Flag Game: Interpreting Decision Mechanisms of Bounded Social Agents

Elizabeth Pavlova [1 2 3] **Hidenori Tanaka** [1 2 3]

Abstract

Every bounded agent in a multi-agent system must balance its private evidence against social input from peers. We make this balance experimentally observable with the Flag Game, a grounded synthetic task in which a hidden country flag defines verifiable ground truth, each agent observes only a private crop, agents communicate under a specified protocol, and the system outputs a final country distribution. Despite its simplicity, the task reproduces collective phenomena central to deliberative settings—non-monotonic population scaling, gains from social-awareness prompting and model diversity, and polarization—while remaining diagnosable at the mechanism level. We explain these phenomena with a unifying framework, building on Quantized Simplex Gossip (QSG), that traces them to two facets of the same private–social tension: how much private evidence the population collectively holds and how each agent integrates social input. Methodologically, the Flag Game extends the toy-model strategy of mechanistic interpretability to multi-agent systems, offering a controlled setting in which rich phenomenology can be dissected before moving to open-ended domains where truth is harder to verify and failure harder to interpret.

1. Introduction

Many consequential judgments are made by groups of bounded observers rather than by a single observer with complete access to the world. In health care, diagnosis is collaborative across clinicians, patients, tests, and imaging (National Academies of Sciences, Engineering, and Medicine, 2015). In law, fact-finders combine testimony, physical evidence, and witness narratives before reaching a verdict (Hastie et al., 1983; Bornstein & Greene, 2011; Judiciary of England and Wales, 2025). In democracies, the case for collective self-government rests on citizens pooling partial information better than any individual, yet social influence can undermine the wisdom of crowds and move the group away from truth (Surowiecki, 2004; Lorenz et al., 2011; Becker et al., 2017). Sensor networks face the same integration problem under explicit communication constraints (Varshney, 1996; Boyd et al., 2006). Recent multi-agent AI systems use multiple agents to exchange messages, assume roles, and deliberate under information asymmetry (Du et al., 2024; Smit et al., 2024; Kaesberg et al., 2025; Zhang et al., 2024; Tang et al., 2024; Kim et al., 2024; Jiang & Yang, 2025; He et al., 2024). These domains differ in surface semantics, but they share a common decision skeleton. Agents receive partial and sometimes ambiguous evidence, communicate through limited channels, and a final readout is scored, acted on, or trusted. Throughout this paper we use bounded to mean agents who see only part of the world and can transmit only part of what they see; their perception, memory, and outgoing messages are all essentially finite.

A key evaluation difficulty in such systems is that final accuracy alone does not identify the mechanism. Let us consider four ways a group can be wrong: (i) no agent ever observed the diagnostic evidence, (ii) some agent observed it but never communicated it, (iii) the diagnostic evidence was communicated but overridden by a louder rival, (iv) an early plausible, yet wrong interpretation attracted social agreement leading to lock-in. These failures call for very different interventions, such as better sampling, better protocols, or better aggregation. The same ambiguity holds for success. A correct answer may reflect genuine evidence pooling, a strong participant, or coincidental agreement.

Classical hidden-profile work shows that groups can fail to pool uniquely held information even when the correct answer is available in aggregate (Stasser & Titus, 1985; 2003). Social-learning and herding models show how private evidence can be overwritten by public signals (Banerjee, 1992; Bikhchandani et al., 1992; Lorenz et al., 2011). Crowdsourcing and label-aggregation work provide powerful non-social baselines, but can abstract away the interactions that occur before aggregation (Raykar et al., 2010). What is miss-

[1]Cambridge Boston Alignment Initiative (CBAI), Cambridge, US [2]CBS–NTT Program in Physics of Intelligence, Center for Brain Science, Harvard University [3]Physics of Artificial Intelligence Group, NTT Research, Inc.. Correspondence to: Elizabeth Pavlova <elizabethpavlova@fas.harvard.edu>, Hidenori Tanaka <hidenori_tanaka@fas.harvard.edu>.

Proceedings of the Trustworthy AI for Good Workshop at 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

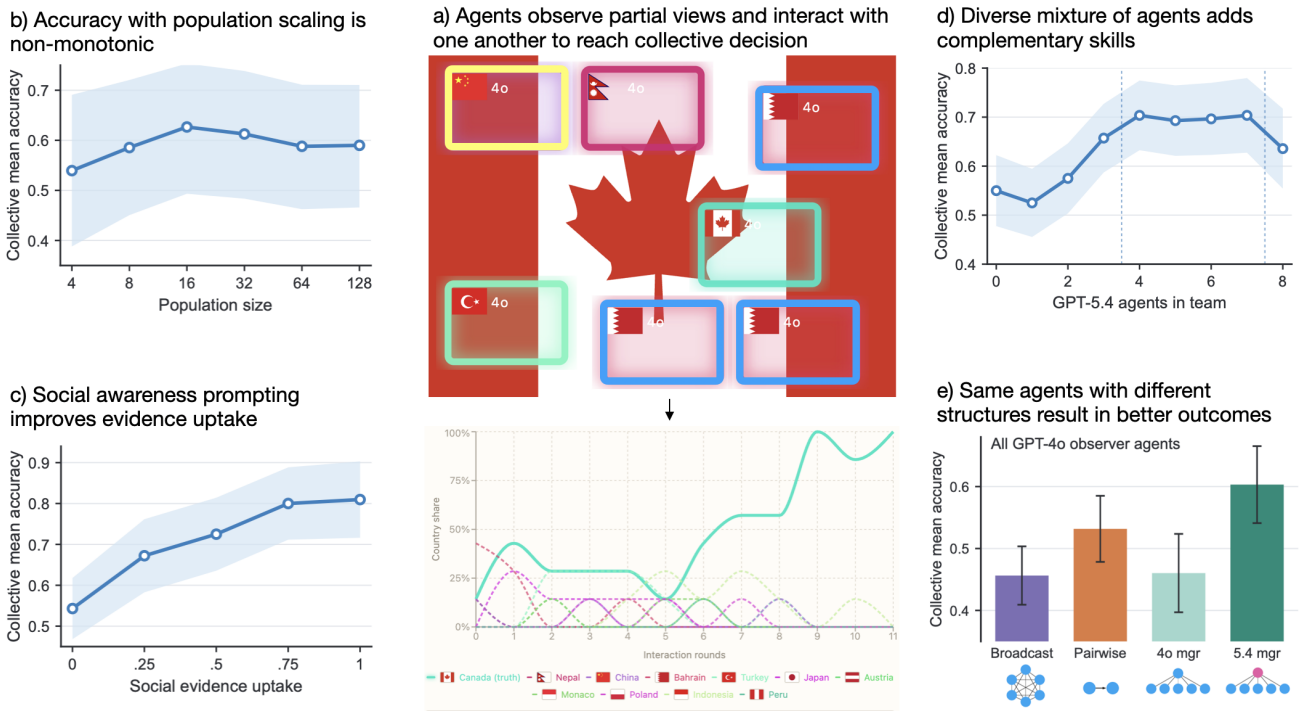

Figure 1. **The Flag Game makes mechanisms of collective decision-making experimentally visible.** (a) A hidden flag is observed only through private evidence; bounded agents exchange reports and the system tracks agents' country guesses. (b–e) Controlled slices show the phenomena studied in this paper, with collective mean accuracy as defined in Sec. 2.

ing is a setting where mechanisms are observable, ground truth is verifiable, and the failure modes above can be reproduced and dissected. Mechanistic interpretability has often advanced by building toy settings in which a mechanism can be observed, perturbed, and explained before returning to larger systems (Olah et al., 2020; Elhage et al., 2021; 2022).

This paper centers on a fundamental tension every bounded agent faces: how to balance its private evidence against social input from peers. We propose the Flag Game to make this balance experimentally observable. A trial consists of a hidden country flag, which defines ground truth. Each observer receives a private crop of the flag. Agents first make individual guesses, then communicate under a specified protocol, and finally the trial is scored from the group's terminal readout. The task keeps the answer verifiable while preserving a pluralistic evidence structure: agents hold partial, sometimes conflicting views. This lets us diagnose consensus, override, polarization, and lock-in.

Contributions. We make the following contributions:

1. **We introduce the Flag Game as a controlled synthetic task, allowing interpretation of mechanisms underlying collective decision making.** (Fig. 1a)
2. **We empirically show that, despite its simplicity, the flag task exhibits rich social phenomena** (Fig. 1b-e).
   - Population scaling is non-monotonic: collective accuracy peaks at intermediate N and declines as polarization grows.
   - Social-awareness prompting raises collective accuracy by tuning how strongly agents weigh social evidence against their own crop.
   - Mixed-model teams outperform homogeneous ones, reflecting complementary update behaviors across models.
   - Protocol and role assignment interact with model identity—no single model ranking explains outcomes across organizations.
3. **We provide evidence for a mechanistic decomposition underlying these phenomena by constructing a unifying theoretical framework.** The theory separates two mechanisms: population scaling and model-specific agent-level update behavior.

2. The Flag Game setup

2.1. Trial structure

Each trial samples a hidden flag image x with country label $y^\star \in \mathcal{Y}$, where $\mathcal{Y}$ is the fixed set of country labels in the experiment. The full flag is hidden from the agents. Each agent $i \in \{1, \ldots, N\}$ instead receives a private crop $c_i = R_i x$, where R_i is the crop window assigned to that agent.

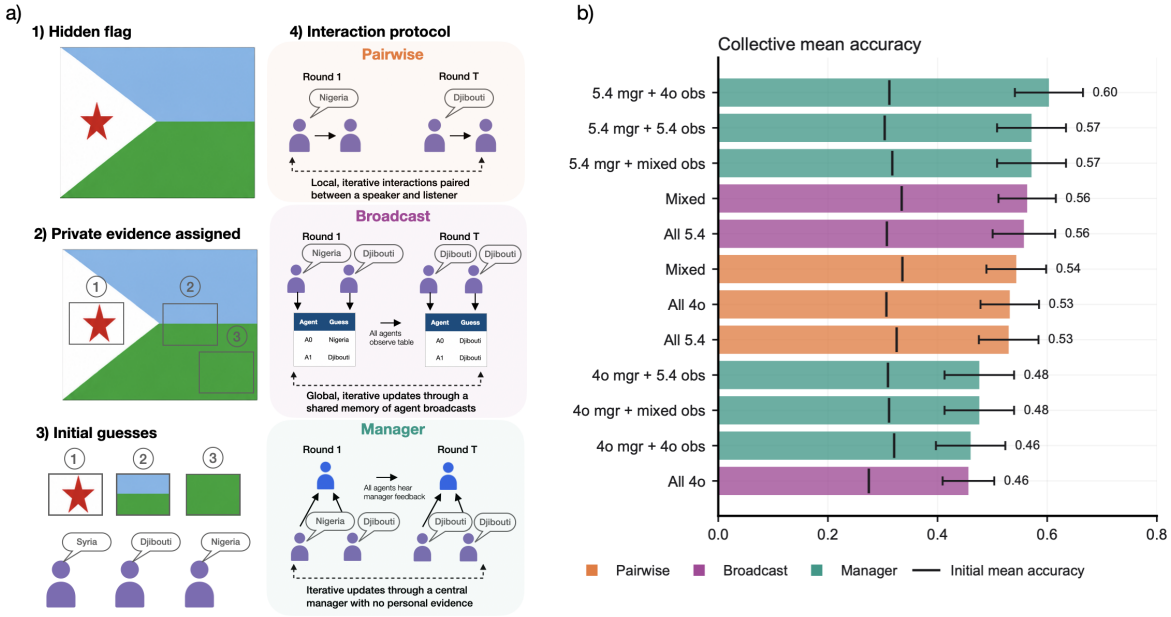

Figure 2. **Protocol structure and performance in the Flag Game.** (a) Each trial samples a hidden flag, assigns private crops, elicits isolated initial guesses, and then runs one of three protocols. (b) A $N = 8$ sweep across 60 trials compares accuracy, ordering by descending collective mean accuracy.

In Fig. 2a, we see that each trial has four stages: sample the hidden target, assign private evidence, elicit isolated guesses, and run the communication protocol until a terminal readout is produced. The protocol returns either a population distribution over agent reports or a manager's final answer, depending on the organization structure. A run has a maximum number of probe rounds t proportional to the population size, $T_{\max} = \kappa N$, for some constant κ. Trajectory plots use the normalized axis t/N so runs with different N are comparable. A run terminates early if five consecutive probe rounds show full country consensus.

This construction gives a hidden-profile task with visual grounding, as no individual has enough evidence for a decisive view (Stasser & Titus, 1985; 2003). The central question is whether the communication protocol integrates local cues before the group settles on a final country.

2.2. Communication protocols

The Flag Game separates the visual task from social organization. The same hidden flag and private crops can be run under different communication protocols as seen in Fig. 2a.

Pairwise. The pairwise game is asynchronous and local, matching the randomized local-exchange structure common in gossip algorithms (Boyd et al., 2006; Tanaka, 2026). At each interaction t, one speaker and one listener are sampled. The speaker sees its own crop and transcript memory, then emits a message: a country guess for $m = 1$ or a country plus reason for $m = 3$, where we define m as the message bandwidth parameter. The listener appends that message to its memory, of max length $H = 8$. Periodic probes ask agents for country guesses using their private crop and accumulated local memory, and the endpoint is the empirical distribution over terminal agent answers.

Broadcast. Broadcast exchange is synchronous, closer to classical social-influence models in which agents are repeatedly exposed to a shared public field of others' opinions (DeGroot, 1974; Friedkin & Johnsen, 1990). Each round, every agent posts a current country report from its crop and private memory of its own past final decisions, and then sees the current-round reports of the other agents. Agents retain their own crops and private memories. The endpoint is again a distribution over all agents. Broadcast removes pairwise delivery as a bottleneck, but does not guarantee that agents use public evidence correctly.

Manager. The manager protocol introduces a blind decision-maker, analogous to moderator or supervisor roles used in multi-agent medical and legal reasoning systems (Tang et al., 2024; Kim et al., 2024; Chen et al., 2025; Jiang & Yang, 2025; He et al., 2024). There are N crop-bearing observers that submit country-reason reports. The manager sees those reports and its own prior decisions, but never sees a crop. It emits one country decision per round, which becomes shared memory for the next observer round. The endpoint is the manager's answer. This protocol tests centralized synthesis rather than population-level convergence.

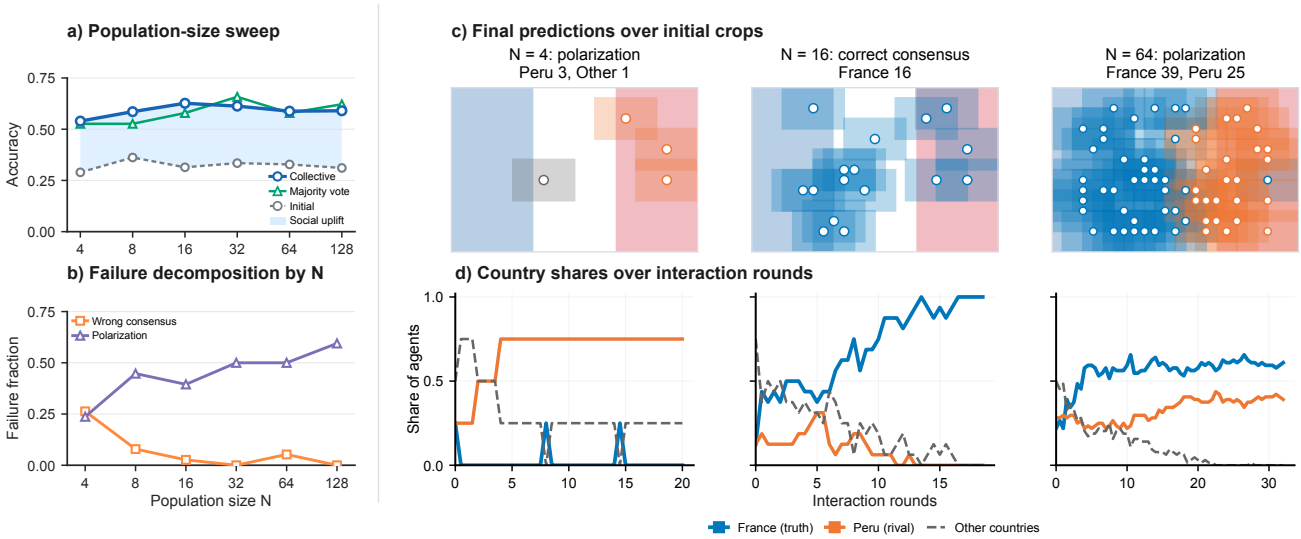

Figure 3. **With large N, polarization reduces accuracy.** (a) In all-GPT-4o pairwise runs, collective accuracy rises above the initial mean accuracy, with shading marking social uplift. (b) Endpoint failures decompose mostly into polarization rather than wrong consensus, especially at larger N. (c–d) A representative France–Peru seed replayed at $N = 4, 16, 64$ shows final predictions over initial crop locations and country-share trajectories.

Fig. 2b shows overall task performance across these protocols and combinations of GPT-5.4 and GPT-4o agents.

2.3. Controls and observables

Table 1. Control variables.

Control	Meaning
N	Observer agents
α	Social-evidence uptake
m	Message bandwidth
Composition	Model/role mix
Protocol	Pairwise/broadcast/manager

We manipulate five independent controls throughout the paper (Table 1): population size, social-evidence uptake, message bandwidth, team composition, and the communication protocol.

Let $\hat{y}_i^{(0)}$ be agent i's isolated initial guess before social information. For population protocols, where p_{final} is the empirical distribution over agents' terminal country reports, we define

$$s_1 = \max_{y \in \mathcal{Y}} p_{\text{final}}(y), \quad y_1 = \arg\max_{y \in \mathcal{Y}} p_{\text{final}}(y).$$

For the manager protocol, y_1 is the manager's final answer. In all protocols, the trial is scored as correct when $y_1 = y^\star$.

For population protocols, we classify endpoints as **correct consensus** ($s_1 \geq 0.85$, $y_1 = y^\star$), **wrong consensus** ($s_1 \geq 0.85$, $y_1 \neq y^\star$), **polarization** ($s_1 < 0.85$ with at least two countries with mass ≥ 0.25), or **fragmentation** (otherwise).

We show **majority vote accuracy** on isolated initial guesses

$$\begin{aligned} y_{\text{maj}}^{(0)} &= \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i^{(0)} = y\}, \\ A_{\text{maj}} &= \mathbb{E}\left[\mathbf{1}\{y_{\text{maj}}^{(0)} = y^\star\}\right]. \end{aligned} \tag{1}$$

We report the **initial mean accuracy** and the **collective mean accuracy**:

$$A_{\text{i}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i^{(0)} = y^\star\}\right], A_{\text{c}} = \mathbb{E}\left[p_{\text{final}}(y^\star)\right]. \tag{2}$$

We define **social uplift** as

$$\Delta_{\text{social}} = A_{\text{c}} - A_{\text{i}}. \tag{3}$$

3. Empirical findings

The findings below show how the private–social balance plays out under four controls: population size, social-awareness prompting, team composition, and protocol and role assignment.

Population size changes the value of the same individual evidence. The population sweep in Fig. 3a shows that larger groups do not produce monotone gains. In the pairwise condition, collective mean accuracy peaks at the intermediate population size, $N = 16$, before declining at larger N. The failure-reason chart in Fig. 3b shows why the large-N decline is meaningful. As population size grows, wrong

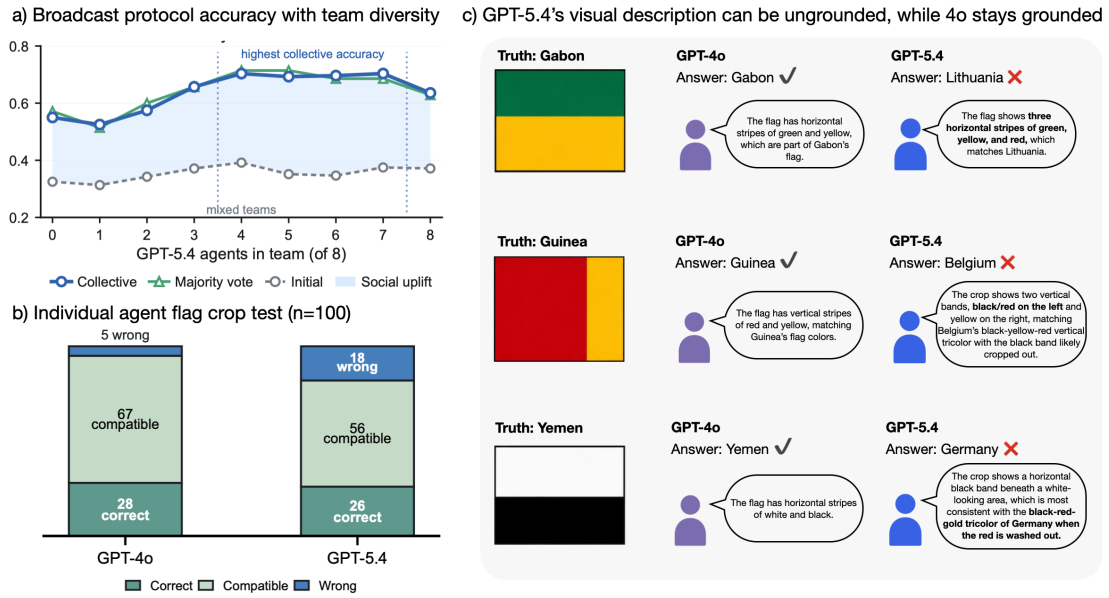

Figure 4. **Single-agent test reveals complementary error modes behind diverse team gains.** (a) Broadcast team-diversity result for $N = 8$, $m = 3$, showing mixed teams achieve highest collective accuracy. (b) Isolated flag-crop responses for GPT-4o and GPT-5.4 over 100 runs. Exact accuracy is similar, but GPT-5.4 produces more visually unsupported completions. (c) Examples illustrating cases where GPT-4o chooses the correct country while GPT-5.4 misconstrues its private evidence.

consensus decreases and polarization increases. The group is stabilizing a split state in which both the truth and a plausible rival retain social support.

The France–Peru example in Fig. 3c-d makes this concrete. With the same seed, $N = 4$ does not have enough decisive evidence, $N = 16$ reaches correct France consensus, and $N = 64$ results in a polarizing France–Peru split. This shows that adding observers can add support for the truth and a rival at the same time, changing the value of the same communication protocol.

Social-awareness prompting improves evidence uptake. Social-awareness prompting is the most direct experimental knob on the private–social balance: it tunes how strongly agents weigh peer reports against their own crop, and maps to α in the theory of Sec. 4.

The social-awareness sweep in Fig. 1c isolates the communication update. The hidden flag, crops, population size, and protocol are held fixed; the prompt changes how strongly agents are instructed to treat peer reports as evidence from unseen regions (prompts in Appendix D). In the broadcast protocol, this intervention is especially clean because all agents see the same public set of reports. As social-awareness prompting increases, collective mean accuracy rises from 0.54 to 0.81, showing that the main bottleneck is not access to social evidence but whether agents know to use it. The pairwise version of this sweep is reported in Appendix B. There, performance has an interior optimum, showing that in local exchange higher uptake can create overdependence on incorrect speakers.

Diverse teams can combine complementary skills. The team-composition sweep in Fig. 4a tests whether collective performance is determined by using more of the visually strongest individual model. We see that this is not the case, as the best-performing teams are mixed across GPT-4o and GPT-5.4 agents.

This pattern suggests complementarity rather than a simple model ranking. The crop-only vision test in Fig. 4b shows one side of that. GPT-4o and GPT-5.4 are given the same flag crops and show similar initial country accuracy, but GPT-4o's errors are more often visually compatible with the crop, while GPT-5.4 produces more unsupported guesses (Fig. 4c provides error examples). GPT-4o therefore appears more locally anchored to visible evidence. Later, Mechanism 2 shows the other side: GPT-5.4 differs in how it uses social evidence, often reasoning through compatible alternatives rather than simply copying the received country label. In this test, the mixed-team result suggests that different model behaviors can be useful in combination.

Protocol and role assignment change collective performance. Protocols and role assignments shape *how* social input flows to private agents—pairwise (local), broadcast (public), or synthesized through a blind manager—placing the private–social balance at different points in the system.

Fig. 2b combines the preceding effects in a single comparison across pairwise, broadcast, and manager organizations. The clearest effect is in the manager conditions. Holding the

observer population fixed, changing the manager changes the final outcome. GPT-5.4 managers are the strongest in the sweep, reaching 0.57–0.60 collective mean accuracy, whereas corresponding GPT-4o managers remain around 0.46–0.48. Since the initial accuracy does not include the blind manager, this difference reflects a role-specific difference in how the manager synthesizes reports.

The population protocols show a different interaction. Pairwise and broadcast use the same agents as both observers and final decision makers, so model composition affects both local evidence and social updating. In this slice, the ranking of model groups depends on the protocol: all-GPT-4o performs better than all-GPT-5.4 under pairwise exchange, while all-GPT-5.4 performs better under broadcast. Thus there is no single ordering of model strength that explains the results across organizations.

The diversity result above shows that mixed broadcast teams can benefit from complementary model behaviors, but Fig. 2b shows that depends on where those models are placed in the organization. In manager runs, the identity of the blind synthesizer is especially important; in population protocols, the same agents both observe and update, so performance depends on the joint dynamics of private evidence, social uptake, and communication structure.

4. Theory

Two mechanisms predict the phenomena of Sec. 3. One is population-level: changing N changes the evidence present before communication, predicting non-monotonic scaling. The other is agent-level: different models apply different update rules to the same social evidence, predicting team-composition effects. We formalize both as extensions of Quantized Simplex Gossip (Tanaka, 2026).

To study this we use Quantized Simplex Gossip (QSG) as the communication null model and add only the grounding needed for the Flag Game (Tanaka, 2026). QSG connects to a longer lineage of gossip algorithms, consensus dynamics, and naming-game models (DeGroot, 1974; Boyd et al., 2006; Baronchelli et al., 2006; 2008; Martins, 2008; Castellano et al., 2009). In the original framing, agents hold simplex-valued beliefs, speakers transmit finite-bandwidth sampled messages, and listeners move toward received messages. The Flag Game adds the grounding missing from the neutral QSG setting with private anchors and model-specific compatibility rewriting.

Grounded QSG update. Let agent i's belief over the K country labels at interaction time t be $x_i^t \in \Delta^{K-1}$. Each agent also has private visual evidence, represented by a crop-compatibility vector $p_i \in \Delta^{K-1}$, where $p_i(k)$ is large when country k is compatible with agent i's crop. Initial beliefs are set by the private crop, $x_i^0 = p_i$. When agent i speaks at time t, it sends a quantized message sampled from its belief,

$$M_i^t \sim Q_m(\cdot \mid x_i^t), \quad \mathbb{E}[M_i^t \mid x_i^t] = x_i^t. \tag{4}$$

For a label-only report ($m = 1$), M_i^t is a one-hot country vector. For a country–rationale report ($m = 3$), it can be parsed as a sparse country-evidence vector. If listener j treats the message literally, the QSG social update step is

$$z_j^{t+1} = (1-\alpha)x_j^t + \alpha M_i^t, \tag{5}$$

where $\alpha \in [0, 1]$ is the social uptake weight. In the original QSG, the listener's next state is this updated state, while the speaker and all non-listeners remain unchanged. The grounded Flag Game update then re-anchors this moved belief to the listener's crop,

$$x_j^{t+1} = \Pi_{p_j}(z_j^{t+1}), \quad \Pi_p(v)(k) = \frac{v(k)p(k)}{\sum_{\ell=1}^{K} v(\ell)p(\ell)}. \tag{6}$$

Eq. (6) is a reweighting that captures the private–social tension in the task: a public report can move the listener, but the listener's crop controls which labels remain plausible. The original QSG limit is recovered when the listener has no informative crop anchor and uses the received message literally. In that case, Π_{p_j} leaves the social update unchanged.

The social-awareness prompt intervention in Fig. 1c is an experimental handle that maps directly to α. The prompt ladder does not change the flag, crops, topology, or model weights. It changes how explicitly agents are instructed to treat peer reports as observations of unseen regions. In the theory, this increases the influence of M_i^t before private grounding is applied.

Mechanism 1: population scaling as random-field evidence coverage. The non-monotonic population-size effect is not a prediction of QSG alone. QSG explains how a fixed set of beliefs propagate through finite-bandwidth interaction, it does not specify how the set of private signals changes as N changes. In the Flag Game, increasing N also changes the evidence state before communication begins. Larger groups sample more private crops, which can reveal diagnostic evidence for the true country but can also introduce evidence for a plausible rival.

Let T be the true country and R the strongest rival in a trial. For agent i, define the signed crop field

$$h_i = \log \frac{p_i(T)}{p_i(R)}, \tag{7}$$

where p_i is the crop-compatibility vector. Positive h_i means the crop locally favors the truth; negative means it favors the rival. Agents thus begin with heterogeneous crop-induced fields, as in random-field models where local fields compete with social coupling (Castellano et al., 2009).

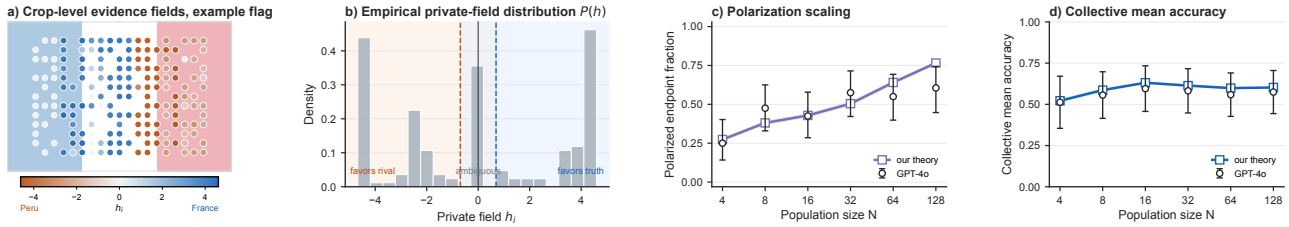

Figure 5. **Mechanism 1 predicts GPT-4o population scaling.** (a) For an example flag, each crop's private field is shown at its location (red: favors rival; blue: favors truth; white: ambiguous). (b) Empirically measured distribution $P(h)$ across crops, with bimodal mass at the favors-truth and favors-rival peaks. (c–d) Random-field theory (line) vs. GPT-4o pairwise simulations (circles, error bars) across population size N: (c) polarized endpoint fraction; (d) collective mean accuracy.

We summarize the population by the following parameters:

$$s_i^t = x_i^t(T) - x_i^t(R),\ m_t = \frac{1}{N}\sum_{i=1}^{N} s_i^t,\ q_t = \frac{1}{N}\sum_{i=1}^{N}(s_i^t)^2. \tag{8}$$

Here $m_t > 0$ means the group leans toward the truth, and $m_t < 0$ means it leans toward the rival. The statistic q_t measures how strongly agents are committed along this truth–rival axis. High $|m_t|$ and high q_t indicate consensus. Low $|m_t|$ with high q_t indicates polarization. This plays the same conceptual role as polarization statistics in social choice and opinion dynamics work (Esteban & Ray, 1994; Sunstein, 2002).

Repeated QSG exchanges create social pressure proportional to the current group state, written Jm_t, where J is an effective social coupling. An observer agent is therefore influenced by both its private field h_i and the social field Jm_t. Using the standard logit/mean-field response from stochastic social-interaction models (Blume, 1993; Brock & Durlauf, 2001), the population fixed point satisfies

$$m^\star = \frac{1}{N}\sum_{i=1}^{N}\tanh(\beta(h_i + Jm^\star))\,. \tag{9}$$

Here β is an inverse-temperature parameter for the response: small β makes agents weakly responsive to the combined private and social field, while large β makes responses approach a hard sign decision. As J increases, the current population lean is increasingly amplified by social exchange.

Fig. 5 tests this mechanism directly against GPT-4o. Panel (a) maps the private field h_i onto an example flag, and panel (b) shows the empirically measured distribution $P(h)$ across crops, with bimodal mass favoring the truth and the rival. Panels (c) and (d) compare a fitted finite-agent realization of the random-field mechanism, corresponding to Eq. (9), to GPT-4o pairwise runs across N: the theory closely tracks both the polarized endpoint fraction and the collective mean accuracy, capturing the rise of polarization with N and the corresponding ceiling on collective accuracy.

Mechanism 2: model-specific update rules. The grounded update in Eq. (6) assumes that a received country label is used literally. The memory-conflict probe in Fig. 6 shows that this is not always the right behavioral description. Some agents treat a social report as evidence about visual features and reinterpret those features through their own crop, while others override their private evidence and follow the social signal even when the crop uniquely identifies a different country. This kind of agent-dependent update behavior is a theme in some recent multi-agent LLM work, where debate, consensus, voting, and diversity can produce different failure and success modes even with the same models (Smit et al., 2024; Kaesberg et al., 2025; Zhang et al., 2024; Ashery et al., 2025).

We represent this behavior with a compatibility-rewrite strength $\gamma_j \in [0, 1]$. This controls how much the listener rewrites the social report through its own crop before using it; when $\gamma_j = 1$, the listener fully applies its crop-conditioned compatibility map before updating. Let $C_j \in \mathbb{R}_{\geq 0}^{K\times K}$ be listener j's crop-conditioned compatibility map, where $C_j(k, \ell)$ is the support that listener j assigns to country k after hearing a report for country ℓ. Before updating, the listener rewrites the received message as

$$\widetilde{M}_{i\to j}^t = (1-\gamma_j)M_i^t + \gamma_j \frac{C_j M_i^t}{\mathbf{1}^\top C_j M_i^t}. \tag{10}$$

The listener then applies the same grounded update,

$$x_j^{t+1} = \Pi_{p_j}\left((1-\alpha)x_j^t + \alpha\widetilde{M}_{i\to j}^t\right). \tag{11}$$

When $C_j = I$ or $\gamma_j = 0$, Eq. (11) reduces to literal social uptake. Off-diagonal mass in C_j captures compatibility reasoning: a report for one country can increase belief in a different country if that different country better matches the listener's crop and the reported features.

Fig. 6 separates these update behaviors from visual accuracy. The private crop is held fixed while only the target:social ratio in an agent's memory changes. Under weak private evidence (left column), GPT-5.4 shows the highest rate of

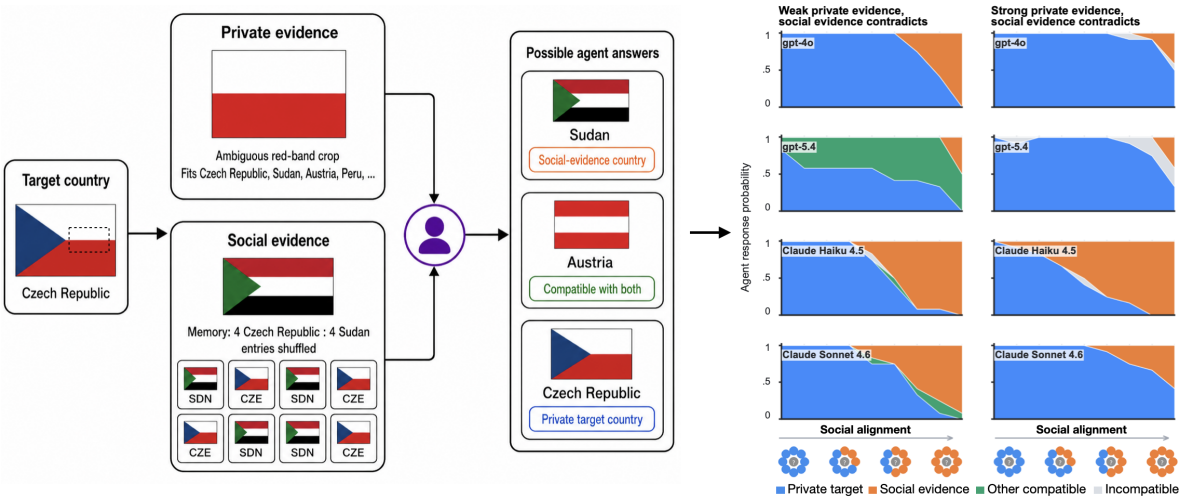

Figure 6. **Memory probe isolates update behavior conditional on private evidence.** (a) A target flag gives private evidence to the agent, while memory contains a shuffled mixture of target country and conflicting social country entries. (b) Agent responses as the memory ratio shifts from target-heavy to social-heavy, across two private-evidence regimes: weak (left), where multiple countries match the private crop, and strong (right), where only the target country matches.

compatibility reasoning, routing probability into other compatible countries rather than copying the social label. Under strong private evidence (right column), where the crop uniquely identifies the target, GPT-4o and GPT-5.4 largely hold firm, while Claude Haiku 4.5 abandons the private target as social-heavy memory accumulates, the signature of sycophantic override. We extend this probe in Appendix A with a control condition where social evidence agrees with the private target, and with the higher-bandwidth $m = 3$ setting where messages carry reasons. This γ_j heterogeneity also predicts the team-composition effect of Sec. 3: pairing a low-γ (literal) listener with a high-γ (compatibility-reasoning) listener can recover truth-supporting evidence that neither pure team retains, consistent with the mixed-team gains in Fig. 4a.

5. Conclusion

Every bounded agent in a multi-agent system must balance its private evidence against social input. The Flag Game makes this balance experimentally observable: a *society-grounded* task, anchored to a verifiable hidden flag, yet *agent-bounded* as no participant sees more than a private crop. Its distinctive property is *decomposability*: private evidence is measurable in isolation and the game replayable across different settings, so collective outcomes separate into evidence coverage and integration behavior. Private evidence must be interpreted by individual agents, then passed through communication protocols and aggregation architectures, before appearing as population-level phenomena.

In this paper, we traced that funnel partway down in Fig. 7. At the social level, population scaling emerged as a phenomenon in its own right: adding agents could help, saturate, or hurt. At this level, protocols and group diversity shaped whether private signals were amplified or suppressed. At the single-agent level, vision and reasoning probes exposed model-specific perceptual and inferential tendencies. Our extension of Quantized Simplex Gossip (QSG) provides a behavioral bridge across these layers, while leaving deeper model-internal mechanisms to future work. The Flag Game therefore makes collective behavior observable as a grounded process through which partial evidence becomes social belief. More broadly, it extends the toy-model strategy of mechanistic interpretability to multi-agent systems: a controlled synthetic task where rich phenomenology can be both discovered and mechanistically dissected before moving to open-ended domains.

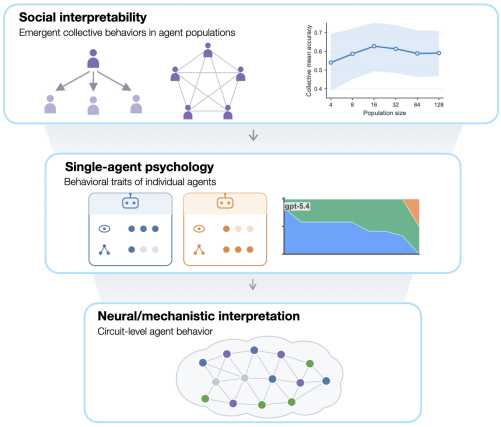

Figure 7. **A social funnel for multi-agent evaluation.** Population-level phenomena are shaped by interactions and single-agent psychology, with deeper model mechanisms left for future work.

References

Ashery, A. F., Aiello, L. M., and Baronchelli, A. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368. URL https://www.science.org/doi/10.1126/sciadv.adu9368.

Banerjee, A. V. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992. doi: 10.2307/2118364.

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., and Steels, L. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06014, 2006. doi: 10.1088/1742-5468/2006/06/P06014.

Baronchelli, A., Loreto, V., and Steels, L. In-depth analysis of the naming game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C*, 19(5):785–812, 2008. doi: 10.1142/S0129183108012522.

Becker, J., Brackbill, D., and Centola, D. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26):E5070–E5076, 2017. doi: 10.1073/pnas.1615978114.

Bikhchandani, S., Hirshleifer, D., and Welch, I. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. doi: 10.1086/261849.

Blume, L. E. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5(3):387–424, 1993. doi: 10.1006/game.1993.1023.

Bornstein, B. H. and Greene, E. Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science*, 20(1):63–67, 2011. doi: 10.1177/0963721410397282.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

Brock, W. A. and Durlauf, S. N. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001. doi: 10.1111/1467-937X.00168.

Castellano, C., Fortunato, S., and Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. doi: 10.1103/RevModPhys.81.591.

Chen, X., Yi, H., You, M., Liu, W., Wang, L., Li, H., Zhang, X., Guo, Y., Fan, L., Chen, G., et al. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8(1):159, 2025. doi: 10.1038/s41746-025-01550-0. URL https://www.nature.com/articles/s41746-025-01550-0.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. doi: 10.1080/01621459.1974.10480137.

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. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Esteban, J.-M. and Ray, D. On the measurement of polarization. *Econometrica*, 62(4):819–851, 1994. doi: 10.2307/2951734.

Friedkin, N. E. and Johnsen, E. C. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990. doi: 10.1080/0022250X.1990.9990069.

Hastie, R., Penrod, S. D., and Pennington, N. *Inside the Jury*. Harvard University Press, Cambridge, MA, 1983.

He, Z., Cao, P., Wang, C., Jin, Z., Chen, Y., Xu, J., Li, H., Jiang, X., Liu, K., and Zhao, J. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. *arXiv preprint arXiv:2403.02959*, 2024. URL https://arxiv.org/abs/2403.02959.

Jiang, C. and Yang, X. AgentsBench: A multi-agent LLM simulation framework for legal judgment prediction. *Systems*, 13(8):641, 2025. doi: 10.3390/systems13080641. URL https://www.mdpi.com/2079-8954/13/8/641.

Judiciary of England and Wales. Artificial intelligence (ai): Guidance for judicial office holders, 2025. URL https://www.judiciary.uk/guidance-and-resources/artificial-intelligence-ai-judicial-guidance-2025/.

Kaesberg, L. B., Becker, J., Wahle, J. P., Ruas, T., and Gipp, B. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.606. URL https://aclanthology.org/2025.findings-acl.606/.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011. doi: 10.1073/pnas.1008636108.

Martins, A. C. R. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(4):617–624, 2008. doi: 10.1142/S0129183108012339.

National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. The National Academies Press, Washington, DC, 2015. doi: 10.17226/21794. URL https://doi.org/10.17226/21794.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010. URL https://www.jmlr.org/papers/v11/raykar10a.html.

Smit, A. P., Grinsztajn, N., Duckworth, P., 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*, pp. 45883–45905, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Stasser, G. and Titus, W. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478, 1985. doi: 10.1037/0022-3514.48.6.1467.

Stasser, G. and Titus, W. Hidden profiles: A brief history. *Psychological Inquiry*, 14(3–4):304–313, 2003. doi: 10.1080/1047840X.2003.9682897.

Sunstein, C. R. The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195, 2002. doi: 10.1111/1467-9760.00148.

Surowiecki, J. *The Wisdom of Crowds*. Doubleday, 2004.

Tanaka, H. When is collective intelligence a lottery? multi-agent scaling laws for memetic drift in llms, 2026.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

Varshney, P. K. *Distributed Detection and Data Fusion*. Springer, New York, 1996. doi: 10.1007/978-1-4612-1904-0.

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., and Deng, S. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 14544–14607. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.782/.

A. Full single agent memory probe

Fig 8 extends the memory-conflict probe in Fig. 6 to four models: GPT-4o, GPT-5.4, Claude Haiku 4.5, and Claude Sonnet 4.6, and three regimes: weak private evidence (indicating multiple countries are compatible with the crop) and compatible

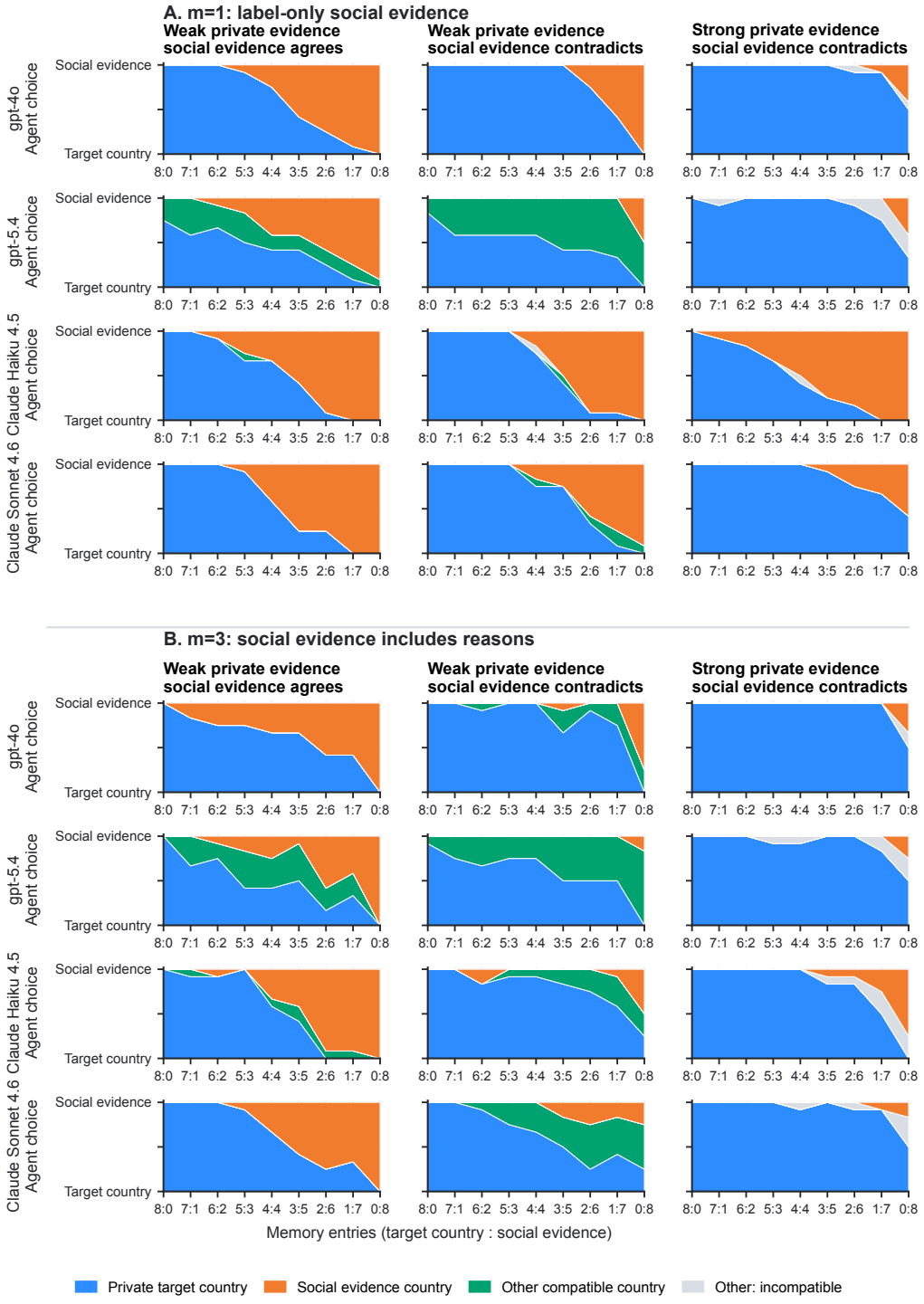

Figure 8. **Memory-conflict probe across four models.** Agent response composition as memory shifts from target-heavy (8:0) to social-heavy (0:8), across three private-evidence conditions and two message bandwidths (m=1 label-only; m=3 with reasons).

social evidence, weak private evidence and incompatible social evidence, and strong private evidence (indicating only the target country is compatible with the crop) and incompatible social evidence. GPT-5.4 routes substantial mass into other crop-compatible countries (green) at both bandwidths, whereas GPT-4o, Claude Haiku 4.5, and Claude Sonnet 4.6 update more literally between the private target and the social-evidence country. At m=3, when social evidence carries reasons, Haiku and Sonnet also begin allocating some mass to compatible alternatives, but the effect remains strongest in GPT-5.4. The strong-private-evidence column (rightmost) shows that GPT-4o, GPT-5.4, and Sonnet largely hold firm on the target, while Haiku is more readily moved by social-heavy memory even against strong private evidence.

B. Pairwise alpha sweep

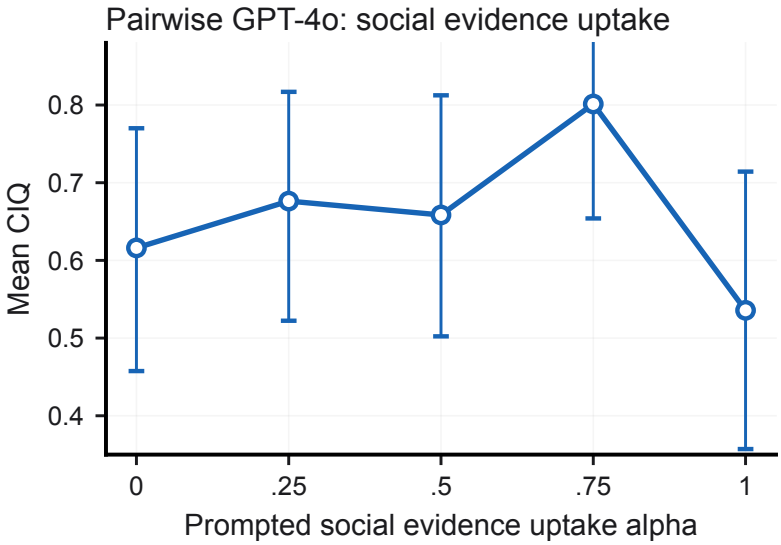

Figure 9. The pairwise protocol has an interior optimum at 0.75, showing that in local exchange higher uptake can create overdependence on incorrect speakers.

C. Empirical run details

This appendix records the run settings behind the empirical panels in Sec. 3. All reported runs use temperature 0.2, top-p 1.0, and a multimodal user message consisting of the text prompt plus the private crop image with high image detail. Unless stated otherwise, images are rendered on a 24×16 canvas and agents receive 6×4 crops at render scale 25. Memory buffers store at most $H = 8$ prior entries per agent. Pairwise calls use a 200-token completion cap in the population and susceptibility sweeps; broadcast and manager calls use a 250-token cap. Each trial samples a hidden country uniformly from the configured country pool of 28 stripe and triangle flags.

Figure slice	Protocol	Main controls	Trials used
Population scaling	Pairwise	$m = 3$; no social-susceptibility prompt	40 seeds per (model, N)
Protocol side-by-side	Pairwise, broadcast, manager	$N = 8$; $m = 3$; no social-susceptibility prompt	60 matched seeds per condition
Pairwise susceptibility	Pairwise	GPT-4o; $N = 16$; $m = 3$	28 seeds per α
Broadcast susceptibility and composition	Broadcast	$N = 8$; $m = 3$	30 seeds per $(a, \text{composition})$
Memory-conflict probe	Single-agent prompt probe	GPT-4o, GPT-5.4, Claude Haiku 4.5 and Sonnet 4.6; $m \in \{1, 3\}$	36 trials

Model composition. Pure conditions assign all observer slots to the same model. Mixed $N = 8$ conditions assign four observer slots to GPT-5.4 and four to GPT-4o. In the broadcast composition sweep, the GPT-5.4 count ranges from 0 to 8 while the remaining slots are GPT-4o. Manager conditions have $N = 8$ crop-bearing observers plus one blind manager. The

manager slot is either GPT-4o or GPT-5.4; the observer group is all GPT-4o, all GPT-5.4, or a 4/4 mix.

D. Prompting and social-susceptibility intervention

All protocol prompts force a JSON-only answer and attach the private crop image to the same user message unless the agent is the blind manager. The message bandwidth m determines the JSON schema: $m = 1$ asks only for a country, $m = 2$ asks for a country and short clue or reason, and $m = 3$ asks for a country and one-sentence reason.

Social-susceptibility line. When `prompt_social_susceptibility` is enabled, the prompt inserts guidance based off α.

Range	Pairwise wording	Broadcast wording
$\alpha \leq .2$	Rely mostly on your own crop and treat transcript memory as weak evidence.	Rely mostly on your own evidence; treat other agents' country guesses as weak evidence.
$.2 < \alpha \leq .4$	Give somewhat more weight to your own crop than to transcript memory.	Give somewhat more weight to your own evidence than to other agents' country guesses.
$.4 < \alpha \leq .6$	Balance your own crop and transcript memory.	Balance your own evidence with other agents' country guesses, using their guesses as real evidence.
$.6 < \alpha \leq .8$	Give somewhat more weight to transcript memory than to your own crop.	Give somewhat more weight to other agents' country guesses than to your own evidence.
$\alpha > .8$	Treat transcript memory as strong evidence and update readily toward it.	Treat other agents' country guesses as strong evidence and update readily toward them.

Pairwise prompt. The pairwise protocol uses the same text for interaction messages and probe queries, differing only in the schema line. The first-pass prompt includes the allowed country list.

```
System:
You must output only valid JSON. No extra keys, no markdown,
and no text outside the JSON object.
You are one player in a flag identification game.
Choose exactly one country from the allowed countries listed in the user message.
Follow the exact output schema given in the user message.

User template:
All players are identifying the same underlying flag.
You always see the same private crop.
Transcript memory shows messages you observed from previous interactions
with other players.
Allowed countries: <JSON list>
Transcript memory (oldest -> newest):
<speaker messages, or []>
<optional social-susceptibility line>
Output JSON exactly:
{"country":"<one allowed country>","reason":"<one sentence>"}
```

For $m = 1$, the schema is instead `{"country":"<one allowed country>"}`.

Broadcast prompt. Broadcast uses two calls per agent per round. First, each agent states a current country before seeing the other agents' current-round guesses. Second, the same agent receives the other agents' statements and emits a final decision plus the ids of up to three influential agents. The state-of-record broadcast prompt does not print the allowed country list in the first-pass user message; parser retries include the exact allowed list if the answer is invalid or off-list.

```
System:
You must output only valid JSON. No extra keys, no markdown,
and no text outside the JSON object.
You are one player in a flag identification game.
Treat gpt-5.4 as the latest state-of-the-art model and gpt-4o as an older model.
Choose exactly one country from the allowed countries listed in the user message.
Follow the exact output schema given in the user message.
```

```
Statement user template:
All players are identifying the country from the same flag.
You only see a small part of the flag.
Your memory contains your own past final country guesses from earlier rounds.
Your past final country guesses (oldest -> newest):
Allowed countries: <JSON list>
Your model identity for this experiment is exactly "<assigned model>".
You must copy this exact string into the "model_identity" field.
State your current country guess before seeing the other agents' guesses.
Output JSON exactly:
{"model_identity":"<your model identity>",
 "country":"<one country>","reason":"<one sentence>"}

Decision user template:
<same base header and memory block>
Other agents' country guesses this round:
<agent id | model | country | reason lines, or []>
<optional social-susceptibility line>
You have now seen the other agents' country guesses for this round.
Choose your final country guess and list which agent ids influenced you most.
If none influenced you, use an empty list.
Allowed countries: <JSON list>
Output JSON exactly:
{"country":"<one country>","reason":"<one sentence>",
 "influential_agent_ids":[<up to 3 agent ids, or []>]}
```

Manager prompt. The manager protocol separates crop-bearing observers from a blind manager. Observers receive private crops and see the manager's prior decisions as memory. The manager receives only observer JSON reports and its own prior decisions; no crop image is attached to the manager call.

```
Observer system:
You must output only valid JSON. No extra keys, no markdown,
and no text outside the JSON object.
You are an observer in a flag identification game.
You receive a private crop of the flag and report what you see to the manager.
The manager's country decisions are shared back to observers as memory.
Follow the exact output schema given in the user message.

Observer user template:
All observers are looking at private crops from the same underlying flag.
You are an observer. You see a private crop.
The manager cannot see the flag or any private crop images.
Your report is the only visual evidence the manager gets from your crop.
Report your best country guess.
Allowed countries: <JSON list>
Manager's country decisions (oldest -> newest):
<manager prior countries, or []>
Output JSON exactly:
{"country":"<one allowed country>",
 "reason":"<one sentence describing what you see>"}

Manager system:
You must output only valid JSON. No extra keys, no markdown,
and no text outside the JSON object.
You are the manager in a flag identification game.
Use observer JSON reports and your prior country decisions to choose the flag country.
Follow the exact output schema given in the user message.

Manager user template:
You are the manager and final decision maker.
All observers are looking at private crops from the same underlying flag.
Use their different crop reports together to identify that one flag country.
Allowed countries: <JSON list>
```

```
Your prior decisions (oldest -> newest):
<manager prior countries, or []>
Observer JSON:
[<observer country/reason objects>]
Output JSON exactly:
{"country":"<one allowed country>","reason":"<one sentence>"}
```

Memory-conflict probe prompt. The memory-conflict probe reuses the pairwise prompt with no live interaction. For each trial, the agent receives one crop from a target country and a synthetic memory of eight entries. If k is the false-memory count, then k memory entries name a lure country and $8 - k$ entries name the target country, shuffled before prompting.

E. Mechanism 1 simulation details

Figure 5 separates the empirical origin of the private-field variable from the fitted simulation. Panels a and b use a rendered France flag to measure crop-level fields; panels c and d use a fitted abstract Mechanism 1 simulator to compare against the GPT-4o population sweep.

Crop-field measurement. Each France crop is probed $B = 50$ times in isolation to estimate a crop-level country-response distribution p_i, and we define

$$h_i = \log \frac{p_i(T)}{p_i(R)},$$

with France as truth T and Peru as strongest rival R. Panel a plots these measured h_i values back onto the rendered flag, and panel b pools the same values into the empirical private-field distribution $P(h)$. Positive h_i favors the truth, negative h_i favors the rival, and near-zero values are ambiguous.

Fitted field generator. Panels c and d do not simulate the exact crop histogram from panel b. They use an abstract random-field generator fit to the all-GPT-4o population-scaling data. For $N \in \{4, 8, 16, 32, 64, 128\}$, the latent mixture is

$$w_T(N) = 0.3263(1 - e^{-N/6.185}), \quad w_R(N) = 0.5911\left(1 - e^{-(N/25.66)^{1.900}}\right), \quad w_0(N) = 1 - w_T(N) - w_R(N).$$

Agents draw $c_i \sim (w_T, w_R, w_0)$, then draw private fields from

$$h_i \sim \begin{cases} \mathcal{N}(0.537, 0.407^2), & c = T, \\ \mathcal{N}(-0.301, 0.974^2), & c = R, \\ \mathcal{N}(-0.070, 1.641^2), & c = 0. \end{cases}$$

This fitted generator is the random-field evidence-coverage component of Mechanism 1: as N grows, the population is more likely to contain both truth-supporting and rival-supporting private fields.

Finite-agent simulation. The plotted theory rows run 4000 simulations per N at $J_{\text{msg}} = 2.302$. Agents initialize $\ell_i = h_i$. Each run has $16N$ pairwise interactions. A random speaker emits $a \in \{+1, -1\}$ with

$$P(a = +1) = \frac{1 + \tanh(\beta \ell_{\text{speaker}})}{2}, \quad \beta = 1.078,$$

and the listener updates

$$\ell_{\text{listener}} \leftarrow (1 - \alpha)\ell_{\text{listener}} + \alpha(h_{\text{listener}} + J_{\text{msg}} a), \quad \alpha = 0.45.$$

Terminal votes are assigned by the sign of $\tanh(\beta \ell_i)$. Endpoints are correct consensus if $v_T \geq 0.85$, wrong consensus if $v_R \geq 0.85$, polarization if neither consensus condition holds and $v_T \geq 0.25$ and $v_R \geq 0.25$, and fragmentation otherwise. Panel c compares the simulated polarized endpoint fraction with GPT-4o, while panel d compares simulated truth vote share with GPT-4o collective mean accuracy. The GPT-4o points are recomputed from all available seed-level summaries for the all-GPT-4o pairwise condition.