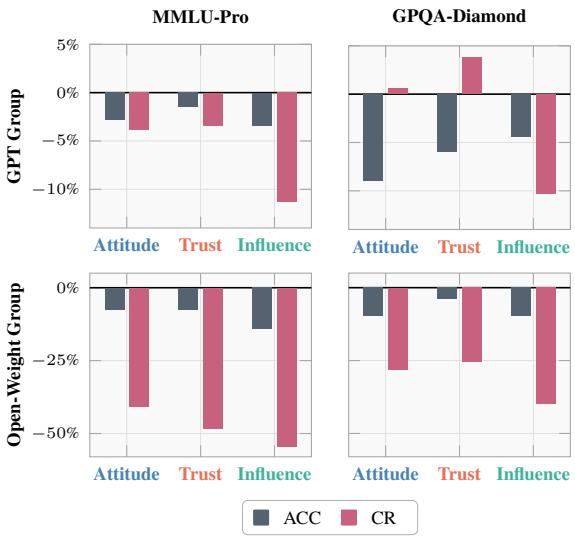


Figure 13: Percentage changes in debate consensus rate (CR) and accuracy (ACC) at $\lambda = 0$ relative to NO-PRIOR, by model group, dataset, and relation type. Solid lines mark zero change.

B.4 Initial-to-Final Answer Transitions

We complement CCR with transitions that describe which initial answer patterns end in correct or wrong consensus. We analyze the open-weight MMLU-Pro team under *Influence*, comparing the all-negative ($\lambda = 0$) and all-positive ($\lambda = 1$) conditions used in Sec. 5.2. We use the paired question-level bootstrap procedure described in App. B.3.

We report three question-level measures. *Wrong-majority correction* is the fraction of initial 2–1 splits with a wrong majority that end in 3–0 correct consensus. *Correct-majority failure* is the fraction of initial 2–1 splits with a correct majority that end in 3–0 wrong consensus. *Wrong Consensus Formation* (WCF) is the fraction of all initially non-unanimous questions, including 1–1–1 splits, that end in 3–0 wrong consensus. Each rate uses the questions with its specified initial answer pattern, determined separately in each condition. The comparison is therefore between all-positive and all-negative relations within each measure. For each measure, we estimate the all-positive minus all-negative difference using 100,000 paired question-level bootstrap samples. The same question indices are used at both endpoints, and the corresponding conditional denominator is recomputed in every sample.

Tab. 1 in Sec. 5.2 reports the transition rates, endpoint differences, and confidence intervals.

Prompt variant	all -1	all $+1$	ΔCR (points)	Paired 95% CI
Original	42.0%	73.9%	$+31.9$	$[+13.9, +49.3]$
Paraphrased	30.8%	70.4%	$+39.6$	$[+22.8, +55.9]$
Reordered	33.3%	76.6%	$+43.3$	$[+25.9, +59.7]$

Table 3: Prompt wording and statement-order check for the open-weight team under *Influence* on the 100-question MMLU-Pro subset. CR is measured among questions with initial disagreement. ΔCR is all-positive minus all-negative, in percentage points. All three formulations produce higher CR under all-positive relations.

C Consensus and Accuracy under Fully Negative Priors

Relative to NO-PRIOR, fully negative priors produce a larger proportional decline in consensus than in accuracy in 10 of the 12 comparisons (Fig. 13). We measure relative percentage change as $100(M_{\lambda=0} - M_{\text{NO-PRIOR}})/M_{\text{NO-PRIOR}}$ for $M \in \{\text{CR}, \text{ACC}\}$. This pattern holds in all six open-weight conditions and all three GPT MMLU-Pro conditions, as well as GPT GPQA-Diamond under *Influence*. The exceptions are GPT GPQA-Diamond under *Attitude* and *Trust*, where CR is slightly above NO-PRIOR while ACC is below it. For open-weight MMLU-Pro, CR decreases by 40.8–54.4%, compared with 7.5–14.1% for ACC. These are proportional changes from each metric's own baseline; the main debate figures instead report percentage-point differences.

D Prompt Robustness to Wording and Statement Order

We test whether all-positive *Influence* still produces more consensus than all-negative *Influence* when the relational wording is paraphrased or the relation statements are reordered.

D.1 Setup and Prompt Variants

We used the first 100 questions from the main experiment's MMLU-Pro sample. The three-agent team comprised Mixtral-8x22B-Instruct-v0.1, Qwen-3-32B, and LLaMA-3.3-70B-Instruct. The debate protocol and decoding settings matched the main experiments. We evaluated six conditions: the original, paraphrased, and reordered *Influence* prompts, each with all three mutual relations set to either -1 or $+1$.

The original condition used the main experiment's *Influence* wording and listed the three mutual-pair blocks in the order A–B, A–C, B–C. The paraphrased condition replaced terms such as "influence score," "follow their lead," and "contrarian influence" with a response-score description in which an agent moves toward or away from another agent's recommendations. Agent identities, pair mappings, statement directions, and numerical signs were unchanged. The reordered condition retained the original text verbatim and changed only the pair-block order to B–C, A–C, A–B.

The relation-definition portion contained 82 words in both the original and paraphrased versions, and the six agent-specific relation statements contained 45 words in both. The variants are matched in word count; model-specific token counts may differ. The question prompts and subsequent-round decision prompts were unchanged across all conditions.

D.2 Evaluation

We report consensus rate (CR) for questions whose three initial answers were non-unanimous. A question counts as a final consensus only when all three final answers are the same option. For each prompt variant, we subtract the all-negative CR from the all-positive CR. We estimate 95% bootstrap percentile intervals using 100,000 paired question-level bootstrap samples. The same resampled question indices are used across all six conditions, and the conditional denominators are recomputed in every sample.

D.3 Results

All-positive relations produce more consensus than all-negative relations with the original, paraphrased, and reordered prompts, and all three paired intervals exclude zero (Tab. 3). On this 100-question MMLU-Pro subset with the open-weight team, the increase in consensus therefore persists under both tested changes. The size of the increase varies with wording and statement order.

E Use of AI Assistants

The manuscript uses AI assistance primarily for language editing and restructuring under the author's supervision. AI assistance was not used to

originate the research idea, generate the experimental data, create the reported numerical results, or perform the evaluation.