FRIEDKIN AND JOHNSEN UNVEIL THE DYNAMICS OF MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

As LLM-based multi-agent debate drives advances in agentic AI, a central question arises: How does collective intelligence emerge from interactions among agents? We introduce a framework to study the underlying dynamics through the Friedkin–Johnsen (FJ) model of opinion formation. Across diverse empirical settings, the FJ model accurately reconstructs the evolution of agents' beliefs while providing interpretable measures of stubbornness and persuasiveness that characterize how agents engage in consensus formation. We systematically compare FJ with alternatives, including DeGroot dynamics and a statistical-physics-inspired model, and find advantages in trajectory reconstruction, consistency with interventions, and interpretability of communication patterns. Analyzing different communication topologies and group sizes, we find that stubborn agents become less prevalent in larger groups, while persuasive agents become more prominent. Our results establish a simple, interpretable framework for understanding how individual interactions shape LLM multi-agent collaboration.

1 INTRODUCTION

When a group of language-model agents agrees on an answer, what has the deliberation accomplished? Several agents may have independently reached the same conclusion, or the group may have followed one agent that held firm. Both processes can produce the same final consensus, but they have different implications about the value and reliability of multi-agent debate. To understand a collective decision, we need to examine how individual beliefs change and the agents's role in it.

Multi-agent systems (MASs) of large language model (LLM) agents have attracted attention for reasoning and collaborative problem solving (Du et al., 2024; Wu et al., 2024; Qian et al., 2024; Chen et al., 2025; Hong et al., 2024). Their potential was recently illustrated by a large-scale, $\geq$ 10,000-agent system proposing a solution to the Navier–Stokes Millennium Prize problem (OpenAI, 2026). At the same time, evidence that multi-agent deliberation reliably outperforms single-agent approaches or static ensembles remains mixed (Smit et al., 2024; Zhang et al., 2025; Tran & Kiela, 2026). Understanding how agents influence a debate is therefore important both for explaining promising results and for identifying when additional agents only amplify an existing view.

Existing accounts have identified phenomena such as biased consensus, while statistical-physics-inspired approaches characterize collective dynamics at the population level (Okawa, 2026) or are limited to binary beliefs (El et al., 2026). These perspectives illuminate group behavior primarily on the macro level, yet, leave a complementary question on the micro level unanswered: On a particular problem, who retains their initial view, who changes their mind, and whose view spreads through the group? Answering it requires a model at the level of individual agents and individual questions. We study deliberations in which agents exchange natural-language messages over multiple rounds and privately report a distribution over the possible answers after each round. This allows us to observe belief trajectories arising from linguistic interaction without making the reported probabilities themselves part of the debate.

Our central finding is that the Friedkin–Johnsen (FJ) model of opinion formation (Friedkin & Johnsen, 1990) reconstructs these belief trajectories remarkably well. Earlier work has used FJ to study belief propagation and security risks in LLM deliberation (Abedini et al., 2026) and motivated a mixture of expert description (Bause et al., 2026). We investigate its broader value as a diagnostic framework for how individual debates unfold and systematically compare it to alternative

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.

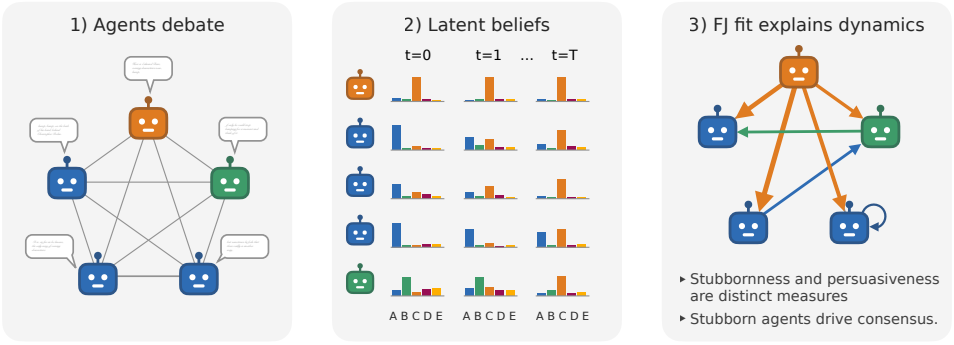


Figure 1: Overview. (1) Agents debate a multiple-choice question by exchanging text only. (2) After each round, every agent privately reports a belief vector, which is never shown to the other agents. (3) A Friedkin–Johnsen fit makes distinct agent roles visible: in this example, one agent retains its initial belief while other agents move toward it.

modeling approaches. FJ represents an agent's next belief as a linear weighted combination of its initial belief, its current belief, and its neighbors' beliefs. Considering the complexity of interactions conducted through natural language, it is surprising that such a simple model with strong constraints can describe most observed trajectories.

A particular advantage of FJ is that it separates two ways an agent can shape a debate. An agent may be *stubborn*, repeatedly giving weight to its own initial belief. Separately, it may be *persuasive* and receive weight in the update of other's beliefs. Fitting FJ separately to each question lets us identify the roles of agents in context and examine the communication patterns they create, including agents that transmit influence between peers and agents whose initial view becomes dominant.

We evaluate FJ across question sets, language models, group sizes, and communication graphs, and compare its trajectory reconstruction with alternative dynamics models, including DeGroot averaging (DeGroot, 1974) and a statistical-physics-inspired model (Okawa, 2026; El et al., 2026). The fitted parameters reveal recurring communication patterns and a change in how agents gain influence as groups grow: In the larger systems we study, agents tend to be less stubborn, while persuasion and the transmission of influence through other agents become more prominent. The distinction between stubbornness and persuasiveness is thus consequential both for explaining an individual debate and for understanding how deliberation changes with group size. Even the rare failures of FJ modeling are informative: A debate that introduces a new answer can depart from dynamics that only combine existing beliefs. In addition, how examine how prompt and graph interventions alter fitted behaviors to demonstrate the broad applicability of the FJ model its potential to guide interventions.

Contributions.

- **A question-level model of deliberation.** We fit FJ separately to individual multi-agent debates and show that its constrained dynamics closely reconstruct agents' belief trajectories across diverse settings. We compare its fit with alternative dynamics models.

- **Distinct routes to influence.** We distinguish an agent's attachment to its initial belief from its influence on other agents, enabling us to characterize stubborn, persuasive, bridging, and dominant roles. In the larger groups we study, stubborn agents become less prevalent and persuasion becomes a more prominent route to influence.

- **Interpretable effects of system changes.** We examine how group size, topology, and interventions on prompts and communication links affect fitted behaviors. We also identify cases that the FJ model cannot readily capture, particularly deliberations that introduce a new answer.

2 RELATED WORK

LLM-based MASs and deliberation. Recent work explores the use of LLMs interacting through structured or unstructured communication to improve reasoning and robustness (Wu et al., 2024; Li et al., 2023). Techniques such as self-consistency (Wang et al., 2023) and chain-of-thought prompting (Wei et al., 2022) demonstrate that aggregating diverse opinions and reasoning can outperform single-pass inference. However, existing methods largely lack a principled framework on how opinions evolve through communication.

Opinion dynamics and social influence models. The Friedkin-Johnsen (FJ) model (DeGroot, 1974; Friedkin & Johnsen, 1990; 2011) has been proposed in the social sciences to understand how group consensus is formed, how stubbornness and belief retention interact with peer influence, and how changes in network structure can affect collective decision-making (Parsegov et al., 2017; Tian & Wang, 2018; Abedini et al., 2026). Recently, it has been used to model opinions co-evolving with platform learning systems (Wu et al., 2026) and creating simulation environments of human opinion formation (He et al., 2026; Wang et al., 2025). Recent work (Yazici et al., 2026) has demonstrated that social models can be used to capture belief formation in MASs, the emergence of influence (Bause et al., 2026), to study the systemic risk induced by stubborn agents (Abedini et al., 2026).

Interpretability of MASs. Research on the interpretability of MASs has explored how agents' behavior can be explained, particularly in the context of collaborative problem-solving and collective decision-making (Wooldridge, 2009; Rosenfeld & Richardson, 2019; Lee et al., 2025). A challenge in multi-agent settings is understanding the emergent behaviors that arise from the complex interactions between agents. Our work adds to this growing body of research by providing a framework for explaining the dynamics in MASs through the lens of the FJ model.

3 THE FRIEDKIN-JOHNSEN MODEL

Our work builds on the empirical observation that the dynamics of LLM-based multi-agent deliberation can be well described by the Friedkin-Johnsen (FJ) model (see also Table 5). This correspondence provides a simple and interpretable framework for characterizing how the beliefs of agents evolve over the deliberation rounds.

Friedkin-Johnsen model. Let each agent $i \in V$ in an agentic network $G = (V, E)$ with $n = |V|$ agents hold a belief $b_i(t) \in \Delta^d$, where $\Delta^d \subset [0,1]^d$ is the d-dimensional simplex representing a probability distribution over potential outcomes, e.g., answers to a multiple-choice question. Each agent i is characterized by an innate belief $b_i(0) = s_i \in \Delta^d$, which may be interpreted as the agent's individual prediction before deliberation.

The belief update at time $t+1$ in the *Friedkin-Johnsen* model (Friedkin & Johnsen, 1990) for belief dynamics is defined as (Abedini et al., 2026):

$$b_i(t+1) = \underbrace{\gamma_i s_i}_{\text{Prior Belief Pull}} + \underbrace{(1-\gamma_i)\alpha_i b_i(t)}_{\text{Belief Retention}} + \underbrace{(1-\gamma_i)(1-\alpha_i)\sum_{j \in N_i} w_{ij} b_j(t)}_{\text{Peer Influence Pull}}, \tag{1}$$

where $N_i = \{j \mid (j,i) \in E(G)\}$ are an agents neighbors, $\gamma_i \in [0,1]$ denotes the attachment to innate beliefs, also called *stubbornness*, $\alpha_i \in [0,1]$ represents the weight given to the previous belief, and $W = [w_{ij}]$ is a row-stochastic influence matrix, where $w_{ij} \geq 0$, $\sum_j w_{ij} = 1$, and $w_{ii} = 0$ (as self-loops are covered by α_i), as well as $w_{ij} = 0$, if $j \notin N_i$.

The FJ parameters provide a natural basis for characterizing the roles of individual agents in the deliberation. In particular, we distinguish between an agents tendency to stick to their own belief through γ_i and $(1-\gamma_i)\alpha_i$, and their susceptibility to external influence, $(1-\gamma_i)(1-\alpha_i)$. However, the interplay between the parameters of different agents also affects the deliberation and its outcome. Therefore, we define several measures based on the FJ parameters and agent beliefs.

Stubbornness. The *stubbornness* of agent i is its γ_i value. It directly determines the weight to its initial opinion at each round.

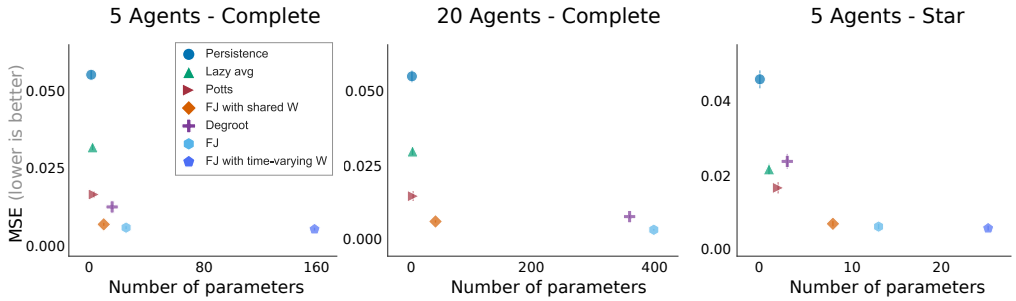


Figure 2: Model complexity and goodness of fit across three experimental settings. Each point shows the number of free parameters and the mean reconstruction MSE, with bootstrapped 95% confidence intervals. The FJ model improves substantially over simpler averaging and Potts dynamics, while the more expensive time-varying W variant yields only modest additional improvements. The time-varying W model is not shown for the 20-agent condition, since its number of free parameters would make it overparameterized relative to the available data.

Persuasion. *Persuasion* ψ_{ji} quantifies the weight agent i assigns to agent j's belief, it is defined as $\psi_{ji} = (1-\gamma_i)(1-\alpha_i)w_{ij}$, where j is the sender and i the receiver. It takes into account both, influence weights w and the agents' susceptibility to external influence.

Effective Persuasiveness. The effective *persuasiveness* of agent i, $\pi_i^{pers} = \sum_{j \neq i} \psi_{ij}$ captures the total influence that agent i has on its neighbors.

Confidence. We define the *confidence* of agent i based on the normalized entropy of its initial belief: $C_i(S) = 1 + \frac{1}{\log d} \sum_{c=1}^{d} s_{i,c} \log s_{i,c}$, where d is the number of possible outcomes and $C_i(S) \in [0,1]$. The *relative confidence* of agent i is defined as $R_i(S) = \frac{C_i(S)}{C_{(n-1)}(S)}$ where $C_{(n-1)}(S)$ is the second most confident agent in the system.

Consensus. Technically, consensus could be defined as all agents belief vectors $b(t)$ being equal for some t. Here, we only care about the answer agents actually believe to be correct. Therefore *consensus* is reached in round t if $\exists j^*$ such that $\forall i \in V(G) : j^* = \arg\max_j b_{ij}(t)$.

Top answer set. We define the *top answer set* in round t as the set of answer options that most of the individual agents of the agentic network believe to be correct: $\text{top}(t) = \{j \in \varrho(t) \mid \nexists j' \in \varrho : \#_{\varrho(t)}(j') > \#_{\varrho(t)}(j)\}$, with $\varrho(t) = \{\!\{\arg\max_j b_{ij}(t) \mid i \in V(G)\}\!\}$.

Stability. We say a consensus or a top answer set is *stable* in round t if it exists, and did not change in subsequent (observed) rounds of deliberation.

Alignment with majority. Whether an agent i initially *aligns* with the majority is a binary variable $al_i(S) = \begin{cases} 1, \text{ if } \arg\max_j(s_{ij}) \in \text{top}(0) \\ 0, \text{ else} \end{cases}$, indicating whether the agents' initial answer is in the initial top answer set.

4 From Multi-Agent Deliberation to FJ Dynamics

Debate protocol. Each debate involves n agents connected by a communication graph and given the same multiple-choice question. Agents receive a system prompt specifying their identity and collaborative role, and optionally additional conditions. In the first round, agents answer independently. In subsequent rounds, agents receive previous messages of their neighbors and produce a new response. Agents are not instructed *how* to reason about their next answer, to avoid steering the debate and enabling us to observe emergent tendencies in their debate.

Belief vectors. In addition to its message, each agent reports a probability distribution over the d answer options, their belief vector $b_i(t) \in \Delta^{d-1}$. These belief vectors form the trajectories analyzed in our experiments. Importantly, agents exchange only text containing reasoning, arguments, and answers. Belief vectors are never shown to other agents.

FJ interpretation. We independently fit an FJ model to each observed deliberation trajectory using the belief vectors reported at every round. The fitted parameters provide a compact, interpretable explanation of how beliefs evolve from the observed initial state. We assess the fit by rolling the model forward from these initial beliefs and measuring how accurately it reconstructs the subsequent trajectory. Thus, we use the FJ model as a diagnostic tool for explaining observed deliberations, rather than for predicting unseen deliberations.

5 Implications of FJ fit for MASs

The empirical fit of the FJ model has several implications, beyond simply providing an interpretable description. Concretely, the FJ model imposes several strong restrictions on the dynamics. If the observed dynamics can be described by the FJ model, then the MAS inherits several properties, such as the tendency toward consensus (under specific conditions), the ability for stubborn agents to overtake the deliberation, and the impossibility of negative influence weights.

Restricted dynamics. Equation 1 restricts how an agent updates its beliefs based on its own belief and the beliefs of neighbors it interacts with. Specifically, all weights w are non-negative, meaning that agents can only move closer to their neighbors opinions (or stick to their own). Additionally, the FJ dynamics only produce mixtures of initial beliefs, excluding everything that was absent initially. This means that the FJ does not capture the creation of new information through interaction, making the fit of the FJ model surprising.

Tendency for consensus. The FJ model predicts a strong tendency towards consensus (i.e. the highest probability outcome being the same for all agents in the system): as long as agents are not totally stubborn and the agentic network allows for sufficient interaction between the agents, the repeated peer influence pull tends to align the agents' beliefs. A tendency for consensus has been observed in LLM-based MASs (Reza, 2026). Here, this tendency follows directly from the models structure and parameters.

Fast convergence. The theoretical properties of the FJ model can also help explain how fast a stable state is reached in deliberation. Fast convergence, for example can be seen especially when the agents are very agreeable or the network is highly connected. Therefore, if the FJ model can accurately describe the dynamics in LLM-based MASs, the convergence speed will depend on the agents stubbornness, as well as the topology.

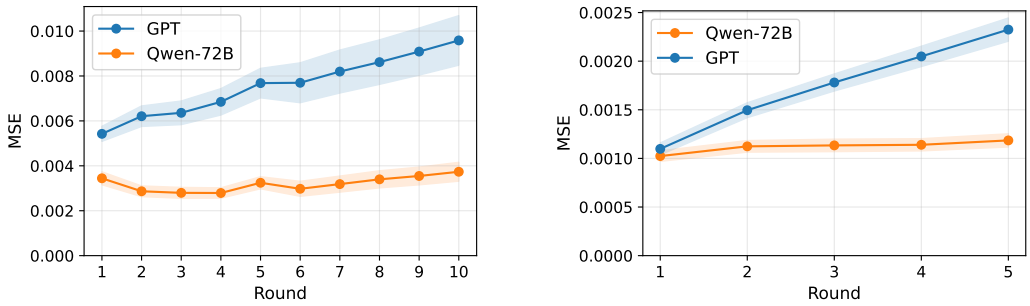
Influence of stubborn agents. In the FJ model, agents with high stubbornness, or even more general, high self-weight, tend to have a higher influence on the outcome, since agents that are more susceptible to external influence will move closer to the stubborn agents' belief each deliberation round. In the context of LLM-based MASs, this has, for example, security implications. Recent work (Abedini et al., 2026) showed that stubborn agents can overtake the whole deliberation. Therefore, the FJ model provides a natural way to describe how stubborn agents can shape the outcome of a deliberation, without having explicit control over the communication.

The above properties are theoretical implications of the FJ model, when applied to LLM-based multi-agent deliberation. We empirically investigate in Section 6, whether these predicted properties are actually reflected in the observed dynamics.

6 Empirical evaluation

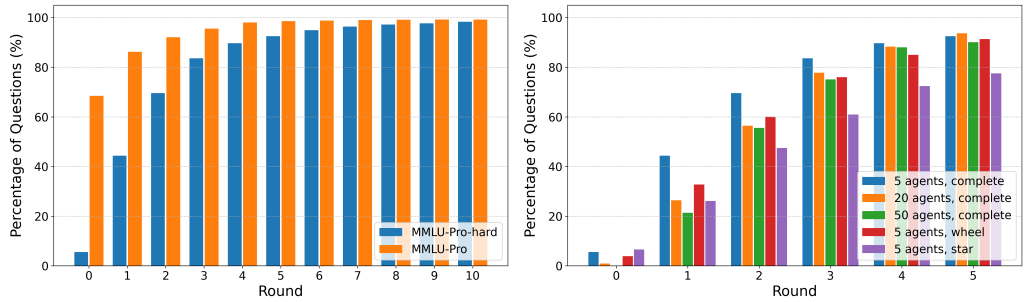
With our empirical analysis, we show that the FJ framework accurately describes the behavior of MASs across settings. We can therefore use it to answer the following questions: 1) what are emergent communication patterns in MASs, 2) what drives consensus in MASs and 3) what are the characteristics of cases that do not fit FJ dynamics? We study these questions under different topologies and with varying number of agents. Lastly, we study the effect of interventions on prompt and graph structure on our metrics of interest: persuasiveness and stubbornness.

Setup. Our experimental evaluation considers the MMLU-Pro (Wang et al., 2024), BBQ (Parrish et al., 2022) and CommonsenseQA (CSQA) (Talmor et al., 2019) dataset. We adopt the setup and subset of 100 questions of CSQA from Abedini et al. (2026) and sample 300 questions from MMLU-Pro and BBQ each (balanced for the categories). For a large part of the MMLU-Pro dataset agents



(a) MMLU-Pro-Hard (b) All datasets

Figure 3: Goodness of fit (MSE) per round. a) MMLU-Pro-Hard, split by model. b) pooled over every question across all datasets and models (excluding MMLU-Pro-Hard).



(a) different difficulty level of questions (b) different number of agents/topologies

Figure 4: We show in which round the system reached a stable consensus (no changes in subsequent rounds). a) We often observe initial stable consensus for MMLU-Pro, while for MMLU-Pro-hard, a consensus is formed later. b) Results for MMLU-Pro-hard: Increasing the number of agents tends to delay consensus. Graph topology also has a big impact on the consensus speed.

agreed on the correct answer in the initial round prior to the deliberation, therefore, we additionally use a subset of 333 hard questions of MMLU-Pro and label it MMLU-Pro-hard (see Appendix G for details). These harder questions are more representative of tasks that warrant the use of MASs and yield more interesting dynamics.

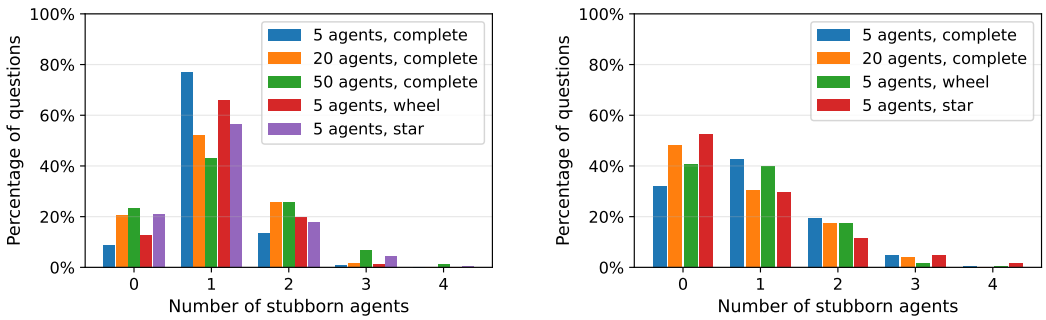
We conduct experiments with three representative language models, *GPT-5.4 Mini*, *Qwen2.5-14B-Instruct*, and *Qwen2.5-72B-Instruct-GPTQ-Int8*, enabling comparison across different model scales.

For the user prompts, communication between agents, and generating prompts for the next deliberation round, we use the setup of Abedini et al. (2026). The prompts can be found in Appendix H. We use $\{5, 20, 50\}$ agents, communicating in a *complete, star, wheel loop, or chain* graph structure over $\{5, 10\}$ rounds. In the main paper, we focus mainly on the results of the two larger models on MMLU-Pro-hard, with a complete graph structure, 5 rounds, and 5 agents, additional results are reported in the appendix. We run most experiments over 3 different seeds, however, due to resource constraints, we do not run all different combinations and settings.

Comparing FJ against alternatives. For comparison, we also fit the following other models to the (task-level) dynamics: the DeGroot model (DeGroot, 1974), and the Ising model by Okawa (2026).

6.1 Evaluating FJ fit

We observe that FJ describes the dynamics well across graph topologies, datasets and models (Tables 5 and 6 in the appendix). The fit remains stable across rounds and the FJ model describes the dynamics well throughout, as can be seen in Figures 3 and 2. The good fit, especially in the early



(a) GPT

(b) Qwen-72B

Figure 5: Distribution of the number of stubborn agents ($\gamma > 0.8$) per system on MMLU-Pro on a subset of hard samples, for GPT (left) and Qwen-72B (right).

rounds, suggests that, interestingly, the agents decide how to weight their neighbors and become stubborn very early in the deliberation.

In rare cases, the FJ model does not accurately describe the behavior of the MAS. For our further analysis in Section 6.3, we set a threshold on the round level goodness of fit, requiring it to be strictly below 0.075 and exclude trajectories that do not meet this threshold. The excluded set only accounts for 2% of all trajectories. The dominant explanation for the bad fit is that the MAS introduces a new answer, not initially given by any agent. This behavior does not conform to FJ dynamics, and explains 88% of the bad fits. It is interesting, that this potentially useful behavior where agents bring up new conclusions during the deliberation is so rare.

6.2 Comparison with alternative dynamics models

To assess whether the strong fit of the FJ model is specific to the FJ dynamics rather than simply a consequence of model flexibility, we compare it against several alternatives with different inductive biases and numbers of free parameters, including persistence, lazy averaging, a deterministic Potts mean-field model, DeGroot dynamics, and restricted and more flexible variants of the FJ model. Each model is fitted independently to the same observed belief trajectories and evaluated by rolling it forward from the observed initial beliefs. Figure 2 shows the resulting trade-off between model complexity and reconstruction error. Full model definitions, fitting details, parameter counts, and numerical results are given in Appendix F.

Notably, `FJ with shared W` also fits well. It gives each agent one influence score that is used by all of its neighbors, while setting $\gamma = 0$. On a complete graph, this reduces the number of influence parameters from $n(n-2)$ to $n-1$, while still outperforming DeGroot in every setting. This is consistent with some agents being influential across the whole system rather than only for specific peers. We still use the full W because it fits better and lets us study routing between individual pairs of agents. If the main goal is good average reconstruction with few parameters, `FJ with shared W` is a useful alternative, while the full W is more robust on complex trajectories.

6.3 Emergent communication

Emergent communication patterns. We investigate communication patterns that arise, specifically in terms of stubbornness and persuasiveness. We say a agent i is stubborn if $\gamma_i > 0.8$. And, we say agent j persuades agent i if $\psi_{ij} \geq 2/|N_i|$. Scaling this threshold by the number of neighbors keeps the metric comparable across graph topologies and sizes. We consider an agent persuasive if it persuades at least one agent.

We observe agent-level communication patterns: an agent can be *neither stubborn nor persuasive*, *stubborn but not persuasive*, *both stubborn and persuasive*, or *not stubborn and persuasive*. Also, we observe that an agent i can function as a *bridge* for a stubborn agent j. Here, agent i is persuaded by stubborn agent j, and agent i persuades at least one other agent $l \neq j$, who was not persuaded

Table 1: Stubborn agents tend win the deliberation, regardless of alignment with majority, contrary to existing work that argues MASs implement majority voting (Choi et al., 2025).

Model	Stubborn Agent	Aligned with majority	Not aligned with majority
GPT	Won	81%	18%
	Lost	0%	0%
Qwen-72B	Won	93%	7%
	Lost	0%	0%

Table 2: Star graph: hub vs. leaf agents (MMLU-Pro-Hard, neutral, 5 agents), GPT.

Agent	Persuasion	Stubbornness	Confidence
Hub (Agent 1)	0.556 ± 0.024	0.276 ± 0.044	0.492 ± 0.024
Leaves (Agents 2-5, pooled)	0.039 ± 0.004	0.294 ± 0.022	0.487 ± 0.012

directly by the stubborn agent j. Lastly, we observe th a stubborn agent can be *dominant*, persuading all other agents in the network. Figure 12 gives examples for each communication pattern.

Consensus is driven by a small group of stubborn agents. We observe that in most cases consensus is reached quickly (Figure 4), which has also been reported in (Yazici et al., 2026; Reza, 2026). Recent work suggests, that the benefits of multi-agent deliberation may largely be attributed to aggregation of opinions through majority voting (Choi et al., 2025). We find that this is not the case: In most deliberations, a single agent is stubborn (Figure 5) and influences the other agents. Specifically, we observe it is extremely rare that a stubborn agent fails to persuade others in a complete graph structure (Figure 6), which can be explained by the theoretical properties of the FJ model.

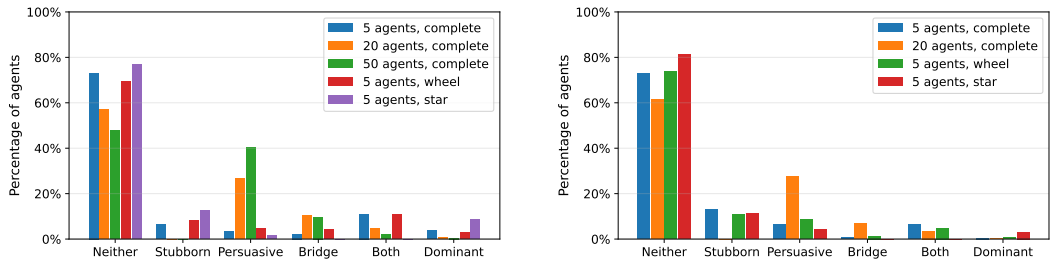
Additionally, we observe that, if there are one or more stubborn agents, the majority vote in the final around always shares a stubborn agents' initial opinion. Strictly speaking, we cannot disentangle who *wins* the deliberation when a stubborn agents' aligns with the majority. However, in Table 1 we observe that in cases where no stubborn agent is aligned with the majority, one of them always convinces the rest and *wins* the deliberation. Clearly, this contradicts the belief that MASs deliberation implements majority voting (Choi et al., 2025).

Agents become less stubborn as the system grows. In Figure 6 we observe that as the number of agents increases, fewer agents are both stubborn and persuasive, and instead a clear increase in agents that are not stubborn but persuasive and agents that act as bridges can be seen.

Stubborn leaves: the graph structure changes the stubbornness dynamics. For the *star* topology, we observe the rare case that stubborn agents disagree (Figure 9a). In general, we observe that agents at leaf nodes tend to be less persuasive, a natural consequence of their restricted communication, and at the same time we observe they are more stubborn than agents at central nodes (see Tables 2 and 4).

6.4 Interventions on the FJ parameters

Apart from intervening on the system by varying the number of agents and topology, we intervene on the prompt of the MAS and observe the effect on persuasiveness and stubbornness. Specifically, we apply the following interventions: we prompt one and two agents to be stubborn, and we prompt an agent to have a balanced, confident, or elaborate communication style. Additional, we intervene on the topology by removing one edge between two agents in the complete graph structure. The results (Table 3 in the appendix), suggest that we can affect agent stubbornness by prompting agents to be more stubborn, and measure this using the FJ model. Note, that the effect is stronger for Qwen-72b than for GPT. The communication style interventions show a less clear effect, which is still stronger for Qwen-72b. Removing the edge between agents 0 and 2, makes both of them less persuasive, while remaining similarly stubborn. Interestingly, this effect is stronger for GPT.



(a) GPT (b) Qwen-72B

Figure 6: As agent group size grows, persuasive agents become more prominent compared to stubborn agents. Agents are categorized as: neither stubborn nor persuasive, only stubborn, only persuasive, bridging for a stubborn agent, both stubborn and persuasive and dominating as a stubborn agent who persuades all others.

7 Discussion

We have found a close fit of Friedkin–Johnsen (FJ) dynamics to individual deliberations: A simple linear model with nonnegative, time-invariant weights describes much of the belief change arising from natural-language exchange. Going beyond a macro description of consensus formation, the model can distinguish agents that retain their initial beliefs from agents whose beliefs carry weight with their peers, exposing stubborn, persuasive, dominant, and bridging roles within particular debates. These roles seem to depend on group size. In larger groups, we found stubborn and persuasive agents to become less prevalent, while bridging roles become more prominent. Such behavioral distinctions make interventions easier to interpret. Changing a prompt or a communication link alter a group's answer potentially by distinct mechanisms.

Limitations. Note that a good fit does not imply that agents internally execute the FJ update rule. A fitted influence weight alone does not establish the causal effect of a particular message. Still, FJ provides a common language for comparing such effects across questions and groups in response to interventions to aid in obtaining causal insights. Our experiments focus on multiple-choice questions and agents' reported beliefs. Open-ended tasks may require a different representation of belief. FJ also cannot generate an answer absent from the initial beliefs. Such cases account for the few poor fit cases, but may be especially informative about when deliberation produces a new possibility rather than redistributing existing opinions.

Future work. The FJ framework could offer deeper insights into mechanisms behind several recent findings. For instance, work on social pressure and rhetorical persuasion (Ko et al., 2026) and self-propagating ideas (Papadopoulos et al., 2026) raises the question whether influence spreads because recipients become more receptive, sources become more persuasive, or influence is relayed through intermediaries. Studies of confidence communication and initial diversity (Zhu et al., 2026) and coordination across network topologies (Grötschla et al., 2025) invite corresponding question-level tests: Do these design choices alter agents' initial beliefs, their attachment to those beliefs, or the influence they exert on peers? Related models and studies suggest useful points of comparison. Question-level influence patterns could be examined alongside work on directional influence in consensus formation (Priyadarshini & Shrivastava, 2026) and the finding that majority voting accounts for much of the performance gain attributed to debate (Zhu et al., 2026). FJ could help distinguish the prevalence of majority-aligned outcomes from the belief changes that produced them. Furthermore, the analytic tractability of the FJ model could enable us to derive the scaling behavior of cascades in the opinion formation (Burkholz et al., 2018; Zhou et al., 2025; Burkholz & Quackenbush, 2021; Xie et al., 2026).

Conclusion. A simple opinion model provides a surprisingly detailed description of belief change in multi-agent debate. By fitting individual questions, we can see who holds firm, who persuades others, and how these roles shift as groups grow. This turns consensus from a single group-level observation into a process whose participants and paths of influence can be studied.

AI USE STATEMENT

We used LLMs to assist with coding, identifying potentially relevant related work, and for text editing. All generated content was reviewed and verified by the authors, who remain responsible for the final code and manuscript.

ETHICS STATEMENT

Understanding multi-agent deliberation is important as such systems are increasingly considered for high-stakes applications. Interpretable models of their interaction dynamics may help identify failure modes, vulnerabilities, and opportunities for manipulation.

REFERENCES

Samira Abedini, Sina Mavali, Lea Schönherr, Martin Pawelczyk, and Rebekka Burkholz. Don't trust stubborn neighbors: A security framework for agentic networks. *CoRR*, abs/2603.15809, 2026. doi: 10.48550/ARXIV.2603.15809.

Franka Bause, Jonas Niederle, Martin Pawelczyk, and Rebekka Burkholz. Multi-agent systems are mixtures of experts: Who becomes an influencer? In *Workshop on Compositional Learning: Safety, Interpretability, and Agents (CompLearn) @ ICML*, 2026. URL https://arxiv.org/abs/2605.25929.

Rebekka Burkholz and John Quackenbush. Cascade size distributions: Why they matter and how to compute them efficiently. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6840–6849, 2021.

Rebekka Burkholz, Hans J Herrmann, and Frank Schweitzer. Explicit size distributions of failure cascades redefine systemic risk on finite networks. *Scientific reports*, 8(1):6878, 2018.

Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. A survey on llm-based multi-agent system: Recent advances and new frontiers in application, 2025. URL https://arxiv.org/abs/2412.17481.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iUjGNJzrF1.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121, 1974.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: Statistical mechanics predicts collective behavior of ai agents, 2026. URL https://arxiv.org/abs/2608.16578.

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990.

Noah E. Friedkin and Eugene C. Johnsen. *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. Cambridge University Press, 2011.

Nigel Goldenfeld. *Lectures on phase transitions and the renormalization group*. CRC Press, 2018.

Florian Grötschla, Luis Müller, Jan Tönshoff, Mikhail Galkin, and Bryan Perozzi. Agentsnet: Coordination and collaborative reasoning in multi-agent llms, 2025. URL https://arxiv.org/abs/2507.08616.

Yulong He, Dutao Zhang, Sergey Kovalchuk, Pengyi Li, and Artem Sedakov. Opinion dynamics and mutual influence with llm agents through dialog simulation. *arXiv preprint arXiv:2602.12583*, 2026.

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*, 2024.

Changgeon Ko, Jisu Shin, Hoyun Song, Huije Lee, Eui Jun Hwang, and Jong C. Park. Social dynamics as critical vulnerabilities that undermine objective decision-making in llm collectives, 2026. URL https://arxiv.org/abs/2604.06091.

Jae Hee Lee, Anne Lauscher, and Stefano V Albrecht. Towards ethical multi-agent systems of large language models: A mechanistic interpretability perspective. *arXiv preprint arXiv:2512.04691*, 2025.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3IyL2XWDkG.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5EtByXq4bX.

OpenAI. On the navier–stokes millennium prize problem. https://openai.com/index/navier-stokes-solution/, September 2026. September 8, 2026.

Vassilis Papadopoulos, McNair Shah, Sam Zimmerman, and Jack Lindsey. Mind viruses: Self-propagating ideas in multi-agent llm systems, 2026. URL https://arxiv.org/abs/2608.10218.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.165.

Sergey E. Parsegov, Anton V. Proskurnikov, Roberto Tempo, and Noah E. Friedkin. Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, 2017. doi: 10.1109/TAC.2016.2613905.

Prisha Priyadarshini and Aryan Shrivastava. Directional influence and consensus formation in multi-agent systems. In *Pluralistic Alignment Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=qKQlfKER2P.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15174–15186, 2024.

Zarreen Reza. The social laboratory: A psychometric framework for multi-agent LLM evaluation. In *Women in Machine Learning Workshop @ NeurIPS 2025*, 2026. URL https://openreview.net/forum?id=XQBG7CMj2O.

Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous agents and multi-agent systems*, 33(6):673–705, 2019.

Andries Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4149–4158. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1421.

Ye Tian and Long Wang. Opinion dynamics in social networks with stubborn agents: An issue-based perspective. *Automatica*, 96:213–223, 2018.

Dat Tran and Douwe Kiela. Single-agent llms outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3913–3923, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.264/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

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. 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*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Michael Wooldridge. An introduction to multiagent systems. *Wiley*, 2009.

Jiduan Wu, Rediet Abebe, and Celestine Mendler-Dünner. Reaching a consensus in predictive loops, 2026. URL https://arxiv.org/abs/2603.12137.

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 conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

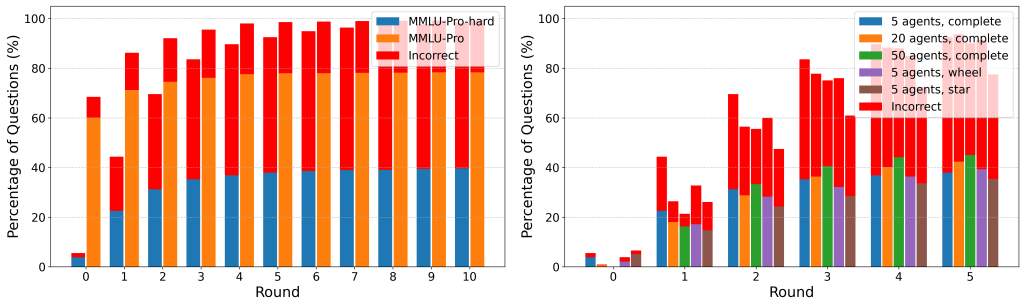
Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in llm-based multi-agent collaboration, 2026. URL https://arxiv.org/abs/2603.04474.

Iris Yazici, Mert Kayaalp, Stefan Taga, and Ali H. Sayed. Opinion consensus formation among networked large language models, 2026. URL https://arxiv.org/abs/2601.21540.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity, 2025. URL https://arxiv.org/abs/2502.08788.

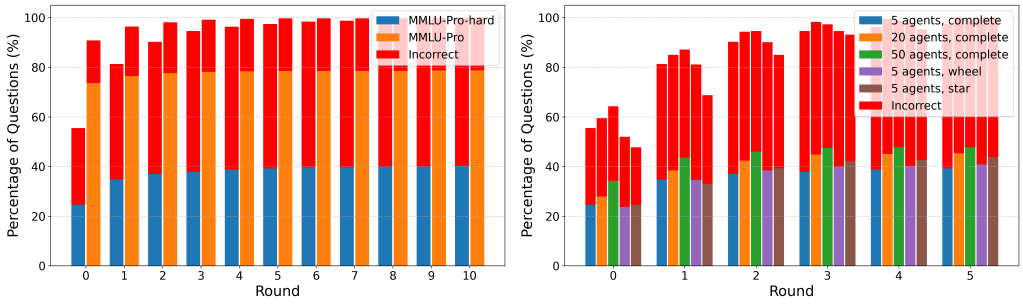
Jialong Zhou, Lichao Wang, and Xiao Yang. Guardian: Safeguarding llm multi-agent collaborations with temporal graph modeling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 7973–8001. Curran Associates, Inc., 2025. doi: 10.52202/085713-0272. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0bc795afae289ed465a65a3b4b1f4eb7-Paper-Conference.pdf.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1694. URL https://aclanthology.org/2026.findings-acl.1694/.



(a) different difficulty level of questions (b) different number of agents/topologies

Figure 7: We show in which round the system reached a stable consensus (no changes in subsequent rounds), split by whether the answer was correct or not. We see that deliberation helps for both hard and easy questions, as well as in different settings. The number of agents and topology also have an effect.



(a) different difficulty level of questions (b) different number of agents/topologies

Figure 8: We show in which round the system reached a stable top answer set (no changes in subsequent rounds), split by whether the answer was correct or not.

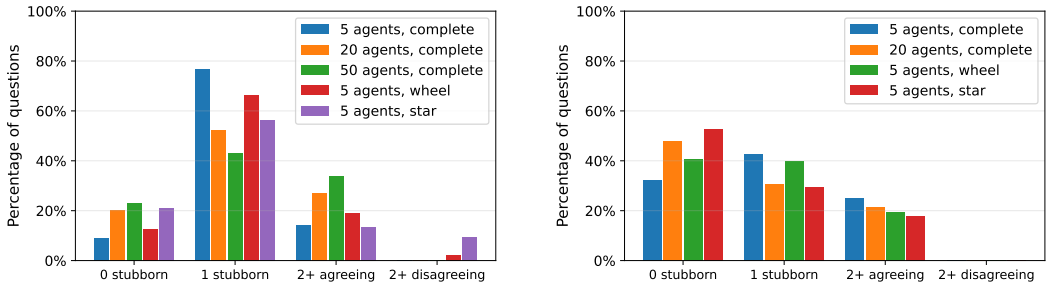
A Differences in Accuracy

We investigate how topology and number of agents affect the deliberation performance by splitting Figure 4 into whether the consensus (/majority) answer was correct or incorrect in Fig. 7(/Fig. 8). We see that more agents tend to slightly help performance, and a more disconnected topology seems to delay consensus.

B Additional Plots for the Experiments

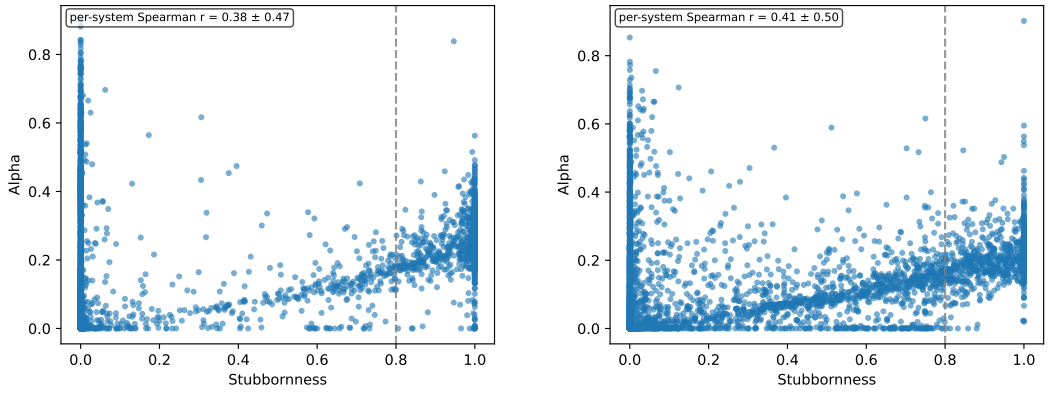
Here, we provide additional plots for our experiments. Fig. 9 shows the distribution of stubborn agents (agreeing or disagreeing with each other). We see that in most cases for GPT 1 agent is stubborn, while for Qwen-72b the agents are less stubborn. In Qwen-72, we also never observe two stubborn agents disagreeing, while we observe some for GPT.

When investigating the α parameter, we observe, that stubborn agents tend to have a non-zero α (Fig. 10), with a clear increasing trend for more stubborn agents. However, for agents that are not stubborn at all, also observe high α values, suggesting two distinct mechanisms.



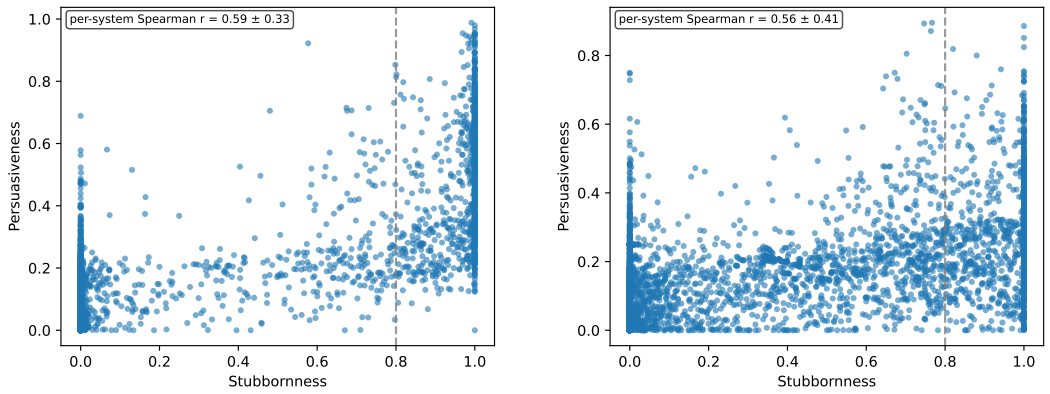
(a) GPT (b) Qwen-72B

Figure 9: Distribution of stubborn agreeing vs. disagreeing agents over questions.



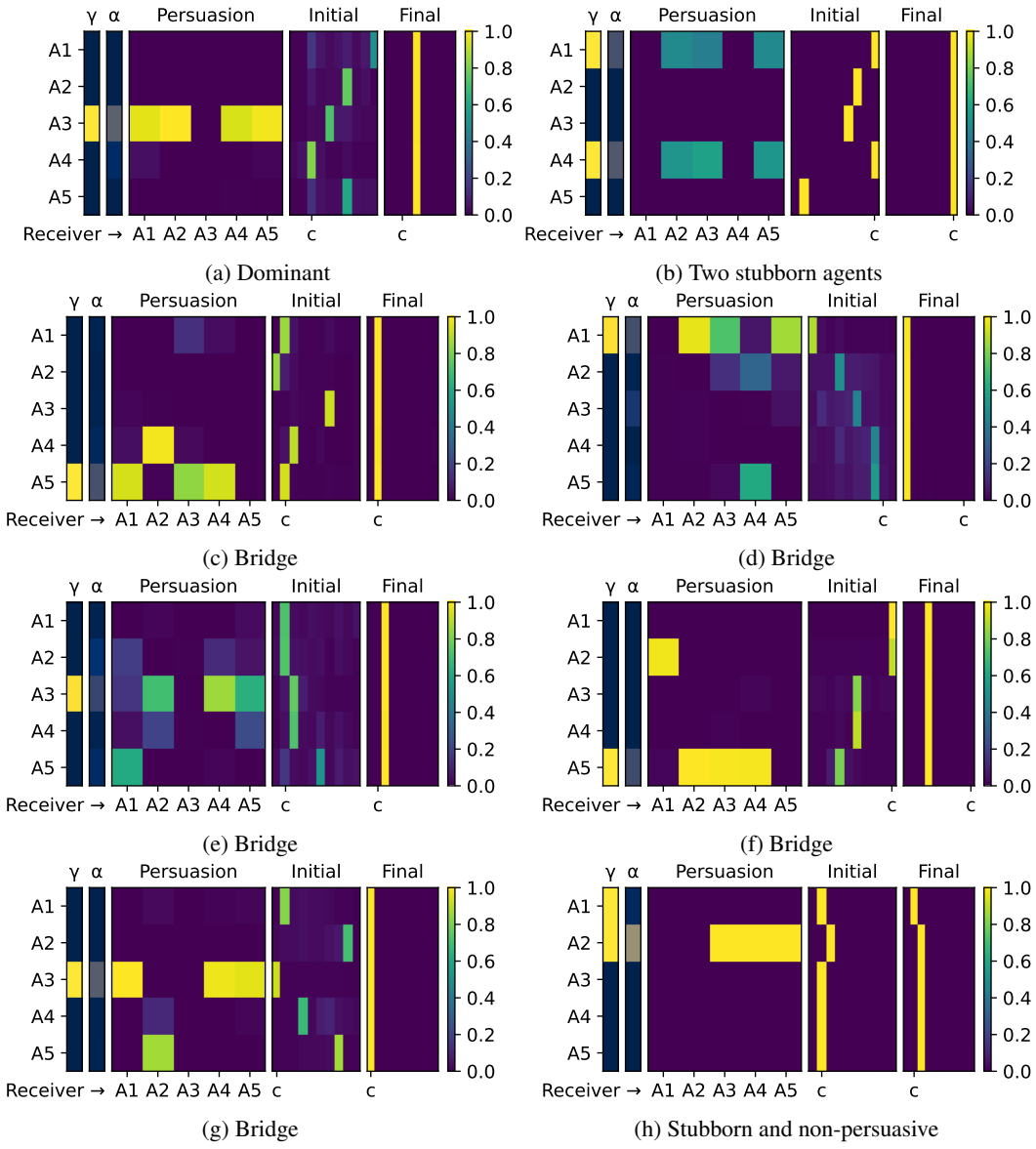
(a) GPT (b) Qwen-72B

Figure 10: Alpha (α) vs. stubbornness (γ) per agent on MMLU-Pro-Hard. We observe a that stubborn agents tend to have a non zero alpha, however, we also see a clear group of non-stubborn agents that are more inert (high alpha). Within systems we observe that stubbornness only weakly tracks inertia.



(a) GPT (b) Qwen-72B

Figure 11: Persuasiveness vs. stubbornness (γ) per agent on MMLU-Pro-Hard, for GPT (left, $r = 0.59 \pm 0.33$) and Qwen-72B (right, $r = 0.56 \pm 0.41$). The dashed line marks our stubbornness threshold ($\gamma = 0.8$). The positive per-system rank correlation shows that stubborn agents are generally also persuasive within a system, and across systems. We observe that for GPT is rare for an agent to be stubborn, without being persuasive at all.



(a) Dominant (b) Two stubborn agents

(c) Bridge (d) Bridge

(e) Bridge (f) Bridge

(g) Bridge (h) Stubborn and non-persuasive

Figure 12: Representative single-system persuasion/belief patterns: gamma/alpha per agent, the sender→receiver persuasion matrix, and initial/final beliefs (correct answer marked ○).

C COMMUNICATION PATTERNS

Figure 12 shows examples of different communication patterns we identified. They show the FJ parameters α and γ for the different agents, as well as the pairwise persuasion matrix. Additionally, the initial and final belief distributions of each agent are shown. In this way we can analyze how the parameters of the agents and their interaction influenced the deliberation result. Especially interesting, are cases like Fig 12a, where a single agent convinces most of the group to adopt an incorrect answer by being stubborn about it, and Fig. 12f, where a stubborn agent bridges the group and convinces the other agents to adopt their answer.

D INTERVENTIONS

Table 3 shows the results of the interventions on GPT and Qwen-72b. We see agent stubbornness indeed can be increased by prompting agents to be more stubborn, and we can measure this using the FJ model. This effect is stronger for Qwen-72b than for GPT. The interventions on communication styles show a less clear effect (still stronger for Qwen-72b). Removing the edge between agents 0 and 2, makes both of them less persuasive, while remaining similarly stubborn. Interestingly, this effect is stronger for GPT.

E COMPUTING RESOURCES REQUIRED FOR REPRODUCING

Experiments for Qwen2.5-14b-Instruct and Qwen2.5-72B-Instruct-GPTQ-Int8 require at least A40 GPU's. Experiments can be run in parallel on multiple GPU's or on a single GPU, depending on GPU ram available per GPU. GPT-5.4-mini, requires access to OpenAI credits but no additional specialized hardware for the researcher.

The time to run each experiment depends on the available hardware and level of parallelization. Given our hardware setup the time to complete one combination of model, dataset and prompt style experiment running on a single A40 GPU is: approximately 2.5 hours for GPT-5.4-mini and approximately 3.5 hours for Qwen-14b. Running Qwen-72b on 4 GPU's takes approximately 19 hours for a single experiment combination.

F DETAILS OF THE MODEL FIT AND COMPARISON

We fit FJ models to the belief vector trajectories over the deliberation rounds, by solving a constrained optimization problem. This approach ensures the constraints of the FJ model, such as row stochasticity of the w matrix are directly enforced. We use KL loss and solve using SQP-GS. We report the quality of our fits over extensive settings in Table 7, as well as in Table 6 for the MMLU pro hard subset. We also report goodness of fit at the round-level in Figure 3. In our extensive settings, we also use different prompts, to create some diversity and show that the FJ model fits well, over different prompt styles (roles, experts, communication styles).

As described in Section 6), a small number of trajectories are not well explained using the FJ framework. We exclude these trajectories from analyses where we interpret FJ parameters.

F.1 COMPARING FJ TO ALTERNATIVES

To demonstrate how well the FJ model captures the emergent dynamics of MAS deliberation, we evaluate its goodness of fit and compare it with several alternative models.

We fit each model separately to every observed belief trajectory and report the mean fit across trajectories. Fit quality is measured by mean squared error, which is bounded per coordinate and directly comparable across models. We prefer MSE over KL divergence because some models assign exactly zero probability to answer options with non-zero observed belief, making the KL divergence infinite and causing a few such cases to dominate the comparison.

Optimization. For the model comparison we optimize parameters with Powell's method using `Scipy`. What matters here is not the best possible fit for a single model but that every model is treated identically, since the figure compares models of very different size. Powell needs no gradients and cannot leave the feasible set, because parameters in $[0, 1]$ come from a sigmoid and each row of W from a softmax over the neighborhood, with non neighbors fixed to zero. At about two seconds per model and trajectory it also keeps the full sweep feasible.

Scoring. Starting from the observed initial beliefs $s_i = b_i(0)$, we roll out the model for all R rounds without feeding observed intermediate beliefs back into the dynamics. We use

$$\mathrm{MSE} = \frac{1}{Rnd} \sum_{t=1}^{R} \sum_{i=1}^{n} \sum_{k=1}^{d} \big(\hat{b}_{ik}(t) - b_{ik}(t)\big)^2,$$

Agent 0 is stubborn

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	-0.012 ± 0.032	+0.065 ± 0.066	+0.041 ± 0.026	**A0**	+0.025 ± 0.023	+0.613 ± 0.049	+0.305 ± 0.034
A1	-0.013 ± 0.029	-0.037 ± 0.060	+0.007 ± 0.024	A1	-0.034 ± 0.020	-0.042 ± 0.058	-0.009 ± 0.032
A2	-0.023 ± 0.030	-0.038 ± 0.059	-0.001 ± 0.026	A2	-0.032 ± 0.018	-0.041 ± 0.051	+0.001 ± 0.033
A3	+0.000 ± 0.033	-0.029 ± 0.064	+0.017 ± 0.025	A3	-0.022 ± 0.018	-0.029 ± 0.055	-0.016 ± 0.037
A4	+0.038 ± 0.033	+0.017 ± 0.066	+0.011 ± 0.026	A4	-0.048 ± 0.018	-0.067 ± 0.059	-0.024 ± 0.035

Agent 0 and Agent 1 are stubborn

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	-0.023 ± 0.030	+0.046 ± 0.063	+0.040 ± 0.027	**A0**	-0.042 ± 0.018	+0.594 ± 0.049	+0.310 ± 0.033
A1	-0.001 ± 0.032	+0.115 ± 0.064	+0.062 ± 0.027	**A1**	-0.051 ± 0.018	+0.588 ± 0.049	+0.312 ± 0.034
A2	+0.001 ± 0.031	+0.009 ± 0.063	+0.021 ± 0.027	A2	-0.065 ± 0.017	-0.041 ± 0.050	+0.009 ± 0.035
A3	-0.010 ± 0.032	-0.014 ± 0.061	-0.007 ± 0.025	A3	-0.057 ± 0.015	-0.006 ± 0.053	+0.007 ± 0.034
A4	-0.017 ± 0.029	-0.063 ± 0.060	-0.016 ± 0.025	A4	-0.062 ± 0.017	-0.004 ± 0.052	-0.016 ± 0.034

Agent 0 has a balanced communication style

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	-0.025 ± 0.030	-0.052 ± 0.060	-0.032 ± 0.025	**A0**	-0.016 ± 0.020	-0.074 ± 0.056	-0.105 ± 0.035
A1	+0.007 ± 0.031	-0.000 ± 0.062	+0.003 ± 0.026	A1	+0.004 ± 0.021	-0.012 ± 0.058	-0.013 ± 0.034
A2	-0.020 ± 0.034	-0.047 ± 0.065	+0.002 ± 0.027	A2	+0.025 ± 0.024	+0.034 ± 0.059	+0.017 ± 0.032
A3	+0.018 ± 0.034	+0.038 ± 0.066	+0.013 ± 0.026	A3	+0.012 ± 0.021	+0.024 ± 0.057	+0.005 ± 0.037
A4	+0.026 ± 0.035	+0.018 ± 0.067	-0.002 ± 0.025	A4	-0.008 ± 0.023	-0.044 ± 0.057	-0.026 ± 0.035

Agent 0 has a confident communication style

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	+0.008 ± 0.032	+0.039 ± 0.063	+0.009 ± 0.024	**A0**	+0.020 ± 0.023	+0.116 ± 0.059	+0.058 ± 0.035
A1	+0.016 ± 0.030	+0.036 ± 0.059	+0.025 ± 0.027	A1	+0.000 ± 0.021	+0.009 ± 0.060	+0.019 ± 0.035
A2	-0.011 ± 0.032	-0.045 ± 0.063	+0.006 ± 0.027	A2	-0.011 ± 0.022	-0.001 ± 0.058	+0.011 ± 0.036
A3	-0.004 ± 0.033	-0.017 ± 0.065	+0.008 ± 0.025	A3	+0.000 ± 0.022	-0.004 ± 0.059	-0.011 ± 0.035
A4	-0.006 ± 0.032	-0.010 ± 0.067	+0.007 ± 0.024	A4	-0.027 ± 0.020	-0.044 ± 0.057	-0.015 ± 0.033

Agent 0 has an elaborate communication style

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	+0.023 ± 0.034	+0.046 ± 0.067	+0.005 ± 0.025	**A0**	+0.039 ± 0.024	+0.023 ± 0.059	+0.006 ± 0.036
A1	+0.005 ± 0.032	+0.008 ± 0.063	+0.016 ± 0.026	A1	-0.004 ± 0.023	-0.050 ± 0.057	-0.007 ± 0.035
A2	-0.008 ± 0.031	-0.051 ± 0.061	+0.003 ± 0.027	A2	+0.000 ± 0.022	+0.003 ± 0.057	+0.005 ± 0.035
A3	-0.029 ± 0.031	-0.035 ± 0.062	-0.012 ± 0.025	A3	-0.009 ± 0.020	-0.050 ± 0.058	-0.001 ± 0.036
A4	+0.013 ± 0.034	-0.002 ± 0.067	-0.002 ± 0.026	A4	-0.002 ± 0.023	-0.024 ± 0.058	+0.006 ± 0.036

Agent 0 and Agent 2 do not have an edge

Agent	Persuasion	Stubbornness	Confidence	Agent	Persuasion	Stubbornness	Confidence
A0	-0.049 ± 0.030	-0.005 ± 0.067	-0.001 ± 0.026	**A0**	-0.040 ± 0.020	-0.006 ± 0.059	-0.013 ± 0.035
A1	+0.038 ± 0.032	+0.004 ± 0.063	+0.012 ± 0.027	A1	+0.022 ± 0.023	-0.022 ± 0.061	-0.007 ± 0.035
A2	-0.063 ± 0.030	-0.031 ± 0.065	+0.005 ± 0.026	**A2**	-0.032 ± 0.021	-0.011 ± 0.059	+0.005 ± 0.035
A3	+0.048 ± 0.035	+0.055 ± 0.065	-0.001 ± 0.025	A3	+0.042 ± 0.019	+0.002 ± 0.056	-0.016 ± 0.033
A4	+0.027 ± 0.031	-0.030 ± 0.064	-0.014 ± 0.027	A4	+0.020 ± 0.022	-0.023 ± 0.059	-0.031 ± 0.034

Table 3: Effect of each intervention vs. neutral baseline, MMLU-Pro-Hard on GPT (left) and Qwen-72b (right). The (directly) influenced agents are **bold**.

Agent	Persuasion	Stubbornness	Confidence
Hub (Agent 0)	0.625 ± 0.028	0.243 ± 0.036	0.615 ± 0.035
Leaves (Agents 1-4, pooled)	0.041 ± 0.003	0.280 ± 0.019	0.607 ± 0.017

Table 4: Star graph: hub vs. leaf agents (MMLU-Pro-Hard, neutral, 5 agents), Qwen.

where $\hat{b}_{ik}(t)$ is the predicted belief of agent i in option k at round t. Across our setups, $R \in \{5, 10\}$, $n \in \{5, 10, 20\}$, and $d \in \{5, 10\}$.

Counting parameters. We report identifiable free parameters. Since each row of W sums to one, agent i contributes $|\mathcal{N}_i| - 1$ free interaction parameters. The softmax parameterization used by the optimizer is redundant, but these extra coordinates are not counted.

Data. We use the hard subset of MMLU-Pro, 333 questions with 10 answer options each, in the neutral condition. Every setting is run with two language models, `gpt-5.4-mini` and `Qwen2.5-72B-Instruct`. We fit three settings.

- complete graph, 5 agents, 10 rounds, 3 seeds
- complete graph, 20 agents, 5 rounds, 1 seed
- star graph, 5 agents, 5 rounds, 1 seed

We pool the fits of the two language models. Separating them does not qualitatively change the results.

Results. Figure 2 compares the number of parameters and the mean MSE for each setting. Error bars show bootstrapped 95% confidence intervals over trajectories. Models should be compared within a setting and not across settings, because the settings differ in the number of rounds and agents, and the MSE of a rollout grows with the number of rounds it has to predict. Table 5 summarizes the remaining conditions.

Model overview. Let $b_i(t)$ denote the belief vector of agent i at round t, $s_i = b_i(0)$ its initial belief, and $\mathcal{N}_i$ its set of neighbors.

`Persistence`: every agent keeps its initial belief, $b_i(t) = s_i$ for all t. **0 parameters.**

`Lazy avg`: agents interpolate between their current belief and the uniform average of their neighbors, $b_i(t+1) = \lambda b_i(t) + (1-\lambda)|\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i} b_j(t)$. **1 parameter.**

`Potts`: we use the deterministic mean-field analogue of the Ising model, inspired by the formulation of Okawa (2026) and Goldenfeld (2018), extended to d answer options: $b_i(t+1) = \mathrm{softmax}\big[\lambda \bar{b}_i(t) + \eta s_i\big]$, where $\bar{b}_i(t) = |\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i} b_j(t)$. Here, λ controls conformity and η the strength of the initial preference, and both are shared by all agents. The inverse temperature is absorbed into the two couplings, since only their products with it are identifiable from a trajectory. Unlike the remaining models, this model is nonlinear and is not a restriction of FJ. **2 parameters.**

`Degroot`: pure averaging with a freely fitted influence matrix, $b_i(t+1) = \sum_{j \in \mathcal{N}_i} w_{ij} b_j(t)$. $\sum_i (|\mathcal{N}_i| - 1)$ **parameters.**

`FJ with shared W`: influence depends only on who is being listened to, not on the pair. Each agent j has one prestige value π_j, and agent i assigns $w_{ij} = \exp(\pi_j) / \sum_{k \in \mathcal{N}_i} \exp(\pi_k)$. Each agent keeps its own α_i, and $\gamma_i = 0$. Only relative prestige matters, so the n values give $n - 1$ free parameters on a complete graph and fewer on a sparse one. $2n - 1$ **parameters (for the complete graph).**

`FJ`: the Friedkin–Johnsen model of Eq. 1, with one α_i and γ_i per agent and a freely fitted row of W. $2n + \sum_i (|\mathcal{N}_i| - 1)$ **parameters.**

`FJ with time-varying W`: the same model, but with a separate $W(t)$ for each round while α and γ remain fixed. $2n + R \sum_i (|\mathcal{N}_i| - 1)$ **parameters.**

Table 5: Model fit per setting: number of free parameters, mean MSE over trajectories, and the percentile-bootstrap 95% CI of the mean. Best (lowest) MSE per setting in bold.

Model	#Params	MSE ↓ ($\times 10^{-3}$)	95% CI
Complete graph, 5 agents, 10 rounds			
Persistence	0	55.22	[53.78, 56.65]
Lazy avg	1	31.75	[30.79, 32.69]
Potts	2	16.36	[15.22, 17.48]
FJ with shared W	9	6.67	[6.21, 7.16]
DeGroot	15	12.33	[11.65, 13.02]
FJ	25	5.61	[5.16, 6.09]
FJ with time-varying W	160	5.11	[4.67, 5.59]
Complete graph, 20 agents, 5 rounds			
Persistence	0	54.90	[52.85, 56.94]
Lazy avg	1	29.39	[28.08, 30.75]
Potts	2	14.08	[12.44, 15.80]
FJ with shared W	39	5.47	[5.08, 5.88]
DeGroot	360	7.12	[6.60, 7.66]
FJ	400	2.64	[2.38, 2.91]
Star graph, 5 agents, 5 rounds			
Persistence	0	45.92	[43.49, 48.32]
Lazy avg	1	21.71	[20.25, 23.19]
Potts	2	16.63	[15.13, 18.19]
FJ with shared W	8	6.96	[6.28, 7.68]
DeGroot	3	23.78	[21.81, 25.80]
FJ	13	6.20	[5.53, 6.90]
FJ with time-varying W	25	5.76	[5.14, 6.42]

G MMLU-PRO-HARD

For the MMLU-Pro-Hard subset we use in our empirical analysis, we consider a subset of 333 hard questions of MMLU-Pro. Specifically, we prompt GPT-5.4-mini five times and measure entropy over the independent model outputs. We rank the questions based on entropy, shuffle tied groups, and pick a subset of questions that were answered most inconsistently by the model.

H ADDITIONAL INFORMATION ON PROMPTS USED

All agents share a base system prompt Abedini et al. (2026), that instructs them to solve the given task, exchange reasoning and their answer, as well as a distribution over their belief in the answer options. For the different modes (*communication styles*, *roles*, and *experts*) the corresponding prompt for the specific communication style etc. is appended to the base system prompt.

Base System Prompt

###Instruction###
Your task is to participate in a multi-round discussion to answer the user's question. Always keep the user's question in mind.
The process will occur in stages:
1. You will receive the user's question. Provide your independent reasoning and initial answer.
2. You will then receive the reasoning and answers provided by other agents.
3. Review the full discussion context and provide your subsequent reasoning and answer.
Your final response must always include your complete reasoning and your definitive answer.

Prompt	Model	Graph	MSE
neutral	*GPT*	Compl.	0.00881 ± 0.00163
		Wheel	0.00813
		Star	0.00893
	Qwen-72B	Compl.	0.00317 ± 0.00030
		Wheel	0.00334
		Star	0.00410
neutral (20 ag.)	*GPT*	Compl.	0.00339
	Qwen-72B	Compl.	0.00178
neutral (50 ag.)	*GPT*	Compl.	0.00339
experts	*GPT*	Compl.	0.00910 ± 0.00216
	Qwen-72B	Compl.	0.00308 ± 0.00012
Interv.: balanced	*GPT*	Compl.	0.00858
	Qwen-72B	Compl.	0.00303
Interv.: stubborn	*GPT*	Compl.	0.00884
	Qwen-72B	Compl.	0.00124
Interv.: confident	*GPT*	Compl.	0.00944
	Qwen-72B	Compl.	0.00285
Interv.: elaborate	*GPT*	Compl.	0.00800
	Qwen-72B	Compl.	0.00308
Interv.: 2 stubborn	*GPT*	Compl.	0.00806
	Qwen-72B	Compl.	0.000889
Interv.: no edge	*GPT*	Incompl.	0.0108
	Qwen-72B	Incompl.	0.00422

Table 6: Goodness of fit of FJ model on different graph topologies, models, prompt styles and number of agents, for MMLU-Hard subset. The complete graph structure with expert and neutral prompts is evaluated over 3 seeds and 95% CI is reported.

Dataset	Prompt	Model	MSE
MMLU-Pro	neutral	*GPT*	0.00248 ± 0.00094
		Qwen-14B	0.00200 ± 0.00007
		Qwen-72B	0.000978 ± 0.000152
	experts	*GPT*	0.00208 ± 0.00040
		Qwen-14B	0.00183 ± 0.00028
		Qwen-72B	0.000936 ± 0.000234
	c-styles	*GPT*	0.00221 ± 0.00050
		Qwen-14B	0.00200 ± 0.00065
		Qwen-72B	0.000880 ± 0.000345
	roles	*GPT*	0.00235 ± 0.00081
		Qwen-14B	0.00227 ± 0.00030
		Qwen-72B	0.00115 ± 0.00046
BBQ	neutral	*GPT*	0.000378 ± 0.000088
		Qwen-14B	0.00118 ± 0.00007
		Qwen-72B	0.000883 ± 0.000130
	experts	*GPT*	0.000337 ± 0.000085
		Qwen-14B	0.00151 ± 0.00048
		Qwen-72B	0.000971 ± 0.000794
	c-styles	*GPT*	0.000448 ± 0.000044
		Qwen-14B	0.00106 ± 0.00010
		Qwen-72B	0.000841 ± 0.000131
	roles	*GPT*	0.000430 ± 0.000130
		Qwen-14B	0.00170 ± 0.00032
		Qwen-72B	0.000828 ± 0.000330
CSQA	neutral	*GPT*	0.00249 ± 0.00029
		Qwen-14B	0.00569 ± 0.00209
		Qwen-72B	0.00265 ± 0.00023
	experts	*GPT*	0.00260 ± 0.00025
		Qwen-14B	0.00420 ± 0.00108
		Qwen-72B	0.00240 ± 0.00002
	c-styles	*GPT*	0.00248 ± 0.00023
		Qwen-14B	0.00449 ± 0.00043
		Qwen-72B	0.00226 ± 0.00031
	roles	*GPT*	0.00274 ± 0.00037
		Qwen-14B	0.00601 ± 0.00137
		Qwen-72B	0.00226 ± 0.00061

Table 7: Goodness of fit of FJ model on different datasets, models and prompt styles.