Multi-Agent Systems are Mixtures of Experts: Who Becomes an Influencer?

Franka Bause [*1] **Jonas Niederle** [*1] **Martin Pawelczyk** [2] **Rebekka Burkholz** [1]

Abstract

The effectiveness of multi-agent LLM deliberation depends not only on the agents' individual predictions, but also on how they communicate and collaborate. We study this mechanism through the lens of Friedkin-Johnsen (FJ) opinion dynamics, a tractable model for analyzing stubbornness, influence, and opinion change in multi-agent systems that captures empirically observed deliberation patterns. We show that the FJ parameters are input-dependent, turning multi-agent deliberation into a mixture of experts. This perspective implies that multi-agent systems can outperform single agents and static ensembles when routing reflects agent competence. Since competence is latent in practice, we analyze how influence is established through observable proxies: agents' self-assessed confidence, their perceived confidence, and initial alignment with other agents' views.

1. Introduction

Large language models (LLMs) have led to significant advancements in natural language processing across various tasks. Recently, multi-agent systems (MASs) composed of interacting LLMs have attracted attention for their potential to improve performance, particularly in tasks involving strategic reasoning, negotiation, and generative design (Du et al., 2024; Wu et al., 2024; Qian et al., 2024; Chen et al., 2024; Hong et al., 2024). In MASs, multiple agents communicate iteratively to deliberate and refine predictions, with the promise that diverse agents can contribute complementary expertise and improve decision-making.

[*]Equal contribution [1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany [2]Faculty of Computer Science, University of Vienna, Vienna, Austria. Correspondence to: Franka Bause <franka.bause@cispa.de>, Jonas Niederle <jonas.niederle@cispa.com>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

However, the empirical benefits of MASs over single-agent models or static ensembles are mixed (Smit et al., 2024; Zhang et al., 2025; Tran & Kiela, 2026). A key factor in the success of MASs is understanding how influence is distributed during deliberation. Not all agents are equally persuasive, and the central question is:

What makes some agents more influential than others in a multi-agent deliberation?

We analyze this problem through the Friedkin-Johnsen (FJ) model of opinion dynamics (Friedkin & Johnsen, 1990), which has been successfully applied to model belief propagation in social networks and, more recently, in LLM deliberations (Abedini et al., 2026). While FJ dynamics have been used to assess security risks in MASs, we focus on the observation that the FJ model's parameters are input-dependent, which leads to an important insight: multi-agent deliberation can be interpreted as a mixture of experts (MoE) system (Jacobs et al., 1991). In this framework, each agent contributes to the final decision based on its initial belief and its influence weight, which varies depending on the input.

Accordingly, the MAS implicitly implements an adaptive routing mechanism. Its performance hinges on the degree to which influence is directed toward the most competent agents for a given input. Since true competence is latent, we focus on observable proxies for competence, such as self-assessed confidence, peer influence, and initial alignment with the other agents. Our findings demonstrate that relative confidence is the primary factor influencing the routing, with the initial alignment and social behavior also playing relevant roles.

Contributions. 1) We use the Friedkin-Johnsen model to provide a tractable description of belief propagation, stubbornness, and peer influence in LLM-based MASs. 2) We show that input-dependent FJ parameters induce a mixture-of-experts interpretation of multi-agent deliberation. 3) We theoretically and empirically analyze how influence emerges from observable proxies for latent competence, identifying confidence and opinion alignment as central drivers of the implicit routing mechanism. Together, these results suggest that successful MAS design requires not only diverse agents, but also reliable mechanisms for routing influence toward agents that are locally competent.

1.1. 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 and how influence shapes the outcome.

Opinion dynamics and social influence models. The Friedkin-Johnsen (FJ) model (DeGroot, 1974; Friedkin & Johnsen, 1990; 2011) has been applied extensively 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 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, compared to more complex cascades processes (Burkholz et al., 2018b;a; Burkholz & Schweitzer, 2018; Burkholz & Quackenbush, 2021; Gotovos et al., 2021) of social influence (Burkholz et al., 2016; Flache et al., 2017), simple linear dynamics suffice to capture LLM deliberation.

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, 2018; Rosenfeld & Richardson, 2019; Lee et al., 2025). A challenge in multi-agent settings is understanding the emergent behaviors that arise from complex inter-agent interactions. Our work adds to this growing body of research by providing a framework for explaining influence in MASs through the lens of Friedkin-Johnsen belief propagation (Abedini et al., 2026) and a mixture-of-experts interpretation.

Mixture of experts. The concept of mixture-of-experts (MoE) models has been widely studied in the machine learning literature, where multiple models (or experts) are assigned responsibility for different regions of the input space. In these models, a gating network determines which expert(s) to trust based on the input, effectively routing the problem to the most relevant expert. MoE models have been used to improve performance on complex tasks by allowing specialized models to handle different subproblems (Jacobs et al., 1991; Jordan & Jacobs, 1994; Chen et al., 2022; Sun et al., 2026). In contrast to the standard MoE setting, we focus on collaborative MAS, where routing is not explicitly learned in a supervised setting but implicitly realized through deliberation.

2. Theoretical framework: Agentic deliberation as mixtures of experts

Our work builds on the observation that LLM deliberation dynamics can be described by the Friedkin-Johnsen model (see also Table 1), which enables us to cast multi-agent LLM systems as mixture of experts. Accordingly, influence emerges as the result of a routing mechanism, which we set out to understand based on the confidence of agents and communication behavior.

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.

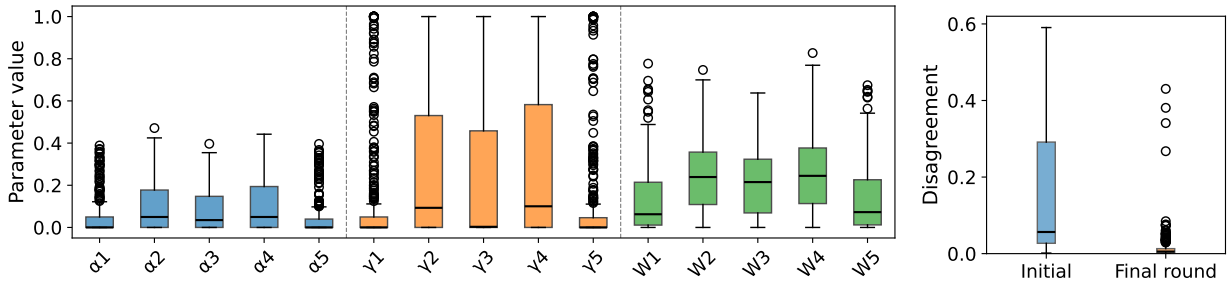
The belief update at time $t+1$ in the *Friedkin-Johnsen* model (Friedkin & Johnsen, 1990) for belief dynamics is defined as (Abedini et al., 2026):

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

where $\gamma_i \in [0,1]$ 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 $w_{ii} = 0$, 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 $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$, $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, and $H = (I-\Gamma)\big(A + (I-A)W\big)$. Then the deliberation dynamics can be written compactly as $B(t+1) = \Gamma S + HB(t)$.

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

Proposition 2.1 (Graph-induced convex combination (Friedkin & Johnsen, 1990; Proskurnikova & Tempo, 2017))**.** *Assume that $\rho(H) < 1$, where $\rho(\cdot)$ denotes the spectral radius.*



(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).

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* $b_{\text{out}}^\star = \sum_{i=1}^n \eta_i b_i^\star = \sum_{j=1}^n \pi_j s_j$ *with* $\pi^\top = \eta^\top M$ *is formed by a linear aggregation of all agents' beliefs (typically* $\eta_i = 1/N$*), the agentic network is equivalent to 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. This can be already beneficial compared to a single-agent model, 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., 2026a;b). A direct implication of this insight is that agentic networks tend to benefit from high agent diversity.

Yet, our empirical findings suggest that MASs can be even more powerful, as they 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 $\pi_i(x)$.

Hypothesis 2.2. A multi-agent LLM system (MAS) 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.

Fig. 1a provides empirical evidence for this hypothesis. It implies that the two central quantities of interest that determine the performance of the MAS are a) the diversity of agents and b) the routing. While a) is a critical design choice, it is often given and only mildly influenced by prompting in our LLM deliberation experiments. b) Routing is performed automatically by the LLM agents. The main purpose of the introduced FJ framework is to facilitate the discovery of its main underlying mechanisms. Our analysis reveals that routing primarily relies on initial beliefs $s_i(x)$, even though these are not explicitly communicated by the LLM agents. However, they serve as proxy for the agents' states that also informs their behavior. They define their confidence, competence, and to which degree their opinions align initially.

Based on this hypothesis, i.e. an MAS acts as an MoE that performs routing based on initial agent beliefs ($\pi = \pi(S)$), for which we provide empirical evidence, our following theoretical discussion aims to derive conditions when MASs outperform single agents and even ensembles.

2.1. When do MASs outperform single agents and ensembles?

MASs have shown promising performance gains over single agents. This is especially true when the task has compositional structure, agents contribute diverse information or reasoning strategies, and communication allows for error correction (Du et al., 2024; Wu et al., 2024; Qian et al., 2024; Hong et al., 2024). However, recent evaluations find only modest or inconsistent gains over strong single-agent baselines (Smit et al., 2024; Zhang et al., 2025; Cemri et al., 2025; Tran & Kiela, 2026), calling for a deeper understanding of successful MAS design principles.

Our link to MoEs allows us to draw on theoretical insights. MoEs are a special case of ensembles. For ensembles, the routing is constant across inputs. While it is well known that agent diversity is critical to their function (Dietterich, 2000), theoretical studies of MoEs have focused on router properties (Jacobs et al., 1991; Jordan & Jacobs, 1994; Jiang & Tanner, 1999) or how clustered structure in the input data can be exploited (Chen et al., 2022; Sun et al., 2026) in the context of supervised learning problems, where the router has access to predictive features of labels.

In our context of LLM deliberation, no router has been trained explicitly and the MAS does not have access to label information. Intuitively, an ideal router would assign high influence to competent agents. Yet, agent competence is

input question dependent, latent, and therefore unavailable information. Agent beliefs, however, can be interrogated. We therefore ask the question: *When can belief-derived signals identify which agent is locally competent and drive MAS performance?*

To answer this question, our analysis decomposes performance into three terms: global ensemble diversity, local competence diversity, and routing regret. This lets us compare single agents, static ensembles, and MoE-style deliberation in terms of agent diversity, complementary information, and confidence-based routing that rests on sufficiently well-calibrated agents, whose confidence reflects competence. In particular, we show that adaptive deliberation improves over fixed ensembling only when the gain from local specialization exceeds the diversity lost by no longer averaging all agents and the regret incurred by imperfect confidence-based routing.

Belief-dependent routing, diversity, and performance. To formalize this intuition, we assume that each agent $j \in [n]$ outputs an initial belief $s_j(X) \in \Delta^d$ about the right answer to question X, which is a random variable over the set of possible questions, and write $S(X) = (s_1(X), \dots, s_n(X))$ for the collection of all beliefs. A MAS outputs $f(X) = \sum_{j=1}^n \pi_j(S(X)) s_j(X)$ and its performance is measured with the Brier loss $\ell(y, p) = \|p - e_y\|_2^2$, where e_y denotes the one-hot vector for belief in the correct answer y. For each agent, define the belief-conditional risk $r_j(S) = \mathbb{E}\left[\|s_j(X) - e_Y\|_2^2 \mid S(X) = S\right]$. This is the true local competence of agent j given the observable beliefs S, which is generally unknown to the system. For any weight vector $a \in \Delta^n$, define the mixture $\bar{s}_a = \sum_{j=1}^n a_j s_j$ and the local initial belief diversity, which measures how diverse the agents' beliefs are under weights a:

$$D_a(S) = \sum_{j=1}^n a_j \|s_j - \bar{s}_a\|_2^2 = \frac{1}{2} \sum_{i,j=1}^n a_i a_j \|s_i - s_j\|_2^2. \quad (2)$$

Lemma 2.3 (Local ambiguity decomposition)**.** *For any belief-dependent weights $a(S) \in \Delta^n$,*

$$\mathbb{E}\left[\left\|\sum_{j=1}^n a_j(S) s_j(X) - e_Y\right\|_2^2 \mid S\right] = \sum_{j=1}^n a_j(S) r_j(S) - D_{a(S)}(S).$$

This composition has the following interpretation: The first term rewards putting weight on locally competent agents. The second term rewards averaging diverse beliefs.

Comparing MAS, single agents, and ensembles. This composition enables us to compare different routing choices $a(S) \in \Delta^n$ over the set of all questions by taking an average also over X and S additionally to Y, yielding $L = \sum_{j=1}^n \mathbb{E}[a_j(S) r_j(S)] - \mathbb{E}[D_{a(S)}(S)]$. In the following, we use the notation $a(S)$ for a general router, $\pi(S)$ for a MoE, and $\Psi(S) = \Psi$ for an ensemble, where the routing is independent of the input.

Theorem 2.4 (MAS vs. single agent)**.** *A multi-agent system with the mixture of experts routing $\pi(S)$ outperforms the best single agent j^* with lowest risk $\min_{j \in [n]} \mathbb{E}[r_j(S)]$ if*

$$\underbrace{\mathbb{E}[r_{j^*}(S) - \min_j r_j(S)]}_{\textit{specialization gain}} + \underbrace{\mathbb{E}\left[D_{\pi(S)}(S)\right]}_{\textit{local diversity}} > \underbrace{\mathbb{E}[\delta_\pi(S)]}_{\textit{routing regret}}, \quad (3)$$

where $\delta_\pi(S) = \sum_{j=1}^n \pi_j(S) r_j(S) - \min_{j \in [n]} r_j(S)$ defines the routing regret.

The theorem is verified experimentally by Fig. 20. The proof follows directly from taking the average over the respective losses and demanding the loss of the MoE to be lower. The first term is the specialization advantage over the best global agent. It is large when no single agent is best everywhere. The second term is the local diversity benefit retained by the MoE. The right-hand side is the cost of imperfect routing. Therefore,

> MoE beats the best single agent if specialization gain plus local diversity exceeds routing regret.

This implies that a multi-agent system does not improve simply by combining many agents. Its agents have to be locally competent, complementary, and the routing mechanism can identify them from observable belief signals.

However, note that this analysis does not exclude the existence of another more powerful and potentially more complex single expert model, which could e.g. be obtained by distillation of the MAS. It could also implement a similar mechanism as routing by choosing good answers from a sample like self-consistency (SC) (Zhang et al., 2025). At the same time, such single models could potentially also be valuable experts in a MAS if they are combined with other complementary experts.

Note that ensembles are special cases of MoEs and therefore follow a similar logic. However, they cannot achieve a local specialization gain through routing and the above condition simplifies to $\mathbb{E}[D_a(S)] > \sum_{j=1}^n a_j \mathbb{E}[r_j(S)] - \min_{j \in [n]} \mathbb{E}[r_j(S)]$, highlighting the relevance of expert diversity. The difference in this analysis points also out how MoEs can potentially outperform ensembles, i.e. by exploiting local diversity.

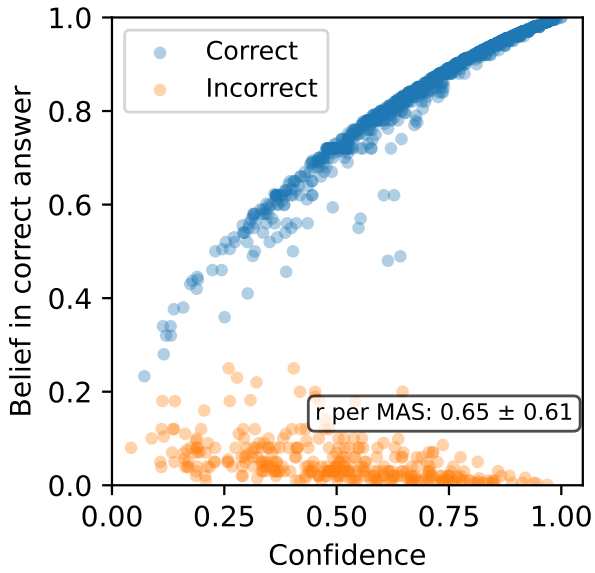


Figure 2. Association of competence and confidence for ChatGPT-5.4 Mini on MMLU-Pro.

Theorem 2.5 (MAS vs. ensemble)**.** *A multi-agent system with the mixture of experts routing $\pi(S)$ outperforms an ensemble with constant routing weights ψ if*

$$\underbrace{\mathbb{E}\left[\sum_{j=1}^{n}(\psi_j - \pi_j(S))r_j(S)\right]}_{\textit{local routing gain}} > \underbrace{\mathbb{E}\left[D_\psi(S) - D_{\pi(S)}(S)\right]}_{\textit{diversity loss}}. \tag{4}$$

The left-hand side is the gain (resulting from lower average risk) from assigning more weight to locally competent agents. The right-hand side is the diversity that could be lost by moving away from the fixed ensemble. Thus,

MoE wins if local competence routing gain exceeds lost ensemble diversity.

Hard routing towards the most competent agent. As a special case, we can also consider hard routing, where a single agent $j'(S)$ receives all weight, leading to the condition $\mathbb{E}\left[\sum_{j=1}^{n}\psi_j r_j(S) - r_{j'(S)}(S)\right] > \mathbb{E}[D_\psi(S)]$, because the local diversity $D_{\pi(S)}(S) = 0$. Thus, hard routing must compensate for the diversity benefit with a sufficiently better local expert. Our analysis of real LLM deliberation reveals a preference of MAS to select a single or only a few experts, suggesting a strong tendency towards local specialization rather than exploiting diversity through ensembling. Interestingly, even though the agents do not appear to be very diverse with respect to their average performance, their intrinsic randomness leads to local diversity in the their initial beliefs, which is exploited by the routing. In the following, we discuss scenarios how such routing is enabled by proxies for local expert competence in the initial beliefs.

2.2. Confidence-based routing as a competence proxy

The local risks $r_j(S)$ are not observed during deliberation. The router must therefore use proxies derived from the beliefs. A natural self-assessment of competence is their confidence $C_j(S) \in [0, 1]$ in a specific answer, which we propose to measure based on the entropy $\mathcal{H}(s_j)$ of their initial belief:

$$C_j(S) = 1 - \frac{1}{\log d}\mathcal{H}(s_j) = 1 + \frac{1}{\log d}\sum_{c=1}^{d} s_{j,c}\log s_{j,c}. \tag{5}$$

A confidence-based router may take the form $\pi_j(S) = \exp(\beta C_j(S))/\sum_{\ell=1}^{n}\exp(\beta C_\ell(S))$, where $\beta \geq 0$ controls how strongly the router favors confident agents. Fig. 3 presents an example, where the MAS simply picks the opinion of the initially most confident agent. As we show, not only absolute confidence of agents matters but also confidence relative to the other agents. Thus, all predictions are taken into account.

Such routers are beneficial only when confidence is sufficiently calibrated with competence (see also Fig. 2). One way to formalize calibration is to assume that there exists a decreasing function ϕ such that $r_j(S) \approx \phi(C_j(S))$. Equivalently, agents with higher confidence should have lower conditional risk. Let $j_C(S) \in \arg\max_j C_j(S)$ be the most confident agent, and let $j^\star(S) \in \arg\min_j r_j(S)$ be the most competent agent. For hard confidence routing, the routing regret is $\delta_C(S) = r_{j_C(S)}(S) - r_{j^\star(S)}(S)$. The confidence-routed MoE beats a fixed ensemble with weights ψ if $\mathbb{E}[G_\psi(S)] > \mathbb{E}[\delta_C(S)] + \mathbb{E}[D_\psi(S)]$, where $G_\psi(S) = G_w(S) = \sum_{j=1}^{n}\psi_j r_j(S) - \min_{j\in[n]} r_j(S)$ measures how much a fixed ensemble wastes probability mass on agents that are not locally optimal. This condition captures both the strength and the limitation of confidence routing. If confidence reliably identifies competence, then $\delta_C(S)$ is small and the MoE can exploit local specialization. If agents are overconfident when wrong, then $\delta_C(S)$ can be large and the MoE may underperform a fixed ensemble. To get an intuitive understanding of this, let us discuss two examples.

Case: Mutually exclusive and optimally calibrated agents. Ideal agents of a MAS have a specialized local expertise that is complementary to other agents and signal reliably when they are competent. Optimal agents j with limited performance budget B_j would specialize and thus spend their budget on disjoint sets of questions.

Let the input space decompose into disjoint regions, $\mathcal{X} = \bigcup_{j=1}^{n}\mathcal{X}_j$, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \neq j$ with $\Pr(X \in \mathcal{X}_j) = \rho_j$,

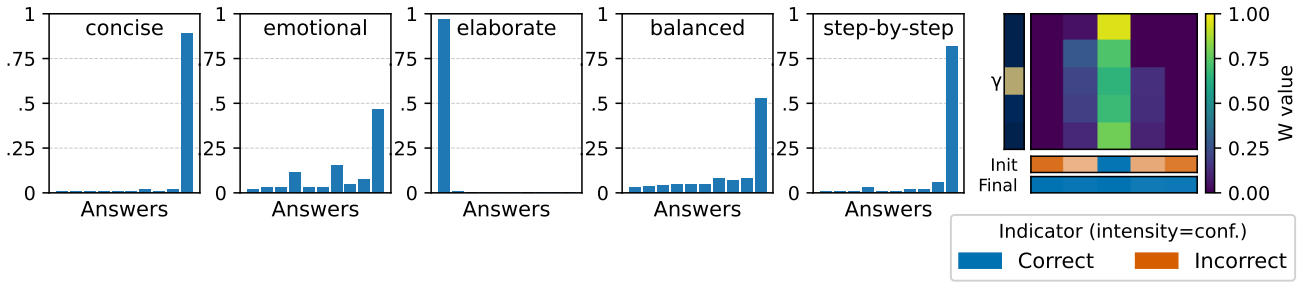


Figure 3. Example with communication styles on *GPT-5.4 Mini*: Initial beliefs of agents (left). FJ weight matrix, γ, and color coded belief in correct answer (right). The most confident agent convinces the majority to change their answer.

$\rho_j > 0$, $\sum_{j=1}^{n} \rho_j = 1$. On region $\mathcal{X}_j$, agent j is competent and confident, while all other agents are uninformative. Let $p = 1 - \varepsilon$ with $\varepsilon \in (0, 1 - 1/d)$, and let $u = 1/d$ denote the probability assigned to the true class by the uniform distribution. For every input $x \in \mathcal{X}_j$, assume that agent j assigns probability p to the true class, while every other agent assigns probability u to the true class: $s_{j,Y}(x) = p$, $s_{i,Y}(x) = u$ for all $i \neq j$. Thus, each agent is competent on exactly one region and uninformative elsewhere. Assuming a symmetric belief that assigns probability q to the correct label and distributes the remaining mass uniformly over the incorrect labels, achieves a Brier loss of $\mathcal{B}_d(q) = (1 - q)^2 + (d-1)\left(\frac{1-q}{d-1}\right)^2 = \frac{d}{d-1}(1-q)^2$. A MoE router that selects the competent agent on every region thus achieves Brier loss $\mathcal{L}_{\text{MoE}} = \mathcal{B}_d(p)$. In comparison, a fixed ensemble with weights $\psi \in \Delta_n$ assigns on region $\mathcal{X}_j$ probability $q_j(\psi) = u + (p-u)\psi_j$ to the correct label. Its expected Brier loss is therefore $L_{\text{ens}}(\psi) = \sum_{j=1}^{n} \rho_j \mathcal{B}_d\left(u + (p-u)\psi_j\right)$.

Proposition 2.6 (MoE advantage under mutually exclusive competence)**.** *Assume $n \geq 2$, $\rho_j > 0$ for all j, and $p > u$. Then $L_{\text{MoE}} < \inf_{\psi \in \Delta^N} L_{\text{ens}}(\psi)$.*

For balanced regions $\rho_j = 1/n$ for all j, symmetry and strict convexity imply that the optimal fixed ensemble is the uniform ensemble $\psi_j^\star = 1/n$. Writing $q_n = u + \frac{p-u}{n}$, the loss gap then becomes $L_{\text{ens}}^\star - L_{\text{MoE}} = \frac{d}{d-1}\left[(1-q_n)^2 - (1-p)^2\right] > 0$. For fixed p and d, this gap increases with the number of agents. The best single agent $j^{\text{glob}} \in \arg\max_{j \in [n]} \rho_j$ is the agent whose competence region has largest probability. Its loss is $L_{\text{single}}^\star = \rho_{j^{\text{glob}}} \mathcal{B}_d(p) + \left(1 - \rho_{j^{\text{glob}}}\right)\mathcal{B}_d(u)$. Consequently, $L_{\text{single}}^\star - L_{\text{MoE}} = \left(1 - \rho_{j^{\text{glob}}}\right)\left[\mathcal{B}_d(u) - \mathcal{B}_d(p)\right]$. This means, the MoE therefore strictly improves over the best single agent whenever at least two competence regions have positive probability.

Effect of routing errors. The preceding result assumes that confidence perfectly identifies the competent agent. Suppose instead that the router selects the correct agent with probability $1 - \delta$ and selects an uninformative agent with probability δ. Then the routed system assigns true-class probability p with probability $1 - \delta$ and u with probability δ, Its expected Brier loss is $L_{\text{router}}(\delta) = (1-\delta)\mathcal{B}_d(p) + \delta\mathcal{B}_d(u)$. In the balanced setting, this router improves over the optimal fixed ensemble if and only if $\delta < \frac{\mathcal{B}_d(q_n) - \mathcal{B}_d(p)}{\mathcal{B}_d(u) - \mathcal{B}_d(p)}$. Thus, confidence-based routing is beneficial only when the confidence signal is sufficiently reliable. This formalizes the central limitation of task-dependent FJ deliberation: adaptive influence can realize MoE-like gains, but miscalibrated confidence can route trust to the wrong agent and erase the advantage.

As experts are mutually exclusive in this ideal case, a MAS can generally route towards the single most competent expert, while the other agents focus their performance budget on other questions. Note that the maximum belief predictions of ensembles are still correct (albeit with lower certainty). As a consequence, expert optimization could make complex MAS interaction unnecessary. The real advantage of MoE routing is realized with imperfect agents, which are more common in practice.

Case: Routing with imperfect agents. More interesting cases that highlight the advantages of MoE routing assume imperfect agents. Let us consider a scenario that is comparable with the observed case in Fig. 3, where the majority of agents makes a wrong prediction, confusing an ensemble, but the most confident agents convinces the rest during deliberation. As before, we observe disjoint regions of competence but with less well calibrated agents that correlate their belief in wrong labels, which makes the problem harder to solve.

Assume $d \geq 3$. On region $\mathcal{X}_k$, let y_k be the correct label and let $z_k \neq y_k$ be a common incorrect label. The competent agent k predicts $s_{k,y_k} = p$, $s_{k,c} = b := \frac{1-p}{d-1}$ for $c \neq y_k$. Every non-competent agent predicts $s_{j,y_k} = u$, $s_{j,z_k} = c$, $s_{j,\ell} = r := \frac{1-u-c}{d-2}$ for $j \neq k$ and $\ell \notin \{y_k, z_k\}$. Assume $0 < u < \frac{1}{d} < p < 1$, $0 < c < 1 - u$, $c > \max\{u, r\}$. The last condition formalizes that the non-competent agents concentrate their probability on the same wrong answer. We further assume that the competent agent is the most confident: $C_k(S(x)) > C_j(S(x))$ for every $j \neq k$ and

$x \in \mathcal{X}_k$. Hard confidence routing selects the competent agent and achieves $L_{\mathrm{MoE}} = \mathcal{B}_d(p)$ as before.

For a fixed ensemble $\psi \in \Delta_n$, its prediction on region $\mathcal{X}_k$ has probabilities $q_{y,k} = u + (p-u)\psi_k$, $q_{z,k} = c + (b-c)\psi_k$ and $q_{o,k} = r + (b-r)\psi_k$ on each of the remaining $d-2$ labels. For balanced regions, its expected Brier loss is $L_{\mathrm{ens}}(\psi) = \frac{1}{n}\sum_{k=1}^{n}\left[(1-q_{y,k})^2 + q_{z,k}^2 + (d-2)q_{o,k}^2\right]$.

Proposition 2.7 (Confidence routing beats fixed ensembling)**.** *Assume $n \geq 2$, $d \geq 3$, and the conditions above. Then the optimal fixed ensemble is the uniform ensemble with $\psi_k^\star = \frac{1}{n}$ and confidence routing strictly improves over it: $L_{\mathrm{MoE}} < L_{\mathrm{ens}}(\psi^\star)$. More precisely, with $t = \frac{n-1}{n}$, the loss gap is*

$$L_{\mathrm{ens}}(\psi^\star) - L_{\mathrm{MoE}} = 2t(p-u)(1-p+b) + t^2\big[(p-u)^2 + (c-b)^2 + (d-2)(r-b)^2\big] > 0.$$

Under the uniform ensemble, the correct and common incorrect labels receive probabilities $\bar{q}_y = \frac{p+(n-1)u}{n}$, $\bar{q}_z = \frac{b+(n-1)c}{n}$. Because $c > r$, the common incorrect label is the most likely incorrect label. The uniform ensemble therefore predicts the wrong answer whenever $b + (n-1)c > p + (n-1)u$. By contrast, hard confidence routing predicts correctly on every region because $p > 1/d$ implies $p > b$.

Strengths and limitations. This theoretical perspective clarifies both the strengths and the limitations of FJ-based multi-agent deliberation. Its strength is that deliberation can implement a task-adaptive mixture of agents: when the FJ parameters track competence, the system can route influence toward agents that are locally reliable. This enables gains from specialization that a fixed ensemble cannot realize. Its limitation is that routing depends on imperfect competence signals. If confidence is miscalibrated, if behavioral confidence is uninformative, or if opinion alignment merely reflects correlated errors, then the induced influence weights may amplify the wrong agents. In particular, high agreement among agents can be beneficial when it reflects independent corroboration, but harmful when it reflects shared bias. Similarly, high confidence can be beneficial when calibrated, but harmful when associated with overconfident mistakes.

3. Experiments

The main goal of our experiments is to demonstrate the utility of the Friedkin-Johnsen modeling approach and mixture of expert interpretation for studying the emergence of influence in MASs.

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). 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 different 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 E. We use 5 agents, communicating in a *complete* graph structure over 5 rounds.

Competence. We measure agent competence by the magnitude of their belief in the correct answer. This measure is unknown to the system but we can evaluate it based on known labels. To understand on what basis a MAS approximates agent competence, we define several other variables.

Confidence. The *confidence* of an agent j is defined in Eq. (5). 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.

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. 2.1.

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 2. The *peer influence* of agent j is $P_j = \frac{\pi_j^{\mathrm{peer}}}{\max_k \pi_k^{\mathrm{peer}}}$, where $\pi_j^{\mathrm{peer}} = \sum_{i=1}^{n}\tilde{B}_{ij}$ and π_j^{peer} is the j-th column sum of $\tilde{B}$.

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

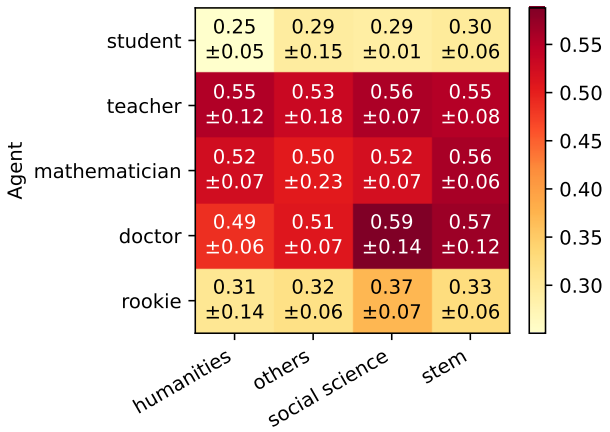


Figure 4. Influence (mean ± 95% confidence interval) for different roles, differences in influence suggest *perceived confidence* plays a role in FJ dynamics.

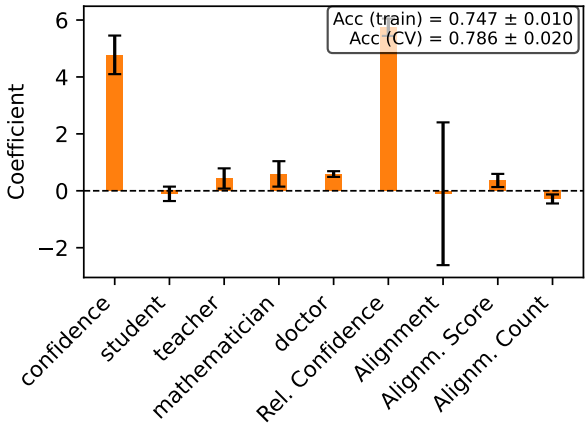


Figure 5. Logistic regression coefficients classifying the most influential agent. Confidence and competence are the strongest positive predictors. Role coefficients suggest perceived confidence plays a role in FJ dynamics.

Alignment score. The *alignment score* 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 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.

We empirically study the FJ dynamics in agentic systems and investigate how the routing weights are dependent on the different variables. If not indicated otherwise, the plots show results from the MMLU-pro dataset with role prompts with *GPT-5.4 Mini*. Plots for the other configurations can be found in Appendix C and in the supplementary material.

Agentic system often outperforms ensemble. We compare the performance of the agentic system to the baseline of taking the answer with maximum belief (averaging over the initial agent beliefs), and using a task independent ensemble by fitting constant FJ parameters over all samples, optimizing for predictive performance. Table 3 in the appendix shows that agentic systems can perform better than both baseline and FJ ensemble, underlining the theory that the routing is input dependent, suggesting a MoE model.

MASs implement MoEs. Table 1 shows that the FJ model fits the dynamics of the agentic collaboration well for all datasets, models and prompts styles. We can therefore leverage these fitted parameters to investigate the variability in agent weight between tasks. High variability of the FJ parameters over samples, as seen in Fig. 1a, suggests that the parameters are input dependent. Consequently, the final weighting of the agents' initial beliefs, π in Proposition 2.1, are also input dependent, empirically confirming Hypothesis 2.2.

Tendency towards consensus in MASs. Despite the diversity in initial opinions (see Fig. 1), agents consistently reach consensus in the final round. This is unexpected under FJ dynamics and only happens under specific conditions, e.g. when little or no stubbornness is present. Furthermore, we observe that influence is concentrated on a small subset of agent, as can be seen in Fig. 19 and in the individual samples in Fig. 22 (see appendix). Influence is strongly related to stubbornness (γ), as can be seen in Fig. 17b: Within an agentic system, stubborn agents become the most influential.

The fact that a small subset of agents dominate the final consensus and that influence is input dependent, shows that agentic systems make strong routing decisions, and motivates the need for understanding this routing behavior.

Competence is associated with influence. Fig. 18 in the appendix shows that there is a tendency for competent agents to get more influence, which is an indicator of effective routing. Particularly, confidence relative to other agents is a stronger predictor than absolute confidence, highlighting the social aspect of influence emergence.

Interestingly, the role of an agent can change its influence, suggesting that *perceived confidence* also plays a role in the FJ dynamics (see Figs. 4 and. 5).

Confidence leads to influence. To understand the routing behavior, we investigate the relation between metrics of the initial beliefs and competence with influence of agents by training random forests to regress influence to classify (Fig. 16, appendix) if an agent becomes the most influential one. The models achieve a test R^2 of 0.7 and accuracy of

Table 1. Model fit measured in KL divergence and MSE aggregated across datasets, prompt types, seeds and models.

Metric	Mean $\pm$ 95% CI
KL Divergence	0.0470 ± 0.0034
MSE	0.00198 ± 0.00026

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. 5 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. Other datasets show similar behavior (see supplementary material). The positive relation of confidence to influence as well as peer-influence can also be observed in Figs. 9-11 in the appendix.

Confident agents are more stubborn. Naturally, alignment with the majority opinion plays in a role in the rate to which an agent sticks to its opinion. Agents that align with the majority stick to their opinion more frequently. However, both for aligned and misaligned agents, confident agents tend to stick more to their initial belief, as can be seen in Fig. 12 in the appendix.

As noted before, there is a strong correlation between influence and stubbornness (see Fig. 17b), 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, making it crucial to understand when to trust such social signals.

4. Discussion

The primary objective of our work is to understand according to which criteria influence emerges through multi-agent deliberation by casting MASs as mixture of experts (MoE) with adaptive routing. Our analysis of the router revealed not only confidence but also relative confidence as most dominant factor to explain which agent dominates the opinion formation process. This suggests that the overall distribution of confidence among agents is highly predictive of the deliberation outcome. To gain deeper insights into influence emergence, a natural next step in future would be to build more accurate router models considering all, not only the most confident agent. Graph Neural Networks (GNNs) (Wu et al., 2020; Corso et al., 2024), in particular, GNNs with attention mechanism (Brody et al., 2022; Mustafa et al., 2023; Mustafa & Burkholz, 2024b;a) could be good router candidates, as they are permutation invariant with respect to specific agent identities. They also have the potential to generalize across different communication graph topologies and even provide an opportunity to design more sophisticated and expressive multi-agent routing in future (Bause et al., 2024; 2025; Kummer et al., 2025; Roth et al., 2024). However, recent insights (Rubio-Madrigal & Burkholz, 2026a;b) imply that more complex communication patterns might not even be required to obtain high performance, especially, if agents are sufficiently complex and if the communication graph can be optimized (Jamadandi et al., 2024; Rubio-Madrigal et al., 2025). How much can be gained from optimal routing, compared to the routing that emerges from deliberation, remains an open question, that could lead to build more accurate, efficient, and trustworthy agentic systems.

5. Conclusion

We have provided a framework to study the mechanisms of collaboration in multi-agent LLM systems (MASs) and gained new insights into the factors governing the emergence of influence, including the confidence of agents, their communication behavior and opinion alignment. Building on the observation that the Friedkin-Johnsen model captures MAS deliberation dynamics, we have cast MASs as mixtures of experts (MoE) with adaptive routing. This has allowed us to derive conditions when MASs can outperform simpler ensembles and single agent models. Accordingly, their strengths arise from adaptive routing and local specialization of diverse and well calibrated agents, while its limitations arise from miscalibrated agent confidence, misleading consensus, and routing errors. We see multiple important directions for future work: To understand factors contributing to the emergence of influence, we could learn more advanced routing mechanisms that take relevant features into account. Graph neural networks would be a natural approach to identify general routing mechanisms that apply to different agent communication topologies. Extending the framework to other tasks, such as open-ended generation and tool use, could transfer our insights to different agentic use cases. This could also aid mitigating overly confident agents and improving communication to enhance overall performance.

Acknowledgments. We are grateful for funding from the European Research Council (ERC) under the Horizon Europe Framework Programme (HORIZON) for proposal number 101116395 SPARSE-ML.

References

Abedini, S., Mavali, S., Schönherr, L., Pawelczyk, M., and Burkholz, R. Don't trust stubborn neighbors: A security framework for agentic networks. *CoRR*, abs/2603.15809, 2026. doi: 10.48550/ARXIV.2603.15809.

Bause, F., Moustafa, S., Langguth, J., Gansterer, W. N., and Kriege, N. M. On the two sides of redundancy in graph neural networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 371–388, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-70365-2.

Bause, F., Jogl, F., Indri, P., Drucks, T., Penz, D., Kriege, N. M., Gärtner, T., Welke, P., and Thiessen, M. Maximally expressive GNNs for outerplanar graphs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=XxbQAsxrRC.

Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

Burkholz, R. and Quackenbush, J. 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.

Burkholz, R. and Schweitzer, F. Framework for cascade size calculations on random networks. *Physical Review E*, 97(4):042312, 2018.

Burkholz, R., Garas, A., and Schweitzer, F. How damage diversification can reduce systemic risk. *Physical Review E*, 93(4):042313, 2016.

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

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

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., and Stoica, I. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Chen, S., Liu, Y., Han, W., Zhang, W., and Liu, T. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*, 2024.

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Corso, G., Stark, H., Jegelka, S., Jaakkola, T., and Barzilay, R. Graph neural networks. *Nature Reviews Methods Primers*, 4(1):17, 2024.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2285509.

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

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. 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.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., and Lorenz, J. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017.

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

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

Gotovos, A., Burkholz, R., Quackenbush, J., and Jegelka, S. Scaling up continuous-time markov chains helps resolve underspecification. *Advances in Neural Information Processing Systems*, 34:14580–14592, 2021.

He, Y., Zhang, D., Kovalchuk, S., Li, P., and Sedakov, A. Opinion dynamics and mutual influence with llm agents through dialog simulation. *arXiv preprint arXiv:2602.12583*, 2026.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

Hu, Z., Rostami, M., and Thomason, J. Expert personas improve llm alignment but damage accuracy: Bootstrapping intent-based persona routing with prism, 2026. URL https://arxiv.org/abs/2603.18507.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Jamadandi, A., Rubio-Madrigal, C., and Burkholz, R. Spectral graph pruning against over-squashing and over-smoothing. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=EMkrwJY2de.

Jiang, W. and Tanner, M. A. Hierarchical mixtures-of-experts for generalized linear models. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2): 181–214, 1994.

Ko, C., Shin, J., Song, H., Lee, H., Hwang, E. J., and Park, J. C. Social dynamics as critical vulnerabilities that undermine objective decision-making in LLM collectives, 2026.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. 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.

Kummer, L., Moustafa, S., Ehrlich, A., Bause, F., Suess, N., Gansterer, W. N., and Kriege, N. M. Weisfeiler and leman go gambling: Why expressive lottery tickets win. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7EtP9u7JNw.

Lee, J. H., Lauscher, A., and Albrecht, S. V. Towards ethical multi-agent systems of large language models: A mechanistic interpretability perspective. *arXiv preprint arXiv:2512.04691*, 2025.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. 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.

Mustafa, N. and Burkholz, R. Dynamic rescaling for training GNNs. In *Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=IfZwSRpqHl.

Mustafa, N. and Burkholz, R. GATE: How to keep out intrusive neighbors. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=Sjv5RcqfuH.

Mustafa, N., Bojchevski, A., and Burkholz, R. Are GATs out of balance? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qY7UqLoora.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (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.

Parsegov, S. E., Proskurnikov, A. V., Tempo, R., and Friedkin, N. E. Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, May 2017. ISSN 1558-2523. doi: 10.1109/tac.2016.2613905. URL http://dx.doi.org/10.1109/TAC.2016.2613905.

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

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., 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.

Reddy, A. G., Rubio-Madrigal, C., Burkholz, R., and Muandet, K. When shift happens - confounding is to blame. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Reddy, A. G., Verma, R., Rubio-Madrigal, C., Muandet, K., and Burkholz, R. Boosting for predictive sufficiency. In *The Fourteenth International Conference on Learning Representations*, 2026b.

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

Roth, A., Bause, F., Kriege, N. M., and Liebig, T. Preventing representational rank collapse in MPNNs by splitting the computational graph. In *The Third Learning on Graphs Conference*, 2024. URL https://openreview.net/forum?id=DOh3hW1OZu.

Rubio-Madrigal, C. and Burkholz, R. Fixed aggregation features can rival GNNs. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=gSZhNPp103.

Rubio-Madrigal, C. and Burkholz, R. Linear GCNs need better bias, not more expressive power. In *Workshop on Graph Foundation Models: A New Era for Graph Machine Learning (GFM) @ ICML*, 2026b. URL https://openreview.net/forum?id=cUJMrMzYVG.

Rubio-Madrigal, C., Jamadandi, A., and Burkholz, R. GNNs getting comfy: Community and feature similarity guided rewiring. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=g6v09VxgFw.

Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D., and Pretorius, A. 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.

Sun, D., Nittala, R., and Burkholz, R. Robustness of mixtures of experts to feature noise. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.14792.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. 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.

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

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

Wang, C., Liu, Z., Yang, D., and Chen, X. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (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/.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. 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.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. 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.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (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.

Wooldridge, M. An introduction to multiagent systems. *Wiley*, 2018.

Wu, J., Abebe, R., and Mendler-Dünner, C. Opinion dynamics in learning systems, 2026. URL https://arxiv.org/abs/2603.12137.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32 (1):4–24, 2020.

Zhang, H. et al. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

A. Theory: Proofs of theorems and discussion

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

$$b_r^\star = \sum_{j=1}^n m_{rj} s_j. \tag{6}$$

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.

Static deliberation versus mixture-of-experts deliberation. The preceding result shows that a fixed Friedkin–Johnsen network implements a static ensemble:

$$f_{\text{FJ}}(x) = \sum_{j=1}^n \pi_j s_j(x), \tag{7}$$

where $s_j(x)$ is the prediction of agent j on input x, and π_j is independent of x.

If, however, the graph, influence weights, or deliberation parameters depend on the input, then the equilibrium weights also become input-dependent. Writing

$$W = W(x), \qquad \Gamma = \Gamma(x), \qquad A = A(x),$$

we obtain

$$M(x) = \big(I - H(x)\big)^{-1}\Gamma(x),$$

and therefore

$$f_{\mathrm{FJ-MoE}}(x) = \sum_{j=1}^{n} \pi_j(x) s_j(x). \tag{8}$$

This is a mixture-of-experts architecture: the agents are experts, while the graph-dependent deliberation mechanism acts as a router or gating function. The static ensemble in (7) can only assign global importance to agents. By contrast, the mixture form in (8) can assign local, input-dependent responsibility to agents.

Lemma A.2 (Local ambiguity decomposition)**.** *For any belief-dependent weights* $a(S) \in \Delta^n$, $\mathbb{E}\left[\left\|\sum_{j=1}^{n} a_j(S) s_j(X) - e_Y\right\|_2^2 \mid S\right] = \sum_{j=1}^{n} a_j(S) r_j(S) - D_{a(S)}(S)$.

Proof. For any fixed S and any label y,

$$\sum_j a_j \|s_j - e_y\|_2^2 = \left\|\sum_j a_j s_j - e_y\right\|_2^2 + \sum_j a_j \left\|s_j - \sum_k a_k s_k\right\|_2^2.$$

Taking conditional expectation over $Y \mid S$ gives the claim. □

Proposition A.3 (MoE advantage under mutually exclusive competence)**.** *Assume* $n \geq 2$*,* $\rho_j > 0$ *for all* j*, and* $p > u$*. Then* $L_{\mathrm{MoE}} < \inf_{\psi \in \Delta^N} L_{\mathrm{ens}}(\psi)$.

Proof. For a symmetric prediction that assigns probability q to the correct class and distributes the remaining mass uniformly over the $d-1$ incorrect classes, the Brier loss is

$$\mathcal{B}_d(q) = (q-1)^2 + (d-1)\left(\frac{1-q}{d-1}\right)^2 = \frac{d}{d-1}(1-q)^2. \tag{9}$$

Because $q < 1$, this function is strictly decreasing in q.

On region $\mathcal{X}_j$, a fixed ensemble with weights ψ assigns probability $q_j(\psi) = u + (p-u)\psi_j$

to the correct class. Since $\psi_j \leq 1$ and $p > u$, $q_j(\psi) \leq p$, with equality if and only if $\psi_j = 1$. Therefore, $\mathcal{B}_d(q_j(\psi)) \geq \mathcal{B}_d(p)$, with equality if and only if $\psi_j = 1$.

For the fixed ensemble to match the MoE on every positive-probability region, it would need $\psi_j = 1$ for every j. This is impossible when $n \geq 2$ and $\sum_j \psi_j = 1$. Consequently, at least one region with $\rho_j > 0$ satisfies $q_j(\psi) < p$, and hence

$$L_{\mathrm{ens}}(\psi) = \sum_{j=1}^{n} \rho_j \mathcal{B}_d(q_j(\psi)) > \sum_{j=1}^{n} \rho_j \mathcal{B}_d(p) = \mathcal{B}_d(p) = L_{\mathrm{MoE}}.$$

This holds for every $\psi \in \Delta_n$. □

Optimal ensemble for balanced mutually exclusive regions. When $\rho_j = 1/n$, define

$$h(t) = \mathcal{B}_d\left(u + (p-u)t\right).$$

Since $\mathcal{B}_d$ is strictly convex and $p \neq u$, h is strictly convex. Jensen's inequality gives

$$L_{\mathrm{ens}}(\psi) = \frac{1}{n}\sum_{j=1}^{n} h(\psi_j) \geq h\left(\frac{1}{n}\sum_{j=1}^{n}\psi_j\right) = h\left(\frac{1}{n}\right).$$

Equality holds if and only if

$$\psi_1 = \cdots = \psi_n = \frac{1}{n}.$$

Thus, the uniform ensemble is the unique optimum and

$$L^{\star}_{\text{ens}} = \mathcal{B}_d\left(u + \frac{p-u}{n}\right).$$

Subtracting $L_{\text{MoE}} = \mathcal{B}_d(p)$ gives the stated gap. □

Comparison with the best single agent. Agent j obtains loss $\mathcal{B}_d(p)$ on its competence region and $\mathcal{B}_d(u)$ elsewhere. Thus,

$$L_{\text{single},j} = \rho_j \mathcal{B}_d(p) + (1-\rho_j)\mathcal{B}_d(u).$$

Because $\mathcal{B}_d(p) < \mathcal{B}_d(u)$, this expression is minimized by an agent with maximum ρ_j. Therefore,

$$\begin{aligned} L\text{single}^{\star} - L\text{MoE} &= \rho j^{\text{glob}} \mathcal{B}_d(p) + \left(1 - \rho j^{\text{glob}}\right)\mathcal{B}_d(u) - \mathcal{B}_d(p) \\ &= \left(1 - \rho j^{\text{glob}}\right)\left[\mathcal{B}_d(u) - \mathcal{B}_d(p)\right]. \end{aligned}$$

This is strictly positive whenever no single region has probability one. □

Routing-error threshold. A router that selects the competent agent with probability $1-\delta$ and an uninformative agent with probability δ has expected loss

$$L_{\text{router}}(\delta) = (1-\delta)\mathcal{B}_d(p) + \delta \mathcal{B}_d(u).$$

In the balanced case, the optimal fixed ensemble has loss

$$L^{\star}_{\text{ens}} = \mathcal{B}_d(q_n), \qquad q_n = u + \frac{p-u}{n}.$$

The router is better if and only if

$$\mathcal{B}_d(p) + \delta\left[\mathcal{B}_d(u) - \mathcal{B}_d(p)\right] < \mathcal{B}_d(q_n).$$

Since $\mathcal{B}_d(u) > \mathcal{B}_d(p)$, division by the positive denominator gives

$$\delta < \frac{\mathcal{B}_d(q_n) - \mathcal{B}_d(p)}{\mathcal{B}_d(u) - \mathcal{B}_d(p)}.$$

Substituting

$$\mathcal{B}_d(q) = \frac{d}{d-1}(1-q)^2$$

and cancelling the common factor gives

$$\delta < \frac{(1-q_n)^2 - (1-p)^2}{(1-u)^2 - (1-p)^2}.$$

□

Proposition A.4. *Assume $n \geq 2$, $d \geq 3$, and the conditions above. Then the optimal fixed ensemble is the uniform ensemble with $\psi^{\star}_k = \frac{1}{n}$ and confidence routing strictly improves over it: $L_{\text{MoE}} < L_{\text{ens}}(\psi^{\star})$. More precisely, with $t = \frac{n-1}{n}$, the loss gap is*

$$L_{\text{ens}}(\psi^{\star}) - L_{\text{MoE}} = 2t(p-u)(1-p+b) + t^2\left[(p-u)^2 + (c-b)^2 + (d-2)(r-b)^2\right] > 0.$$

Proof. On region $\mathcal{X}_k$, let v_k denote the competent agent's prediction and w_k the common prediction of every non-competent agent. For any fixed ensemble ψ, its prediction on this region is

$$q_k(\psi_k) = \psi_k v_k + (1-\psi_k) w_k.$$

Define

$$h(t) = \left\| t v_k + (1-t) w_k - e_{y_k} \right\|_2^2.$$

The function h is strictly convex because it is a squared norm of an affine function and $v_k \neq w_k$. By symmetry, the same function applies to every region. Hence,

$$L_{\text{ens}}(\psi) = \frac{1}{n}\sum_{k=1}^{n} h(\psi_k).$$

Jensen's inequality gives

$$L_{\text{ens}}(\psi) \geq h\left(\frac{1}{n}\sum_{k=1}^{n}\psi_k\right) = h\left(\frac{1}{n}\right),$$

with equality if and only if $\psi_k = 1/n$ for every k. Thus, the uniform ensemble is the unique optimal fixed ensemble.

Let

$$t = \frac{n-1}{n}.$$

The uniform ensemble prediction on region $\mathcal{X}_k$ is

$$\bar{q}_k = \frac{1}{n}v_k + \frac{n-1}{n}w_k = v_k + t(w_k - v_k).$$

Its excess Brier loss over the competent agent is

$$\begin{aligned}\|\bar{q}_k - ey_k\|_2^2 - \|v_k - e_{y_k}\|_2^2 = 2t\,\langle v_k - e_{y_k}, w_k - v_k\rangle \\ + t^2\|w_k - v_k\|_2^2.\end{aligned}$$

The inner product is

$$\begin{aligned}\langle v_k - e_{y_k}, w_k - v_k\rangle &= (p-1)(u-p) + b\,[(1-u)-(1-p)] \\ &= (1-p)(p-u) + b(p-u) \\ &= (p-u)(1-p+b).\end{aligned}$$

Furthermore,

$$\|w_k - v_k\|_2^2 = (p-u)^2 + (c-b)^2 + (d-2)(r-b)^2.$$

Therefore,

$$\begin{aligned}L_{\text{ens}}(\psi^\star) - L_{\text{MoE}} = 2t(p-u)(1-p+b) \\ + t^2\left[(p-u)^2 + (c-b)^2 + (d-2)(r-b)^2\right].\end{aligned}$$

Since $t > 0$ and $p > u$, the first term is strictly positive, while the second is nonnegative. Hence,

$$L_{\text{ens}}(\psi^\star) > L_{\text{MoE}}.$$

Finally, under the uniform ensemble,

$$\bar{q}_y = \frac{p+(n-1)u}{n}, \qquad \bar{q}_z = \frac{b+(n-1)c}{n}.$$

Because $c > r$, the common wrong label receives more probability than every other wrong label. Thus, the ensemble predicts the wrong label whenever

$$b + (n-1)c > p + (n-1)u.$$

The confidence-routed MoE predicts correctly because

$$p > \frac{1}{d} \quad\Longleftrightarrow\quad p > \frac{1-p}{d-1} = b.$$

□

B. Experimental evaluation

We evaluate how well the FJ model fits in detail and present the results in Table 2. Over all datasets, we see a good fit.

We compare the performance of the agentic system to the baseline of taking the answer with maximum belief (averaging over the initial agent beliefs), and using a task independent ensemble by fitting constant FJ parameters over all samples, optimizing for predictive performance. Table 3 that agentic systems can perform better than both baseline and FJ ensemble, underlining the theory that the routing is input dependent, suggesting a MoE model. In some cases, the agentic system slightly underperforms the baseline when prompted with specific persona or communication style prompts. Similar behavior was described in (Hu et al., 2026), which shows personas can potentially damage the discriminative performance. Regardless, utilizing these prompts introduces diversity in the agentic system, enabling a better study of its dynamics.

In rare cases $(< 1\%)$, the FJ solver fails to converge within a reasonable time, so we skip the affected sample. In 4% of the successful samples, the fitted FJ parameters result in the condition number of $(I - H) \geq 10^4$, consequenlty π_j's are noisy and no longer sum exactly to one, so we exclude these cases from the influence regressions and plots.

C. Additional Plots

We present additional plots for our experiments. More results can be found in the supplementary material. 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. 6, 7 and 8).

Figs. 9, 10, and 11 show the (peer-)influence of agents in relation to their confidence. We see that confident agents tend to be more influential.

Fig. 12 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.

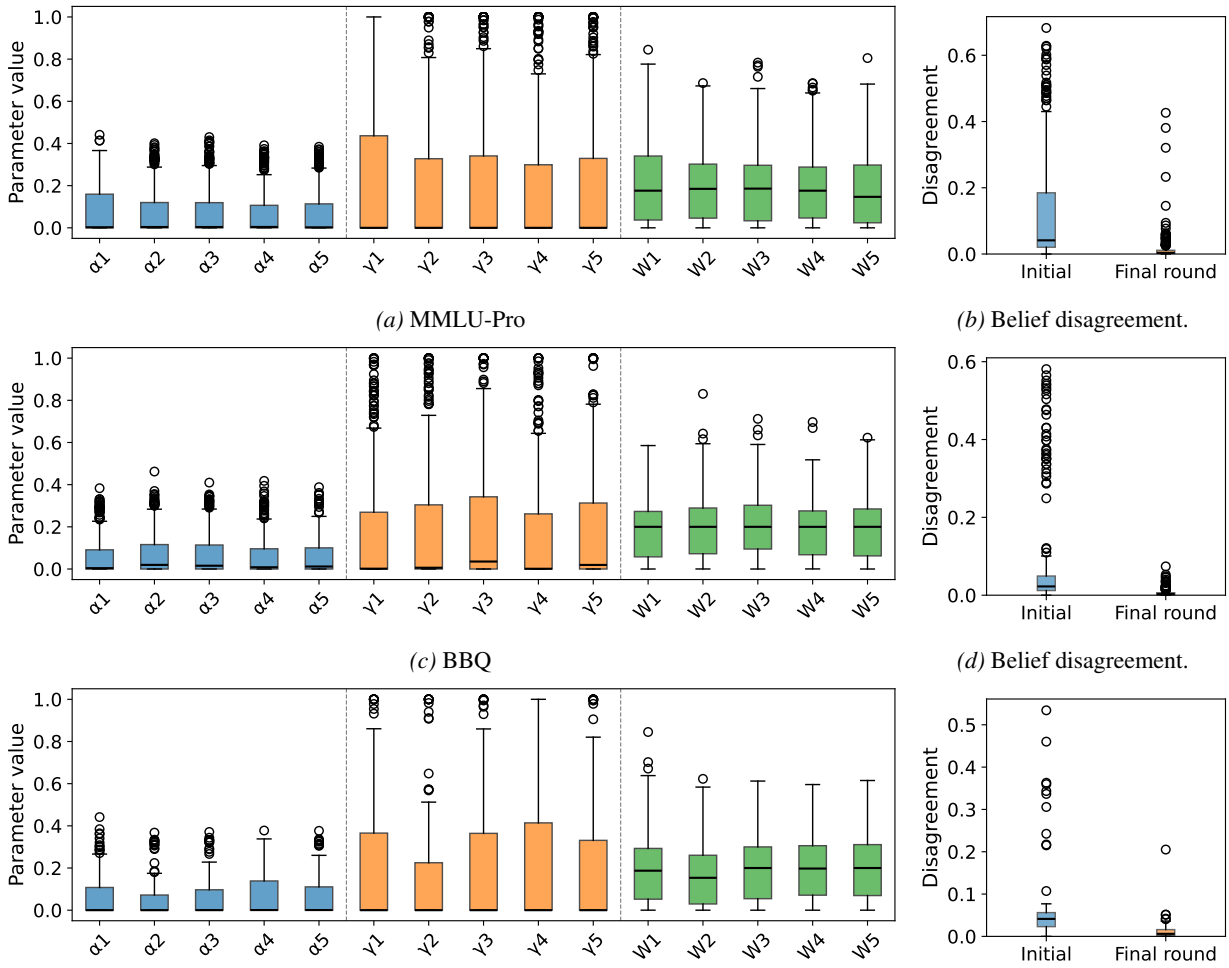
Figs. 13, 14, and 15 show the impact of different diversifying prompts on the peer influence of the agents, grouped by question (super-)category on dataset MMLU-Pro. The influence of an agent seems to be driven also by perceived confidence, and roles seem to have a relatively big effect that is consistent over different models.

Fig. 16 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. 17. 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).

Figs. 20 and 21 show the number of cases where Eq. 3 of Theorem 2.4 is satisfied for individual samples, compared to when it does not, and whether the MAS outperforms the best single agent. Theorem 2.4 states that, if Eq.3 is satisfied, the MAS should outperform the best single agent, so large values in the upper left and lower right entries of the confusion matrix point to Theorem 2.4 holding. We can see, that in most cases, the equation is satisfied, and routing wins for *GPT-5.4 Mini*. For *Qwen2.5-72B-Instruct-GPTQ-Int8* the MAS loses more often, but in most of these cases Eq. 3 is not satisfied.

Figs. 22 and 23 show weight heatmaps for different samples of the MMLU-Pro dataset (with communication style prompts). We can see that the weight matrices are task dependent.



(a) MMLU-Pro

(b) Belief disagreement.

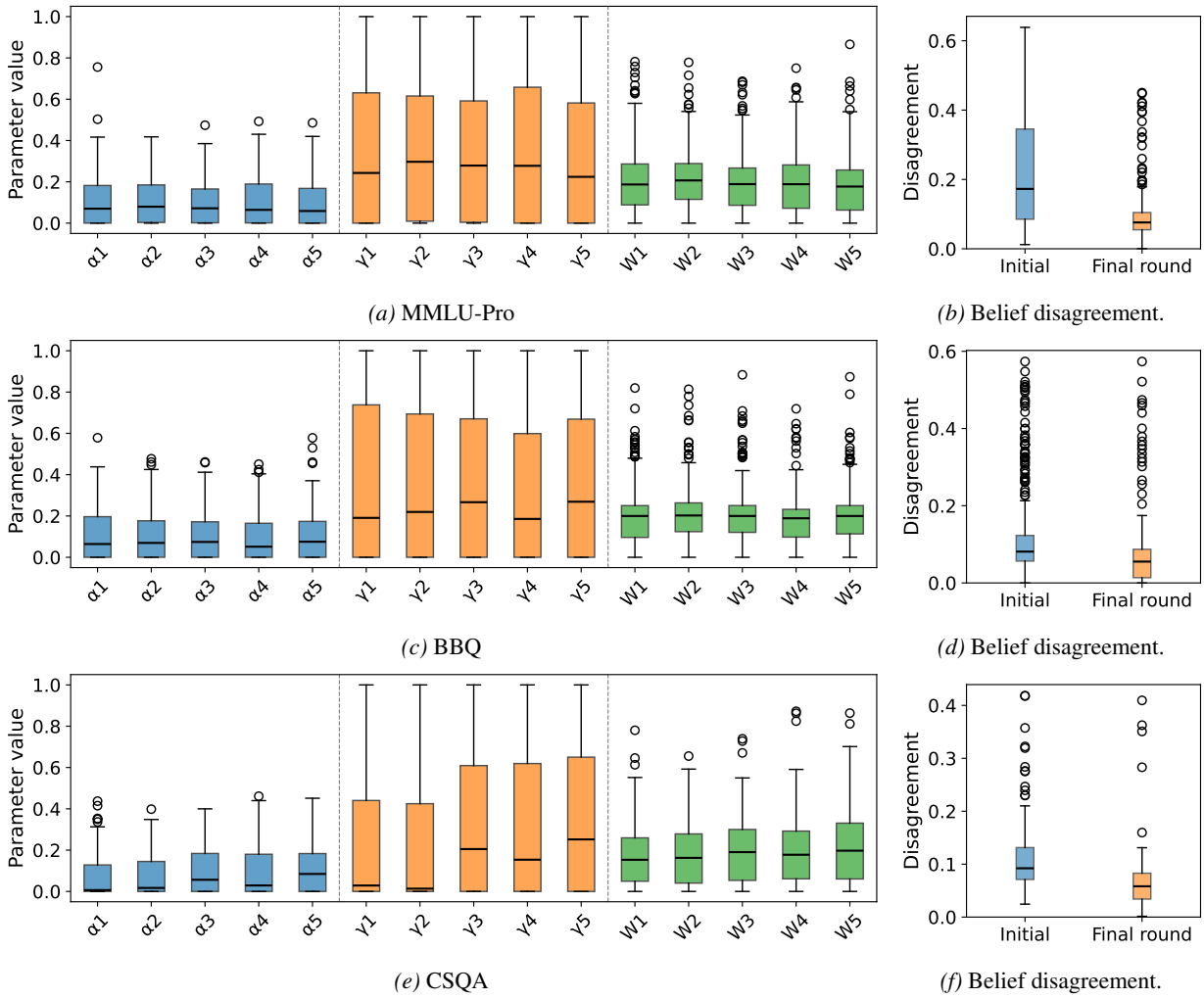
(c) BBQ

(d) Belief disagreement.

(e) CSQA

(f) Belief disagreement.

Figure 6. 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).

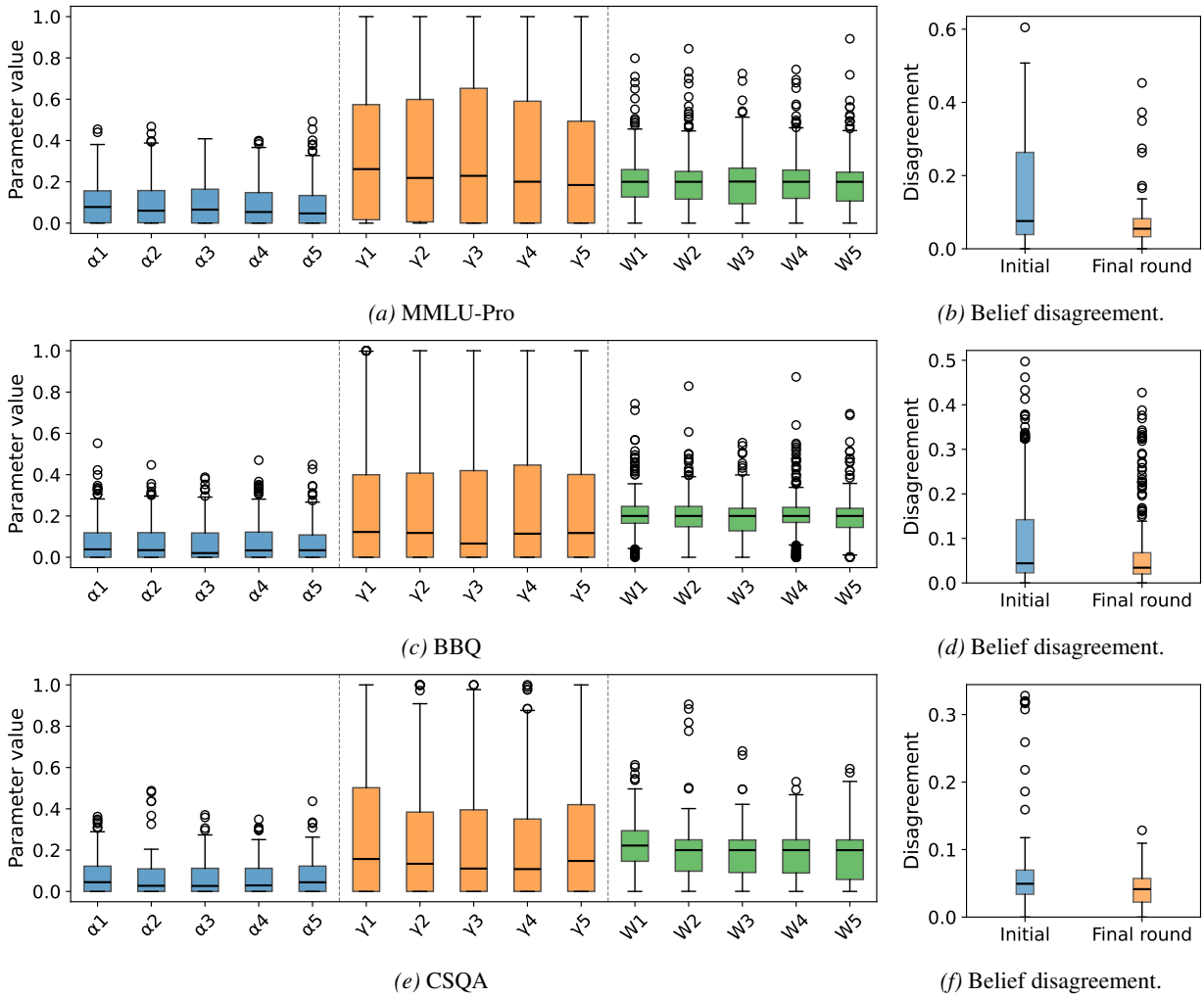


(a) MMLU-Pro *(b)* Belief disagreement.

(c) BBQ *(d)* Belief disagreement.

(e) CSQA *(f)* Belief disagreement.

Figure 7. 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).

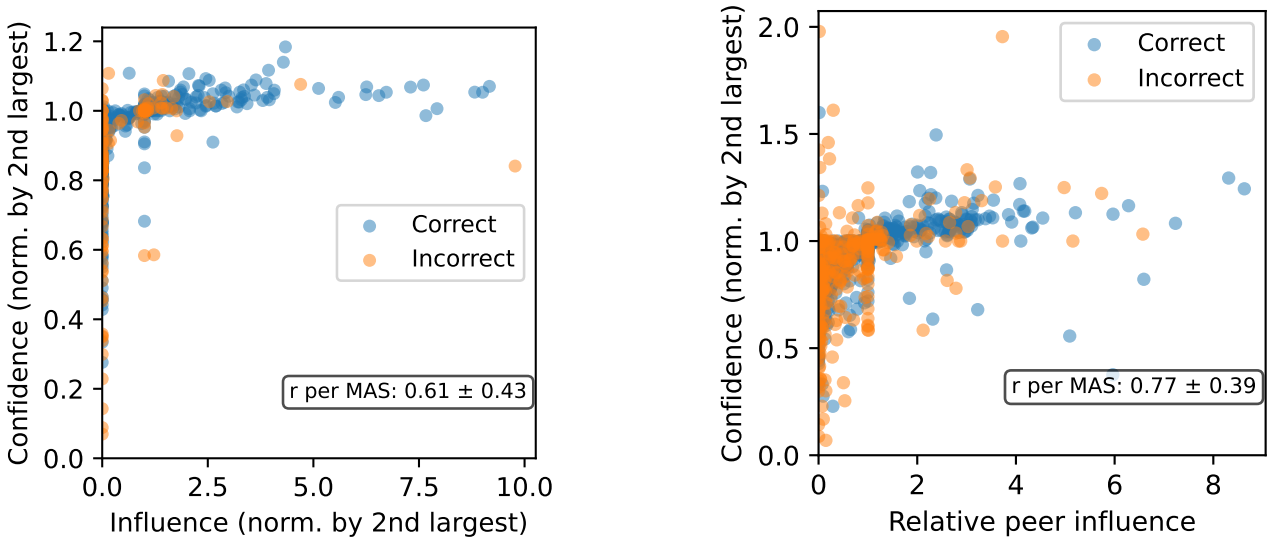


(a) MMLU-Pro *(b)* Belief disagreement.

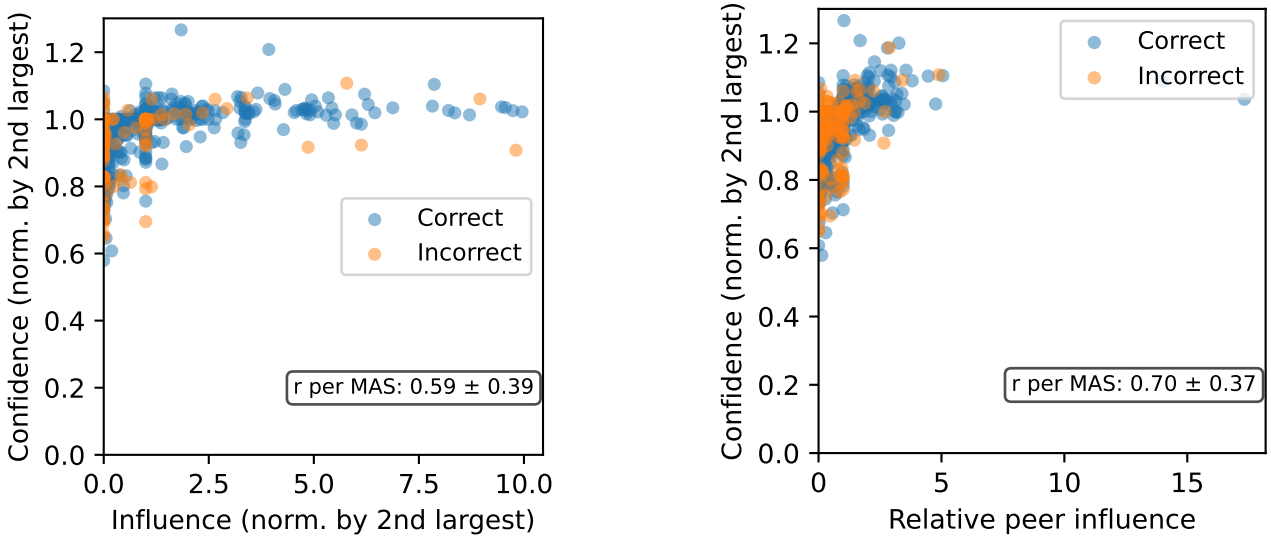
(c) BBQ *(d)* Belief disagreement.

(e) CSQA *(f)* Belief disagreement.

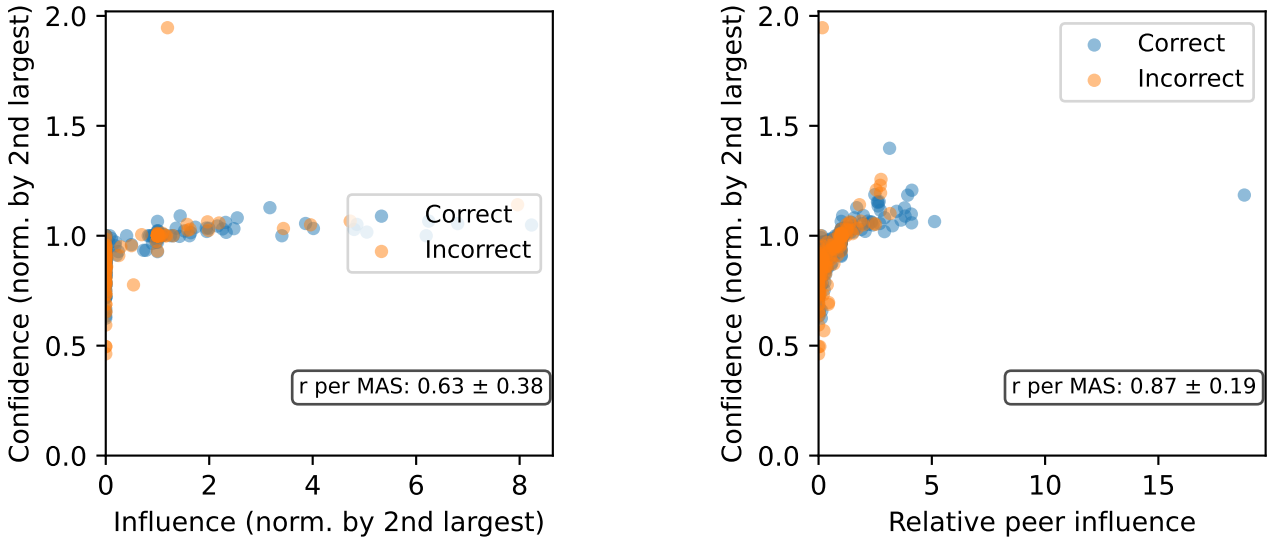
Figure 8. Parameter variability over all samples no special prompts for *Qwen2.5-72B-Instruct-GPTQ-Int8*, receiving weights (W1-5) are averaged over senders (left). Tendency towards consensus for the corresponding dataset (right).



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

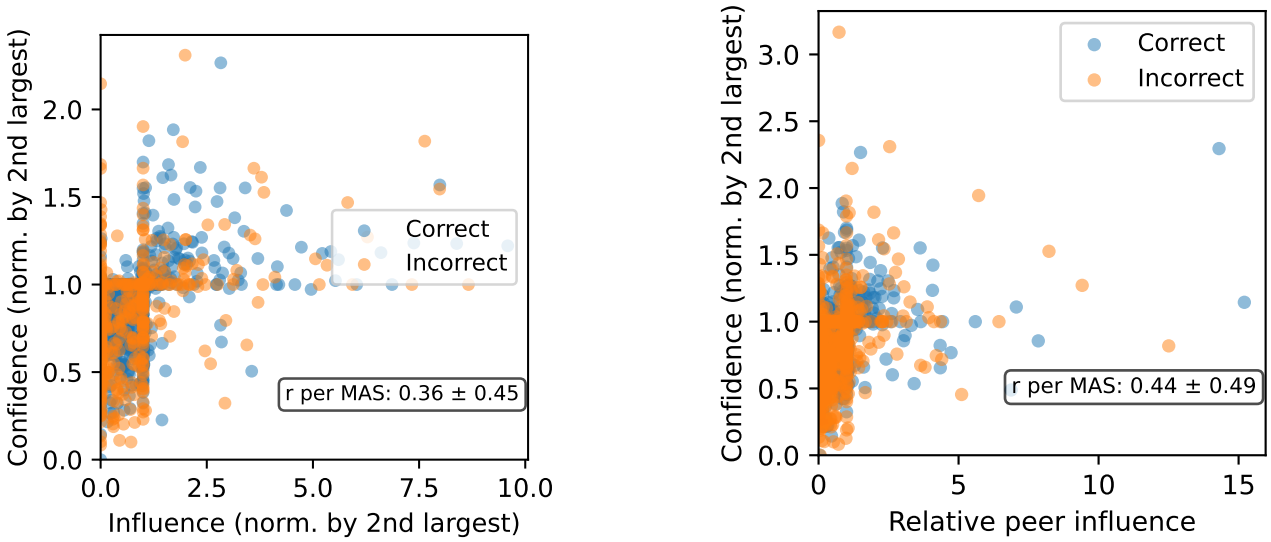


(b) BBQ: Confidence vs. (Peer-)Influence

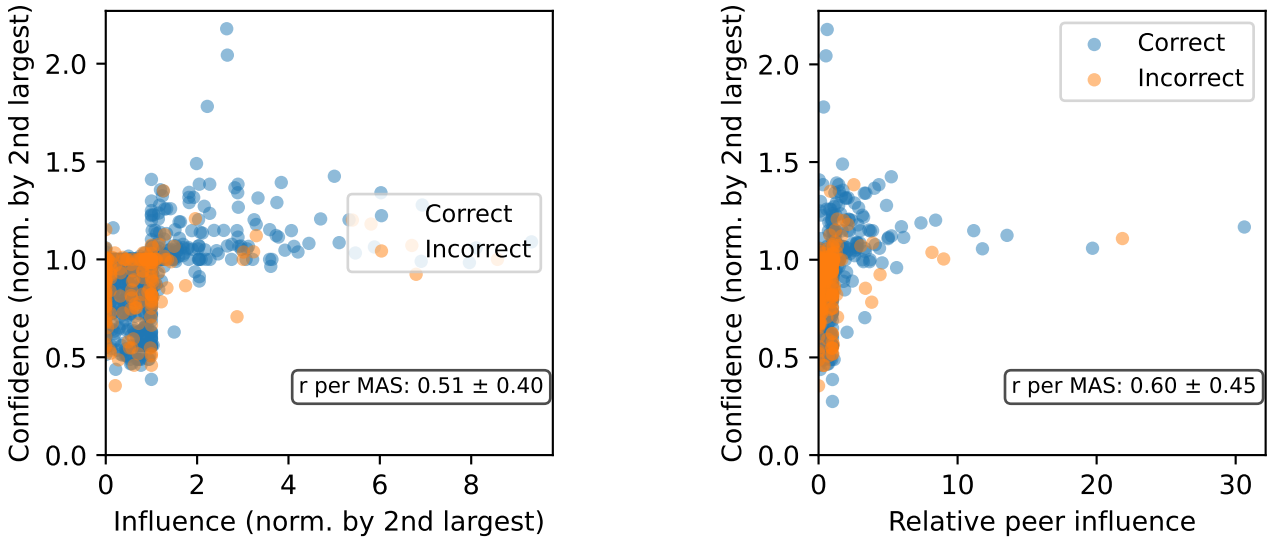


(c) CSQA: Confidence vs. (Peer-)Influence

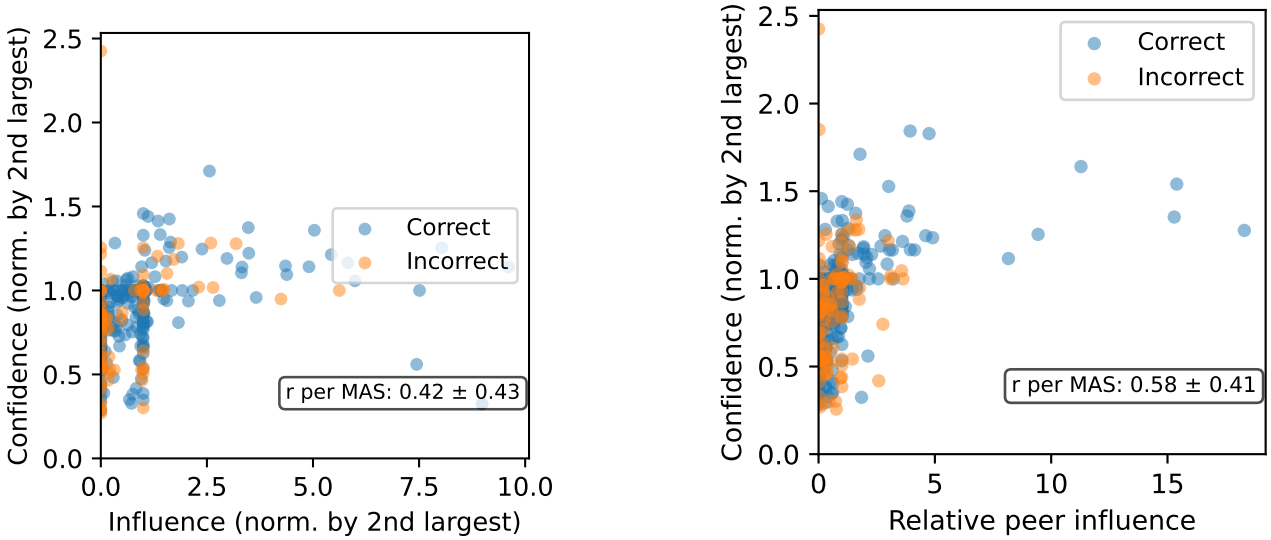
Figure 9. Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *GPT-5.4 Mini*.



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

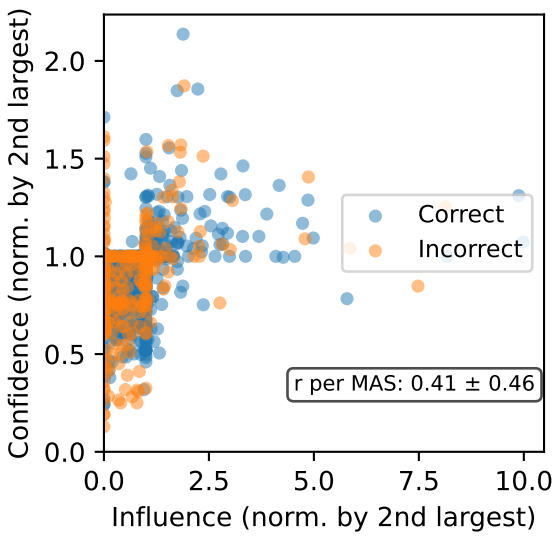


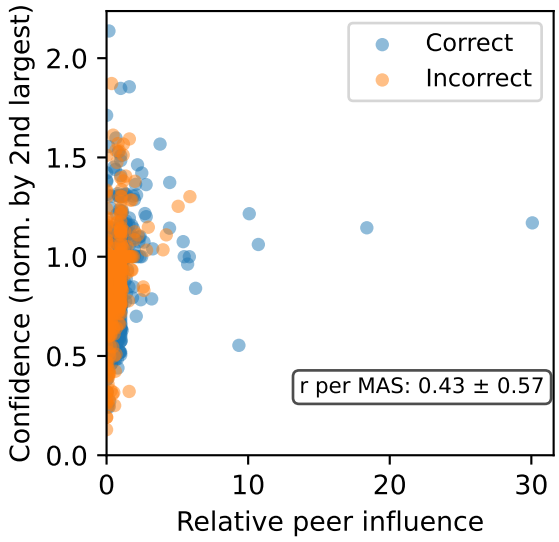
(b) BBQ: Confidence vs. (Peer-)Influence



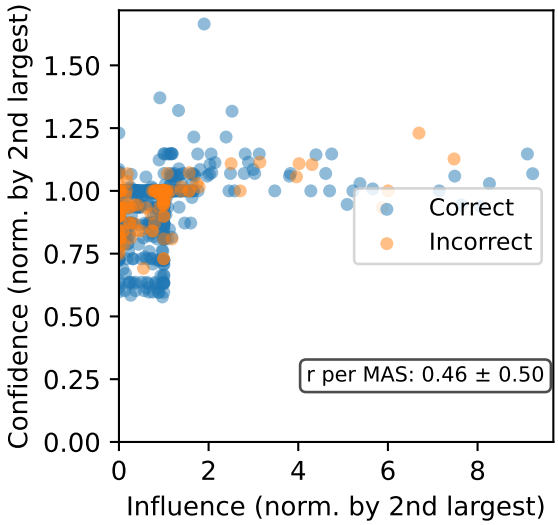
(c) CSQA: Confidence vs. (Peer-)Influence

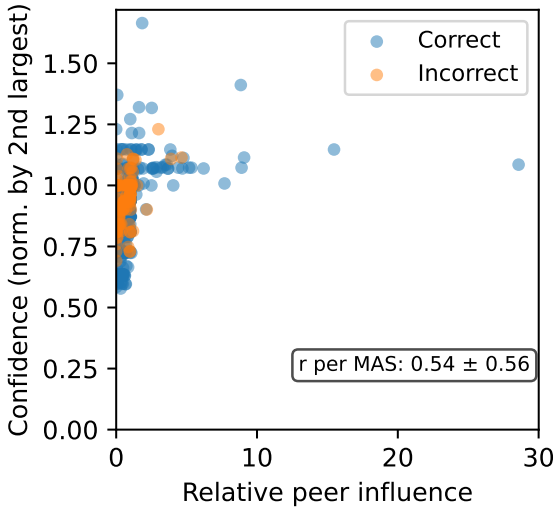
Figure 10. Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *Qwen2.5-14B-Instruct*.



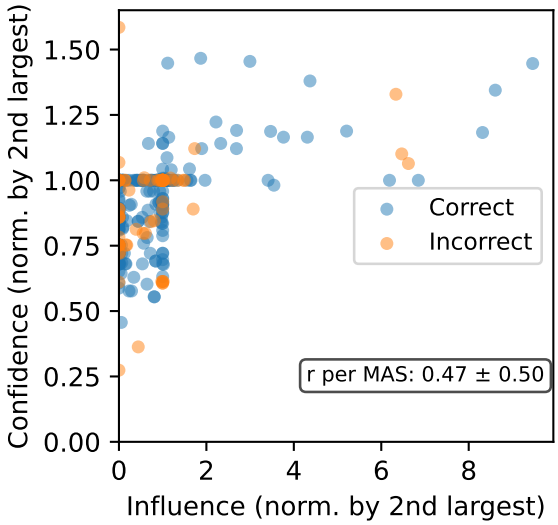


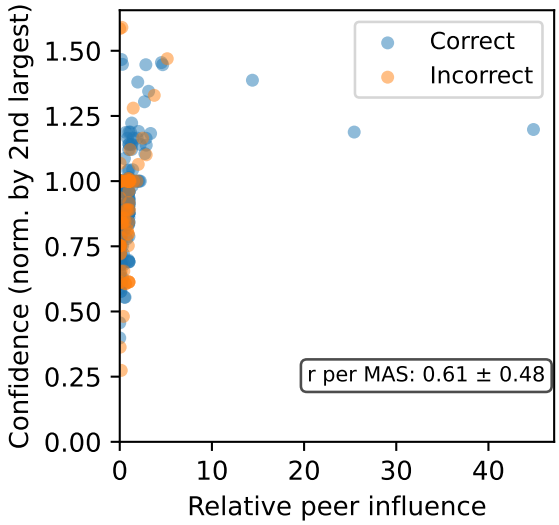
(a) MMLU-Pro: Confidence vs. (Peer-)Influence





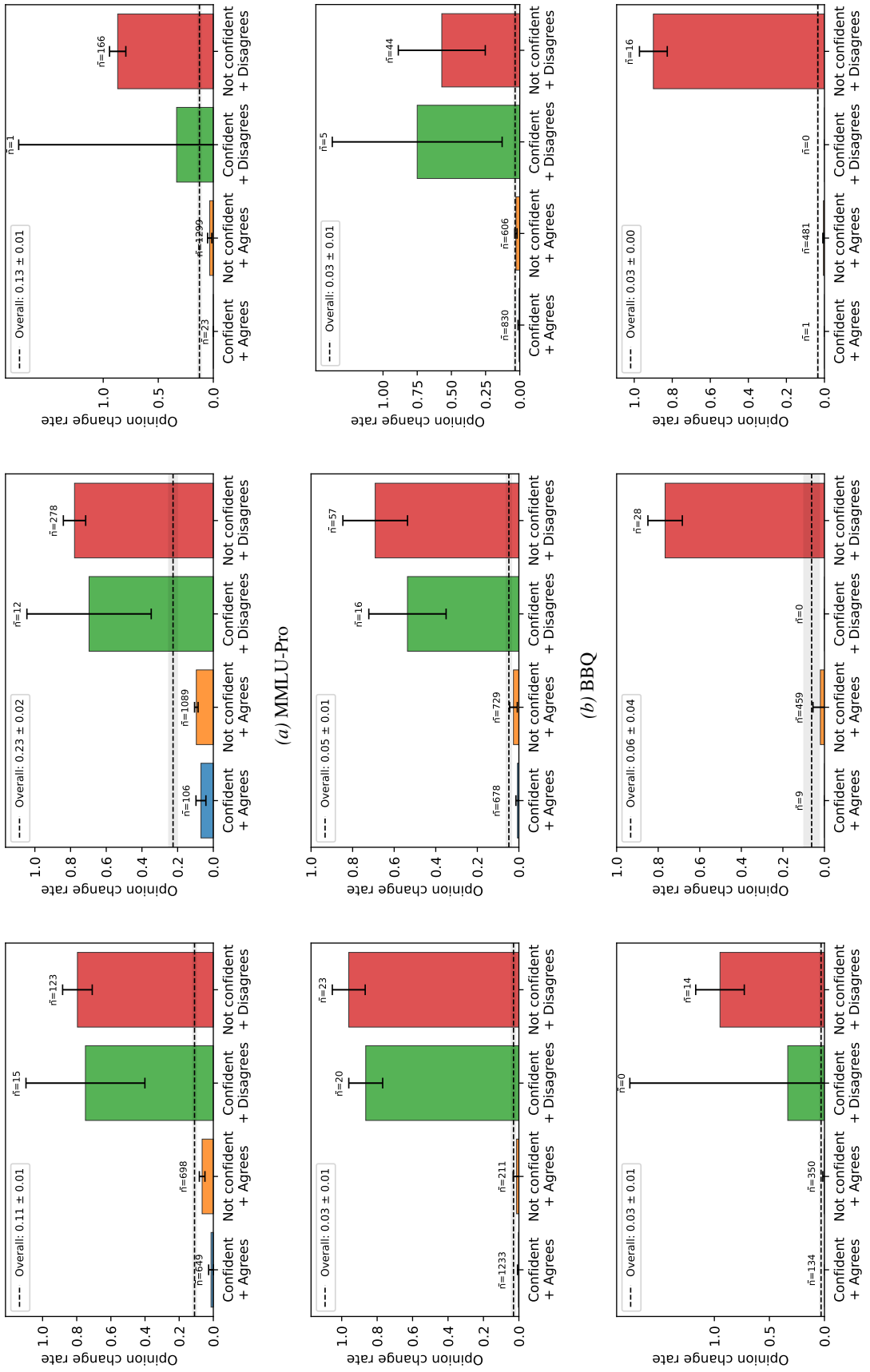
(b) BBQ: Confidence vs. (Peer-)Influence





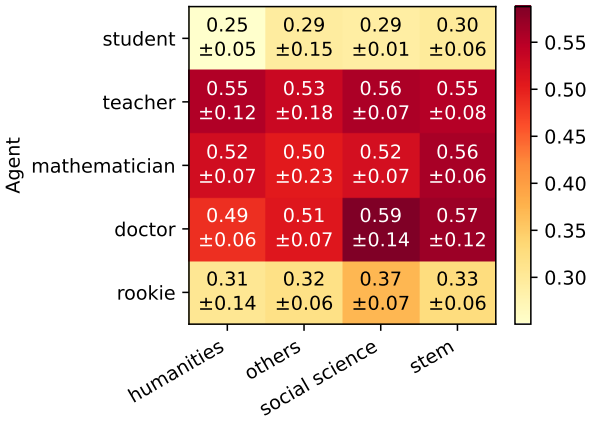
(c) CSQA: Confidence vs. (Peer-)Influence

Figure 11. Confidence vs. Influence (left) and confidence vs. peer influence (right) with no special prompts for *Qwen2.5-72B-Instruct-GPTQ-Int8*.

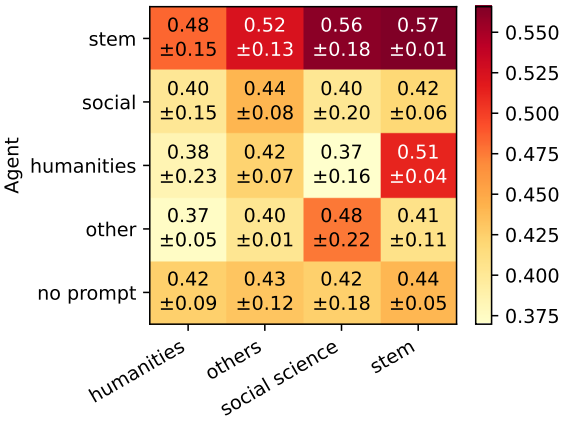


(a) MMLU-Pro (b) BBQ (c) CSQA

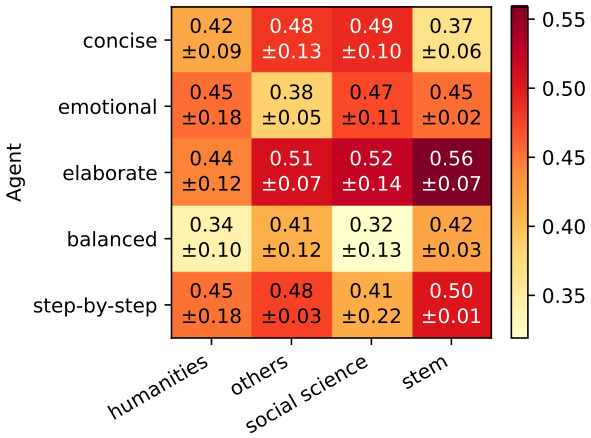
Figure 12. Probability of changing answer to the majority answer for cases where the agent is confident, $p(\text{change}|\text{confident} >= 0.75, \text{agree_with_mean} = True)$, $\bar{n}$ indicates average sample size over seeds rounded to nearest integer. With no special prompts for *GPT-5.4 Mini* (left), *Qwen2.5-14B-Instruct* (middle), and *Qwen2.5-72B-Instruct-GPTQ-Int8* (right).



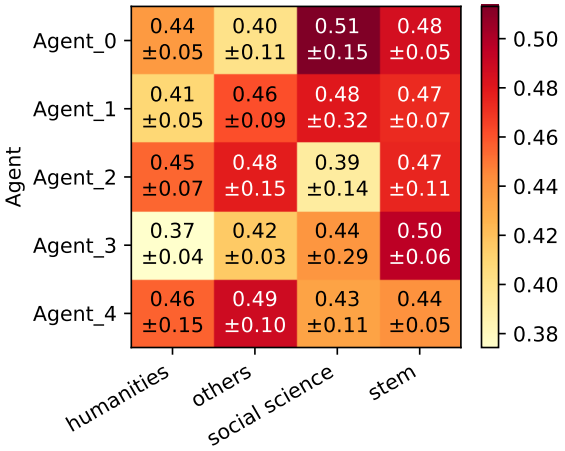
(a) Peer influence for different roles.



(b) Peer influence for different experts.

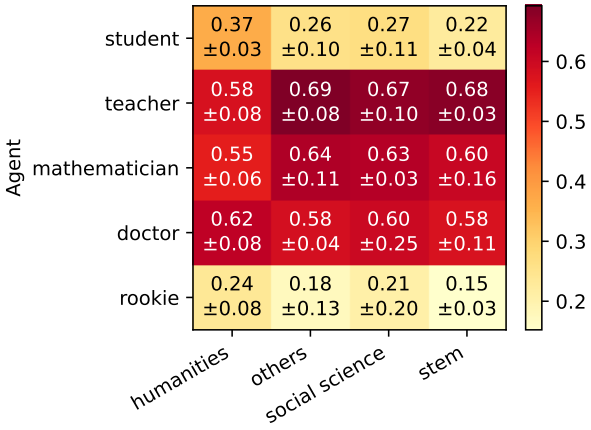


(c) Peer influence for different communication styles.

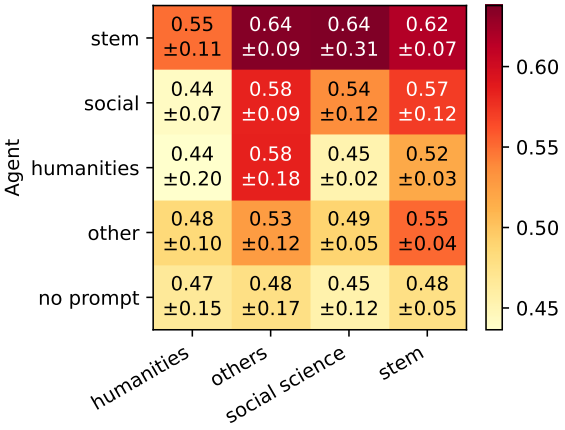


(d) Peer influence without special prompts.

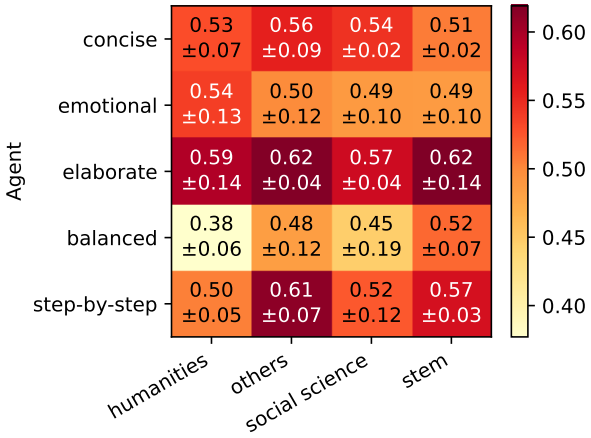
Figure 13. Results for *GPT-5.4 Mini* on the MMLU-Pro dataset. Roles seem to have an the biggest influence on perceived confidence as seen by the weights agents receive from the others.



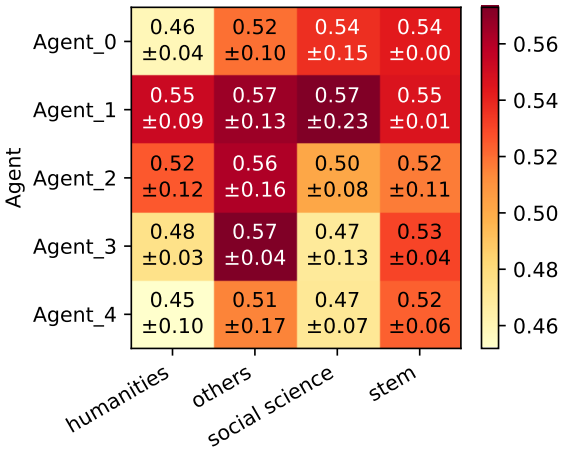
(a) Peer influence for different roles.



(b) Peer influence for different experts.

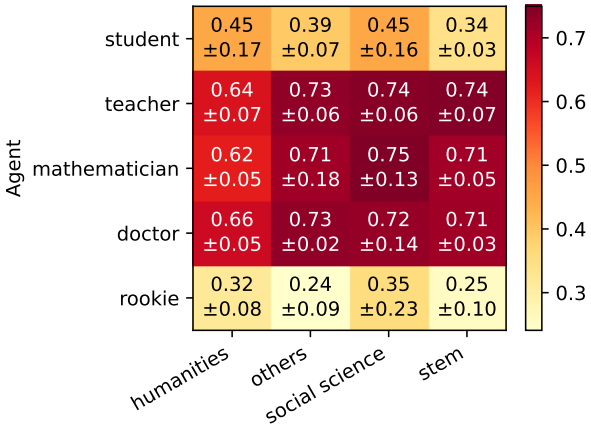


(c) Peer influence for different communication styles.

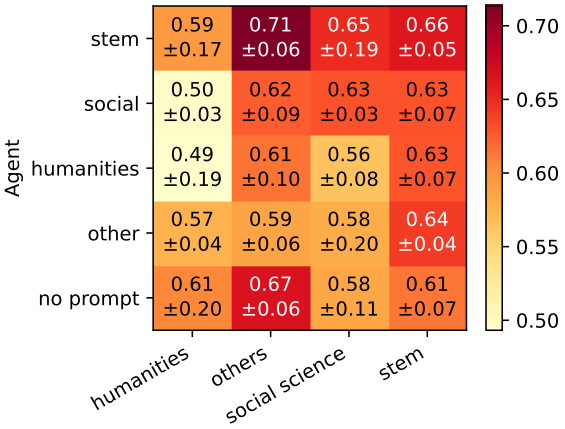


(d) Peer influence without special prompts.

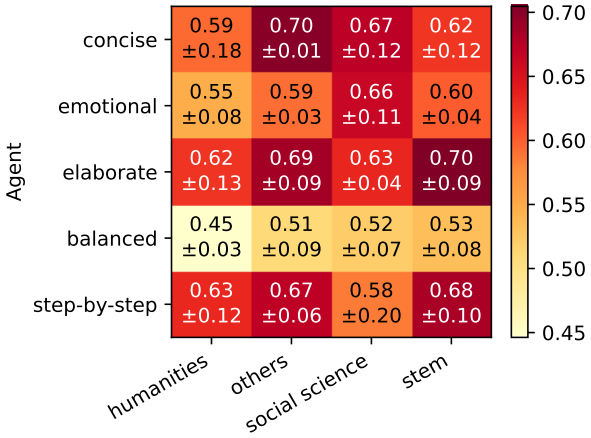
Figure 14. Results for *Qwen2.5-14B-Instruct* on the MMLU-Pro dataset. The same tendencies as for *GPT-5.4 Mini* can be seen.



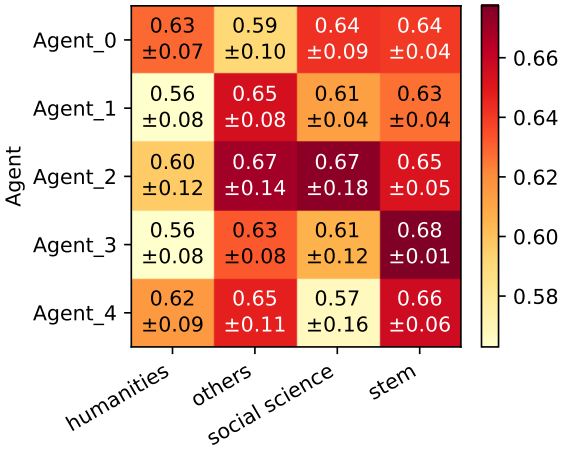
(a) Peer influence for different roles.



(b) Peer influence for different experts.



(c) Peer influence for different communication styles.



(d) Peer influence without special prompts.

Figure 15. 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.

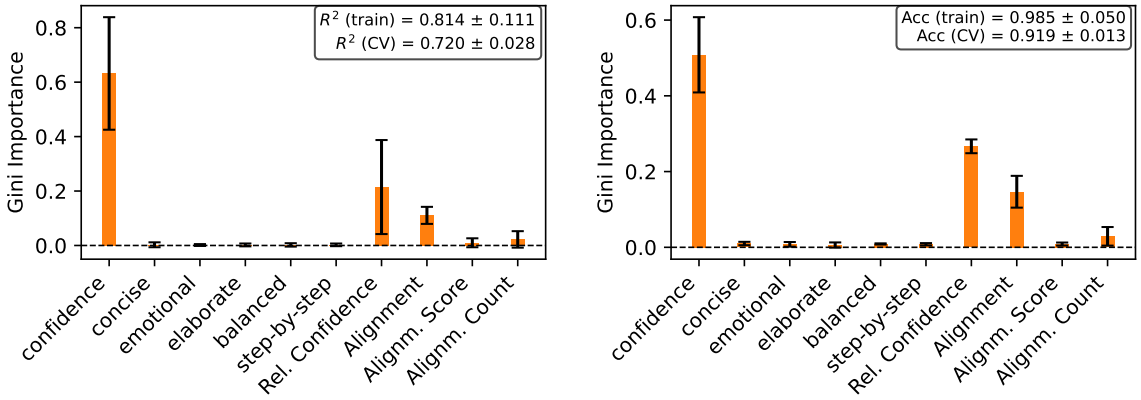
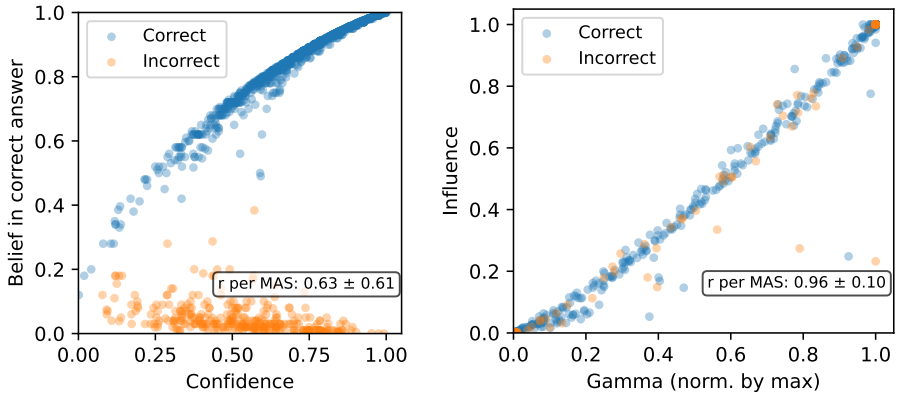


Figure 16. Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. Coefficients for a random forest regression predicting influence (a) and a random forest classifying the most influential agent in a MAS (b), with reported CV R^2 and Acc respectively. The high CV score, especially on the classification task (b) suggests the defined variables are highly predictive of the FJ dynamics.



(a) Confidence. *(b)* Influence vs. γ.

Figure 17. Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. When relating confidence and relative confidence to competence (a), we see a clear trend: confidence agents are more competent. We report Spearman's ranked correlation per MAS. (b) An agents influence is related to its stubbornness γ.

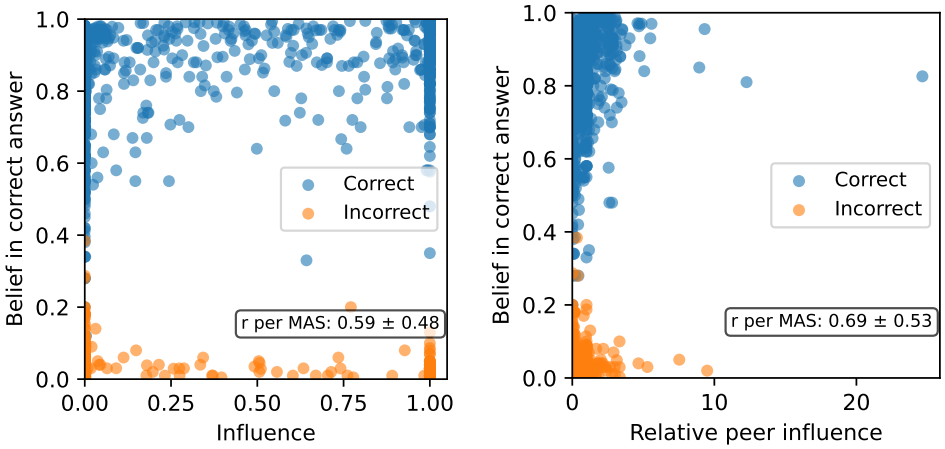


Figure 18. Results for the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts. More competent agents gain more influence (a) and relative influence (b), we again report Spearman's ranked correlation per MAS.

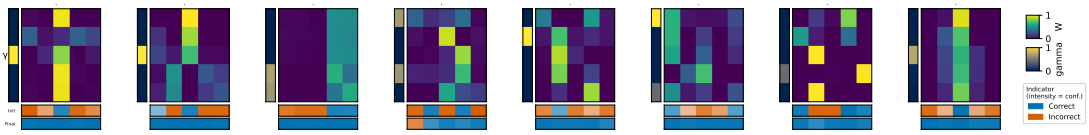
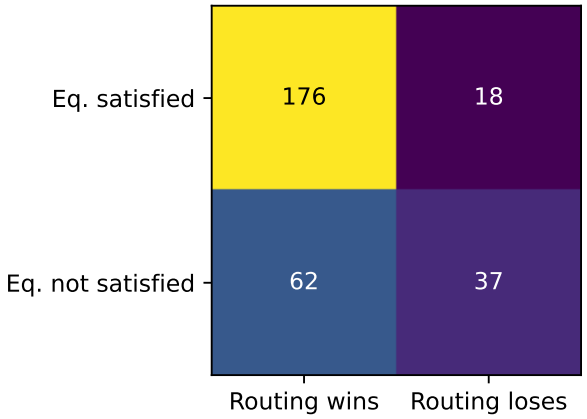
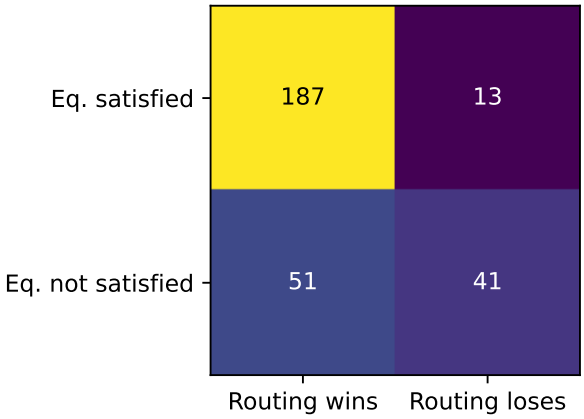


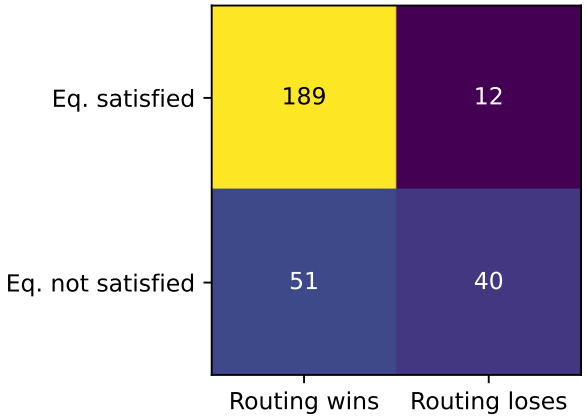
Figure 19. Exemplary questions of MMLU-Pro with communication style prompts with *GPT-5.4 Mini*, where the MAS outperforms the optimal ensemble. Weight heatmaps with indicators whether an agent was right (blue) or wrong (orange) initially (top) and in the final round (bottom), showing the γ (stubbornness) of each agent on the side (yellow being highest).



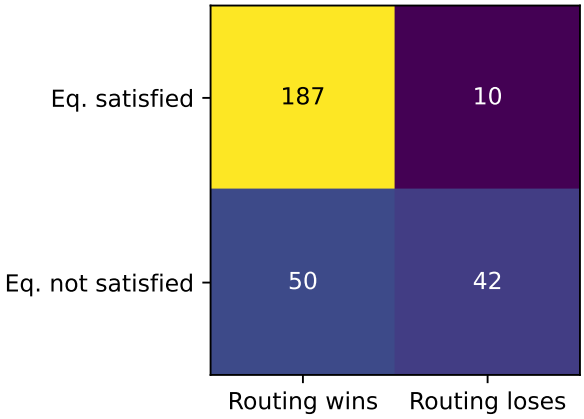
(a) Roles



(b) Communication styles

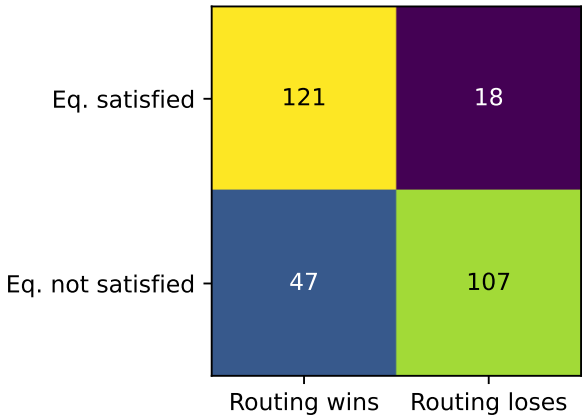


(c) Experts

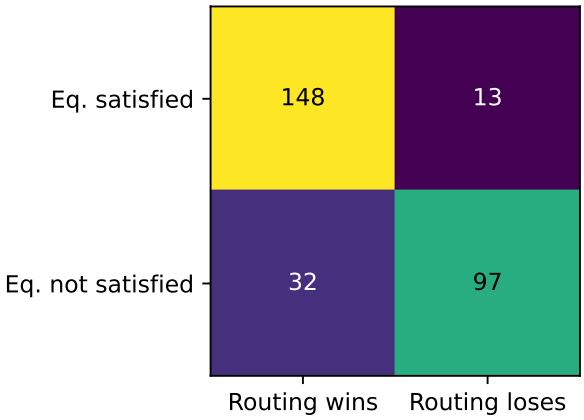


(d) Neutral

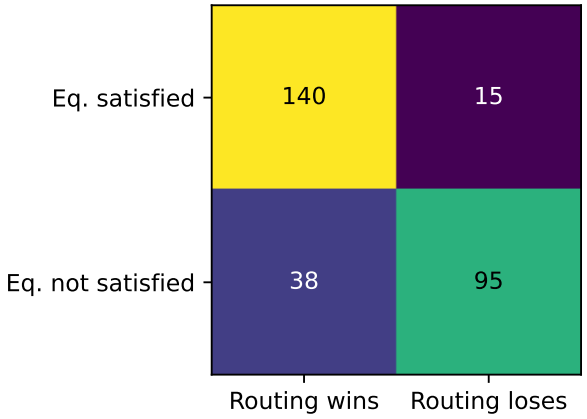
Figure 20. Confusion matrices showing empirical result of whether per sample version of Eq. 3 holds (Theorem 2.4) for different prompt styles of GPT on MMLU-Pro, averaged over seeds. On average, MoE routing outperforms a single agent if Eq. 3 holds.



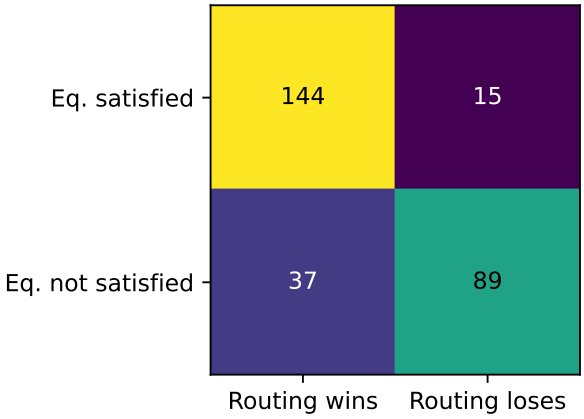
(a) Roles



(b) Communication styles



(c) Experts



(d) Neutral

Figure 21. Confusion matrices showing empirical result of whether per sample version of Eq. 3 holds (Theorem 2.4) for different prompt styles of *Qwen2.5-72B-Instruct-GPTQ-Int8* on MMLU-Pro, averaged over seeds. On average, MoE routing outperforms a single agent if Eq. 3 holds.

Table 2. Model fit (KL divergence and MSE) across datasets, prompt types, and models.

Dataset	Prompt	Model	KL	MSE
MMLU-Pro	neutral	*GPT-5.4 Mini*	0.0829 ± 0.0147	0.00247 ± 0.00097
		Qwen2.5-14B-Instruct	0.0503 ± 0.0013	0.00201 ± 0.00004
		Qwen2.5-72B-Instruct	0.0282 ± 0.0040	0.000972 ± 0.000176
	experts	*GPT-5.4 Mini*	0.0714 ± 0.0104	0.00206 ± 0.00040
		Qwen2.5-14B-Instruct	0.0467 ± 0.0044	0.00182 ± 0.00032
		Qwen2.5-72B-Instruct	0.0268 ± 0.0068	0.000935 ± 0.000226
	comm. styles	*GPT-5.4 Mini*	0.0740 ± 0.0096	0.00219 ± 0.00052
		Qwen2.5-14B-Instruct	0.0505 ± 0.0148	0.00200 ± 0.00067
		Qwen2.5-72B-Instruct	0.0263 ± 0.0077	0.000859 ± 0.000326
	roles	*GPT-5.4 Mini*	0.0786 ± 0.0212	0.00227 ± 0.00090
		Qwen2.5-14B-Instruct	0.0575 ± 0.0061	0.00225 ± 0.00032
		Qwen2.5-72B-Instruct	0.0320 ± 0.0117	0.00114 ± 0.00049
BBQ	neutral	*GPT-5.4 Mini*	0.0420 ± 0.0093	0.000373 ± 0.000100
		Qwen2.5-14B-Instruct	0.0403 ± 0.0043	0.00117 ± 0.00006
		Qwen2.5-72B-Instruct	0.0238 ± 0.0005	0.000876 ± 0.000135
	experts	*GPT-5.4 Mini*	0.0423 ± 0.0075	0.000324 ± 0.000091
		Qwen2.5-14B-Instruct	0.0541 ± 0.0076	0.00151 ± 0.00053
		Qwen2.5-72B-Instruct	0.0239 ± 0.0152	0.000963 ± 0.000827
	comm. styles	*GPT-5.4 Mini*	0.0459 ± 0.0052	0.000436 ± 0.000063
		Qwen2.5-14B-Instruct	0.0354 ± 0.0039	0.00105 ± 0.00011
		Qwen2.5-72B-Instruct	0.0221 ± 0.0016	0.000837 ± 0.000125
	roles	*GPT-5.4 Mini*	0.0479 ± 0.0039	0.000422 ± 0.000135
		Qwen2.5-14B-Instruct	0.0508 ± 0.0017	0.00169 ± 0.00031
		Qwen2.5-72B-Instruct	0.0214 ± 0.0057	0.000817 ± 0.000305
CSQA	neutral	*GPT-5.4 Mini*	0.0513 ± 0.0040	0.00247 ± 0.00019
		Qwen2.5-14B-Instruct	0.0754 ± 0.0146	0.00570 ± 0.00197
		Qwen2.5-72B-Instruct	0.0364 ± 0.0034	0.00264 ± 0.00021
	experts	*GPT-5.4 Mini*	0.0517 ± 0.0015	0.00256 ± 0.00027
		Qwen2.5-14B-Instruct	0.0599 ± 0.0106	0.00414 ± 0.00110
		Qwen2.5-72B-Instruct	0.0331 ± 0.0019	0.00238 ± 0.00004
	comm. styles	*GPT-5.4 Mini*	0.0516 ± 0.0042	0.00244 ± 0.00020
		Qwen2.5-14B-Instruct	0.0647 ± 0.0086	0.00443 ± 0.00047
		Qwen2.5-72B-Instruct	0.0316 ± 0.0026	0.00222 ± 0.00026
	roles	*GPT-5.4 Mini*	0.0534 ± 0.0037	0.00272 ± 0.00032
		Qwen2.5-14B-Instruct	0.0770 ± 0.0121	0.00592 ± 0.00133
		Qwen2.5-72B-Instruct	0.0306 ± 0.0049	0.00223 ± 0.00063

Table 3. Accuracy of baseline, FJ ensemble, and agent system, across datasets, prompts, and models.

Dataset	Prompt	Model	Baseline	FJ Ens.	MAS
MMLU-Pro	neutral	*GPT-5.4 Mini*	0.804 ± 0.013	0.820 ± 0.013	$\mathbf{0.826 \pm 0.027}$
		Qwen2.5-14B-Instruct	0.569 ± 0.027	0.577 ± 0.034	$\mathbf{0.604 \pm 0.023}$
		Qwen2.5-72B-Instruct	0.695 ± 0.036	0.694 ± 0.075	$\mathbf{0.712 \pm 0.016}$
	experts	*GPT-5.4 Mini*	0.816 ± 0.005	0.819 ± 0.009	$\mathbf{0.835 \pm 0.008}$
		Qwen2.5-14B-Instruct	0.637 ± 0.019	0.633 ± 0.034	$\mathbf{0.640 \pm 0.016}$
		Qwen2.5-72B-Instruct	0.728 ± 0.043	0.729 ± 0.007	$\mathbf{0.742 \pm 0.037}$
	comm. styles	*GPT-5.4 Mini*	0.797 ± 0.012	0.809 ± 0.029	$\mathbf{0.824 \pm 0.029}$
		Qwen2.5-14B-Instruct	0.591 ± 0.023	$\mathbf{0.614 \pm 0.029}$	0.606 ± 0.016
		Qwen2.5-72B-Instruct	0.727 ± 0.019	0.736 ± 0.017	$\mathbf{0.743 \pm 0.015}$
	roles	*GPT-5.4 Mini*	0.805 ± 0.023	0.806 ± 0.029	$\mathbf{0.824 \pm 0.004}$
		Qwen2.5-14B-Instruct	0.557 ± 0.014	0.576 ± 0.018	$\mathbf{0.596 \pm 0.009}$
		Qwen2.5-72B-Instruct	0.700 ± 0.040	0.706 ± 0.045	$\mathbf{0.720 \pm 0.060}$
BBQ	neutral	*GPT-5.4 Mini*	0.916 ± 0.014	$\mathbf{0.916 \pm 0.028}$	0.916 ± 0.024
		Qwen2.5-14B-Instruct	0.866 ± 0.028	0.867 ± 0.015	$\mathbf{0.876 \pm 0.021}$
		Qwen2.5-72B-Instruct	0.882 ± 0.002	$\mathbf{0.888 \pm 0.016}$	0.875 ± 0.015
	experts	*GPT-5.4 Mini*	0.932 ± 0.017	0.932 ± 0.012	$\mathbf{0.934 \pm 0.012}$
		Qwen2.5-14B-Instruct	0.880 ± 0.017	$\mathbf{0.886 \pm 0.009}$	0.870 ± 0.012
		Qwen2.5-72B-Instruct	0.836 ± 0.015	$\mathbf{0.858 \pm 0.031}$	0.840 ± 0.006
	comm. styles	*GPT-5.4 Mini*	0.906 ± 0.025	0.903 ± 0.015	$\mathbf{0.916 \pm 0.014}$
		Qwen2.5-14B-Instruct	0.865 ± 0.005	0.882 ± 0.020	$\mathbf{0.887 \pm 0.018}$
		Qwen2.5-72B-Instruct	0.857 ± 0.020	$\mathbf{0.875 \pm 0.034}$	0.870 ± 0.025
	roles	*GPT-5.4 Mini*	0.904 ± 0.018	$\mathbf{0.918 \pm 0.018}$	0.915 ± 0.018
		Qwen2.5-14B-Instruct	0.874 ± 0.034	0.866 ± 0.026	$\mathbf{0.898 \pm 0.028}$
		Qwen2.5-72B-Instruct	0.860 ± 0.009	$\mathbf{0.887 \pm 0.013}$	0.882 ± 0.018
CSQA	neutral	*GPT-5.4 Mini*	$\mathbf{0.726 \pm 0.013}$	0.722 ± 0.010	0.722 ± 0.017
		Qwen2.5-14B-Instruct	0.775 ± 0.011	$\mathbf{0.782 \pm 0.013}$	0.782 ± 0.018
		Qwen2.5-72B-Instruct	$\mathbf{0.803 \pm 0.012}$	0.799 ± 0.028	0.793 ± 0.011
	experts	*GPT-5.4 Mini*	0.723 ± 0.014	$\mathbf{0.733 \pm 0.014}$	0.730 ± 0.025
		Qwen2.5-14B-Instruct	0.787 ± 0.014	$\mathbf{0.803 \pm 0.014}$	0.797 ± 0.029
		Qwen2.5-72B-Instruct	0.806 ± 0.017	$\mathbf{0.816 \pm 0.013}$	0.803 ± 0.016
	comm. styles	*GPT-5.4 Mini*	0.730 ± 0.025	$\mathbf{0.733 \pm 0.052}$	0.730 ± 0.025
		Qwen2.5-14B-Instruct	0.777 ± 0.014	$\mathbf{0.787 \pm 0.038}$	0.763 ± 0.057
		Qwen2.5-72B-Instruct	0.799 ± 0.047	0.806 ± 0.013	$\mathbf{0.806 \pm 0.036}$
	roles	*GPT-5.4 Mini*	0.727 ± 0.029	$\mathbf{0.737 \pm 0.014}$	0.720 ± 0.000
		Qwen2.5-14B-Instruct	0.777 ± 0.014	0.793 ± 0.052	$\mathbf{0.803 \pm 0.014}$
		Qwen2.5-72B-Instruct	$\mathbf{0.809 \pm 0.027}$	0.806 ± 0.030	0.799 ± 0.003

D. Computing resources required for reproducing

Experiments for Qwen2.5-14b-Instruct and Qwen2.5-72B-Instruct-GPTQ-Int8 require A100 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 A100 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.

E. 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.

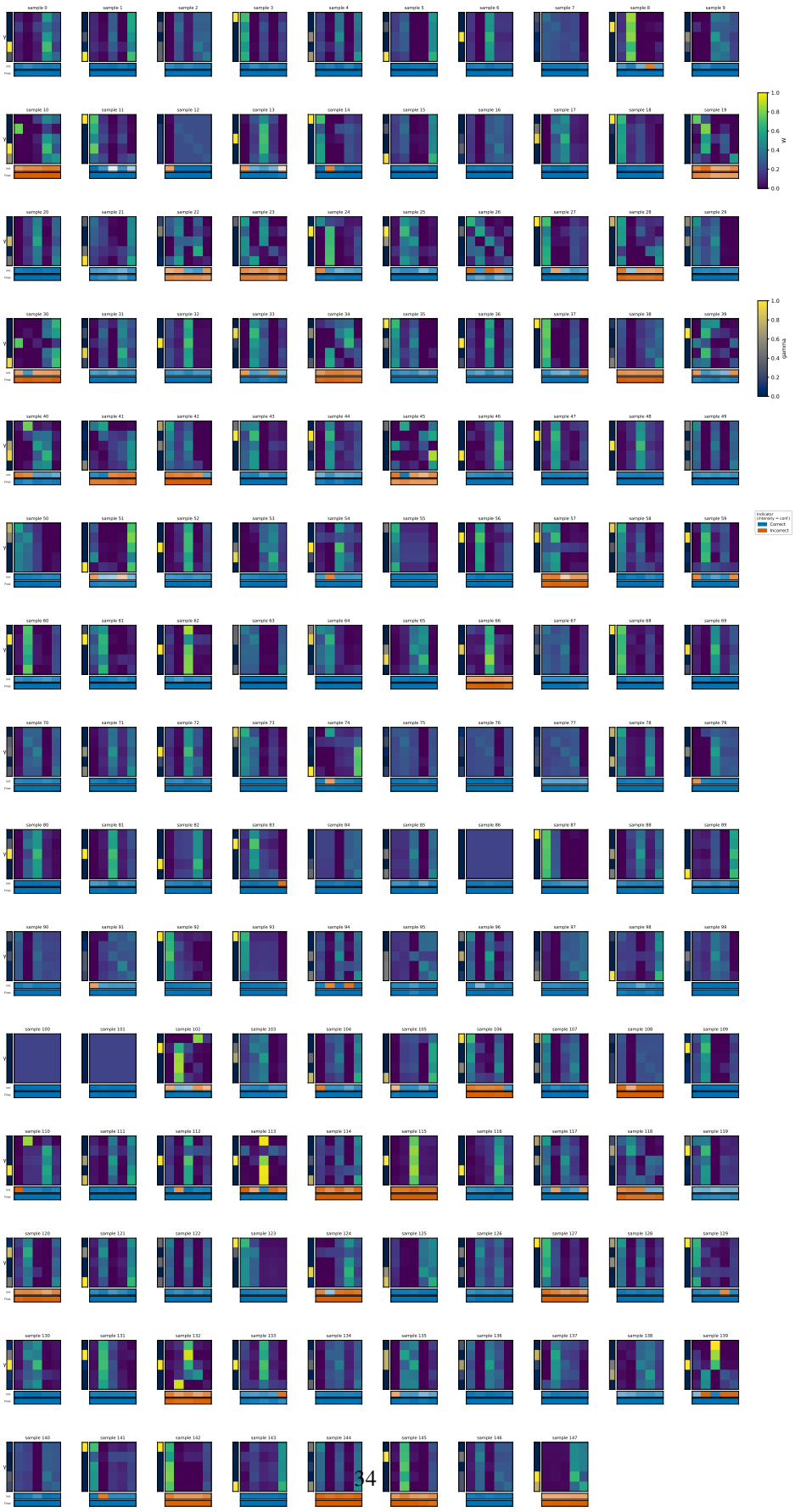


Figure 22. Weight heatmaps for samples from the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts, with

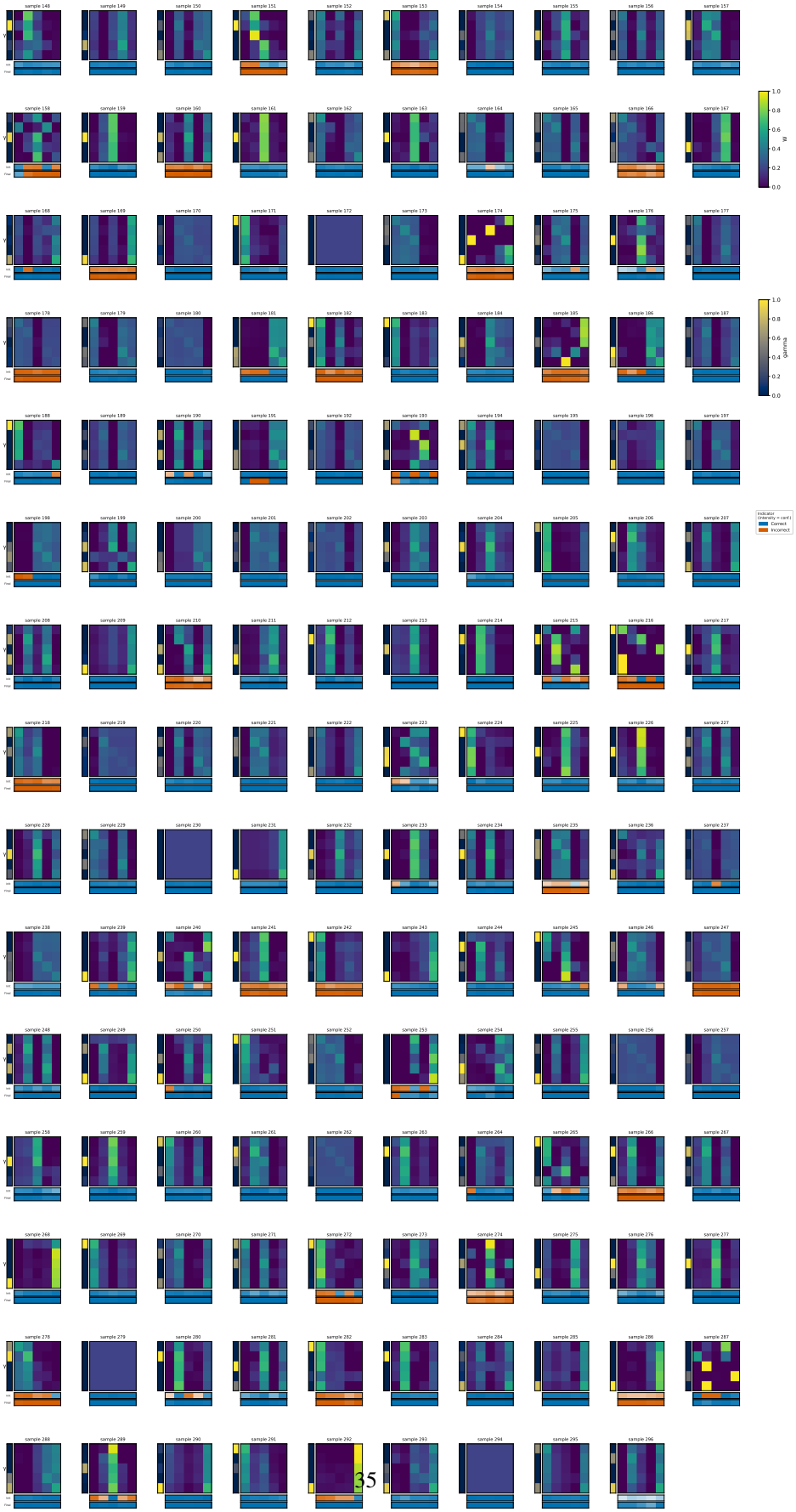


Figure 23. Weight heatmaps for samples from the MMLU-Pro dataset with *GPT-5.4 Mini* and communication style prompts, with

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