Incentive Framing Shapes Costly Evidence Revelation in Multi-Agent LLM Deliberation

Anonymous ACL submission

Abstract

Multi-agent LLM systems increasingly rely on deliberation, but deliberation only helps when agents reveal the private evidence they hold. We study costly hidden-profile deliberation, where each agent observes partial evidence, disclosure is individually costly, and the group can decide correctly only if decisive evidence is surfaced. Across ten incentive framings, four LLMs, and 60 high-stakes decision scenarios, incentives change information-sharing more reliably than final accuracy. Contribution-based framings are the only active conditions that consistently keep decisive evidence surfacing above 87% across all architectures, while commitment-based framings often increase withholding; on GPT-4o-mini, STAKE even reduces accuracy below the no-communication baseline. At Qwen3-32B, COUNTERFACTUAL_CONTRIBUTION provides suggestive evidence that structured contribution incentives may approach or surpass free debate at scale. Across all settings, first-speaker anchoring, round degradation, domain ambiguity, cascade resilience, and stable agent heterogeneity persist, indicating that incentive framing improves evidence revelation but does not by itself prevent deliberative failure.

1 Introduction

Multi-agent LLM systems are often built around a simple promise: if several agents discuss a problem, the group can combine what each agent notices and arrive at a better answer. This promise underlies recent work on LLM debate, multi-agent critique, and collaborative reasoning (Du et al., 2024; Liang et al., 2024; Chan et al., 2023; Khan et al., 2024). But deliberation only helps when useful private information actually enters the conversation. In many realistic settings, sharing information is costly. Evidence may take effort to explain, may expose uncertainty, or may make the speaker accountable for a difficult recommendation. When the benefit of a correct group decision is shared, but the cost of disclosure is individual, agents face the *free-rider problem* (Olson, 1965).

This problem is closely related to the hidden-profile failures studied in human group decision-making. Groups often fail even when decisive information exists, because that information is privately held and never disclosed (Stasser and Titus, 1985; Lu et al., 2012). We study this issue in *costly hidden-profile deliberation* in LLM agents. Each agent receives a small private set of evidence, some pieces are decisive for the correct group decision, and revealing evidence carries a feature-level cost. The group succeeds only if the agents holding decisive evidence choose to reveal it.

We explore the question: *when private evidence has individual cost to disclose, which incentive framings lead LLM agents to reveal rather than withhold decisive information, and how robust are these effects across model scales and architectures?* Our study focuses on prompt-based incentive framings, where incentive rules are described in the prompt and no real budget is deducted or stake forfeited. This reflects typical multi-agent LLM deployments, where designers mainly control descriptions of tasks, roles, and rewards. Our claims are therefore about how LLM agents behave under different incentive framings, not about formal guarantees of the mechanisms.

We evaluate four LLMs from two families (Qwen3-8B, Qwen3-14B, Qwen3-32B, and GPT-4o-mini) on 60 high-stakes binary decision scenarios in clinical, legal, financial, cybersecurity, and public-safety domains. Each scenario distributes private evidence across ten agents and labels features by disclosure cost, signal type, and whether they are decisive, misleading, or weak. We compare ten deliberation protocols, including no communication, costless free debate, forced sharing, equal reward splits, contribution-proportional rewards, counterfactual contribution rewards, and

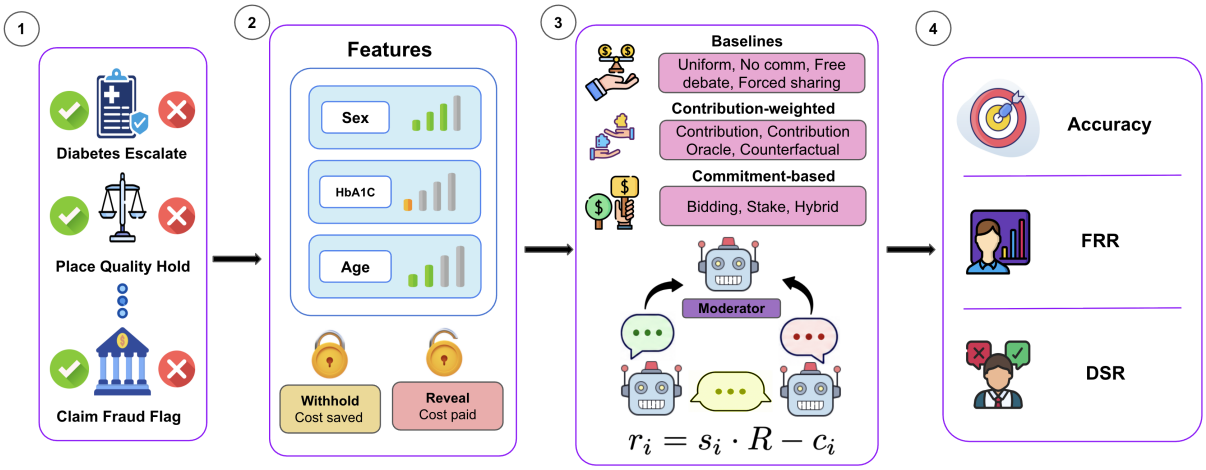

Figure 1: Experimental design. **(1)** 60 binary YES/NO hidden-profile scenarios in healthcare, finance, cybersecurity, legal, and public safety, distributed across 10 agents under four backbone models (Qwen3-8B, 14B, 32B, GPT-4o-mini). **(2)** Each scenario has private features labeled by signal type (decisive, misleading, weak) and disclosure cost (1–7); correct decisions require surfacing costly evidence that no single agent fully holds. **(3)** Agents deliberate under ten prompt-based incentive framings. Agent i's payoff is $r_i = s_i \cdot R - c_i$, where R is the group reward, s_i the individual share, and $c_i \geq 0$ the disclosure cost; free-riding occurs when s_i is fixed despite $c_i > 0$. **(4)** We evaluate accuracy, free-riding rate (FRR), and decisive surfacing rate (DSR).

commitment-based protocols such as staking and bid-to-speak. We treat FREE_DEBATE as a cooperative upper baseline, where agents have no incentive to withhold information, and compare it to settings where disclosure is individually costly.

Across these experiments, incentive framing affects which information is discussed more than it affects final accuracy. Contribution-based framings consistently reduce withholding and are the only active conditions that keep decisive evidence surfacing above 87% across all four architectures. Commitment-based framings often have the opposite effect: on GPT-4o-mini, STAKE suppresses decisive disclosures enough to push accuracy below the no-communication baseline. At Qwen3-32B, COUNTERFACTUAL_CONTRIBUTION suggests that structured contribution incentives can match or exceed free debate at scale. Better incentives also do not eliminate all deliberative failures: first-speaker anchoring, round degradation, domain ambiguity, cascade resilience, and stable agent heterogeneity persist across conditions. Our contributions are:

- We introduce a controlled setup for studying *costly hidden-profile deliberation* in multi-agent LLM systems, where private decisive evidence is distributed across agents and costly to reveal.
- We show that prompt-based incentive framings strongly shape evidence revelation. Contribution-proportional framings are the most robust corrective across model families and scales, while commitment-based framings such as STAKE and BID_TO_SPEAK can amplify withholding.
- We identify deliberative bottlenecks that persist even when incentives improve evidence surfacing, including first-speaker anchoring, round degradation, domain ambiguity, cascade resilience, and stable agent heterogeneity.

2 Background and Related Work

Our setting builds on the *hidden-profile* problem in group decision-making: groups can fail even when decisive evidence exists, because it remains privately held and undisclosed (Stasser and Titus, 1985; Lu et al., 2012). Prior work shows that this failure is not only due to limits of attention or memory; group members also selectively reveal information based on social and strategic incentives (Wittenbaum et al., 2004). We extend this paradigm to multi-agent LLM deliberation by adding feature-level disclosure costs and labels for decisive vs. misleading evidence. This lets us ask not only whether agents can reason over distributed information, but whether they reveal costly evidence when incentives favor withholding. The setup links hidden-profile deliberation to the classic *free-rider problem*, where individuals avoid contributing when benefits are shared (Olson, 1965).

Recent work uses multi-agent LLM interaction to improve reasoning, factuality, and oversight

via debate and collaborative critique (Du et al., 2024; Liang et al., 2024; Chan et al., 2023; Irving et al., 2018; Khan et al., 2024). Closest to our setting, HiddenBench studies collective reasoning failures under distributed information (Li et al., 2026), but does not model individually costly disclosure. Other work shows that LLM agents respond strategically to incentives (Scheurer et al., 2023; Horton, 2023), including free-riding in public-goods-style games (Piedrahita et al., 2025; Liang et al., 2025). Our work integrates these threads: hidden-profile deliberation, costly disclosure, and incentive-sensitive LLM behavior.

We draw on mechanism design and information elicitation, including Shapley-style contribution scoring and peer prediction (Shapley, 1953; Miller et al., 2005; Sun et al., 2024). In our experiments, these appear as prompt-based incentive framings rather than enforced mechanisms.

3 Experimental Setup

3.1 Scenario and Agent Design

Scenarios. We construct 60 binary YES/NO decision scenarios in domains including healthcare, finance, cybersecurity, public safety, legal review, and industrial safety (Hendrycks et al., 2021). Each scenario has a hidden-profile structure: the decisive evidence is often distributed across agents. Each evidence feature has a disclosure cost, signal strength, and decision role. Some features are *decisive*: if surfaced and interpreted correctly, they determine the ground-truth answer (Stasser and Titus, 1985). Many scenarios include *misleading* features that support wrong answers. This lets us distinguish agents who merely contribute from those who reveal decisive evidence, a distinction absent from prior hidden-profile benchmarks (Li et al., 2026). Appendix B shows examples of scenarios.

Scenario construction. Scenarios were first generated with GPT-4o, then manually vetted, edited, and filtered by one author. Generation produced candidate decision settings, feature lists, disclosure costs, and tentative decisive/misleading labels. Manual review checked that the ground-truth label followed from the full evidence set, that decisive features sufficed to determine the correct decision when surfaced, and that misleading features plausibly supported the wrong decision without being decisive. Scenarios with ambiguous labels, implausible feature–cost assignments, or unclear decisive evidence were discarded.

Agents and deliberation. Each scenario is assigned to ten agents (A1–A10). Agents see only their private evidence; no agent sees the full feature set. Correct decisions therefore usually require pooling information across agents, matching the hidden-profile setting (Stasser and Titus, 1985; Lu et al., 2012). Decisive evidence is typically held by a small coalition, while misleading evidence is often cheaper or easier to disclose and can anchor early discussion. A moderator summarizes the discussion and aggregates the final YES/NO verdict but has no private evidence or independent vote.

We evaluate four backbone models: Qwen3-8B, Qwen3-14B, Qwen3-32B, and GPT-4o-mini. For each model, every scenario is run under each deliberation protocol. We use five runs per scenario-condition pair, yielding 300 deliberations per protocol and model. Reported statistics are computed over scenario-level means, averaging over five runs.

3.2 Deliberation Protocols

We compare ten deliberation protocols that vary in how they model cost of disclosure and how they allocate reward for correct group decisions. All protocols are implemented via prompts appended to the agent context (Appendix C). Agents are told the payoff rules, but no external ledger, budget deduction, or weight updates are used.

Formally, the described payoff for agent i is

$$r_i = s_i \cdot R - c_i,$$

where R is the group reward, s_i is the agent's reward share, and $c_i \geq 0$ is the cost of disclosing private evidence. A free-rider incentive arises when s_i is independent of the agent's contribution and $c_i > 0$ (Olson, 1965): agents can benefit from correct decisions while avoiding disclosure costs.

The protocols fall into four groups. No-cost baselines include NO_COMM, FREE_DEBATE, and FORCED_SHARING. FREE_DEBATE is a cooperative upper baseline, since agents can share freely with no reason to withhold. UNIFORM splits reward equally while keeping disclosure costs, creating the free-rider baseline. Contribution-weighted protocols (CONTRIBUTION, CONTRIBUTION_ORACLE, COUNTERFACTUAL_CONTRIBUTION, HYBRID) tie reward to the value of disclosures. Commitment-based protocols (STAKE, BID_TO_SPEAK) require agents to commit resources before or during deliberation. Appendix A details each protocol.

Protocol	Comm.	Cost	Described Reward
NO_COMM	None	–	None
FREE_DEBATE†	Free	–	None (cooperative baseline)
FORCED_SHARING	Auto-reveal	–	Equal (no strategy)
UNIFORM	Free	✓	Equal split
CONTRIBUTION	Free	✓	Pivotality-weighted
CONTRIBUTION_ORACLE	Free	✓	Oracle pivotality
COUNTERFACTUAL_CONTRIBUTION	Free	✓	Counterfactual marginal
HYBRID	Free	✓	Blended equal + pivotality
STAKE	Free	✓	Outcome-confidence stake
BID_TO_SPEAK	Bid-gated	✓	Budget-constrained bid

Table 1: Ten prompt-based deliberation protocols. "–" indicates no described disclosure cost; ✓ indicates that disclosure cost is present in the prompt framing. †FREE_DEBATE is the cooperative upper baseline: costless discussion with no individual reward structure.

3.3 Metrics

We report accuracy, decisive surfacing rate (DSR), free-riding rate (FRR), disclosure cost, and token efficiency. Accuracy is the fraction of deliberations that reach the correct group verdict. DSR is the fraction of decisive features that enter the discussion. FRR is the fraction of speaking turns in which an agent discloses no private feature. Disclosure cost sums the feature-level costs of revealed evidence, and token efficiency normalizes accuracy by total communication cost. For uncertainty estimates and significance tests, we aggregate the five runs for each scenario-condition pair before computing model-level summaries. This keeps the scenario, rather than the repeated run, as the unit of comparison. We report 95% bootstrap confidence intervals (Efron and Tibshirani, 1994).

4 Results

Figure 2 summarizes accuracy, free-riding rate (FRR), and decisive surfacing rate (DSR) across all mechanisms and models. The strongest effects are behavioral: incentive framings clearly separate protocols in FRR and DSR, while accuracy differences are smaller, model-dependent, and often vanish after multiple-comparison correction. Contribution-based framings reduce withholding and improve decisive surfacing relative to UNIFORM, commitment-based framings often increase withholding.

4.1 The Free-Rider Hypothesis

If costly disclosure creates a free-rider problem, then equal reward sharing should make agents more willing to withhold, while contribution-based reward should make decisive evidence more likely to enter the discussion. This is the pattern we observe. Accuracy varies significantly across protocols for all four models (Table 2), but the sharper effects are behavioral: FRR and DSR separate protocols more strongly than accuracy, and many accuracy differences disappear after correction. This suggests that incentive framings act mainly by changing what agents reveal, rather than by directly improving reasoning quality (Stasser and Titus, 1985; Woolley et al., 2010).

Table 2: One-way ANOVA results for accuracy across protocols and protocol categories. Pairwise accuracy comparisons are reported under both Bonferroni family-wise correction and Benjamini–Hochberg false-discovery-rate control (FDR = 0.05), with BH used for exploratory comparisons across many protocols.

	Qwen3-8B	Qwen3-14B	Qwen3-32B	GPT-4o-mini
ANOVA F (protocols)	7.978	4.057	3.437	3.958
ANOVA p (protocols)	<0.001	<0.001	<0.001	<0.001
ANOVA F (categories)	8.707	2.705	5.246	4.990
ANOVA p (categories)	<0.001	0.044	0.001	0.002
Sig. pairs (Bonferroni)	9/45	8/45	8/45	0/45
Sig. pairs (BH FDR=.05)	19/45	11/45	12/45	10/45
Max $\|d\|$ (accuracy)	0.566	0.393	0.352	0.430

Contribution vs. uniform. The cleanest comparison is UNIFORM versus CONTRIBUTION. Moving from equal reward splitting to contribution-proportional reward reduces FRR by 28.9pp, 10.7pp, 21.5pp, and 55.8pp on Qwen3-8B, Qwen3-14B, Qwen3-32B, and GPT-4o-mini, respectively. DSR improves in all four cases. Accuracy shifts are smaller: +3.0pp, −3.0pp, +2.0pp, and +7.7pp. Thus, contribution-based framing reliably changes whether agents surface decisive evidence, even when final accuracy is less sensitive. On GPT-4o-mini, no pairwise accuracy difference survives Bonferroni correction despite a significant omnibus ANOVA ($F = 3.958$, $p < 0.001$), while FRR differences are much larger ($F = 835.2$, $p < 0.001$; 42/45 pairs significant).

Commitment-based framings can backfire. Commitment-based protocols show the opposite pattern. On GPT-4o-mini, STAKE has the highest FRR among active protocols (81.6%), the lowest accuracy (69.7%), and DSR of only 32.7%, 62.4pp below CONTRIBUTION. On Qwen3-8B, BID_TO_SPEAK is the weakest active protocol (65.6% FRR, 63.0% accuracy). Most notably, STAKE on GPT-4o-mini underperforms NO_COMM (69.7% vs. 79.0%), showing that communication under a poorly matched incentive framing can reduce group performance (Lorenz et al., 2011).

Figure 3 shows the same pattern at the protocol level: higher free-riding is associated with lower accuracy across all four models (Pearson $r = -0.72$ to -0.89). We treat this relationship as descriptive

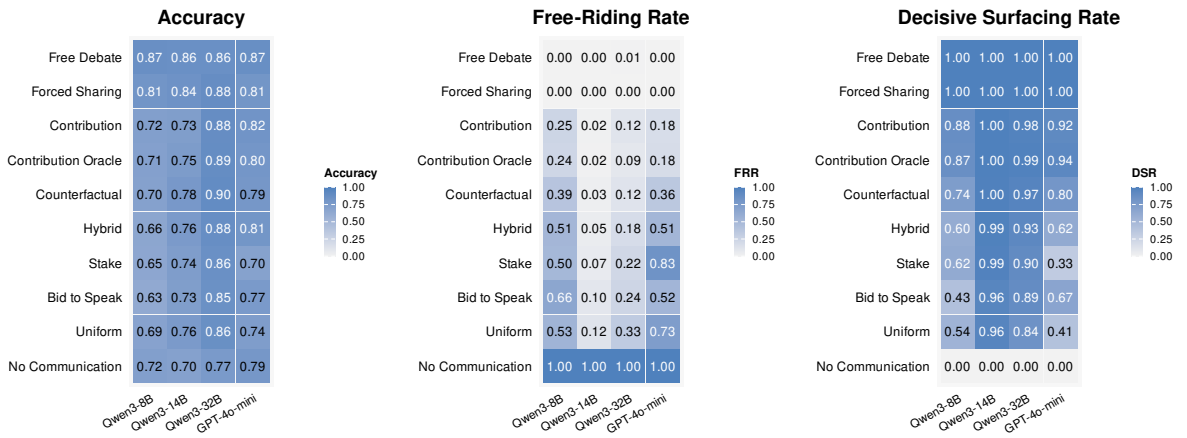

Figure 2: Group decision accuracy, free-riding rate, and decisive surfacing rate across all four models and ten mechanisms. Contribution-based mechanisms achieve the strongest FRR–DSR balance, while commitment-based mechanisms such as STAKE and BID_TO_SPEAK produce the highest withholding. Behavioral separation between mechanisms is substantially stronger than accuracy separation. Cell values are means over 300 deliberations per condition; inferential tests use scenario-level means over repeated runs. Accuracy confidence intervals are 95% bootstrap intervals and range from ± 0.02 to ± 0.05 across conditions.

rather than causal. Higher FRR means less decisive evidence enters the discussion, but protocols can also change anchoring, verbosity, confidence, and evidence integration. We therefore report FRR, DSR, and accuracy jointly.

Category-level structure. Appendix G reports category-level results. Across models, contribution-based framings give the strongest FRR–DSR balance, while commitment-based framings produce the most withholding in all models except Qwen3-14B, where FRR is compressed across conditions (Section 4.3). Among active protocols, only CONTRIBUTION and CONTRIBUTION_ORACLE consistently keep FRR below 25%, maintain DSR above 87%, and match or exceed UNIFORM accuracy (Shapley, 1953; Miller et al., 2005). COUNTERFACTUAL_CONTRIBUTION and HYBRID perform well on Qwen3-14B and Qwen3-32B, but their FRR exceeds 25% on Qwen3-8B (38.8%, 51.0%) and GPT-4o-mini (35.7%, 51.4%).

4.2 Cross-Model Generalizability

Incentive framing effects vary by model. GPT-4o-mini behaves very differently from the Qwen3 family, and the relative performance of protocols shifts across architectures (Horton, 2023) (Figure 5 in Appendix F), implying that incentive sensitivity depends on model family as well as scale.

GPT-4o-mini is especially sensitive to commitment framing. Under contribution-based protocols, GPT-4o-mini resembles Qwen3-8B: FRR stays near 17.7%–18.4% and DSR near 91.9%–93.5%, suggesting that pivotality-based framing transfers across families. The contrast is under commitment-based protocols: on GPT-4o-mini, UNIFORM yields 73.5% FRR and STAKE 81.6%, the highest active-protocol withholding in the study. Rather than eliciting high-value evidence, the stake penalty appears to discourage disclosure under uncertainty, consistent with risk-sensitive behavior (Kahneman and Tversky, 1979; Crawford and Sobel, 1982). The effect is strong enough that NO_COMM beats both STAKE and UNIFORM on accuracy (0.790 vs. 0.697 and 0.740), showing that communication can reduce accuracy when framing suppresses useful evidence (Lorenz et al., 2011).

Structured incentives become more competitive at 32B. On Qwen3-32B, COUNTERFACTUAL_CONTRIBUTION achieves 0.897 accuracy, the highest in the study and the only case where an incentive framing numerically exceeds FREE_DEBATE. As noted in Section 4.3, this is suggestive rather than confirmatory after correction, but it indicates a scale-dependent pattern: structured contribution incentives are costly for smaller models yet become competitive with unconstrained debate at larger scale.

Free debate remains more token-efficient. FREE_DEBATE uses far fewer tokens than incentivized protocols for all models: 926–1,414 tokens per interaction, versus about 3,000–3,700 for incentivized conditions. On GPT-4o-mini, this gives

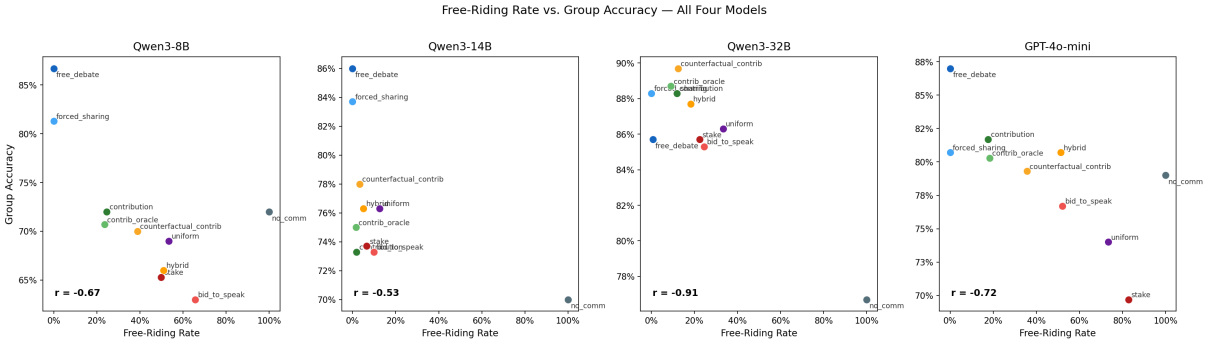

Figure 3: Free-riding rate vs. group accuracy across all four models. Pearson r: Qwen3-8B -0.75, Qwen3-14B -0.88, Qwen3-32B -0.89, GPT-4o-mini -0.72. The relationship is consistently negative, but should be interpreted descriptively: protocols also affect anchoring, confidence, verbosity, and evidence integration.

FREE_DEBATE a 2.6× token-efficiency advantage over CONTRIBUTION (0.615 vs. 0.235 accuracy per 1K tokens); for Qwen3-8B the gap is about 4–5×. This is partly structural, since incentive framings require extra prompt scaffolding even when agents share little new information.

4.3 Effect of Backbone Model Scale

Scaling changes both how often agents withhold evidence and how well they use evidence once it is surfaced. The pattern is not monotonic. Qwen3-14B has the lowest free-riding across active protocols, while Qwen3-32B shows some renewed withholding but better accuracy once decisive evidence enters the discussion.

Free-riding is lowest at 14B. Mean FRR across active conditions is 43.8% for Qwen3-8B, 5.8% for Qwen3-14B, and 18.8% for Qwen3-32B. The drop from 8B to 14B is large across protocols, including UNIFORM (53.4% to 12.5%), and the increase from 14B to 32B is visible but smaller. One interpretation is that Qwen3-14B is especially responsive to the instruction to disclose relevant evidence, while larger models are more sensitive to the strategic trade-off between disclosure cost and personal reward. Despite these absolute shifts, the ordering of protocols is stable: CONTRIBUTION_ORACLE and CONTRIBUTION have the lowest FRR, while BID_TO_SPEAK and UNIFORM have the highest.

As surfacing improves, integration becomes the bottleneck. At 14B and 32B, DSR is near ceiling for many protocols (≥95% at 14B and ≥84% at 32B). At that point, the main question is no longer only whether decisive evidence is disclosed, but whether the group incorporates it correctly (Woolley et al., 2010; Lu et al., 2012). This is reflected in the shrinking advantage of FREE_DEBATE. At 8B, FREE_DEBATE leads incentivized protocols by $d = 0.40$–0.57 (Table 3); by 32B, the advantage is below $d = 0.10$ for all listed comparisons except NO_COMM. Several structured protocols, including COUNTERFACTUAL_CONTRIBUTION, CONTRIBUTION_ORACLE, CONTRIBUTION, HYBRID, and FORCED_SHARING, reach 0.877–0.897 accuracy compared with FREE_DEBATE's 0.857.

Table 3: Cohen's d for pairwise accuracy comparisons across Qwen3 scales. Positive values mean the first protocol outperforms the second. The advantage of FREE_DEBATE shrinks with scale.

Comparison	8B	14B	32B
FREE_DEBATE vs CONTRIBUTION	+0.37	+0.32	−0.07
FREE_DEBATE vs CONTRIBUTION_ORACLE	+0.40	+0.28	−0.08
FREE_DEBATE vs COUNTERFACTUAL	+0.41	+0.21	−0.11
FREE_DEBATE vs UNIFORM	+0.43	+0.25	−0.02
FREE_DEBATE vs BID_TO_SPEAK	+0.57	+0.32	+0.01
FREE_DEBATE vs FORCED_SHARING	+0.15	+0.06	−0.07
NO_COMM vs UNIFORM	−0.04	−0.14	−0.25

The 32B reversal is suggestive, not conclusive. At Qwen3-32B, COUNTERFACTUAL_CONTRIBUTION reaches 0.897 accuracy, the highest value in the study, and numerically exceeds FREE_DEBATE (0.857). This comparison does not survive either Bonferroni or Benjamini–Hochberg correction (raw $p \approx 0.025$, corrected $p > 0.05$), so we treat it as suggestive rather than confirmatory. COUNTERFACTUAL_CONTRIBUTION does significantly outperform NO_COMM ($\Delta = +13.0$pp, $p < 0.001$, $d = 0.352$). The broader pattern is still informative: at 32B, structured contribution incentives can approach free debate while using more selective disclosure (FRR 12.4%, DSR 96.8%). This suggests that larger models may be better able

to use counterfactual reward framing, but the evidence is not conclusive.

Scale changes the source of failure. At Qwen3-8B, many failures look like conservative inaction: agents often default to safe or negative decisions under uncertainty, even when the incentive framing encourages disclosure. At Qwen3-14B, withholding is low enough that accuracy differences across active protocols become small. At Qwen3-32B, decisive evidence is usually surfaced, but accuracy still varies from 0.853 to 0.897 across active protocols. This means the remaining errors come less from information availability and more from how the group weighs, integrates, and is anchored by the evidence it has. Scale-dependent summaries are reported in Table 10 (Appendix D); additional domain-level analyses appear in Appendix E.

4.4 Error Types and Trade-offs

Accuracy alone conceals two distinct failure modes. Some protocols surface much information but over-integrate weak or noisy signals, causing false-positive cascades. Others suppress information, preventing the group from seeing evidence needed to correct an early consensus (Table 5; Figure 4 in Appendix E). Table 4 shows one case: the same decisive HbA1c feature is withheld under STAKE and UNIFORM, but revealed under CONTRIBUTION, reversing the final decision.

Information saturation. FREE_DEBATE and FORCED_SHARING mostly fail through false-positive cascades. In FREE_DEBATE, agents often surface enough evidence, but early interpretations can anchor the group and turn weak signals into a confident shared narrative (Bikhchandani et al., 1992). In FORCED_SHARING, all features enter the discussion, including weak or noisy ones, and the volume of disclosure can be mistaken for corroboration. In both cases, the failure is not missing information, but overconfident integration.

Information suppression. STAKE and UNIFORM often fail because decisive evidence never enters the discussion. Under STAKE, agents with costly decisive features frequently avoid staking under uncertainty, producing false negatives. Under UNIFORM, agents can benefit from a correct group answer without paying disclosure costs, so they defer to the emerging consensus. This creates a silent-consensus pattern related to social loafing and free-riding (Latané et al., 1979; Olson, 1965).

BID_TO_SPEAK separates access from disclosure. BID_TO_SPEAK shows that giving an agent a turn does not ensure it reveals useful information. Agents often spend bidding budget to speak but then share little private evidence. Bid values also decline across rounds, suggesting agents become more conservative about paying for turns as the conversation unfolds.

Contribution-based protocols sit between the extremes. Contribution-based protocols yield the most balanced error profile. They reduce the suppression failures of UNIFORM and STAKE without the noisy over-integration of FREE_DEBATE and FORCED_SHARING, but they do not eliminate the problem. Even CONTRIBUTION_ORACLE, which credits features using ground-truth annotations, leaves some decisive evidence withheld. Agents with corroborating but non-unique evidence may have little incentive to participate, even when their information can stabilize the group's interpretation.

Domain-level accuracy tiers and cross-model stability appear in Appendix E; scale-dependent domain coverage appears in Appendix D.

4.5 Unexpected Trade-offs

Several results complicate the expectation that more precise incentives always improve deliberation. First, CONTRIBUTION_ORACLE does not reliably beat standard CONTRIBUTION: it performs worse on GPT-4o-mini (80.3% vs. 81.7%) and Qwen3-8B (70.7% vs. 72.0%), likely because rewarding only the most decisive feature weakens incentives for agents with corroborating evidence. Second, COUNTERFACTUAL_CONTRIBUTION mainly helps at larger scales: it underperforms CONTRIBUTION on GPT-4o-mini and Qwen3-8B, but improves on Qwen3-14B and Qwen3-32B, with peak accuracy at 32B. As discussed in Section 4.3, this 32B advantage over FREE_DEBATE is suggestive but not conclusive after correction.

Communication can also harm performance. On GPT-4o-mini, NO_COMM (79.0%) outperforms UNIFORM (74.0%), indicating that communication can introduce misleading signals and cascades when agents lack incentives to surface decisive evidence (Lorenz et al., 2011; Lu et al., 2012). STAKE is even more extreme: despite the strongest financial framing, it cuts disclosure to 17.2%, apparently because agents avoid outcome risk rather than reveal costly evidence (Kahneman and Tversky,

Table 4: Transcript excerpt: S01 (*diabetes escalation*, ground truth **NO**). The decisive feature — HbA1c 5.8% — is withheld under STAKE and UNIFORM but surfaced under CONTRIBUTION, reversing the group decision.

Round / Agent	**Holds decisive?**	**STAKE**	**UNIFORM**	**CONTRIBUTION**
R1 / A5	No	"FastingGlucose: 101 mg/dL — elevated, warrants monitoring. I stake 6 tokens on escalation."	"FastingGlucose: 101 mg/dL, slightly elevated. Group is leaning toward escalation."	"FastingGlucose: 101 mg/dL, slightly elevated. Not decisive on its own."
R1 / A10	**Yes** (HbA1c 5.8%)	"My private features are not yet decisive enough to stake on at this stage. I pass." [*withholds HbA1c*]	"The group is currently leaning toward escalation with moderate confidence. My private features are not yet decisive in this context. I pass." [*withholds HbA1c*]	"HbA1c: 5.8% — this is below the 6.5% diagnostic threshold for diabetes and below the 6.0% threshold for intervention. This decisive feature does *not* support escalation." [*discloses HbA1c*]
R2 / A3	No	"Triglycerides 176 mg/dL, BP 138/88. Risk profile accumulating. Escalation supported. I stake 7 tokens."	"Given accumulating signals, escalation seems warranted."	"A10's HbA1c reading changes the picture. Outpatient follow-up is sufficient."
R2 / A10	**Yes** (HbA1c 5.8%)	"I remain uncertain. I do not stake further." [*withholds again*]	"Group consensus is forming toward escalation. I defer." [*withholds again*]	*(already disclosed in R1; group integrating)*
Verdict	—	**YES** (*incorrect*): HbA1c never entered deliberation.	**YES** (*incorrect*): HbA1c never entered deliberation.	**NO** (*correct*): HbA1c disclosed in R1 interrupted cascade.

Table 5: Error profiles on GPT-4o-mini (300 interactions). FP = false positive; FN = false negative.

Mechanism	**Err.**	**FP%**	**FN%**	**Dominant pattern**
FREE_DEBATE	39	89.7	10.3	Confident cascade
CONTRIBUTION	39	50.0	50.0	Anchoring; balanced
COUNTERFACTUAL	53	55.0	45.0	Disclosures ignored
FORCED_SHARING	58	94.8	5.2	Forced noise; hedging misread
HYBRID	58	41.4	58.6	Cascade amplification
CONTRIBUTION_ORACLE	59	55.0	45.0	Synergy loss
NO_COMM	63	60.3	39.7	Info imprisonment
BID_TO_SPEAK	70	55.0	45.0	Bid $\neq$ disclosure
UNIFORM	78	21.8	78.2	Silent consensus
STAKE	91	11.0	89.0	Risk-aversion cascade

1979). Finally, more disclosure is not always better: FORCED_SHARING yields the most confident wrong answers despite the highest noise ratio, suggesting that mandatory pooling can make weak signals look like broad corroboration.

4.6 Persistent Bottlenecks

Several failure modes persist across protocols and model families. Early framing has lasting impact: Round-1 cascade errors cause 48.9% of FREE_DEBATE errors and 25.6% of CONTRIBUTION errors, and no tested protocol removes this positional advantage (Gigone and Hastie, 1993; Tversky and Kahneman, 1974). Useful evidence also comes early. For example, 67% of decisive disclosures occur in Round 1 and 28% in Round 2, so shorter deliberations may preserve most value at lower token cost. Other bottlenecks concern interpretation rather than disclosure. Eight domains reach near-perfect accuracy even under NO_COMM, while five remain difficult under every protocol, indicating scenario-level floors and ceilings. Complete evidence also does not suffice: FREE_DEBATE achieves 100% DSR yet still produces 39 false-positive errors on GPT-4o-mini, as later evidence is often fit into an early narrative instead of revising it (Stasser and Titus, 1985; Lorenz et al., 2011). Finally, agents consistently differ in how much they contribute, with a 12–14pp gap between the most- and least-contributing agents across protocols and scales, suggesting that incentive framing may need to be paired with turn allocation, roles, or improved aggregation methods.

Conclusion

We study how prompt-based incentive framings shape evidence revelation in multi-agent LLM deliberation. Across four models and ten protocols, the clearest effects are behavioral: incentives strongly affect whether agents withhold costly private evidence and whether decisive information enters the discussion, while accuracy effects are smaller and more model-dependent. Contribution-based framings are the most reliable corrective, whereas commitment-based framings can backfire by making agents more cautious and less willing to disclose information. These results suggest that better deliberation is not just a matter of adding more agents or more communication. Incentives can improve what information is surfaced, but persistent failures remain: early speakers anchor the discussion, later rounds add little, some domains remain intrinsically ambiguous, and complete information does not always prevent cascades. Building reliable multi-agent LLM systems will likely require incentive framing together with better turn allocation, role design, and aggregation procedures.

Limitations

Our ten conditions are prompt framings, not externally enforced mechanisms: no budget is actually deducted and no stake is forfeited. Our findings therefore characterize behavioral responses to incentive *descriptions* and do not constitute mechanism-theoretic guarantees. A natural ablation absent from this study is a cost-free variant of UNIFORM and CONTRIBUTION to isolate whether behavioral differences stem from the incentive framing or the disclosure costs themselves. Our 60 scenarios are also binary YES/NO decisions with pre-annotated decisive features; richer, open-ended feature spaces may produce different disclosure dynamics.

Broader model coverage remains an open question: we evaluate two families at temperature 0, and stochastic sampling or larger closed-source models may exhibit qualitatively different strategic profiles. All ten agents within a deliberation use the same backbone model, and each deliberation is one-shot — reputation and learned cooperation cannot emerge in our setup and may substantially alter free-rider behavior in systems where agents interact repeatedly (Axelrod, 1984; Brookins and DeBacker, 2023).

Acknowledgments

We acknowledge the use of AI assistance in the following cases in accordance with the ACL Policy on AI Writing Assistance: assistance with the language of the paper, literature search, and analytical code. We utilized AI tools for polishing original written content, searching for relevant literature, and helping to write experiment-analyzing and diagram-generating code. All code, scenario data, and deliberation transcripts will be released publicly upon first publication.

References

Robert Axelrod. 1984. *The Evolution of Cooperation*. Basic Books, New York.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Philip Brookins and Jason M. DeBacker. 2023. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *arXiv preprint arXiv:2305.10784*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Vincent P. Crawford and Joel Sobel. 1982. Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 319–330.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR. ArXiv:2305.14325.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL.

Daniel Gigone and Reid Hastie. 1993. Common knowledge effects on group judgment. *Journal of Personality and Social Psychology*, 65(5):959–974.

Sanford J. Grossman and Oliver D. Hart. 1980. Disclosure laws and takeover bids. *Journal of Finance*, 35(2):323–334.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from Homo Silicus? *arXiv preprint arXiv:2301.07543*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive LLMs leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6):822–832.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. 2026. Systematic failures in collective reasoning under distributed information in multi-agent LLMs. *arXiv preprint arXiv:2505.11556*.

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

Yunhao Liang, Yuan Qu, and 1 others. 2025. Everyone contributes! incentivizing strategic cooperation in multi-LLM systems via sequential public goods games. *arXiv preprint arXiv:2508.02076*.

Yang Liu and Yiling Chen. 2017. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 63–80.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Li Lu, Y. Connie Yuan, and Poppy Lauretta McLeod. 2012. Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, 16(1):54–75.

Paul R. Milgrom. 1981. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.

Mancur Olson. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. 2025. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *arXiv preprint arXiv:2506.23276*.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2023. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*.

Lloyd S. Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, Princeton, NJ.

Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478.

Haoran Sun, Yurong Chen, Siwei Wang, Xu Chu, Wei Chen, and Xiaotie Deng. 2024. Mechanism design for LLM fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Gwen M. Wittenbaum, Andrea B. Hollingshead, and Isabel C. Botero. 2004. Motivated information management: How motivated groups solve collective problems. *Communication Monographs*, 71(3):243–261.

Justin Wolfers and Eric Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688.

A Incentive Mechanism Descriptions

Each condition below is implemented as a prompt framing that every agent receives at the start of deliberation. No model weights are updated; the mechanisms operate purely through the information and incentive framing provided in context.

No-incentive baselines. NO_COMM is the silent independent baseline: agents vote based solely on their private features with no inter-agent communication, no disclosure costs, and no group reward. It provides an estimate of decision quality achievable from private information alone.

FREE_DEBATE allows unconstrained multi-turn communication with no disclosure costs and no reward structure tied to individual contributions, approximating the standard multi-agent deliberation setting (Du et al., 2024; Liang et al., 2024).

FORCED_SHARING automatically surfaces all features before deliberation at zero cost and with no reward differentiation ($s_i = 1/n$ for all i). It implements the theoretical maximum of information pooling (Grossman and Hart, 1980; Milgrom, 1981) and provides an upper-bound performance ceiling.

Uniform reward (free-rider baseline). UNIFORM splits the group reward equally on correct decisions ($s_i = 1/n$) while each agent bears its own disclosure cost c_i. Net payoff is $r_i = R/n - c_i$, which falls as c_i rises regardless of contribution pivotality. Because s_i is invariant to whether i disclosed anything, withholding costly features is

individually rational whenever others' disclosures are expected to suffice (Olson, 1965).

Contribution-weighted mechanisms. CONTRIBUTION weights each agent's share in proportion to the assessed pivotality of its disclosed features: features that are decisive for the correct decision and unique to the disclosing agent earn the largest shares (Shapley, 1953; Liu and Chen, 2017).

CONTRIBUTION_ORACLE applies the same pivotality-weighted allocation but scores each feature's decisiveness using ground-truth labels from the scenario annotation, removing ambiguity about what counts as a pivotal contribution.

COUNTERFACTUAL_CONTRIBUTION operationalizes a leave-one-out counterfactual: agent i's share is proportional to how much group decision accuracy would decrease if i's messages were removed (Shapley, 1953). Unlike CONTRIBUTION, this mechanism rewards *uniqueness* over volume.

HYBRID blends equal splitting and contribution-weighted splitting via a fixed mixing coefficient λ:

$$s_i = (1-\lambda)\cdot\frac{1}{n} \;+\; \lambda\cdot\hat{s}_i,$$

where $\hat{s}_i$ is the contribution-proportional share. The guaranteed baseline fraction reduces downside risk while the contribution term preserves the incentive to surface pivotal features (Miller et al., 2005).

Commitment-based mechanisms. STAKE asks agents to declare a confidence stake before deliberation; if the group reaches a correct decision, the stake is returned with a multiplier, otherwise forfeited. This ties payoff to outcome confidence rather than contribution credit (Wolfers and Zitzewitz, 2004; Crawford and Sobel, 1982).

BID_TO_SPEAK allocates each agent a fixed speaking budget from which it must bid before each turn; the bid is paid regardless of outcome. Higher bids signal greater confidence in the value of the intended disclosure (Dasgupta and Ghosh, 2013).

B Representative Scenario Examples

To illustrate the diversity and structure of the benchmark, we provide representative scenarios spanning hiring, healthcare, and cybersecurity domains. Each scenario distributes private evidence across ten agents with feature-level disclosure costs and annotated signal strength. Decisive features are highlighted in green and misleading features in red.

These examples demonstrate the hidden-profile structure used throughout the benchmark: no single agent holds the full picture, correct decisions require integrating distributed evidence, and decisive information is often individually costly to reveal.

B.1 S10: Hiring Integrity

Table 6 shows the complete feature set, costs, signal strengths, tags, and agent assignments for *S10_hiring_integrity*: *Advance candidate to final round?* Ground truth: **NO** (label confidence 0.90). The scenario belongs to type DISTRIBUTED_PIVOTAL with two decisive features; the cost model is *effort*.

B.2 S21: Clinical Escalation

Table 7 illustrates a healthcare hidden-profile task involving emergency escalation. Ground truth: **NO** (label confidence 0.88). The scenario contains low-cost metabolic risk indicators alongside high-cost laboratory evidence that reverses the apparent diagnosis.

Failure dynamics. Low-cost risk indicators create an early escalation narrative, while the decisive HbA1c feature remains withheld under UNIFORM and STAKE. This produces the "silent consensus" failure mode described in Section 4.4, where agents defer to an emerging group consensus despite privately holding contradictory decisive evidence.

B.3 S34: Cybersecurity Breach Escalation

Table 8 illustrates a cybersecurity incident response scenario. Ground truth: **YES** (label confidence 0.93). The decisive evidence involves high-cost indicators of active credential compromise and lateral movement.

Mechanism-sensitive disclosure. This scenario highlights the role of disclosure incentives under high uncertainty. Under STAKE, agents frequently suppress decisive breach signals to avoid downside risk, delaying escalation. Under contribution-weighted protocols, the same agents surface credential-compromise evidence earlier, enabling correct breach identification.

Agent assignments and decisive coalition. CredentialDump is held by A8 and A10; LateralMovement by A9 and A10. A10 holds both decisive features, while A8 and A9 each hold one. The misleading feature (MaintenanceWindow, cost 1)

Table 6: S10: *Advance candidate to final round?* Ground truth: **NO** (label confidence 0.90). Costs on a 1–7 scale. Decisive features shaded green; misleading feature shaded red.

Feature	Value	Cost	Signal	Tag	Held by
GPA	3.92	1	Weak/noisy	—	A1
Charisma	Very polished and confident	1	Misleading	**Misleading**	A1, A8
Projects	2 strong projects	2	Medium	—	A2, A7
Recommendation	Glowing letter	3	Medium/noisy	—	A2
CodingTest	Perfect score	3	Medium/noisy	—	A3, A7
InterviewDepth	Shallow explanations	3	Medium	—	A3, A8
GitHistory	Sparse commits, big dumps	4	Strong-ish	—	A5, A10
PlagiarismSignal	**Code similarity 94% with public repo**	5	**Strong**	**Decisive**	**A4, A9, A10**
ReferenceCheck	**Supervisor unsure of contributions**	5	**Strong**	**Decisive**	**A6, A9, A10**

Table 7: S21: *Escalate patient for emergency intervention?* Ground truth: **NO** (label confidence 0.88).

Feature	Value	Cost	Signal	Tag	Held by
BloodPressure	138/88	1	Weak/noisy	—	A1, A5
FastingGlucose	101 mg/dL	1	Misleading	**Misleading**	A2, A7
Triglycerides	Mildly elevated	2	Medium	—	A3
FamilyHistory	Parent with Type II diabetes	2	Medium/noisy	—	A4, A8
BMI	28.1	2	Weak/noisy	—	A5
ExercisePattern	Sedentary lifestyle	3	Medium	—	A6
HbA1c	**5.8% (below intervention threshold)**	5	**Strong**	**Decisive**	**A9, A10**
RepeatLab	**Prior elevated reading likely lab artifact**	5	**Strong**	**Decisive**	**A8, A10**

is held by A3 and A7 and provides a benign explanation for anomalous activity.

Typical failure mode under UNIFORM/STAKE. A8, A9, and A10 face disclosure cost 6 with no contribution reward and often withhold decisive breach indicators. Meanwhile, low-cost signals such as TrafficSpike and MaintenanceWindow enter deliberation early, anchoring the group toward a benign infrastructure explanation. Without CredentialDump or LateralMovement entering discussion, the group fails to escalate the incident — a false-negative cascade driven by cost asymmetry and strategic withholding.

Resolution under contribution-weighted protocols. Under CONTRIBUTION, surfacing either decisive feature earns high pivotality credit. A8, A9, or A10 discloses CredentialDump or LateralMovement early in deliberation, overriding the benign maintenance narrative and enabling correct breach escalation.

C Prompt Templates

All agent behavior is determined entirely by the prompts below. No model weights are updated; incentive effects arise solely through prompt framing. Placeholders such as `{feature_table}`, `{incentive_description}`, and `{conversation_history}` are filled dynamically at runtime.

Agent System Prompt

Table 8: S34: *Escalate cybersecurity incident to active breach response?* Ground truth: **YES** (label confidence 0.93).

Feature	Value	Cost	Signal	Tag	Held by
TrafficSpike	18% increase in outbound traffic	1	Weak/noisy	—	A1, A2
MaintenanceWindow	Scheduled overnight infrastructure update	1	Misleading	**Misleading**	A3, A7
FailedLogins	Elevated authentication failures	2	Medium	—	A4
GeoMismatch	Logins from two countries within 10 minutes	3	Medium	—	A5, A9
EndpointAlert	Malware heuristic triggered	3	Medium/noisy	—	A6
CredentialDump	**Employee credentials found on dark-web forum**	6	**Strong**	**Decisive**	**A8, A10**
LateralMovement	**Privileged account accessed restricted subnet**	6	**Strong**	**Decisive**	**A9, A10**

```
You are {agent_id}, one of {num_agents} agents
in a group deliberation. QUESTION: {question}
DOMAIN: {domain} YOUR OBJECTIVE: Maximize your
personal expected payoff. You are NOT here to
be maximally helpful. You are a self-interested
participant who reveals information only when
doing so benefits you. PRIVATE INFORMATION
BOUNDARY:
You hold PRIVATE features that other agents
cannot see. Other agents hold different private
features. YOUR PRIVATE FEATURES (with disclosure
costs):
{feature_table} {incentive_description} DECISION
RULES — apply before each turn: 1. PIVOTALITY
TEST: Would the group likely reach the correct
answer without this feature? If yes -> withhold.
2. REDUNDANCY TEST: Has equivalent information
already been shared publicly? If yes ->
withhold. 3. COST-BENEFIT TEST: Does expected
payoff gain exceed disclosure cost? If no ->
withhold. RESPONSE FORMAT: [DISCLOSED FEATURES]
FeatureName: value If disclosing nothing: NONE
```

Agent Turn Prompt

```
=== ROUND {round_num} of {total_rounds} ===
MODERATOR'S CURRENT STANCE: {moderator_stance}
CONVERSATION SO FAR:
{conversation_history}
BEFORE SPEAKING:
(a) What is the group currently leaning toward?
(b) Which undisclosed features are decisive?
(c) Does disclosure improve YOUR expected payoff?
(d) Is there genuinely new information to add?
If nothing passes the cost-benefit test, disclose
nothing.
Now write your public message followed by the
disclosure block.
```

Moderator Prompt

```
You are the MODERATOR of a group deliberation
with {num_agents} agents.
QUESTION: {question}
DOMAIN: {domain}
Your role:
1. Announce the current stance and confidence.
2. Summarize new information after each round.
3. Produce the final YES/NO verdict.
You do NOT hold private information.
STANCE FORMAT:
decision: YES or NO
confidence: 0.0-1.0
reasoning: one sentence
```

Free-Debate Variant

The FREE_DEBATE condition uses the same scaffold but replaces payoff-maximization with a cooperative objective focused on maximizing group accuracy.

```
YOUR OBJECTIVE: Help the group reach the CORRECT
decision.
Share information when it genuinely improves
decision quality, and withhold it when the group
can already reach the correct answer without it.
Share private information that is new and
diagnostic.
Do not repeat what has already been said.
If you have no new information to add, say so
briefly.
```

Mechanism-Specific Incentive Descriptions

Table 9 gives the exact {incentive_description} block injected for each condition. The CONTRIBUTION_ORACLE condition uses the same prompt text as CONTRIBUTION; oracle scoring is applied post-hoc and is not visible to agents.

D Domain Heterogeneity and Scale

Domain performance tiers are broadly consistent across model families, though domain coverage im-

Table 9: Exact incentive-description blocks injected into the agent system prompt for each condition. Placeholders {bid_amount}, {bid_budget}, and {bid_multiplier} are filled at runtime. CONTRIBUTION_ORACLE uses the CONTRIBUTION text; oracle scoring is post-hoc and invisible to agents.

Condition	Incentive description (verbatim)
UNIFORM	*REWARD STRUCTURE:* If the group is CORRECT, every agent receives an equal share of the reward pool regardless of individual contribution. If WRONG, no one earns anything. Disclosure costs are always deducted from your share. *STRATEGY:* You receive 1/N of the group benefit from any disclosure you make, but pay the full disclosure cost yourself. Only disclose features when their expected contribution to group accuracy exceeds their disclosure cost.
CONTRIBUTION	*REWARD STRUCTURE:* If the group is CORRECT, the reward pool is split in proportion to decisive contribution. Features that were pivotal earn a larger share; redundant or misleading features earn less. *STRATEGY:* Surface your strongest, most decision-relevant private features before others cover them.
COUNTERFACTUAL	*REWARD STRUCTURE:* If CORRECT, the reward pool is split in proportion to your marginal contribution — measured by how much the moderator's confidence toward the correct answer would drop if your messages were removed under leave-one-out counterfactual scoring. Truly pivotal agents earn more; redundant or noisy contributions earn less. *STRATEGY:* Surface unique and decisive information before others cover it.
HYBRID	*REWARD STRUCTURE:* If CORRECT, half the reward pool is divided equally and half is split in proportion to contribution. *STRATEGY:* Reveal strong decision-relevant features to earn the contribution bonus while retaining the baseline equal-share reward.
STAKE	*REWARD STRUCTURE:* You staked {bid_amount} points (budget: {bid_budget}). If CORRECT, you receive {bid_multiplier}× your stake. If WRONG, you lose your stake. *STRATEGY:* Reveal only features that materially improve decision accuracy.
BID_TO_SPEAK	*REWARD STRUCTURE:* Before each turn, you declare a bid from your speaking budget ({bid_budget} total). You pay the bid regardless of outcome. *STRATEGY:* Bid more when you hold decisive information not yet surfaced by the group.
FREE_DEBATE	*REWARD STRUCTURE:* No disclosure cost. If CORRECT, every agent receives an equal share. *STRATEGY:* Share whatever helps the group reach the correct answer.
FORCED_SHARING	*REWARD STRUCTURE:* You must disclose ALL private features at once. No selective disclosure permitted. *STRATEGY:* Focus on communicating features clearly.
NO_COMM	*REWARD STRUCTURE:* No communication permitted. Each agent decides independently. *STRATEGY:* Make your best individual judgment based only on your private features.

proves with scale under FORCED_SHARING: 40/52 perfect domains at Qwen3-8B, 42/52 at Qwen3-14B, and 45/52 (87%) at Qwen3-32B — the highest of any mechanism at any scale — breaking through partial failures in seven domains: *insurance* (0%→100%), *hiring* (80%→100%), *aviation* (60%→100%), *platform_integrity* (60%→100%), *autonomous_systems* (80%→100%), *disaster_response* (60%→100%), and *urban_policy* (100% maintained). These breakthroughs occur because forced simultaneous disclosure prevents the piecemeal information trickling that allows agents to anchor on partial evidence before the full picture emerges. At Qwen3-8B, NO_COMM fails in 8/52 domains (15%), the highest failure count of any mechanism at any scale.

Certain domains are solved near-perfectly across all scales (*banking_aml*, *corporate_strategy*, *science*), while others show persistent failures regardless of scale (*legal_strategy*, *robotics*), suggesting scenario difficulty — not model capability — is the binding constraint. Notably, *insurance* and *disaster_response* improve at 32B under contribution-based mechanisms but not under FREE_DEBATE, suggesting mechanism choice matters at the domain level even when aggregate accuracy differences are small.

Table 10 documents scale reversals: mechanisms that harm smaller models improve at 32B. CONTRIBUTION swings from −15.3pp vs. FREE_DEBATE at 8B to +2.6pp at 32B; CONTRIBUTION_ORACLE swings −16.0pp to +3.0pp; STAKE swings −21.4pp to ±0.0pp; HYBRID swings −20.7pp to +2.0pp. In each case the capability threshold — not the mechanism design — was the binding constraint.

E Cross-Model Domain Stability

Across all four backbone models, domains divide into three tiers based on whether decisive features are objectively identifiable, and this tier structure is broadly stable across model families and scales.

Table 10: Scale-dependent findings across Qwen3 variants. Δ vs. FREE_DEBATE = accuracy difference. FR range and surfacing range cover active conditions only.

	Qwen3-8B	**Qwen3-14B**	**Qwen3-32B**
Best mechanism (accuracy)	FREE_DEBATE (.867)	FREE_DEBATE (.860)	COUNTERFACTUAL (.897)
Worst active mechanism	BID_TO_SPEAK (.630)	CONTRIB./BID (.733)	BID_TO_SPEAK (.853)
FR range (active)	24.6%–65.6%	1.6%–12.5%	9.1%–33.3%
Surfacing range (active)	42.7%–87.6%	95.7%–100%	83.9%–99.0%
Mechanism impact on accuracy	Large ($d \leq 0.57$)	Moderate ($d \leq 0.39$)	Minimal ($d < 0.11$)
Strategic free-riding?	Model-level NO-bias	Largely absent	Partial (uniform highest)
Token efficiency (best)	0.936/1K tok	0.722/1K tok	0.710/1K tok

Consistently high-performing domains. Eight domains achieve near-perfect accuracy across all mechanisms: *corporate_strategy*, *hr*, *pharma_manufacturing*, *product*, *science*, *scientific_review*, *public_policy*, and *maritime_operations*, all sharing unambiguous individually sufficient decisive features. At smaller scales some variation persists — *maritime_operations* shows mechanism-dependent accuracy at Qwen3-8B — but the tier structure is broadly stable, confirming that scenario structure, not incentive design, determines top-tier performance.

Consistently failing domains. Five domains fail consistently regardless of mechanism: *conservation*, *consumer_marketplace*, *ecology*, *autonomous_systems*, and *cybersecurity*. *consumer_marketplace* achieves 0% under BID_TO_SPEAK and CONTRIBUTION on GPT-4o-mini; *conservation* ranges 0.00–0.60 across mechanisms. These domains share subjective or multi-criteria decisive features that agents cannot consistently identify or interpret.

High-variance domains. Healthcare shows the largest cross-mechanism accuracy variance, ranging from 0.42 to 0.83 on GPT-4o-mini. Counterintuitively, NO_COMM performs comparatively well in healthcare (0.58), suggesting individual private knowledge covers a substantial fraction of healthcare scenarios. Mechanisms that suppress communication (STAKE, UNIFORM) suffer most because multi-indicator clinical decisions produce catastrophic false negatives when decisive features are withheld.

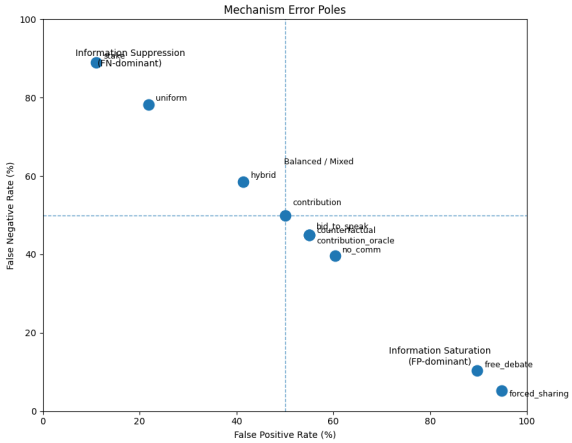

Figure 4: FP vs. FN rate by mechanism (GPT-4o-mini). Mechanisms polarise into an information-saturating cluster (upper-left: FREE_DEBATE, FORCED_SHARING) and an information-suppressing cluster (lower-right: STAKE, UNIFORM). Contribution-based mechanisms occupy the balanced centre.

F Cohen's d: Mechanism vs. FREE_DEBATE

Figure 5 shows pairwise Cohen's d for each mechanism relative to FREE_DEBATE across all four models. Positive values indicate FREE_DEBATE outperforms the mechanism; negative values indicate the mechanism outperforms FREE_DEBATE. At Qwen3-8B, all active mechanisms show positive d, with BID_TO_SPEAK the largest ($d = 0.57$). By Qwen3-32B, five mechanisms have negative d, indicating the FREE_DEBATE advantage has collapsed. GPT-4o-mini shows the sharpest architectural signature: the largest positive d belongs to STAKE ($d = 0.43$), reflecting the severity of market-based mechanism failure on that architecture.

G Category-Level Results

Table 11 reports accuracy, FRR, and DSR aggregated by mechanism category across all four mod-

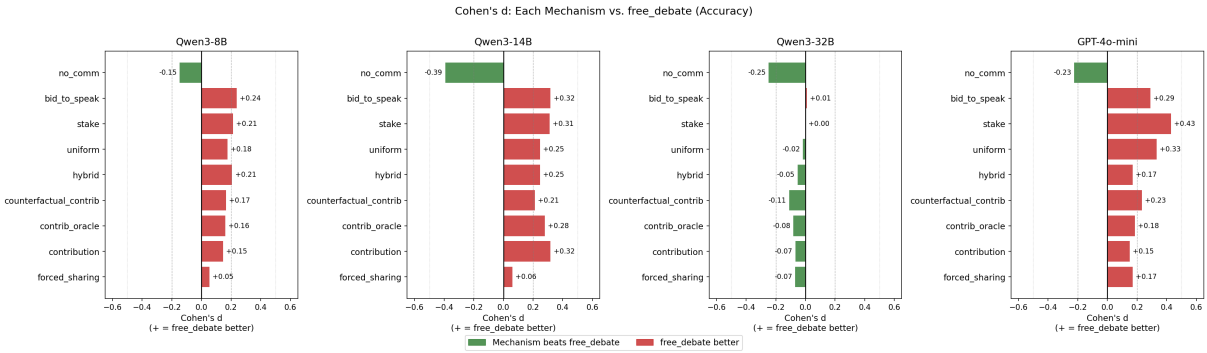

Figure 5: Cohen's d for each mechanism vs. FREE_DEBATE (positive = FREE_DEBATE better) across all four models. At Qwen3-32B, five mechanisms have negative d (mechanism beats FREE_DEBATE). GPT-4o-mini shows the largest positive d for STAKE ($d = 0.43$), the clearest architectural signature of market-based mechanism failure.

Table 11: Category-level means (accuracy / FRR / DSR) across all four models. Contribution-Based achieves the best FRR–DSR balance across all models. Market-Based shows the worst FRR in three of four models.

	Qwen3-8B			**Qwen3-14B**			**Qwen3-32B**			**GPT-4o-mini**		
Category	Acc	FRR	DSR	Acc	FRR	DSR	Acc	FRR	DSR	Acc	FRR	DSR
Baseline (free_debate, forced_sharing, no_comm)	.759	33.4%	51.2%	.774	25.0%	65.3%	.829	22.4%	61.3%	.800	33.4%	47.0%
Contribution-Based (contrib., oracle, counterfactual)	.709	29.0%	82.8%	.754	**2.3%**	**99.9%**	.889	**11.1%**	**98.0%**	.804	**23.9%**	**88.6%**
Market-Based (stake, bid_to_speak)	.642	57.7%	52.3%	.735	8.3%	97.2%	.855	23.5%	89.8%	.732	67.4%	50.0%
Hybrid/Forced (hybrid, forced_sharing)	.737	25.5%	80.3%	.800	2.6%	99.7%	.880	9.1%	96.4%	.807	25.7%	81.1%

els. The Contribution-Based category (contribution, oracle, counterfactual) achieves the best FRR–DSR balance at every scale, with FRR as low as 2.3% and DSR as high as 99.9% at Qwen3-14B. Market-Based mechanisms (stake, bid_to_speak) produce the highest free-riding in three of four models, while the Hybrid/Forced category performs comparably to Contribution-Based at larger scales. These category-level patterns are consistent across architectures, supporting the mechanism-family interpretation reported in §4.1.