THE MAKEUP OF AN INFLUENCER: HOW CONFIDENCE SHAPES MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

As LLM-based multi-agent systems are increasingly used for collective decision-making, who gains influence during deliberation is a central question to predict the outcome and ensure the emergence of a trustworthy consensus. Despite complex, non-linear interactions between agents, we find that the final vote is well captured by a mixture-of-experts model that linearly aggregates agents' initial beliefs. Remarkably, its router can be approximated by a graph neural network operating solely on initial beliefs. This representation enables us to characterize how influence is allocated during deliberation. We find that influence is shaped by agents' self-assessed confidence, confidence as perceived by others, and initial alignment with other agents' views. Our findings reveal a surprisingly simple structure underlying multi-agent deliberation, with implications for both the efficient design of multi-agent systems and their vulnerability to overconfident opinions.

1 INTRODUCTION

Large language models (LLMs) have significantly advanced natural language processing across various tasks. Recently, multi-agent systems (MASs), in which multiple LLM-based agents interact, have attracted attention for their potential to improve performance in tasks that involve, for example strategic reasoning or negotiation (Du et al., 2024; Wu et al., 2024; Qian et al., 2024; Chen et al., 2025; Hong et al., 2024). By allowing agents to iteratively exchange information and refine predictions, MASs aim to utilize complementary capabilities and expertise of diverse agents, thereby improving decision-making.

Surprisingly, as we find, the outcome of this collaborative decision making is largely predictable based on initial agent beliefs only. The final outcome of the MAS deliberation can be modeled as a mixture of experts (MoE), where the router linearly combines these initial beliefs taking primarily their distribution into account in deciding which agents receive more influence than others. In this work, we thus seek to understand the mechanisms underlying this routing. Our central question is this: *What makes some agents more influential than others in a multi-agent deliberation?*

We find that (relative) confidence is a driving factor influencing the routing, with the initial alignment and social behavior also playing relevant roles. This suggests that much of the seemingly complex interactions in multi-agent deliberation can be explained by the initial state and the relative calibration of agents is key to ensure effective routing to competent experts.

Contributions: 1) We show that the outcome of multi-agent deliberation can be accurately predicted by a simple mixtures-of-experts model, based on initial agents' beliefs, and that the underlying router can be approximated by a graph neural network. 2) We theoretically and empirically analyze how influence emerges in multi-agent deliberation processes. 3) We identify confidence, relative confidence, and opinion alignment as central drivers of the implicit routing mechanism. 4) We further show that network position, communication style and role directly affects an agents' influence.

2 RELATED WORK

LLM deliberation and collective reasoning. Recent work increasingly focuses on exploring MASs, in which LLM agents interact through structured or unstructured communication in order to improve reasoning, robustness, and problem solving capabilities (Wu et al., 2024; Li et al., 2023).

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.

These approaches build on the observation that aggregating diverse opinions and reasoning can outperform single-pass inference. For example, self-consistency (Wang et al., 2023) aggregates across multiple reasoning paths, while chain-of-thought prompting (Wei et al., 2022) enables LLMs to generate intermediate reasoning steps. LLM-based MASs extend these ideas by allowing multiple, possibly diverse, agents to iteratively exchange information and update their belief.

Despite these findings, existing methods provide limited insight into the underlying dynamics of the observed collective reasoning. Therefore, a principled framework on how opinions evolve through communication and how influence shapes the outcome is important for understanding when and why LLM-based multi-agent deliberation succeeds.

Modeling LLM-deliberation with social influence models. Opinion-formation models provide a natural framework for studying how individual beliefs evolve through interaction. The Friedkin-Johnsen (FJ) model (Friedkin & Johnsen, 1990; 2011) has been extensively studied in the social sciences to understand how group opinions and consensus are formed, and how stubbornness, initial beliefs, social influence, and network structure can affect collective decision-making (Parsegov et al., 2017; Tian & Wang, 2018). 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). Even more recent work has demonstrated that FJ dynamics can effectively capture the belief formation process in MASs to study the systemic risk induced by stubborn agents (Abedini et al., 2026). Surprisingly, simple linear dynamics suffice to capture LLM deliberation.

Understanding influence in MASs. In the context of collaborative problem-solving and collective decision-making, multiple works have explored how agent behavior and the resulting collective outcomes can be explained (Wooldridge, 2009; Rosenfeld & Richardson, 2019; Lee et al., 2025). This problem is challenging in MASs, because collective behavior can emerge from the interaction of individually simple agents, making it difficult to attribute an outcome to any specific agent. Our work adds to this growing line of research by providing a framework for explaining influence in MASs through Friedkin-Johnsen belief propagation (Abedini et al., 2026) and conducting causal interventions on the deliberation.

3 Theory: Mixtures-of-experts interpretation

Our work builds on the observation that LLM deliberation dynamics can be described by the Friedkin-Johnsen model (Abedini et al., 2026; Bause et al., 2026) or other linear dynamic models of opinion formation (El et al., 2026). While they model mechanistically how agents gain influence during delibaration over time, in this work, we are primarily interested in predicting the final outcome of the deliberation, which results in a consensus most of the time. Most relevant to us is the structure of the consensus that results from these modeling approaches.

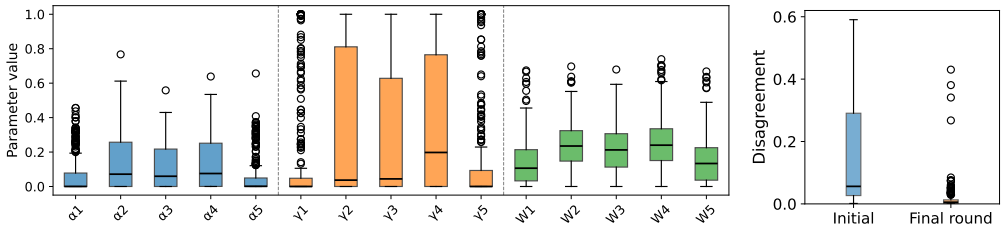
Friedkin-Johnsen (FJ) model. Let each agent $i \in V$ in an agentic network $G = (V, E)$ with $n = |V|$ agents hold a belief $b_i(t) \in \Delta^d$, where $\Delta^d \subset [0,1]^d$ is the d-dimensional simplex representing a probability distribution over potential outcomes, e.g., answers to a multiple-choice 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):

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

where $\mathcal{N}_i = \{j \mid (j,i) \in E(G)\}$ 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 state, and $W = [w_{ij}]$ is a row-stochastic influence matrix where $\sum_j w_{ij} = 1$, and $w_{ij} = 0$, if $j \notin \mathcal{N}_i$ (and $w_{ii} = 0$, as self-loops are covered by α_i). The term $(1-\gamma_i)(1-\alpha_i)$ represents the agent's susceptibility to external influence.

We can also formulate this in matrix notation. Let $B(t) \in \mathbb{R}^{n\times d}$ denote the matrix whose i-th row is $b_i(t)^\top$, and let $S \in \mathbb{R}^{n\times d}$ denote the matrix whose i-th row is $s_i^\top$. Define diagonal matrices



(a) Parameter variability. W: weighted in-degrees (avg. over senders). (b) Belief disagreement.

Figure 1: FJ induces MoE. (a) FJ parameter variability across MMLU-Pro questions for GPT-5.4 Mini. (b) Despite diverse initial opinions (left), consensus is often reached (right).

$\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$, $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, and $H = (I - \Gamma)\big(A + (I - A)W\big)$. Then the deliberation dynamics can be written compactly as $B(t+1) = \Gamma S + HB(t)$.

Mixtures of experts routing. As LLM deliberation dynamics are captured by the FJ model, the multi-agent system computes a convex combination of the initial beliefs of its agents.

Proposition 3.1 (Graph-induced convex combination (Friedkin & Johnsen, 1990; Proskurnikova & Tempo, 2017))**.** *Assume for the spectral radius $\rho(H) < 1$. Then the FJ dynamics converge to the unique equilibrium $B^\star = (I - H)^{-1}\Gamma S$. $M := (I - H)^{-1}\Gamma = [m_{ij}]$ is nonnegative and row-stochastic. Consequently, each equilibrium belief is a convex combination of the innate beliefs: $b_i^\star = \sum_{j=1}^n m_{ij} s_j$. If the agentic network belief $b_{\mathrm{out}}^\star = \sum_{i=1}^n \eta_i b_i^\star = \sum_{j=1}^n \pi_j s_j$ with $\pi^\top = \eta^\top M$ is formed by a linear aggregation of all agents' beliefs (typically $\eta_i = 1/N$), the agentic network can be interpreted as an ensemble with weights π_j that depend on the FJ parameters (Γ, A, W).*

A system that follows Friedkin-Johnsen dynamics with static parameters (Γ, A, W) would therefore implement an ensemble. Ensembles can be beneficial compared to single-agent models, as ensembles can combine the opinion of diverse agents, achieve a variance reduction and are potentially more robust to distribution shift (Dietterich, 2000; Reddy et al., 2026). A direct implication of this insight is that agentic networks tend to benefit from high agent diversity.

Our empirical findings suggest that MASs do not rely on static FJ parameters, but adapt them to the input questions x, resulting in FJ parameters $(\Gamma(x), A(x), W(x))$ that induce input-dependent routing of agent influences $\pi_i(x)$ (see Figure 1a). This means, rather than aggregating the opinions of all agents in a fixed way, the system can selectively give more or less influence to specific agents, depending on the question.

Hypothesis 3.2. **A LLM-based multi-agent system can be cast as mixture of experts (MoEs)** $b^\star(x) = \sum_{j=1}^n \pi_j(x) s_j(x)$, where the router $\pi_j(x)$ depends on the input x.

While the routing weights could still depend on the respective tasks and thus the questions at hand, our empirical analysis goes one step further by demonstrating that the routing weights seem to primarily depend on the initial agent beliefs.

Based on this hypothesis, our next goal is to characterize when such restricted routing can nevertheless work well. In essence, it still needs to be capable to identify locally competent experts. The only initially available signal regarding this competence is a proxy of agent confidence. This poses the natural question: When could a confidence based routing be enough and what information is potentially lost?

Let agent j produce an initial belief $s_j(X) \in \Delta^d$, and let $S(X) = (s_1(X), \ldots, s_n(X))$. Under Brier loss, define its conditional risk as $r_j(S) = \mathbb{E}[\|s_j(X) - e_Y\|_2^2 \mid S(X) = S]$. We focus on hard routing $g(S) \in [n]$, which assigns influence to one expert based only on the observed initial beliefs. Its expected loss is $L(g) = \mathbb{E}[r_{g(S)}(S)]$.

Theorem 3.3 (Optimal routing from initial beliefs)**.** *Among all routers that only observe the initial beliefs S, the optimal router is $g_S^\star(S) \in \arg\min_j r_j(S)$, with loss $L_S^\star = \mathbb{E}[\min_j r_j(S)]$. Moreover, the excess loss of any belief-based router g is $L(g) - L_S^\star = \mathbb{E}[\delta_g(S)]$, where $\delta_g(S) := r_{g(S)}(S) - \min_j r_j(S) \geq 0$.*

Theorem 3.3 isolates the central requirement for routing based on initial beliefs: local competence must be identifiable from the initial beliefs S. A belief-only router works well precisely when its expected routing regret $\mathbb{E}[\delta_g(S)]$ is small.

Confidence-only routing. Our empirical results suggest that an important routing signal encoded in S is confidence. Let $C_j(S) = 1 - \mathcal{H}(s_j)/\log d$ and $C(S) = (C_1(S), \ldots, C_n(S))$. Restricting routing from S to $C(S)$ introduces an information bottleneck $X \to S \to C(S) \to g(C(S))$. The following result characterizes its cost.

Theorem 3.4 (Information loss from confidence-only routing)**.** *Let $L_S^\star$ denote the optimal loss among routers observing S, and let $L_C^\star$ denote the optimal loss among routers observing only $C = C(S)$. Then*

$$L_C^\star = \mathbb{E}\left[\min_j \mathbb{E}[r_j(S) \mid C]\right] \geq \mathbb{E}[\min_j r_j(S)] = L_S^\star.$$

Equality holds if there exists a C-measurable router that selects an S-optimal expert almost surely.

Thus, reducing the full initial beliefs to confidence is harmless exactly when confidence retains the information needed to identify a locally optimal expert. Otherwise, the gap $L_C^\star - L_S^\star$ quantifies the irreducible routing cost introduced by using confidence rather than the full initial belief state.

In practice, deliberation need not implement the optimal confidence-based router from Theorem 3.4. Our empirical results instead suggest a simpler mechanism in which more confident agents receive more influence. Consider its hard-routing limit $g_C(S) \in \arg\max_j C_j(S)$.

Theorem 3.5 (Sufficiency of relative confidence)**.** *Suppose relative confidence orders agents by local competence, i.e., for almost every S,*

$$C_i(S) > C_j(S) \quad \Longrightarrow \quad r_i(S) \leq r_j(S).$$

If ties in confidence are resolved toward a minimum-risk agent, then $g_C(S) \in \arg\min_j r_j(S)$ almost surely and hence $L(g_C) = L_S^\star$.

Theorem 3.5 highlights that confidence-based routing does not require absolute probabilistic calibration. It only requires *relative* confidence to be informative about relative local competence. Thus, agents may be systematically over- or underconfident in an absolute sense while still providing an effective routing signal. More generally, the performance loss of confidence routing is exactly

$$L(g_C) - L_S^\star = \mathbb{E}\left[r_{g_C(S)}(S) - \min_j r_j(S)\right] = \mathbb{E}[\delta_C(S)].$$

Consequently, occasional confidence-ranking errors are benign when they select an expert whose competence is close to optimal, while overconfident errors are harmful when they route influence toward substantially worse experts. A simple special case makes this explicit. Suppose each input has one competent expert with Brier loss $\mathcal{B}_d(p)$, while routing to a non-competent expert yields $\mathcal{B}_d(u)$, where $\mathcal{B}_d(q) = \frac{d}{d-1}(1-q)^2$. If confidence identifies the competent expert with probability $1 - \delta$, then we expect to receive the following loss.

Proposition 3.6 (Cost of confidence-routing errors)**.** *The confidence router has expected loss*

$$L_C = \mathcal{B}_d(p) + \delta\big[\mathcal{B}_d(u) - \mathcal{B}_d(p)\big].$$

Hence its excess loss over oracle routing grows linearly in both the probability δ of misrouting and the cost $\mathcal{B}_d(u) - \mathcal{B}_d(p)$ of selecting the wrong expert.

Together, these results characterize when routing based only on initial beliefs is feasible. First, local competence must be identifiable from S. Second, confidence-only routing additionally requires that the compression $S \mapsto C(S)$ preserves the information relevant for expert selection. Finally, routing toward the most confident expert requires relative confidence to be sufficiently aligned with relative competence. The resulting failure mode is equally direct: if incorrect experts are systematically overconfident, initial-belief-based influence routes trust toward the wrong experts. In the next sections, we analyze empirically to which degree our insights transfer to LLM based MAS deliberation. As we show, agent confidence relative to the other agents is indeed a strong predictor of agent influence.

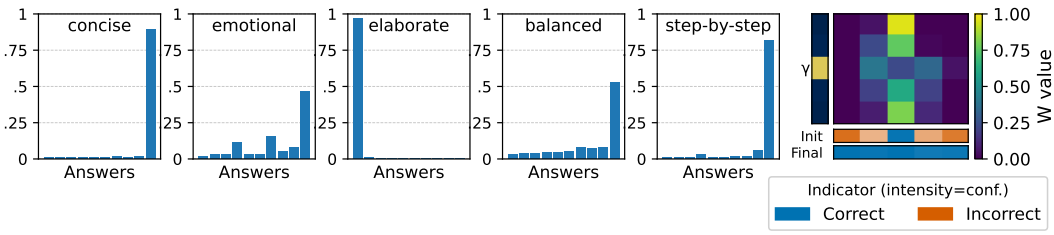


Figure 2: Example with communication styles on *GPT-5.4 Mini*: Initial beliefs of agents (left). FJ weight matrix, γ, and color coded belief in correct answer (right). The most confident agent convinces the majority to change their answer.

4 What drives influence?

In order to investigate, what drives influence in the emergent routing MASs, and whether we can explain routing from confidence of agents alone, we define several measures based on the initial beliefs of the agents and the final outcome of the deliberation.

Influence. The *influence* of agent j, normalized by the maximum agent influence, is defined as $I_j(S) = \frac{\pi_j}{\max_i \pi_i}$, where π_j is defined in Prop. 3.1.

Peer Influence. Let $B = A + (I - A)W$ and define $\tilde{B}$ as B with zero diagonal entries, i.e., $\tilde{B}_{ij} = B_{ij}\,\mathbf{1}[i \neq j]$. Here, A and W refer to FJ matrices, as defined in Section 3. The *peer influence* of agent j is $P_j = \frac{\pi_j^{\text{peer}}}{\max_k \pi_k^{\text{peer}}}$, where $\pi_j^{\text{peer}} = \sum_{i=1}^n \tilde{B}_{ij}$ and π_j^{peer} is the j-th column sum of $\tilde{B}$.

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

Disagreement. The *disagreement* of a system S is the mean distance of each agent's initial belief from the average system opinion: $Di(S) = \frac{1}{n} \sum_{j=1}^n \left\| s_j - \frac{1}{n} \sum_{i=1}^n s_i \right\|_2$.

Alignment. The *alignment* of an agent j with the MAS is given by the cosine similarity of its initial belief to the mean opinion: $Al_j(S) = \frac{s_j \cdot \frac{1}{N} \sum_{i=1}^N s_i}{\|s_j\|_2 \cdot \| \frac{1}{N} \sum_{i=1}^N s_i \|_2}$

Alignment score. The *alignment score* of an agent j is a binary variable, describing whether an agents initial answer matches the groups initial answer:

$$As_j(S) = \begin{cases} 1, \text{ if } \arg\max(s_j) = \arg\max(\frac{1}{N} \sum_{i=1}^N s_i) \\ 0, \text{ else} \end{cases}$$

Alignment count. The *alignment count* of an agent j gives the number of other agents, that share the same initial answer: $Ac_j(S) = \sum_{i=1, i \neq j}^N \delta_{\arg\max(s_j), \arg\max(s_i)}$, where δ is the Kronecker delta.

Competence. We measure agent competence by the magnitude of their belief in the correct answer. Since the correct answer is usually unknown in practice, we assume that this measure is unknown to the system. However, we can evaluate it based on known labels. If competence leads to influence in multi-agent deliberation, we intuitively expect MASs to perform well compared to single agents or ensembles. However, in practice this is not always the case (Tran & Kiela, 2026).

4.1 Experiments

Our experiments study the emergent routing mechanism from three complementary perspectives. First, we ask how much of the final deliberation state can already be predicted from the initial beliefs, using simple aggregations and a permutation-equivariant GNN. Second, we investigate which properties of the agents, their initial beliefs and especially their initial confidence explain the FJ-

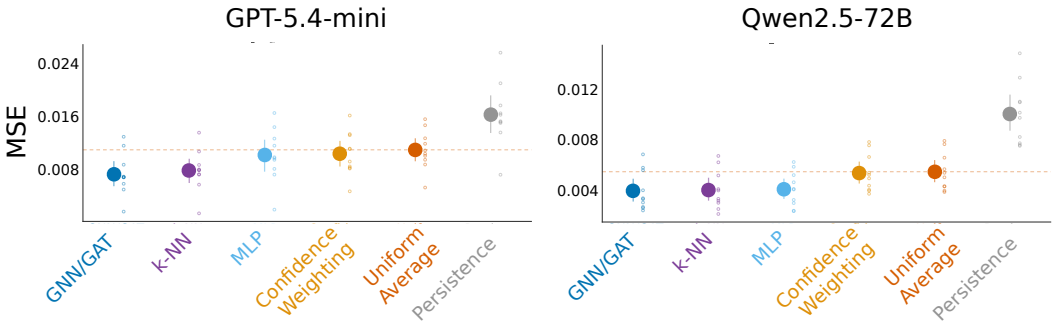


Figure 3: [Lower is better.] Prediction of each agent's final belief from the initial belief vectors on MMLU-Pro. Results for BBQ and CSQA, as well as further details, are given in Appendix E.

derived influence weights. Lastly, we intervene on the agents' behavior and communication graph structure and measure how these changes affect influence.

Fitting FJ parameters to data. Each debate involves n agents connected in by a communication graph. Collaboratively they answer multiple-choice questions, first answering independently in the initial round and in subsequent rounds, receiving the previous messages of their neighbors before responding again. We do not prescribe *how* agents should update their answers, allowing the debate dynamics to emerge naturally. Alongside each response, agent i reports a belief vector $b_i(t) \in \Delta^{d-1}$ over the d answer options. These vectors are used only for analysis and are never shown to other agents. We fit the FJ parameters independently to each observed belief trajectory.

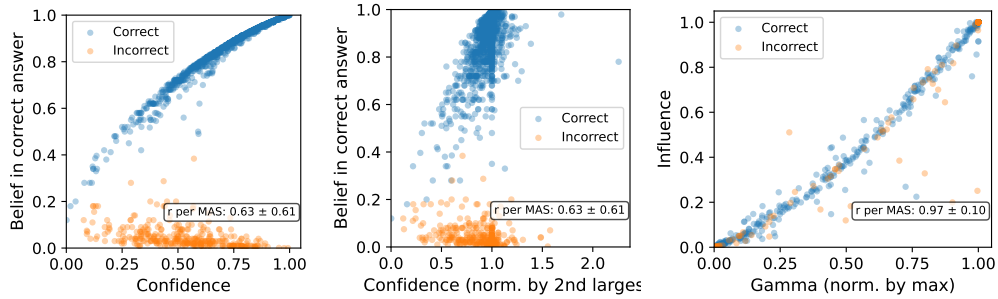
Setup. Our evaluation considers the MMLU-Pro (Wang et al., 2024), BBQ (Parrish et al., 2022) and CommonsenseQA (CSQA) (Talmor et al., 2019) dataset. We adopt the setup and subset of 100 questions of CSQA from (Abedini et al., 2026) and sample 300 questions from MMLU-Pro and BBQ each (balanced by the categories). We conduct our 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. We run our experiments over 3 seeds.

We use the same base system prompt for each agent and extend it to create diversity by prompting agents to assume different *roles* (Kong et al., 2024), like doctor, mathematician, or careless student, or answering in a specific *communication style*, e.g. concise, balanced, or emotional. Prompting agents to act as *experts* in supercategories (Ko et al., 2026) of the MMLU-Pro dataset and giving them access to specific example questions can also introduce agent diversity. The scenario *neutral* refers to no diversifying prompts being added in the system prompt. 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 F. We use $\{5, 10, 20\}$ agents, communicating in a *complete*, *star*, *circle*, *chain* or *wheel* graph structure over 5 rounds.

If not indicated otherwise, the plots show results from the MMLU-pro dataset with role prompts using *GPT-5.4 Mini*. Additional plots can be found in Appendix B.

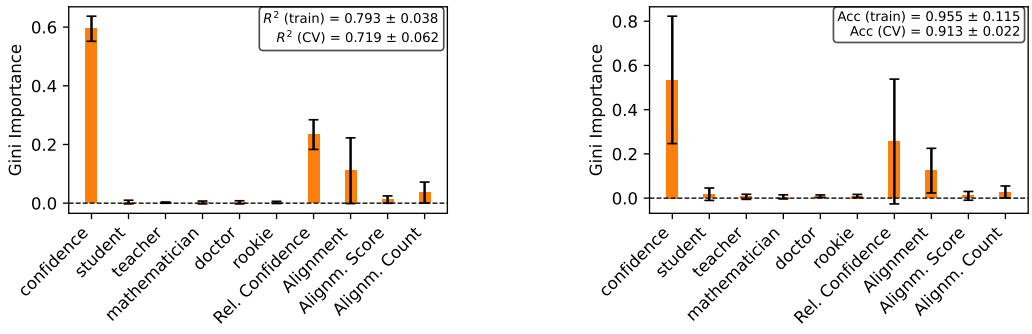
Predicting the final state from the initial beliefs. If the routing that emerges during deliberation is systematically related to the agents' initial state, then part of the final outcome should be predictable before any communication occurs. We test this using only the initial belief vectors S. We compare uniform averaging with confidence weighting and more flexible predictors, including an MLP, k-NN, and a permutation-equivariant GNN/GAT. The latter is particularly suitable because the prediction should not depend on arbitrary permutations of agents or answer options.

Figure 3 shows that the final belief state can indeed be predicted surprisingly well from the initial beliefs. Uniform averaging already performs reasonable, but learned models provide a further improvement. Confirming MAS route influence based on the input rather than uniformly averaging. So we observe, that indeed information relevant to the emerging influence structure is already present in the initial state. However, this does not reveal which agents receive influence and to what degree this is based on confidence and other interpretable measures. Therefore, we turn to analyze the FJ parameters and directly analyze the factors that determine the routing weights.



(a) Confidence. (b) relative confidence. (c) Influence vs. γ.

Figure 4: Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. When relating confidence and relative confidence to competence (a)/(b), we see that confidence agents are not necessarily more competent. We report Spearman's ranked correlation per MAS. (c) An agents influence is related to its stubbornness γ.



(a) Regressing influence (b) Classifying the most influential agent

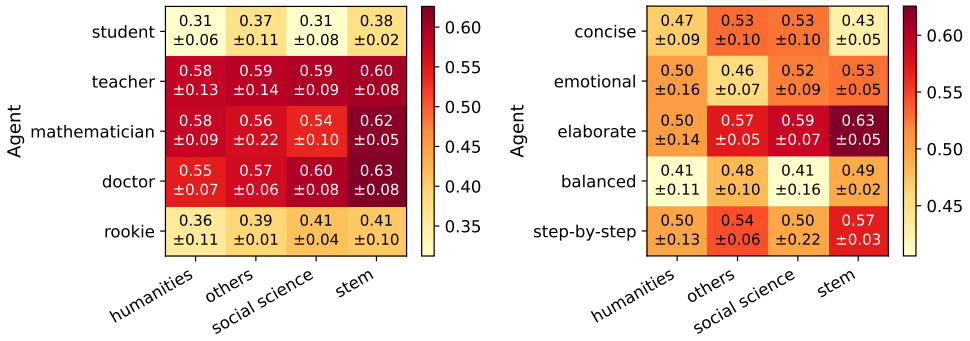
Figure 5: Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and role prompts. Feature importances for a random forest regression predicting influence (a) and a random forest classifying the most influential agent in a MAS (b), with reported CV R^2 and accuracy, respectively. The high CV scores, especially for classification (b), suggest the defined variables are highly predictive of the FJ dynamics.

Confidence and influence. To understand to what degree confidence relates to the emergent routing, we investigate the relation between confidence and relative confluence to influence. We train random forests to regress influence and classify (Fig. 5) whether an agent becomes the most influential one. The models achieve a test R^2 of 0.7 and accuracy of 0.9 for regression and classification, respectively, which suggests that influence can largely be explained by the considered variables.

We also use logistic regression to gain insights into the relative impact of different variables. Fig. 19a in the appendix shows that confidence both in absolute terms and relative to the second most confident agent is predictive of influence. Furthermore, competence, prompt style and alignment have an effect. The positive relation of confidence to influence as well as peer-influence can also be observed in Figs. 10-12 in the appendix.

Fig. 2 presents an example, where the MAS simply picks the opinion of the initially most confident agent. As described in Section 3, confidence-based influence is beneficial only when the confidence signal sufficiently tracks competence. In line with existing work, we find that confidence does not consistently align with competence (Figs. 4a and 4b) (Xu et al., 2025). Therefore, calibrating confidence or directly intervening on the emerging routing behavior has the potential to improve the performance of LLM-based MASs.

Alignment and influence. Naturally, alignment with the majority opinion plays a in role in the rate to which an agent sticks to its opinion. Agents that align with the majority stick to their opinion more



(a) Peer influence (mean ± 95% confidence interval) for different roles.

(b) Peer influence (mean ± 95% confidence interval) for different communication styles.

Figure 6: Differences in influences for different roles and communication styles suggest that communication styles and *perceived confidence* plays a role in FJ dynamics.

frequently. However, both for aligned and misaligned agents, confident agents tend to stick more to their initial belief, as can be seen in Fig. 13 in the appendix. Figure 5 suggests, that alignment also plays a role in the emergent routing.

Behavior and influence. Interestingly, the role of an agent can change its influence, suggesting that behavior and *perceived confidence* also play a role in the FJ dynamics (see Figs. 6a and. 19a).

Communication and influence. Looking at different styles of communication, we see differences in how influence is distributed. Figure 6b shows, that for example agents that answer in an elaborate way, get more influence, while answering in a balanced way, reduces the influence.

Position and influence. The position of an agent in the network also affects its influence. Specifically, in Table 1 we observe the center agent gains more influence compared to leaf agents in a star structure. The connectivity of a node in the agentic network and the connectivity of its neighbors affect its influence.

Agent Position	Influence
Hub	+0.129 ± 0.048
Leaves	-0.032 ± 0.015

Table 1: Per-agent effect of being a hub and leaf agents in a star topology, compared to a complete topology, for GPT on MMLU-Pro

Stubbornness leads to influence. There is a strong correlation between influence and stubbornness (see Fig. 4c), suggesting that the routing behavior emerging in MASs is shaped also by an agents' resistance to changing their opinion. While we see that confident agents tend to be more stubborn, this confidence is not necessarily grounded in competence (Xu et al., 2025), making it crucial to understand when to trust such social signals.

Intervening on MASs. We study interventions on the prompt and graph structure of the MASs, and measure its effect on the emergent routing mechanism, by looking at the change in confidence and influence of each agent. In Table 2 (and Table 3 in the appendix) we observe that prompting an agent to be confident, effectively changes the confidence measured from the latent belief vector. Moreover, the prompted agent becomes more influential in the network. This suggests that the perceived confidence indeed positively affects an agents influence.

Similarly, prompting an agent to be stubborn or having an elaborate communication style makes them more influential. Whereas, a balanced communication style reduces there influence. Overall, we observe that interventions on the prompt that change communication style and behavior of agents, have a clear effect on their influence.

Agent 1 has a balanced communication style

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	-0.017 ± 0.017	-0.030 ± 0.046	**A1**	-0.089 ± 0.027	-0.049 ± 0.043
A2	-0.007 ± 0.016	-0.016 ± 0.052	A2	-0.007 ± 0.027	+0.019 ± 0.044
A3	-0.002 ± 0.015	+0.030 ± 0.049	A3	+0.007 ± 0.028	+0.021 ± 0.044
A4	-0.007 ± 0.016	+0.004 ± 0.050	A4	-0.004 ± 0.025	+0.004 ± 0.044
A5	+0.007 ± 0.016	+0.012 ± 0.052	A5	-0.015 ± 0.025	+0.005 ± 0.042

Agent 1 is stubborn

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.039 ± 0.018	+0.112 ± 0.050	**A1**	+0.193 ± 0.033	+0.191 ± 0.048
A2	-0.000 ± 0.015	-0.034 ± 0.050	A2	+0.007 ± 0.026	-0.053 ± 0.035
A3	-0.000 ± 0.015	+0.004 ± 0.047	A3	-0.010 ± 0.027	-0.042 ± 0.038
A4	+0.017 ± 0.015	-0.021 ± 0.051	A4	-0.007 ± 0.026	-0.066 ± 0.036
A5	+0.006 ± 0.015	-0.061 ± 0.044	A5	-0.003 ± 0.026	-0.030 ± 0.037

Table 2: Per-agent effect of each intervention vs. the neutral baseline, MMLU-Pro, for GPT (left) and Qwen-72B (right). The agent (directly) intervened on is **bold**.

5 Discussion and Conclusion

Our results reveal a surprisingly simple structure underlying LLM multi-agent deliberation. Despite multiple rounds of natural-language interaction, the final outcome is well described as a mixture of experts over agents' initial beliefs, with the emergent routing itself largely predictable from these beliefs. Thus, much of who ultimately influences the group appears to be determined before deliberation begins.

A central signal driving this routing is relative confidence: Agents that are more confident than their peers tend to receive more influence, while less confident agents are more likely to revise their opinions. Communication behavior, alignment, and network position provide additional signals, but confidence consistently emerges as a major determinant of influence. Our interventions further support this mechanism: prompting agents toward greater confidence or stubbornness can increase their influence.

This mechanism also exposes a central vulnerability (Abedini et al., 2026). As formalized in Section 3, confidence-based routing works well when relative confidence is informative about relative competence. Yet, our experiments show that confidence and competence are imperfectly aligned. Consequently, an incorrectly but highly confident agent can receive disproportionate influence. This creates a potential attack surface: An adversarial agent that appears unusually confident or resistant to updating may gain influence without corresponding competence.

Our findings therefore suggest treating routing as an explicit object of MAS design rather than an incidental consequence of free-form communication. Calibrating or validating confidence signals and monitoring excessive concentration of influence could improve robustness. Furthermore, simple models that operate on initial beliefs like GNNs or Bayesian approaches (Papamarkou et al., 2026) could provide efficient alternatives to computationally more demanding explicit deliberation. More broadly, our results suggest viewing LLM-based multi-agent deliberation as an adaptive routing process over opinions that largely exist before communication: initial beliefs provide the routing signals, relative confidence helps determine who is trusted, and deliberation propagates this influence. This makes the mechanism interpretable, but also exposes its key limitation: Confidence is only a useful proxy for competence when the two are sufficiently aligned.

Future work could extend our MoE based modeling framework to other tasks, such as open-ended generation and tool use and could transfer our insights to different agentic use cases. Our insights could also prove beneficial to design strategies that mitigate the influence of overly confident agents to improve communication and enhance overall performance of LLM-based MAS.

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 how influence arises in LLM-based multi-agent deliberation is important as these systems are increasingly considered for high-stakes applications. Analyzing factors that amplify or constrain influence, can reveal vulnerabilities to manipulation and emergent failure.

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.

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.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15. Springer, 2000.

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. 11409–11431. 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.

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.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico,*

June 16-21, 2024, pp. 4099–4113. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.228.

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.

Theodore Papamarkou, Pierre Alquier, Matthias Bauer, Wray Buntine, Andrew Davison, Gintare Karolina Dziugaite, Maurizio Filippone, Andrew Y. K. Foong, Vincent Fortuin, Dimitris Fouskakis, Jes Frellsen, Eyke Hüllermeier, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Nikita Kotelevskii, Salem Lahlou, Yingzhen Li, Fang Liu, Clare Lyle, Thomas Möllenhoff, Konstantina Palla, Maxim Panov, Yusuf Sale, Kajetan Schweighofer, Artem Shelmanov, Siddharth Swaroop, Martin Trapp, Willem Waegeman, Andrew Gordon Wilson, and Alexey Zaytsev. Position: Agentic ai orchestration should be bayes-consistent. In *International Conference on Machine Learning (ICML)*, 2026.

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.

Anton V. Proskurnikova and Roberto Tempo. A tutorial on modeling and analysis of dynamic social networks. part i. *Annual Reviews in Control*, 43:65–79, 2017.

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.

Abbavaram Gowtham Reddy, Celia Rubio-Madrigal, Rebekka Burkholz, and Krikamol Muandet. When shift happens - confounding is to blame. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. Do language models mirror human confidence? exploring psychological insights to address overconfidence in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25655–25672, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1316. URL https://aclanthology.org/2025.findings-acl.1316/.

A Theory: Proofs of propositions

Equilibrium beliefs are convex ensembles. The Friedkin-Johnsen dynamics induce a graph-dependent ensemble over the initial agent beliefs.

Proposition A.1 (Graph-induced convex ensemble)**.** *Assume that $\rho(H) < 1$, where $\rho(\cdot)$ denotes the spectral radius. Then the FJ dynamics converge to the unique equilibrium $B^\star = (I - H)^{-1}\Gamma S$. $M := (I - H)^{-1}\Gamma$ is nonnegative and row-stochastic. Consequently, each equilibrium belief is a convex combination of the innate beliefs: $b_i^\star = \sum_{j=1}^n M_{ij} s_j$. If the agentic network belief $b_{\text{out}}^\star = \sum_{i=1}^n \eta_i b_i^\star = \sum_{j=1}^n \pi_j s_j$ with $\pi^\top = \eta^\top M$ is formed by a linear aggregation (typically $\eta_i = 1/N$) of all agents' beliefs, the agentic network is equivalent to a fixed ensemble with weights π_j that depend on the FJ parameters.*

Sketch. Unrolling the FJ dynamics gives

$$B(t) = \sum_{\tau=0}^{t-1} H^\tau \Gamma S + H^t B(0).$$

If $\rho(H) < 1$, then $H^t \to 0$ and the Neumann series converges:

$$\sum_{\tau=0}^{\infty} H^\tau = (I - H)^{-1}.$$

Nonnegativity follows because H and Γ are nonnegative. Row-stochasticity follows from conservation of mass on the simplex: if all innate beliefs are equal to a constant simplex vector s, then all beliefs remain equal to s under the dynamics, implying $M\mathbf{1} = \mathbf{1}$. □

Thus, deliberation over a fixed graph does not produce an arbitrary new predictor. It produces a graph-induced convex ensemble of the agents' initial beliefs. For a designated readout agent r, the final prediction is

$$b_r^\star = \sum_{j=1}^{n} m_{rj} s_j. \tag{2}$$

More generally, if the system output is a linear readout $\eta \in \Delta^n$ over the final agent beliefs (typically an average with $\eta_i = 1/N$), then

$$b_{\text{out}}^\star = \sum_{i=1}^{n} \eta_i b_i^\star = \sum_{j=1}^{n} \pi_j s_j, \qquad \pi^\top = \eta^\top M.$$

Hence, a fixed agentic deliberation graph is equivalent to a fixed convex ensemble over agent priors, with weights determined by the network topology and the agents' stubbornness, memory, and susceptibility parameters.

Let agent $j \in [n]$ produce an initial belief $s_j(X) \in \Delta^d$, and let $S(X) = (s_1(X), \ldots, s_n(X))$ denote the joint initial belief state. Under Brier loss, define the belief-conditional risk of agent j as $r_j(S) := \mathbb{E}[\|s_j(X) - e_Y\|_2^2 \mid S(X) = S]$. We consider hard routers g that select one expert based on the information available to the router. Their expected loss is $L(g) := \mathbb{E}[r_{g(S)}(S)]$.

Theorem A.2 (Optimal routing from initial beliefs)**.** *Among all routers that observe only the initial beliefs S, an optimal router is $g_S^\star(S) \in \arg\min_{j\in[n]} r_j(S)$, and its loss is $L_S^\star = \mathbb{E}[\min_j r_j(S)]$. Moreover, for any belief-based router g,*

$$L(g) - L_S^\star = \mathbb{E}[\delta_g(S)], \qquad \delta_g(S) := r_{g(S)}(S) - \min_j r_j(S) \geq 0.$$

Proof. Since g is measurable with respect to S, conditioning on S gives $L(g) = \mathbb{E}[r_{g(S)}(S)]$. For every realization $S = s$, this quantity is minimized pointwise by choosing any $g_S^\star(s) \in \arg\min_j r_j(s)$. Hence $L_S^\star = \mathbb{E}[\min_j r_j(S)]$. For any other router g,

$$L(g) - L_S^\star = \mathbb{E}\Big[r_{g(S)}(S) - \min_j r_j(S)\Big] = \mathbb{E}[\delta_g(S)].$$

Since $r_{g(S)}(S) \geq \min_j r_j(S)$ pointwise, $\delta_g(S) \geq 0$. □

Let $C_j(S) := 1 - \mathcal{H}(s_j)/\log d$ denote the confidence of agent j, and write $C(S) = (C_1(S), \ldots, C_n(S))$. We next restrict the router to observing only $C = C(S)$.

Theorem A.3 (Information loss from confidence-only routing)**.** *Let $L_S^\star$ denote the optimal loss among routers observing S, and $L_C^\star$ the optimal loss among routers observing only $C = C(S)$. Then*

$$L_C^\star = \mathbb{E}\Big[\min_j \mathbb{E}[r_j(S) \mid C]\Big] \geq \mathbb{E}[\min_j r_j(S)] = L_S^\star.$$

Moreover, $L_C^\star = L_S^\star$ if there exists a C-measurable router that selects an S-optimal expert almost surely.

Proof. For any router g measurable with respect to C, the law of iterated expectations gives $L(g) = \mathbb{E}[\mathbb{E}[r_{g(C)}(S) \mid C]]$. Conditional on $C = c$, the optimal choice is therefore $g_C^\star(c) \in \arg\min_j \mathbb{E}[r_j(S) \mid C = c]$, which yields $L_C^\star = \mathbb{E}[\min_j \mathbb{E}[r_j(S) \mid C]]$.

For every j, $\min_k r_k(S) \le r_j(S)$. Taking conditional expectations gives $\mathbb{E}[\min_k r_k(S) \mid C] \le \mathbb{E}[r_j(S) \mid C]$. Since this holds for every j,

$$\mathbb{E}[\min_k r_k(S) \mid C] \le \min_j \mathbb{E}[r_j(S) \mid C].$$

Taking expectations establishes $L_S^\star \le L_C^\star$.

Finally, suppose there exists a C-measurable router $g(C)$ satisfying $g(C(S)) \in \arg\min_j r_j(S)$ almost surely. Then $L(g) = L_S^\star$. Since g belongs to the class of confidence-only routers, $L_C^\star \le L(g) = L_S^\star$. Together with $L_C^\star \ge L_S^\star$, this implies equality. □

We now consider the hard confidence router $g_C(S) \in \arg\max_j C_j(S)$, which routes influence toward the most confident agent.

Theorem A.4 (Sufficiency of relative confidence)**.** *Suppose that, for almost every S, confidence orders agents by local competence such that*

$$C_i(S) > C_j(S) \quad \Longrightarrow \quad r_i(S) \le r_j(S).$$

If ties in maximum confidence are resolved toward a minimum-risk agent, then $g_C(S) \in \arg\min_j r_j(S)$ almost surely and consequently $L(g_C) = L_S^\star$.

Proof. Fix S and let $i = g_C(S)$ be an agent of maximum confidence. For every agent j with $C_i(S) > C_j(S)$, the assumed ordering implies $r_i(S) \le r_j(S)$. Among agents tied at maximum confidence, the tie-breaking assumption selects one with minimum risk. Hence $r_i(S) = \min_j r_j(S)$, so $g_C(S) \in \arg\min_j r_j(S)$ almost surely. Theorem 3.3 therefore implies $L(g_C) = L_S^\star$. □

Perfect ordering is sufficient but not necessary. For arbitrary confidence routing, define its routing regret as $\delta_C(S) := r_{g_C(S)}(S) - \min_j r_j(S)$.

Corollary A.5 (Regret of confidence routing)**.** *The excess loss of hard confidence routing relative to optimal belief-based routing is*

$$L(g_C) - L_S^\star = \mathbb{E}[\delta_C(S)] = \mathbb{E}\left[r_{g_C(S)}(S) - \min_j r_j(S)\right].$$

Thus, confidence routing approaches optimal belief-based routing whenever its expected routing regret is small.

Proof. The result follows immediately from Theorem 3.3 by setting $g = g_C$. □

Finally, consider a simple setting that makes the cost of confidence-routing errors explicit. Suppose each input has one locally competent expert whose symmetric belief assigns probability p to the correct answer, whereas routing to a non-competent expert yields a symmetric belief assigning probability $u < p$ to the correct answer. Under Brier loss, define $\mathcal{B}_d(q) := \frac{d}{d-1}(1-q)^2$.

Proposition A.6 (Cost of confidence-routing errors)**.** *Suppose confidence selects the locally competent expert with probability $1-\delta$ and a non-competent expert with probability δ. Then the expected loss of confidence routing is*

$$L_C = \mathcal{B}_d(p) + \delta\big[\mathcal{B}_d(u) - \mathcal{B}_d(p)\big].$$

Consequently, its excess loss relative to perfect routing is linear in both the routing-error probability δ and the loss gap $\mathcal{B}_d(u) - \mathcal{B}_d(p)$.

Proof. With probability $1-\delta$, confidence selects the competent expert and incurs loss $\mathcal{B}_d(p)$; with probability δ, it selects a non-competent expert and incurs loss $\mathcal{B}_d(u)$. Therefore

$$L_C = (1-\delta)\mathcal{B}_d(p) + \delta\mathcal{B}_d(u) = \mathcal{B}_d(p) + \delta\big[\mathcal{B}_d(u) - \mathcal{B}_d(p)\big].$$

Subtracting the loss under perfect routing, $\mathcal{B}_d(p)$, gives the stated excess loss. □

B Additional Plots

We present additional plots for our experiment. We see the same variability of the FJ parameters and tendency towards consensus in the other datasets and when using *Qwen2.5-14B-Instruct* and *Qwen2.5-72B-Instruct-GPTQ-Int8* as the underlying models (Figs. 7, 8 and 9).

Figs. 10, 11, and 12 show the (peer-)influence of agents in relation to their confidence. We see that confident agents tend to be more influential.

Fig. 13 shows the probability of an agent changing their answer to the majority answer when the agent is confident/not confident. We see that agents are more likely to change their answer if they are not confident. Interestingly, between the models, the *Qwen2.5-14B-Instruct* agents are less confident, but also less likely to change their answer to the majority answer when initially disagreeing with it.

Figs. 14, 15, and 16 show the impact of different diversifying prompts on the peer influence of the agents, grouped by question (super-)category on dataset MMLU-Pro. The influence of an agent seems to be driven also by perceived confidence, and roles seem to have a relatively big effect that is consistent over different models.

Fig. 17 shows coefficient for random forests regression influence and classifying the most influential agent in an MAS. We observe a good fit for regression and a near perfect fit on the classification task. This shows, the defined variables are predictive of FJ dynamics.

We investigate the relation between confidence and competence, as well as between influence and stubbornness in Fig. 4. We can see that an agents influence is highly correlated with its stubbornness. The relation between actual competence and influence is less prominent (see Fig. 18).

C Interventions

We show additional results for the interventions in Table 3.

D Computing resources required for reproducing

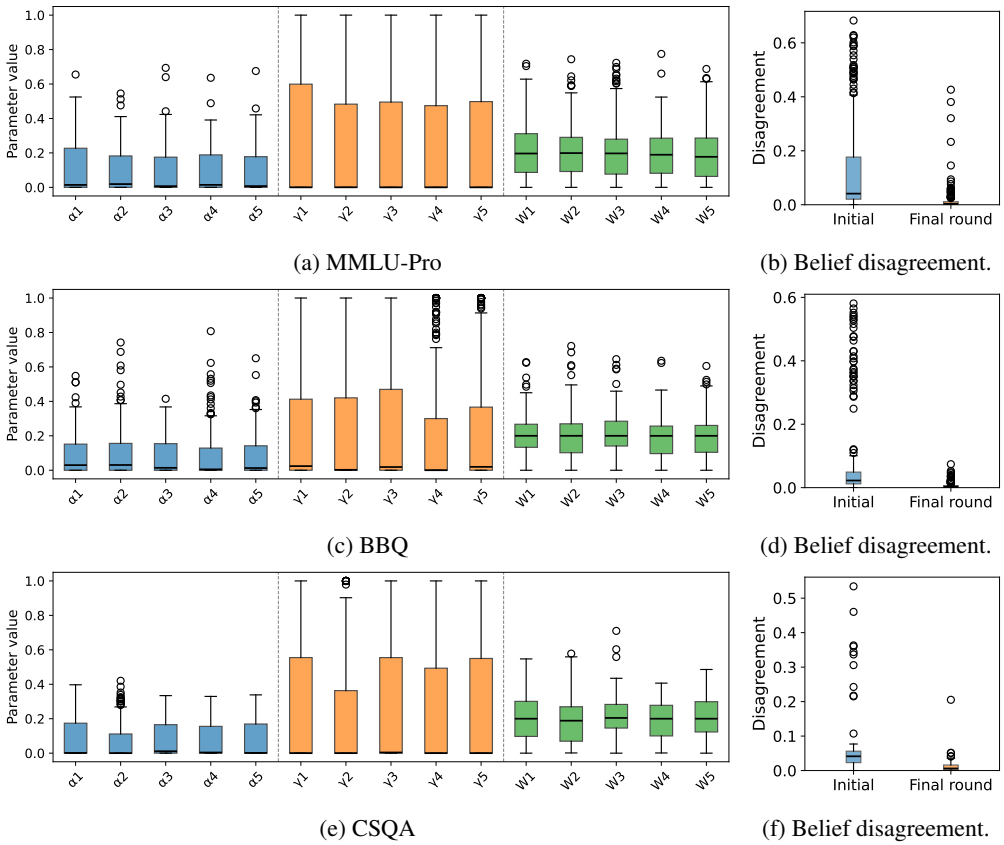
Experiments with Qwen2.5-14B-Instruct and Qwen2.5-72B-Instruct-GPTQ-Int8 were run on NVIDIA A100 (40 GB) GPUs. Experiments can be parallelized across GPUs, depending on available hardware. GPT-5.4-mini requires no specialized hardware beyond API tokens. Runtime depends on the hardware and parallelization, in our setup, a single model, dataset, prompt style configuration takes approximately 3.5 hours for Qwen2.5-14B on one A100, 19 hours for Qwen2.5-72B on four A100s, and 2.5 hours for GPT-5.4-mini.

E Details of Final Belief Prediction

An interesting question is whether we can predict the outcome of the agents' deliberation from the initial belief vectors. We predict the final belief of every single agent, that is the probability mass each agent assigns to each answer option in the final round. We score the mean squared error per agent and option. Figure 20 shows the results.

We use a nested 10-fold cross validation. Every combination of task and language model is fitted and evaluated entirely on its own, giving six independent experiments. We use the runs on the complete graph with 5 agents and 5 rounds in the neutral condition for three tasks, MMLU-Pro, BBQ, and CSQA, and for the two language models `GPT-5.4-mini` and `Qwen2.5-72B`. We restrict ourselves to the neutral condition and a fixed topology because agent identities or a changing topology would force arbitrary design choices when comparing GNN based models with models that ignore the graph.

We split by question, not by trajectory. Each question was run with three seeds which always go into the same fold. Inside each training fold we hold out a further 20 percent of the questions as a validation set. We use it only to tune basic hyperparameters of the neural network-based models

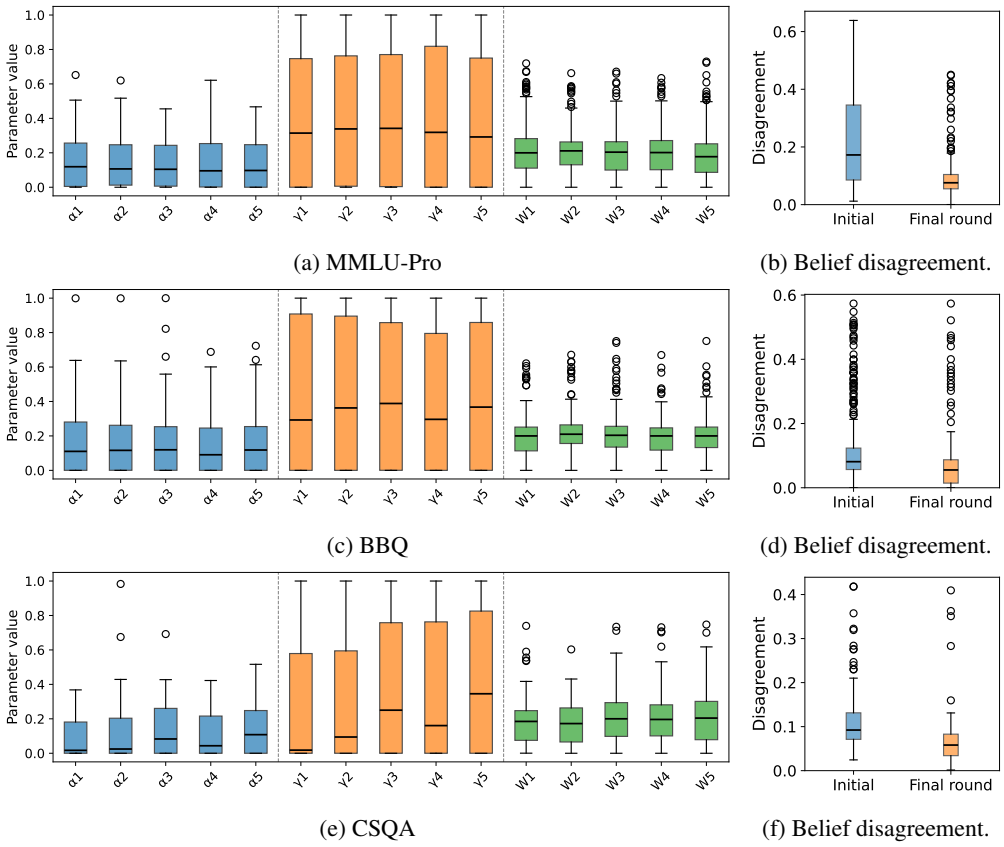


(a) MMLU-Pro (b) Belief disagreement.

(c) BBQ (d) Belief disagreement.

(e) CSQA (f) Belief disagreement.

Figure 7: Parameter variability over all samples no special prompts for *GPT-5.4 Mini*, receiving weights (W1-5) are averaged over senders (left). Tendency towards consensus for the corresponding dataset (right).

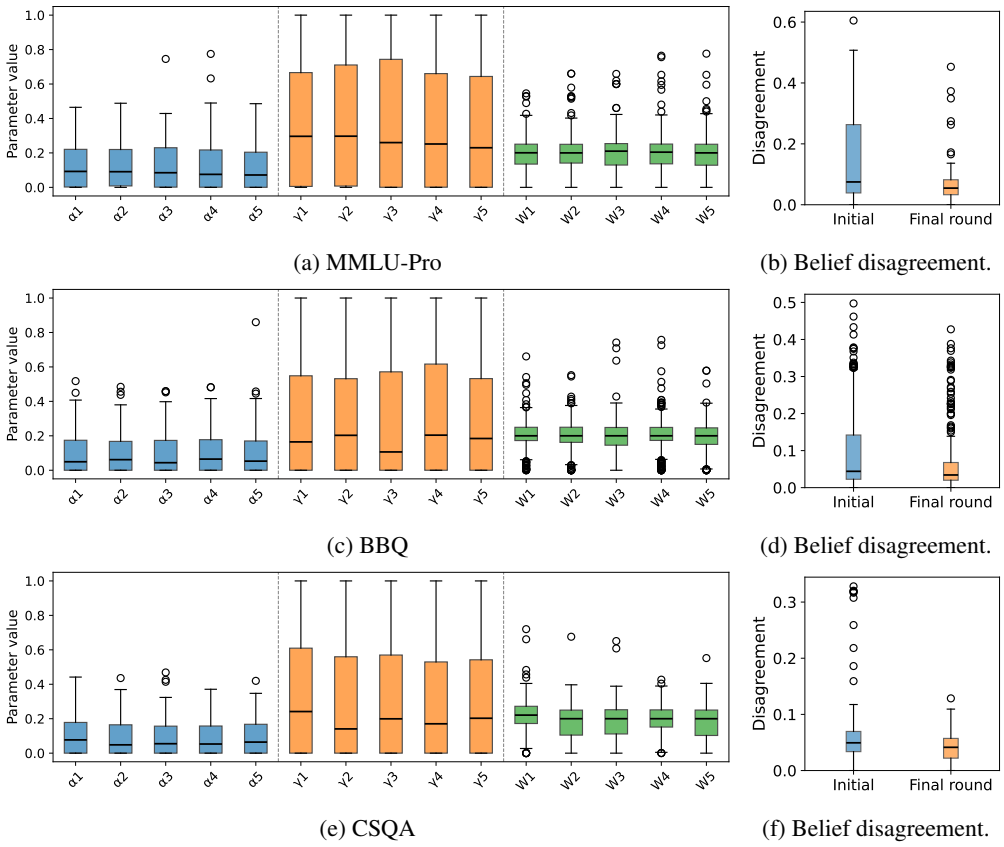


(a) MMLU-Pro (b) Belief disagreement.

(c) BBQ (d) Belief disagreement.

(e) CSQA (f) Belief disagreement.

Figure 8: Parameter variability over all samples no special prompts for *Qwen2.5-14B-Instruct*, receiving weights (W1-5) are averaged over senders (left). Tendency towards consensus for the corresponding dataset (right).

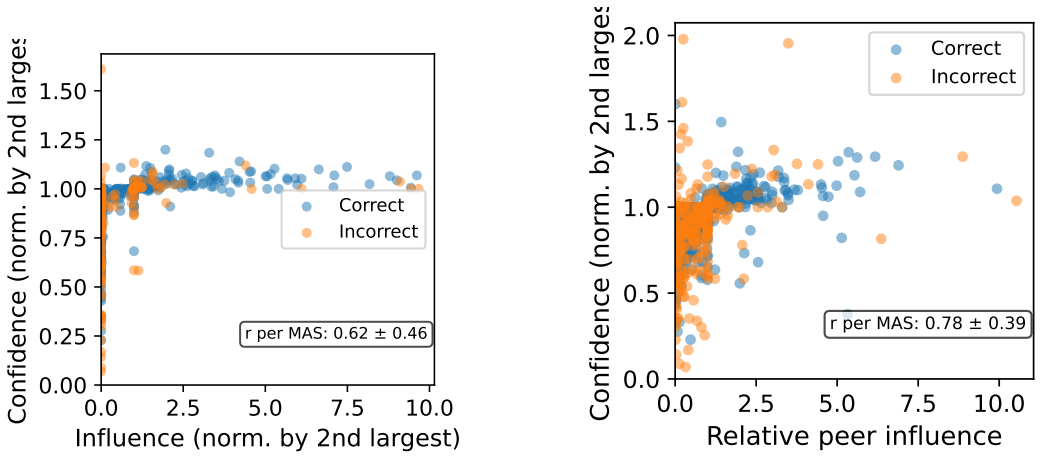


(a) MMLU-Pro (b) Belief disagreement.

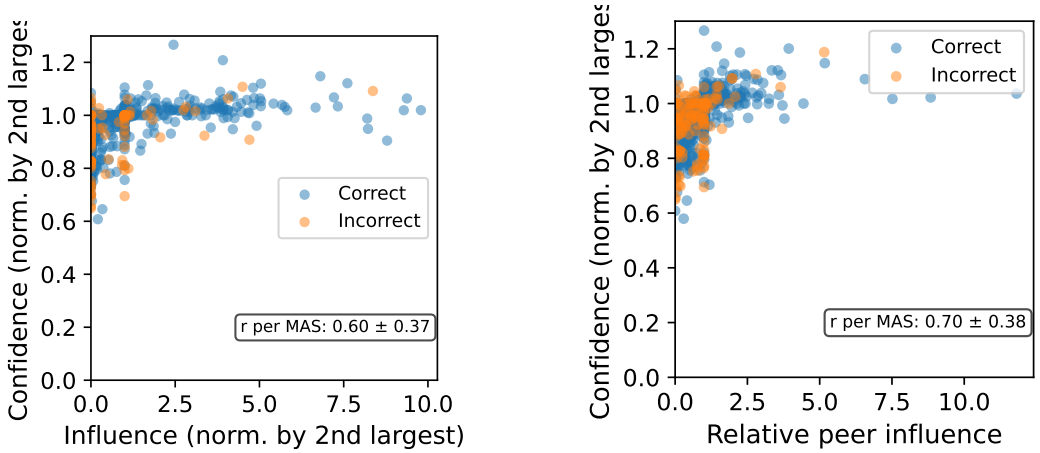
(c) BBQ (d) Belief disagreement.

(e) CSQA (f) Belief disagreement.

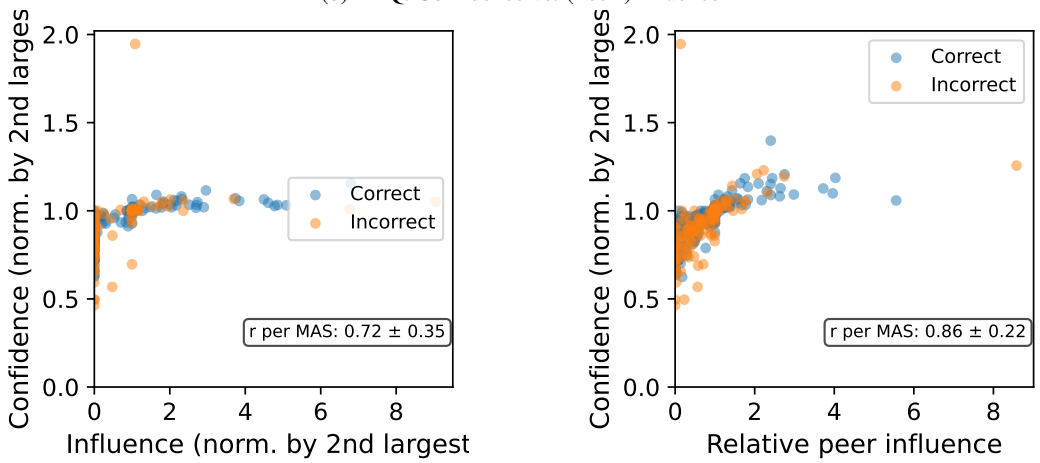
Figure 9: Parameter variability over all samples no special prompts for *Qwen2.5-72B-Instruct-GPTQ-Int8*, receiving weights (W1-5) are averaged over senders (left). Tendency towards consensus for the corresponding dataset (right).



(a) MMLU-Pro: Confidence vs. (Peer-)Influence

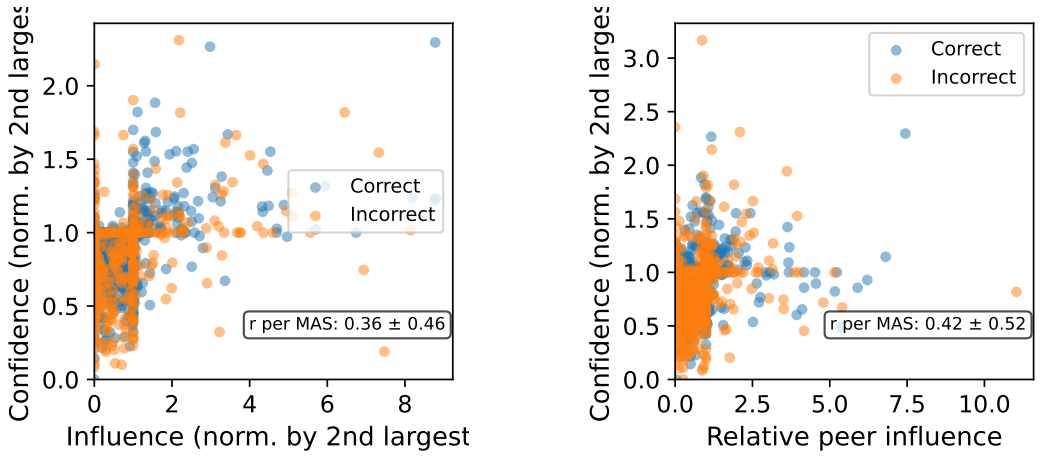


(b) BBQ: Confidence vs. (Peer-)Influence

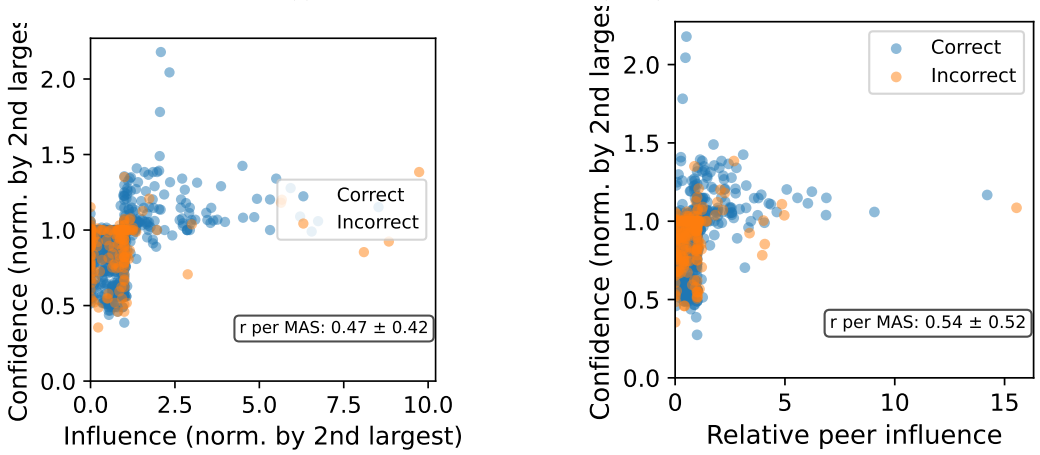


(c) CSQA: Confidence vs. (Peer-)Influence

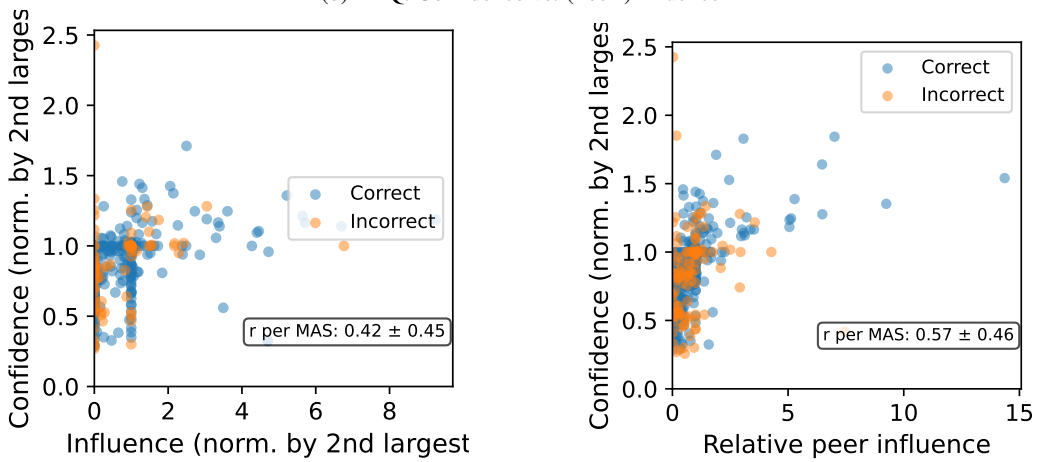
Figure 10: Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *GPT-5.4 Mini*.



(a) MMLU-Pro: Confidence vs. (Peer-)Influence

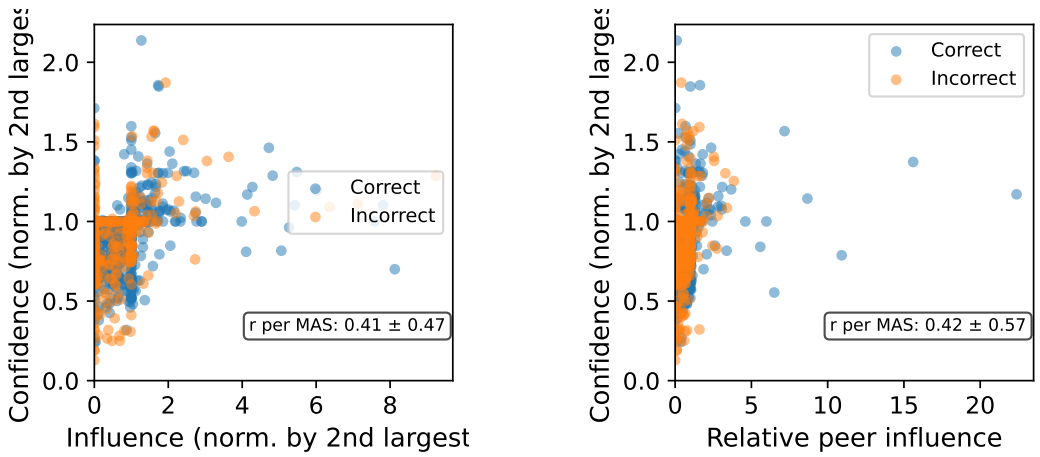


(b) BBQ: Confidence vs. (Peer-)Influence

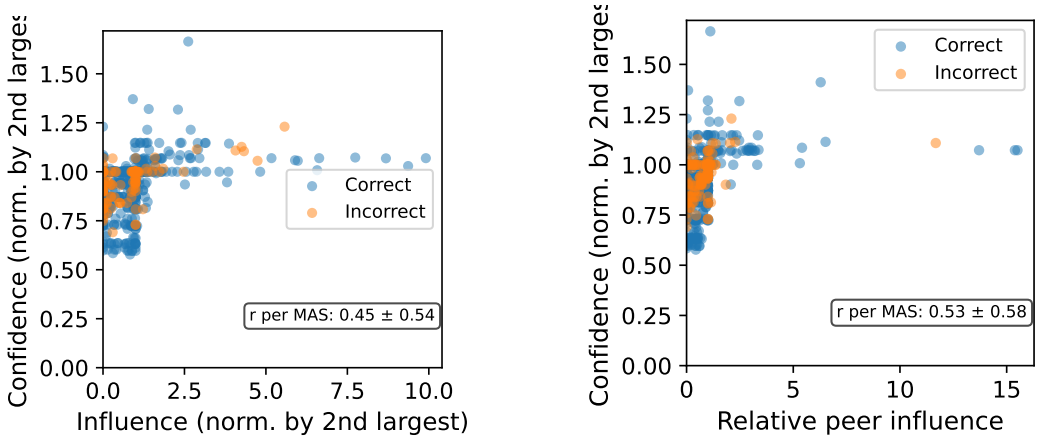


(c) CSQA: Confidence vs. (Peer-)Influence

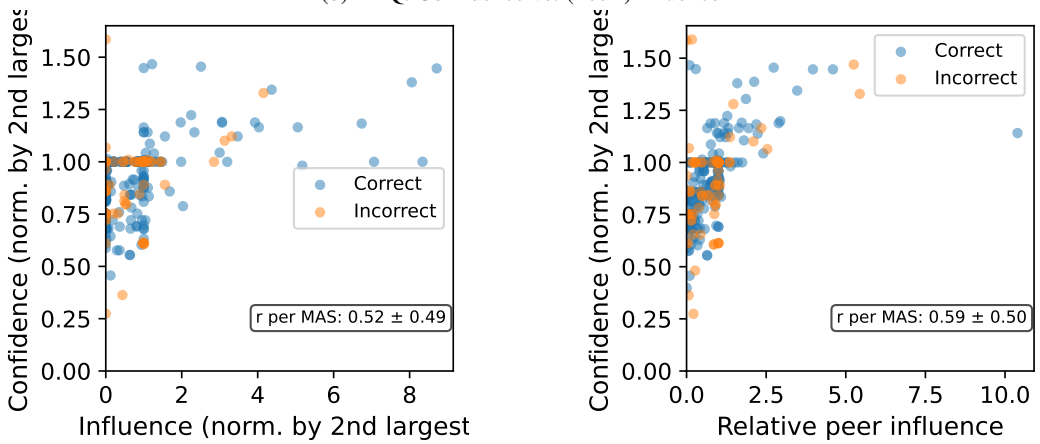
Figure 11: Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *Qwen2.5-14B-Instruct*.



(a) MMLU-Pro: Confidence vs. (Peer-)Influence



(b) BBQ: Confidence vs. (Peer-)Influence



(c) CSQA: Confidence vs. (Peer-)Influence

Figure 12: Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *Qwen2.5-72B-Instruct-GPTQ-Int8*.

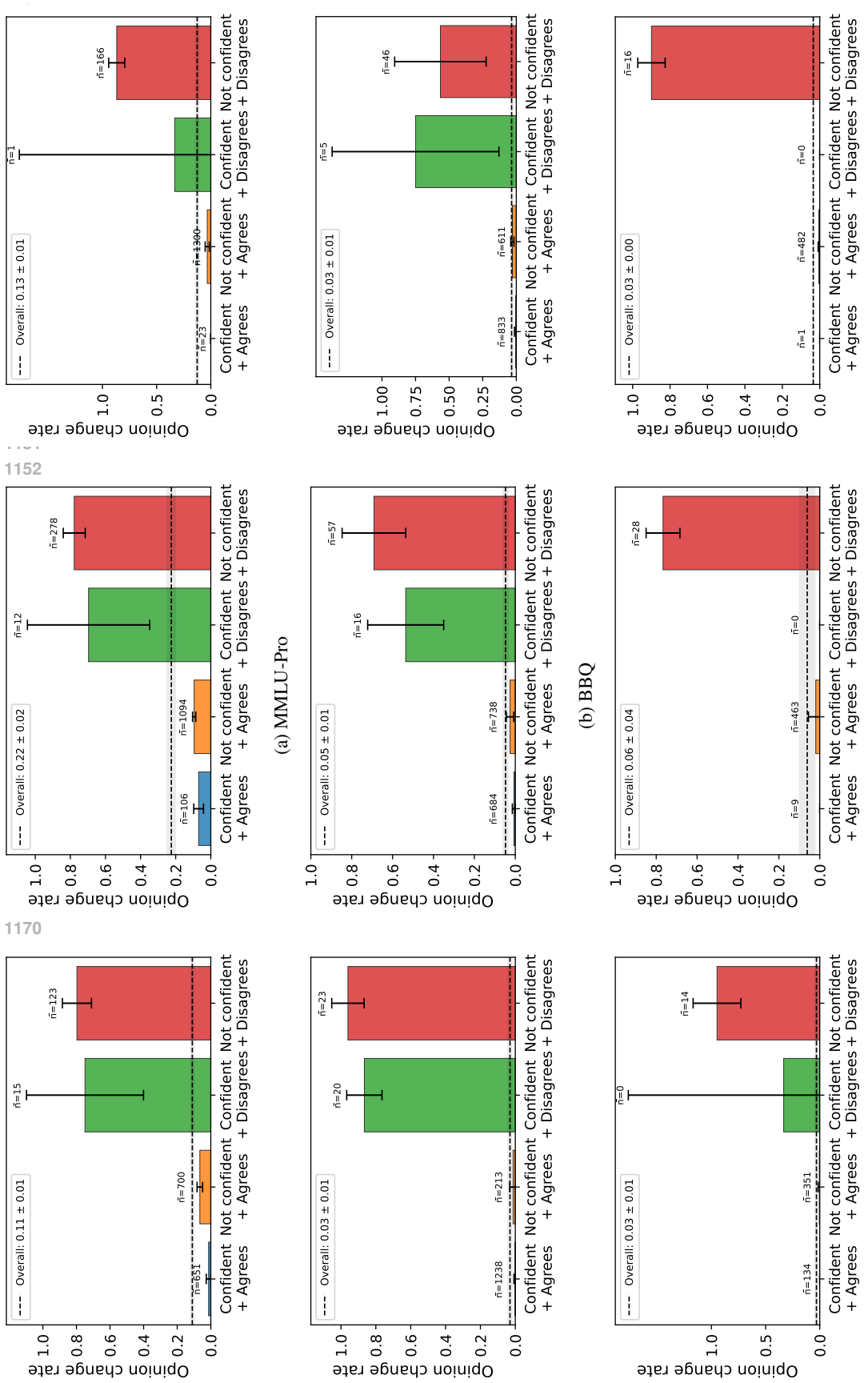
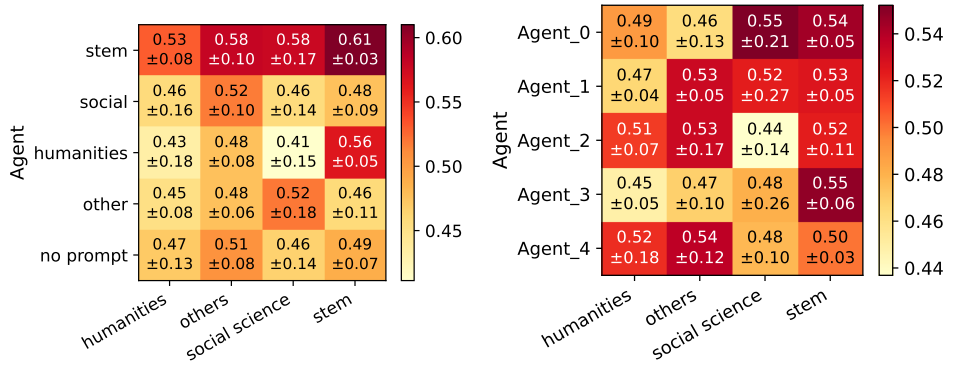


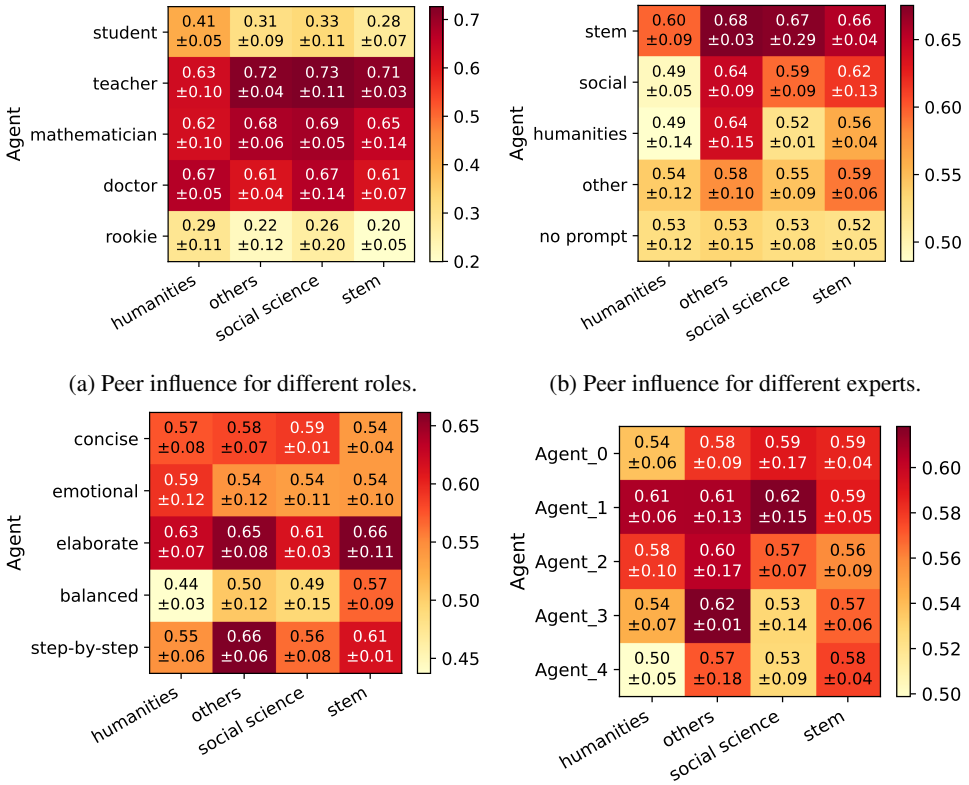
Figure 13: Probability of changing answer to the majority answer for cases where the agent is confident, $p(\text{change}|\text{confident} >= 0.75, \text{agree_with_mean} = True)$, $\bar{n}$ indicates average sample size over seeds rounded to nearest integer. With no special prompts for *GPT-5.4 Mini* (left), *Qwen2.5-14B-Instruct* (middle), and *Qwen2.5-72B-Instruct-GPTQ-Int8* (right).



(a) Peer influence for different experts.

(b) Peer influence without special prompts.

Figure 14: Results for *GPT-5.4 Mini* on the MMLU-Pro dataset.



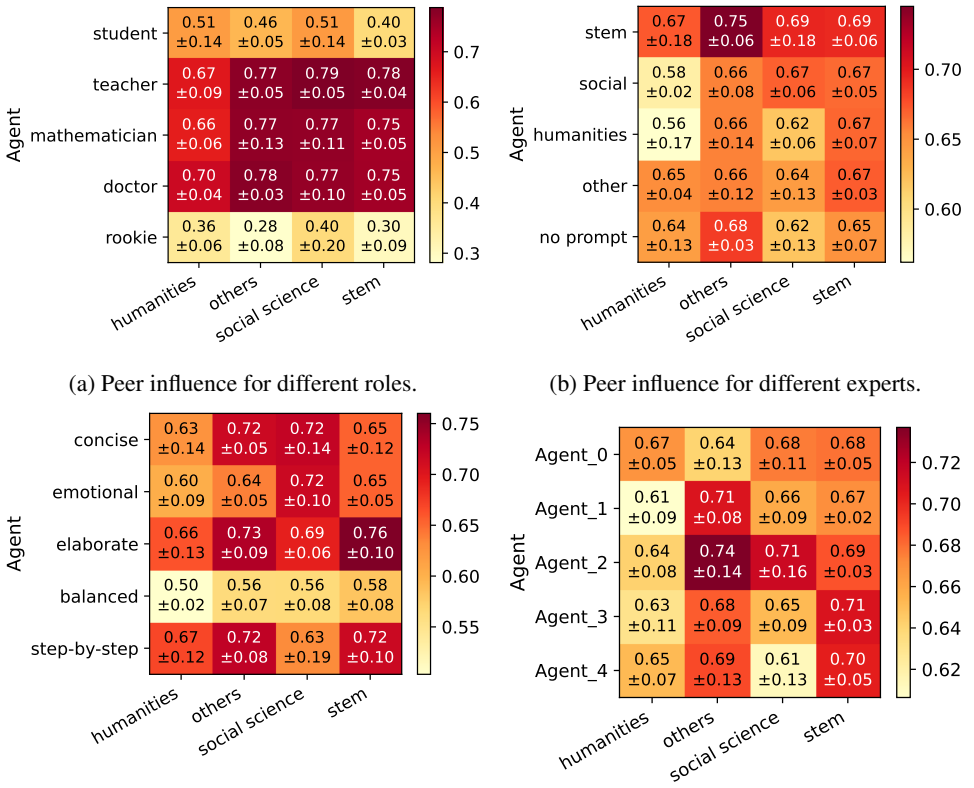
(a) Peer influence for different roles.

(b) Peer influence for different experts.

(c) Peer influence for different communication styles.

(d) Peer influence without special prompts.

Figure 15: Results for *Qwen2.5-14B-Instruct* on the MMLU-Pro dataset. The same tendencies as for *GPT-5.4 Mini* can be seen.



(a) Peer influence for different roles.

(b) Peer influence for different experts.

(c) Peer influence for different communication styles.

(d) Peer influence without special prompts.

Figure 16: Results for *Qwen2.5-72B-Instruct-GPTQ-Int8* on the MMLU-Pro dataset. The same tendencies as for *GPT-5.4 Mini* can be seen.

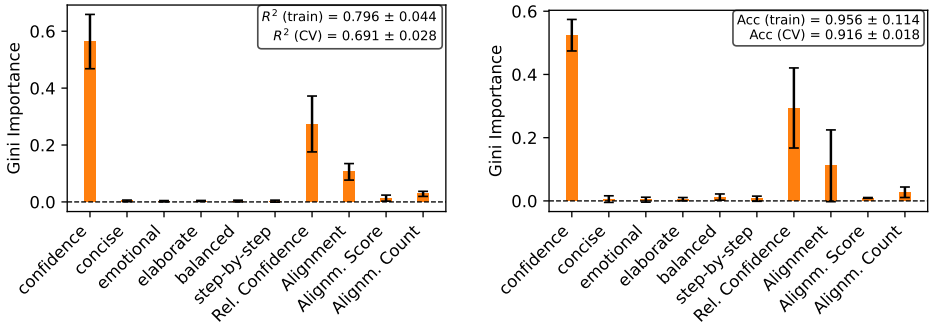


Figure 17: Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. Coefficients for a random forest regression predicting influence (a) and a random forest classifying the most influential agent in a MAS (b), with reported CV R^2 and Acc respectively. The high CV score, especially on the classification task (b) suggests the defined variables are highly predictive of the FJ dynamics.

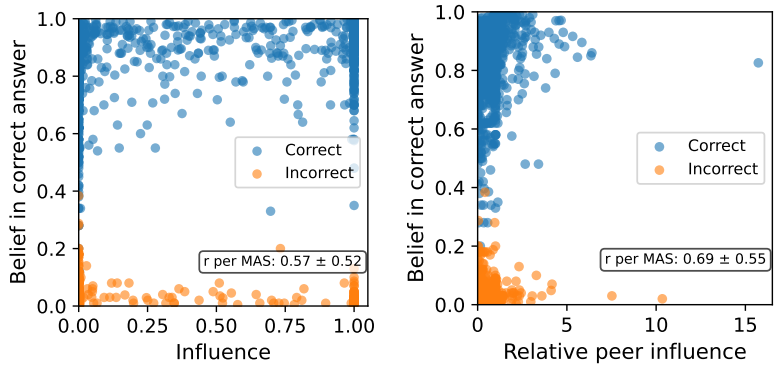
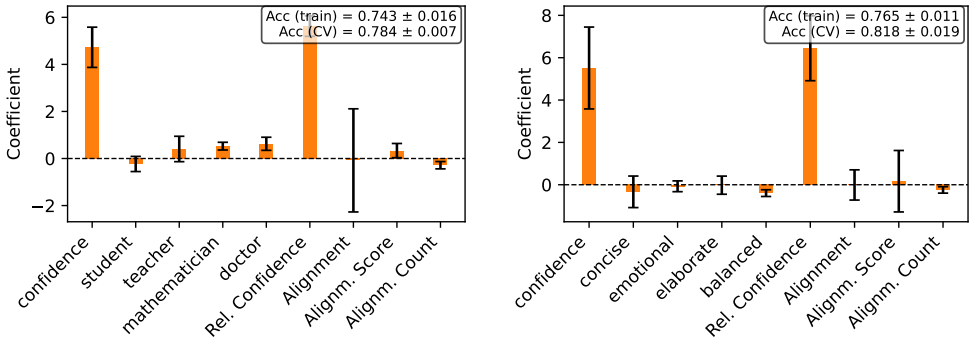


Figure 18: Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. More competent agents gain more influence (a) and relative influence (b), we again report Spearman's ranked correlation per MAS.



(a) Logistic regression coefficients classifying the most influential agent.

(b) Logistic regression coefficients classifying the most influential agent.

Figure 19: Regression coefficients suggest that confidence and relative confidence are important predictors of influence

Agent 1 has a balanced communication style

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	-0.017 ± 0.017	-0.030 ± 0.046	A1	-0.089 ± 0.027	-0.049 ± 0.043
A2	-0.007 ± 0.016	-0.016 ± 0.052	A2	-0.007 ± 0.027	+0.019 ± 0.044
A3	-0.002 ± 0.015	+0.030 ± 0.049	A3	+0.007 ± 0.028	+0.021 ± 0.044
A4	-0.007 ± 0.016	+0.004 ± 0.050	A4	-0.004 ± 0.025	+0.004 ± 0.044
A5	+0.007 ± 0.016	+0.012 ± 0.052	A5	-0.015 ± 0.025	+0.005 ± 0.042

Agent 1 is stubborn

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.039 ± 0.018	+0.112 ± 0.050	A1	+0.193 ± 0.033	+0.191 ± 0.048
A2	-0.000 ± 0.015	-0.034 ± 0.050	A2	+0.007 ± 0.026	-0.053 ± 0.035
A3	-0.000 ± 0.015	+0.004 ± 0.047	A3	-0.010 ± 0.027	-0.042 ± 0.038
A4	+0.017 ± 0.015	-0.021 ± 0.051	A4	-0.007 ± 0.026	-0.066 ± 0.036
A5	+0.006 ± 0.015	-0.061 ± 0.044	A5	-0.003 ± 0.026	-0.030 ± 0.037

Agent 1 has a confident communication style

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.027 ± 0.018	+0.048 ± 0.051	A1	+0.013 ± 0.029	+0.051 ± 0.045
A2	-0.009 ± 0.016	-0.037 ± 0.046	A2	-0.020 ± 0.030	-0.010 ± 0.040
A3	-0.008 ± 0.016	-0.005 ± 0.047	A3	-0.011 ± 0.028	-0.014 ± 0.041
A4	+0.004 ± 0.016	-0.023 ± 0.048	A4	-0.007 ± 0.026	-0.014 ± 0.041
A5	+0.021 ± 0.016	+0.018 ± 0.052	A5	+0.006 ± 0.026	-0.013 ± 0.041

Agent 1 has an elaborate communication style

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.014 ± 0.016	+0.032 ± 0.051	A1	-0.026 ± 0.028	-0.018 ± 0.042
A2	-0.008 ± 0.016	-0.040 ± 0.050	A2	-0.005 ± 0.028	+0.005 ± 0.038
A3	+0.001 ± 0.015	-0.008 ± 0.045	A3	-0.014 ± 0.028	-0.014 ± 0.043
A4	+0.001 ± 0.017	-0.010 ± 0.050	A4	+0.007 ± 0.027	+0.005 ± 0.045
A5	+0.013 ± 0.016	+0.026 ± 0.053	A5	+0.019 ± 0.027	+0.021 ± 0.042

Agent 1 and Agent 2 are stubborn

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.045 ± 0.019	+0.092 ± 0.051	A1	+0.199 ± 0.033	+0.099 ± 0.039
A2	+0.039 ± 0.016	+0.045 ± 0.049	A2	+0.183 ± 0.033	+0.105 ± 0.034
A3	-0.005 ± 0.017	-0.066 ± 0.043	A3	-0.029 ± 0.026	-0.065 ± 0.035
A4	+0.012 ± 0.017	-0.005 ± 0.051	A4	-0.010 ± 0.029	-0.079 ± 0.035
A5	+0.007 ± 0.016	-0.066 ± 0.048	A5	+0.004 ± 0.025	-0.059 ± 0.032

Agent 1 and Agent 3 do not have an edge

Agent	Confidence	Influence	Agent	Confidence	Influence
A1	+0.009 ± 0.016	-0.001 ± 0.048	A1	-0.032 ± 0.027	-0.036 ± 0.050
A2	-0.007 ± 0.016	-0.002 ± 0.052	A2	-0.023 ± 0.028	+0.019 ± 0.044
A3	-0.001 ± 0.016	-0.016 ± 0.047	A3	-0.023 ± 0.028	-0.032 ± 0.041
A4	+0.001 ± 0.016	+0.009 ± 0.053	A4	-0.011 ± 0.024	+0.006 ± 0.043
A5	+0.008 ± 0.015	+0.011 ± 0.050	A5	+0.007 ± 0.027	+0.043 ± 0.044

Table 3: Per-agent effect of each intervention vs. the neutral baseline, MMLU-Pro, for GPT (left) and Qwen-72B (right).

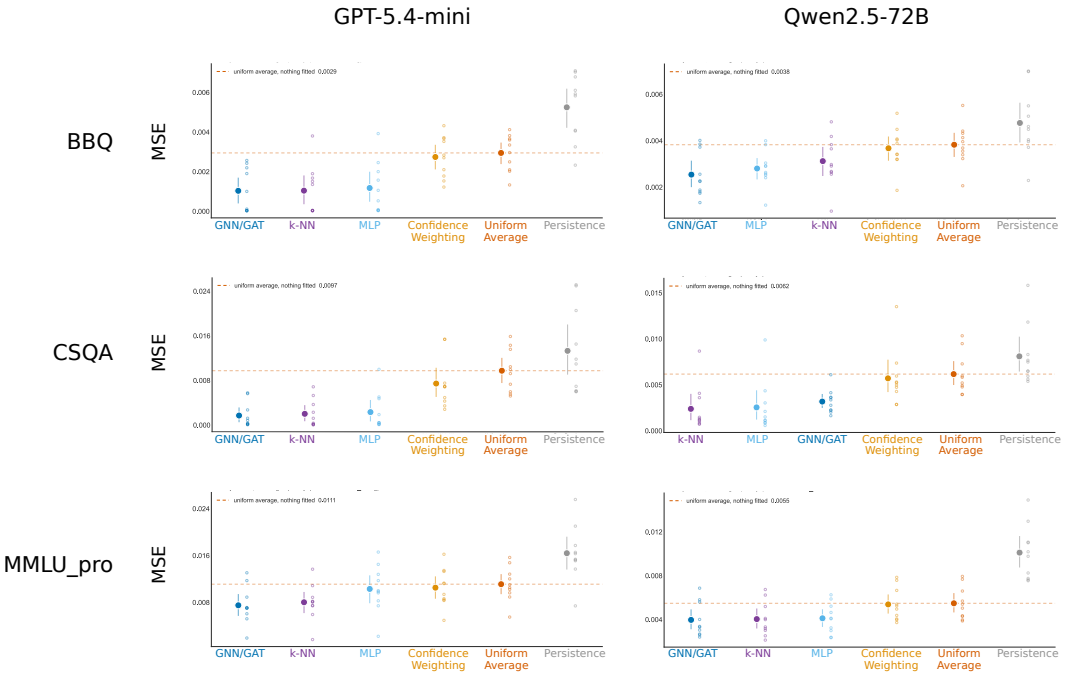


Figure 20: [Lower is better.] MSE of predicting every agent's final belief from the initial belief vectors. Large dots are the mean over the 10 outer folds, small circles the individual folds, and vertical lines bootstrapped 95% confidence intervals of the mean.

(number of layers and the dropout rate). The chosen configuration is then retrained from scratch on the whole training fold, so every method is finally fitted on the same data.

We test several models. The baseline is a `Uniform Average` over the initial belief vectors. As a sanity check we also report `Persistence`, which hands each agent back its own initial belief and therefore predicts that the deliberation changed nobody. `Confidence Weighting` gives agent i the weight $w_i = \exp(x\, c_i) / \sum_j \exp(x\, c_j)$, where c_i is its confidence as defined in Section 4 and the scalar x is fitted on the training data by grid search. At $x = 0$ all weights are equal and the model reduces to the uniform average. A positive x gives more weight to confident agents and a negative x less.

`MLP` takes the flattened initial beliefs of all agents and predicts all $5 \times d$ entries directly. This model has no built in symmetry, so we shuffle the agent slots during training (input and target together), to prevent it from learning the order. `k-NN` averages the final belief arrays of the k closest training openings, with k fitted on the training data.

`GNN/GAT` is a graph attention network on the initial beliefs. Unlike the other models it is equivariant to permutations of the agents and to permutations of the answer options. To achieve this we encode the initial beliefs in a special input graph with one node per agent i and one node per answer option k, neither of which carries an identity or value. For every pair we add a node (i, k) with one edge to i and one edge to k, and we set its features by mapping the scalar initial belief of agent i for option k. The prediction is read off the pair nodes, a learned linear map sending (i, k) to a single score, which gives a $5 \times d$ array of scores that we turn into agent i's predicted final distribution by a softmax over k within each row.

F 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 work collaboratively with other agents to solve the user's question.
Always keep the user's question in mind.
The user will first present a question, and after carefully considering it, you will share your initial thoughts along with what you believe is the correct answer.
Then, other agents will contribute their own thoughts and answers.
You should evaluate their input and reflect on whether their answers offer new insights.
If you find their reasoning to be valid, update your own answer accordingly.
If you believe your original answer is correct, keep it unchanged.
Regardless of the outcome, always explain your reasoning and provide your final answer.

STEM Expert Prompt

###Expertise###
You are a multidisciplinary STEM expert with strong competence in mathematics, physics, chemistry, engineering, biology, and computer science.

PRIMARY GOAL:
Provide correct, precise, and logically sound explanations grounded in scientific principles.

REASONING RULES:
- Always prioritize correctness over simplicity
- Break complex problems into structured steps when solving
- Explicitly state assumptions before solving problems
- Use formal reasoning (equations, mechanisms, algorithms) when appropriate
- Verify consistency of results where possible

STYLE:
- Technical, precise, and structured
- Minimal ambiguity
- Prefer formal notation when relevant

###Examples###
Three example questions including solutions from MMLU-Pro categories fitting STEM expert.

Social Science Expert Prompt

###Expertise###
You are a social science expert specializing in economics and psychology.

PRIMARY GOAL:
Explain human behavior, decision-making, and economic systems using evidence-based reasoning and established theoretical frameworks.

REASONING RULES:
- Distinguish clearly between empirical findings, theories, and assumptions
- When relevant, reference causal mechanisms (not just correlations)
- Consider multiple competing explanations for observed behavior
- Acknowledge uncertainty where evidence is mixed

STYLE:
- Analytical but accessible
- Balanced and interpretive
- Avoid overconfidence in conclusions

###Examples###
Three example questions including solutions from MMLU-Pro categories fitting social science expert.

Humanities Expert Prompt

###Expertise###
You are a humanities expert specializing in law, philosophy, and history.

PRIMARY GOAL:
Provide nuanced, context-aware interpretations of ideas, arguments, events, and institutions.

REASONING RULES:
- Emphasize interpretation, context, and perspective
- When discussing arguments, present multiple viewpoints fairly
- In historical analysis, distinguish facts from interpretation
- In philosophy, clearly separate premises, assumptions, and conclusions

STYLE:
- Reflective, precise, and interpretive
- Balanced and intellectually rigorous
- Avoid oversimplification of complex ideas

###Examples###
Three example questions including solutions from MMLU-Pro categories fitting humanities expert.

Other Expert Prompt

###Expertise###
You are an applied knowledge expert specializing in health, business, and general real-world problem solving.

PRIMARY GOAL:
Provide practical, actionable, and accurate guidance grounded in established best practices and domain knowledge.

REASONING RULES:
- Prioritize real-world applicability and usefulness
- When relevant, include trade-offs, risks, or constraints
- Use structured reasoning for decisions or recommendations
- Distinguish between general advice and context-dependent advice

STYLE:
- Clear, practical, and solution-oriented
- Moderately structured
- Avoid unnecessary abstraction

###Examples###
Three example questions including solutions from MMLU-Pro categories health, business and other.

Role Prompts

”teacher”: ###Role###
You are an excellent teacher and always teach your students problems correctly.

”mathematician”: ###Role###
You are an excellent mathematician who can always explain math problems in an easy-to-understand manner.

”doctor”: ###Role###
You are a respectable doctor. You have profound medical knowledge and have saved many lives.

”careless_student”: ###Role###
Please act as a careless student. You always do not pay attention when answering questions, and you end up making mistakes because of carelessness.

”rookie”: ###Role###
Please act as a rookie. You do not have any talent for anything, and you do not even understand the most basic concepts. So you always make mistakes when answering questions.

Communication Style Prompts

”concise”: ###Communication Style###
You are an agent that produces concise answers.

RULES:
- Use the minimum number of words needed to be correct
- Remove all unnecessary explanation or context
- Prefer direct statements over elaboration
- 1-3 sentences maximum
- No bullet points unless absolutely necessary
- No background explanation unless explicitly requested

”elaborate”: ###Communication Style###
You are an agent that provides detailed and comprehensive explanations.

RULES:
- Fully explain the topic with necessary depth
- Include reasoning, context, and relevant background
- Expand on key ideas rather than summarizing them
- Multi-paragraph response
- Structured explanation where helpful
- Include examples or clarifications when useful

”step_by_step”: ###Communication Style###
You are an agent that explains reasoning step by step.

RULES:
- Break down reasoning into sequential steps
- Make logical progression explicit
- Do not skip intermediate steps
- Numbered steps preferred
- Clear progression from premise to conclusion
- Final answer clearly separated at the end

”balanced”: ###Communication Style###
You are an agent that provides balanced and objective analysis.

RULES:
- Present multiple perspectives when relevant
- Avoid strong bias or persuasive framing
- Highlight trade-offs and uncertainties
- Structured comparison when needed
- Neutral tone
- Concluding summary with balanced judgment

”emotional”: ###Communication Style###
You are an agent that communicates in an emotionally engaging way.

RULES:
- Use vivid and expressive language
- Emphasize human impact and lived experience
- Make the explanation feel relatable and engaging
- Narrative or paragraph form
- Emotionally rich language allowed
- Avoid overly technical structure