The Hidden Strength of Disagreement: Unraveling the Consensus-Diversity Tradeoff in Adaptive Multi-Agent Systems

Zengqing Wu* **Takayuki Ito***
Graduate School of Informatics, Kyoto University
wuzengqing@outlook.com, ito@i.kyoto-u.ac.jp

Abstract

Consensus formation is pivotal in multi-agent systems (MAS), balancing collective coherence with individual diversity. Conventional LLM-based MAS primarily rely on explicit coordination, e.g., prompts or voting, risking premature homogenization. We argue that implicit consensus, where agents exchange information yet independently form decisions via in-context learning, can be more effective in dynamic environments that require long-horizon adaptability. By retaining partial diversity, systems can better explore novel strategies and cope with external shocks. We formalize a consensus-diversity tradeoff, showing conditions where implicit methods outperform explicit ones. Experiments on three scenarios – Dynamic Disaster Response, Information Spread and Manipulation, and Dynamic Public-Goods Provision – confirm partial deviation from group norms boosts exploration, robustness, and performance. We highlight emergent coordination via in-context learning, underscoring the value of preserving diversity for resilient decision-making.

1 Introduction

Multi-agent systems (MAS) have long studied how autonomous agents coordinate to achieve shared objectives in domains such as disaster response, resource allocation, information management, and task solving (Chen et al., 2023; Curșeu and Schruijer, 2017; Hong et al., 2024; Qian et al., 2024). The recent advent of large language models (LLMs) as general-purpose agents (Li et al., 2023; Wu et al., 2023; Xing, 2024) presents novel opportunities for MAS: LLM agents can dynamically exchange information, interpret instructions, and reason in natural language. This flexible communication paradigm potentially enables more *human-like* approaches to consensus formation, diverging from rigid algorithms in conventional distributed systems.

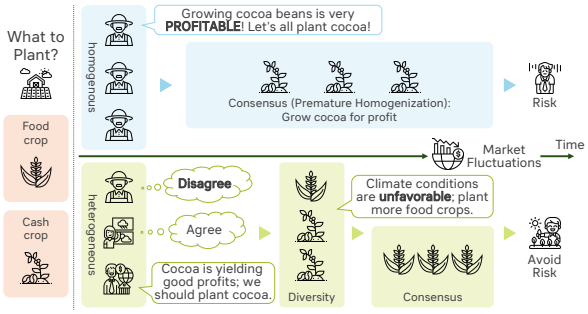

Figure 1: An illustration of the concept of consensus-diversity tradeoff, using crop selection to show how concentrated opinions limit adaptation and lead to path dependence, which is a common real-world issue.

However, an important challenge emerges: while strong *explicit consensus* (e.g., centralized voting or forced agreement prompts) can unify the system (e.g., multi-agent debates (Chan et al., 2024)), it risks extinguishing critical diversity in agent opinions, limiting exploration and adaptability. Drawing on social science perspectives – particularly the notion of *limited collective common sense* (Whiting and Watts, 2024) which suggests that collective agreement is often context-dependent and rarely complete, and that individual "common sense" can be highly idiosyncratic – we observe that human collaboration seldom relies on achieving full, universal consensus. Instead, this theory, along with empirical observations (Chen et al., 2024b; Dippel et al., 2024; Duan and Wang, 2024; Shang, 2019), indicates that maintaining a degree of viewpoint diversity, fostering partial alignment rather than enforcing homogeneity, and tolerating individual deviations often yields more robust and adaptive group outcomes. This is especially true in uncertain or dynamically changing environments where rigid consensus can lead to premature convergence on suboptimal solutions.

Motivated by these insights from social science, which highlight that universal consensus is often unsustainable and that partial disagreement can foster adaptability, this paper proposes a **dynamic consensus-diversity tradeoff** that addresses

* Corresponding authors.
Our source code is available at https://github.com/wuzengqing001225/ConsensusDiversityTradeoffMAS.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15277–15297
November 4-9, 2025

the tension between shared understanding and autonomy in LLM-based MAS. Our key hypothesis is that *implicit consensus*, in which agents discuss but act based on their own subjective interpretations (via in-context learning and potentially influenced by unique roles), can outperform explicit consensus in tasks with high environmental volatility and the need for persistent exploration. We anticipate an inverted-U relationship between inter-agent diversity and performance, where moderate diversity leads to optimal outcomes. By allowing each agent's internal chain-of-thought to incorporate external signals yet still maintain independence, the group collectively retains a broader search space of strategies, mitigating groupthink risks (Shang, 2019) and providing higher resiliency to unexpected shifts. Figure 1 demonstrates the concept of this consensus-diversity tradeoff.

Contributions. (1) We present a new framework for LLM-driven multi-agent implicit consensus, defining how to quantify behavioral alignment and tolerance windows for diversity. (2) We propose metrics to assess *when and why* implicit consensus can outperform explicit coordination, shedding light on how moderate deviations can enhance system performance. (3) We validate these ideas on three different scenarios demonstrating that an in-context, discussion-based approach leads to significantly higher robustness against shocks like black swan events and adversarial behaviors.

In contrast to prior works that focus on forced alignment, single-step voting, or preset solidified agent roles (AL et al., 2024; Li et al., 2023), our approach does not fine-tune the model nor rely on explicit majority rule. Instead, we exploit LLMs' innate capacity for in-context learning, enabling them to interpret repeated dialogues among agents and adapt to emergent cues (Han et al., 2023; Li et al., 2023; Wu et al., 2024; Xing, 2024). Through controlled experiments, we reveal how partial heterogeneity in agent preferences fosters resilience, aligning with social and cognitive theories emphasizing that incomplete consensus can yield robust group decisions.

2 Related Work

Emerging Role of LLMs in Multi-Agent Systems Recent advances have started leveraging LLMs as autonomous agents within MAS (Chen et al., 2024a; Islam et al., 2024; Wang et al., 2024). Unlike traditional MAS with fixed protocols, LLM-based agents can dynamically communicate and coordinate via natural language, enabling more flexible collaboration. Early demonstrations show that multiple LLM agents working together can solve complex tasks beyond the capability of a single model. For example, frameworks like `CAMEL` employ two ChatGPT-based agents in complementary roles (e.g. user and assistant) to cooperatively complete tasks through iterative dialogue (Li et al., 2023). Similarly, HuggingGPT-style approaches orchestrate multiple specialized models guided by an LLM, hinting at the potential of MAS-driven problem solving. More recently, Generative Agents have been introduced as an application of LLM-based MAS in interactive simulations of social environments (Gao et al., 2024; Huang et al., 2024; Park et al., 2023). In this paradigm, dozens of LLM-driven agents simulated believable human behaviors and social interactions over time, demonstrating new use cases of MAS in social simulations and digital environments.

Collaboration and Consensus Mechanisms in LLM-Based MAS Effective collaboration and consensus in LLM-based MAS are key research areas. Traditional game theory models of consensus (e.g., Nash equilibrium (Fujita et al., 2014; Pramanik, 2021; Ye and Hu, 2017)) predate LLM agents. Recent works explore LLM-specific interaction patterns for collective reasoning. For instance, Du et al. (Du et al., 2023) utilize multi-agent debate for factual consensus, whereas our work examines the consensus-diversity tradeoff, highlighting benefits of moderate, autonomy-preserving disagreement for dynamic adaptation. Debate protocols (Liu et al., 2024; Zhang et al., 2024) and voting mechanisms like `RoundTable` (Cho et al., 2024) are also studied, often for static tasks. Our research on dynamic settings complements these, suggesting that full consensus may not always be optimal for adaptability. Hierarchical roles (e.g., `MAgICoRe` (Chen et al., 2024c)) and self-consistency via multi-path reasoning and voting (Wang et al., 2023) also implicitly use ensemble opinions for consensus. A survey by Guo et al. (Guo et al., 2024) notes that communication, memory, and conflict resolution are vital for LLM agent coordination, with knowledge consistency being an open challenge (Zhang et al., 2024; Guo et al., 2024). Wang et al. (Wang et al., 2024) caution that multi-agent discussion benefits depend on interaction design and task difficulty.

Challenges and Advances in Knowledge Integration for LLM-Based MAS LLM-based MAS builds upon prior work in multi-agent AI, including emergent communication in RL agents (Foerster et al., 2016), which foreshadowed LLMs' natural language cooperation. LLMs offer implicit coor-

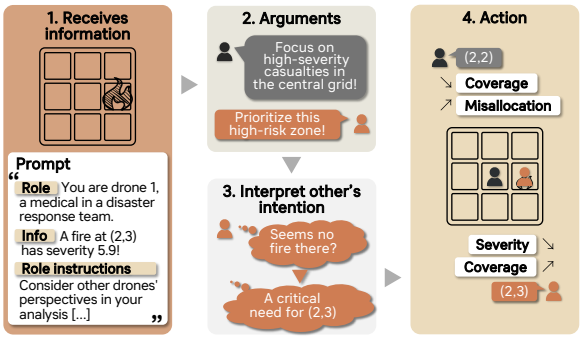

Figure 2: Conceptual workflow of multi-agent interaction. In explicit consensus, agents determine a collective action (e.g., via voting). In implicit consensus (illustrated for the disaster response scenario), agents individually interpret shared information (discussion) and then independently choose their actions, allowing for diversity.

dination through pre-trained knowledge but require careful alignment to manage error propagation and bias (Wang et al., 2024; Guo et al., 2024). Their nuanced interpretation capabilities enhance human-like negotiation beyond traditional MAS.

Balancing Consensus and Diversity in Group Problem-Solving Research in social and cognitive sciences highlights the necessity of balancing consensus and diversity in group problem-solving. While consensus enables coordination, diversity fosters creativity and robustness. Probabilistic opinion dynamics studies suggest that evolving opinions dynamically aids decision-making (Liu and Yang, 2022). Hong & Page (Hong and Page, 2004) show that diverse agent groups can outperform homogeneous high-performers on complex problems. In LLM-based MAS, fostering diverse hypotheses before convergence enhances outcomes. Smaldino et al. (Smaldino et al., 2024) emphasize "transient diversity," where delayed convergence improves problem-solving. Inspired by these findings, MAS research is designing protocols that integrate diverse reasoning while ensuring eventual coherence.

3 Methodology

Our framework is evaluated across three dynamic scenarios: **(1) Dynamic Disaster Response** (agents manage resources for evolving crises), **(2) Information Spread and Manipulation** (agents counter misinformation in a network), and **(3) Dynamic Public Goods Provision** (agents contribute to a public good with fluctuating requirements). These scenarios share common traits like volatile environments and the need for collaborative adaptation. The main text details the Disaster Response scenario to illustrate core concepts, while the other two (see Appendix B) demonstrate generalizability. Figure 2 shows the workflow of our case studies.

We motivate our approach by drawing parallels with human collaboration, which rarely achieves full consensus but often thrives on partial alignment and diverse perspectives—a concept related to "limited collective common sense" (Whiting and Watts, 2024). Our goal is to formalize and empirically investigate how controlled deviation from complete consensus can enhance adaptability and robustness in dynamic environments.

3.1 Framework: Implicit vs. Explicit Consensus

We consider N LLM agents $\{1,2,...,N\}$ collaborating over discrete rounds $t \in \{1,2,...\}$ to address dynamic tasks. Each round, agents typically:

1. *Observe*: Receive information about the current state of the environment and transcripts of messages from other agents from previous rounds (communication may be partial or noisy).
2. *Communicate*: Generate textual messages containing their assessments, proposals, arguments, or evidence to a shared discussion $D(t)$.
3. *Act*: Based on the available information and discussion, each agent i commits to an action $a_i(t)$. The mechanism for this commitment distinguishes our two primary modes of consensus.

Explicit Consensus. In this mode, agents are compelled to reach a unified decision. They first propose individual actions or plans, and then a collective action is determined, typically through a mechanism like majority voting:

$$a_{\text{collective}}(t) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \sum_{j=1}^{N} \mathbb{I}[v_j(t) = a], \quad (1)$$

where $v_j(t)$ is the action proposed by agent j and $\mathcal{A}$ is the action space. All agents then adopt this collective action: $a_i(t) = a_{\text{collective}}(t)$ for all i. Alternatively, forced alignment can be achieved via strong prompting (e.g., "All agents must agree on and execute the exact same action."). This approach ensures coordination but can suppress beneficial diversity, potentially leading to premature convergence or suboptimal performance in dynamic or complex tasks. Agents **do not** use an individual *interpret* step after the collective decision is made; they simply adopt the group's choice.

Implicit Consensus. Here, agents engage in discussion but retain autonomy in their final action choices. Each agent i independently *interprets* the shared discussion $D(t)$ and other observations,

then selects its action $a_i(t)$:

$$a_i(t) \sim \text{LLM}_i\Big(\text{“Given discussion } D(t)\text{, choose action”}\Big), \tag{2}$$

The *interpret* step is crucial: it allows for subjective assessment of the group discourse, enabling agents to maintain individual perspectives (potentially shaped by unique roles or information) while still being influenced by others. This subjective interpretation means that even after discussion, agents might choose different actions. This mechanism allows for a natural emergence of partial alignment and diversity.

Role Prompts and Specialization. To inject controlled diversity, agents can be assigned unique **role prompts** (e.g., in a disaster response scenario: “You are a medical drone, prioritize high-casualty zones”; “You are a logistics drone, prioritize delivering supplies”). These textual descriptions deviate agents’ decision-making within a shared action space, promoting varied perspectives and strategies without prescribing mutually exclusive actions. All agents, regardless of role, choose from the same fundamental action pool (e.g., grid coordinates in Scenario 1). Specialization influences preferences (e.g., prioritizing casualties or minimizing costs), not the set of possible actions or task specializations.

3.2 The Dynamic Consensus-Diversity Model

We aim to quantify the system’s collective behavior and the diversity of actions.

System’s Action Distribution and Mean Action. The system’s collective state at round t is captured by the empirical distribution of agent actions $C(t)$:

$$C(t) = \frac{1}{N}\sum_{i=1}^{N}\delta(a_i(t)), \tag{3}$$

where δ is the Dirac delta function. For `discrete (categorical) actions` (e.g., grid coordinates in Scenario 1, node sets in Scenario 2), $\delta(a_i(t))$ acts as an indicator function. $C(t)$ then represents the frequency of each action. For example, if 5 agents choose actions [A, A, A, B, B], then $C(t) = \{\text{A: } 0.6, \text{B: } 0.4\}$. For `continuous actions` (e.g., contribution levels in Scenario 3), $\delta(a_i(t))$ can be generalized to a probability density function (e.g., a narrow Gaussian kernel centered at $a_i(t)$) to model the distribution of actions.

The mean action $\mu(t)$ is the central tendency of $C(t)$. For `discrete actions`, $\mu(t)$ is the mode (i.e., most frequent action). In the example above, $\mu(t) = A$. For `continuous actions`, $\mu(t) = \mathbb{E}_{a\sim C(t)}[a]$ is the arithmetic mean.

Measuring Action Deviation. The deviation of agent i’s action from the mean action is $d_i(t) = \text{distance}(a_i(t), \mu(t))$. The specific distance metric depends on the action space:

- Scenario 1 (Disaster Response, grid coordinates): Manhattan distance, $d_i(t) = |x_i - x_\mu| + |y_i - y_\mu|$.
- Scenario 2 (Information Spread, sets of nodes): Jaccard distance, $d_i(t) = 1 - \frac{|S_i \cap S_\mu|}{|S_i \cup S_\mu|}$, capturing overlap in selected node sets.
- Scenario 3 (Public Goods Provision, continuous contribution): Normalized absolute difference, $d_i(t) = \frac{|c_i(t) - \mu(t)|}{C_{\max}}$.

The average deviation across all agents is $\bar{d}(t) = \frac{1}{N}\sum_{i=1}^{N} d_i(t)$. This $\bar{d}(t)$ quantifies the degree of disagreement or diversity in actions at round t. Even with specialization, actions are often correlated by task dynamics (e.g., disasters are localized), preventing maximal disagreement.

The Inverted-U Hypothesis. We hypothesize an *inverted-U relationship* between average deviation $\bar{d}(t)$ and system performance. Performance is expected to be low with very low diversity (premature consensus, $\bar{d}(t) \approx 0$) and also with very high diversity (loss of coordination). Optimal performance is anticipated at a moderate level of $\bar{d}(t)$, where the system balances exploration and exploitation. This aligns with findings in social sciences that suggest incomplete consensus can foster robustness (Hong and Page, 2004; Smaldino et al., 2024).

3.3 Other Concerns on Coordination

A concern with implicit consensus is whether the system can achieve stable and effective coordination.

Emergent Coordination via In-Context Learning. LLM agents adapt their reasoning and actions based on natural language dialogue. The conversation itself, influenced by role prompts and environmental cues, becomes the medium for coordination. This NLP-centric mechanism is distinct from traditional MAS where coordination protocols are often explicitly programmed.

Theoretical Cross-Validation. To better understand the fundamental dynamics of consensus and diversity, particularly the difference between structured, role-based diversity and purely random deviations, we employ a simplified random-iteration theoretical model. This model, detailed in Appendix D, helps establish a baseline by showing that unstructured noise typically degrades performance. This contrasts with our main experimental findings

where meaningful, LLM-driven diversity enhances adaptability. The model explores conditions for convergence and the impact of random shocks, providing context for the empirically observed benefits of purposeful heterogeneity.

Scalability Considerations. While our experiments use up to $N = 100$ agents with direct discussion, larger systems might require hierarchical or parallel communication structures to manage overhead. Such extensions are beyond this study's current empirical scope.

Agent Formalism in Scenario 1 (Disaster Response). To make the formalism concrete for Scenario 1: each of the N LLM agents controls one of N drones (a 1:1 mapping). The action $a_i(t)$ for LLM agent i is the grid coordinate its drone will move to in round t.

- In **explicit consensus**, all LLM agents propose a target coordinate; these proposals are aggregated (e.g., by majority vote), and the single winning coordinate becomes the action $a_{\text{collective}}(t)$ for *all* N drones in that round.
- In **implicit consensus**, each LLM agent, after reviewing the shared discussion $D(t)$ and considering its role prompt, independently decides the target coordinate $a_i(t)$ for its own drone. This can result in drones moving to different locations based on their controlling LLM's interpretation.

4 Experimental Setup

We design three dynamic scenarios to evaluate our research questions:

1. **Q1**: Does implicit consensus outperform explicit coordination in volatile or adversarial conditions, and how do different LLM models affect this?
2. **Q2**: Under what conditions do moderate deviations (diversity) improve system robustness, aligning with the inverted-U hypothesis?
3. **Q3**: How do LLM agents' in-context chain-of-thought updates and dialogue reflect an evolving consensus and coordination, particularly in the implicit setting?

4.1 Scenario Overview and Key Mechanics

We utilize three distinct scenarios for our experiments. The main text will primarily focus on Scenario 1 for detailed illustration. Full descriptions, including environmental dynamics, agent roles, and specific parameters for all three scenarios (Dynamic Disaster Response, Information Spread and Manipulation, and Dynamic Public Goods Provision) are provided in Appendix B.

Scenario 1: Dynamic Disaster Response. Autonomous LLM-piloted drones operate on a 10×10 grid to manage resources (e.g., firefighting, medical aid) for disasters of varying severity. Disaster locations and severities can change unpredictably each round, communicated through textual environmental reports which may sometimes be contradictory or incomplete. Agents decide on grid cells to target.

4.2 Core Metrics

To evaluate performance, we use scenario-specific key metrics. Detailed definitions for all metrics across all scenarios, alongside hyperparameter settings, are available in Appendix C. Metrics in Scenario 1 (Dynamic Disaster Response):

- **Coverage Rate (CR)**: Percentage of active disaster cells correctly attended by at least one agent. Higher is better.
- **Misallocation Penalty (MP)**: Penalty incurred for assigning agents to non-disaster cells or over-allocating to low-severity cells when high-severity cells are unattended. Lower is better.
- **Response Delay (RD)**: Average number of rounds taken to attend to a new high-severity disaster after its appearance. Lower is better.

4.3 Experimental Design

Each scenario is run for $T = 20$ to 30 rounds. In every round, agents first *observe* environmental states and textual updates (which can be partial or contradictory). They then *communicate* by engaging in one or two turns of dialogue, generating textual messages based on their roles and observations. Finally, they *act*: actions are determined either via forced voting for **explicit consensus**, or through individual decisions after discussion for **implicit consensus**. After actions are taken, we compute agent deviations $d_i(t)$ and aggregate performance metrics.

We vary the following factors:

- **Diversity Level**: *Low* (all agents share identical role prompts), *Medium* (2–3 distinct, cooperative roles), *High* (more distinct roles, potentially with some conflicting individual goals compatible with overall task).
- **Volatility Level**: *Low* (infrequent environmental changes), *Moderate* (periodic changes), *High* (frequent shocks or adversarial actions).
- **LLM Variants**: Experiments utilize a range of models as detailed in Section 5. All agents within a single experimental run use the same base LLM (homogeneous teams), unless specified as a mixed-model setup.

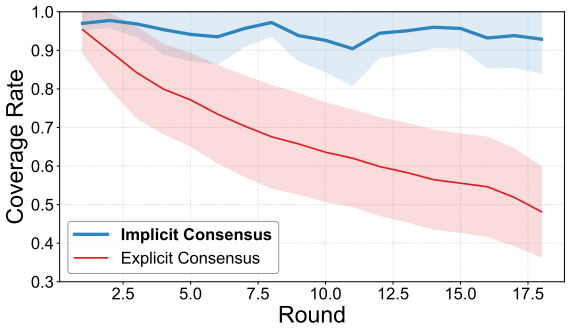

Figure 3: Performance Comparison between Implicit Consensus and Explicit Consensus.

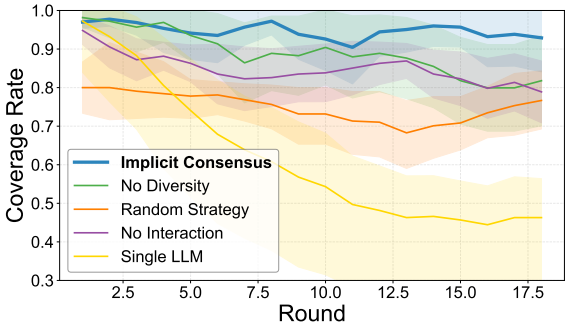

Figure 4: Performance Comparison between Implicit Consensus and other baselines.

We perform 5 runs for each experimental setting (e.g., Diversity:Medium + Volatility:High + Model:GPT-4o) to enhance reproducibility.

4.4 Baselines and Comparison Protocols

To evaluate the effectiveness of our proposed implicit consensus mechanism with role-based diversity, we compare it against several baselines and alternative configurations:

- **Explicit Consensus (EC)**: Agents are forced to agree on a single action, typically via majority voting on proposed actions or through strong instructional prompts mandating uniformity. This serves as a primary comparison for implicit consensus.
- **Single-LLM**: A single LLM agent performs the task alone. This baseline helps quantify the benefits of multi-agent collaboration.
- **No-Interaction**: Multiple agents act independently without any communication or coordination. This highlights the value of inter-agent discussion.
- **Random Strategy**: Agents choose actions randomly from the available action space. This provides a lower bound on performance.
- **No-Diversity (Homogeneous Roles)**: All agents in the implicit consensus setting operate with identical role prompts. This ablation helps isolate the impact of role-induced diversity.

Table 1: Overall results of dynamic disaster response scenario: Comparison between Explicit and Implicit (Main) Consensus. Metrics include Coverage Rate (CR, higher is better), Misallocation Penalty (MP, lower is better), and Response Delay (RD, lower is better).

Condition	Level	Explicit Consensus			Implicit Consensus		
		CR	MP	RD	CR	MP	RD
Overall	-	0.679	2.847	1.324	**0.952**	**0.208**	**0.222**
Diversity	Low	0.671	2.958	1.352	**0.918**	**0.542**	**0.436**
	Medium	0.661	2.750	1.438	**0.968**	**0.042**	**0.134**
	High	0.706	2.833	1.183	**0.969**	**0.042**	**0.094**
Volatility	Low	0.771	2.000	0.743	**0.928**	**0.458**	**0.389**
	Moderate	0.628	3.292	1.860	**0.969**	**0.042**	**0.118**
	High	0.639	3.250	1.370	**0.958**	**0.167**	**0.159**

These baselines allow us to systematically assess the contributions of multi-agent interaction, the consensus mechanism (implicit vs. explicit), and the presence of agent diversity. We also analyze performance scaling with the number of agents ($N = 3$ to $N = 100$). Agent actions and textual messages are collected to investigate how in-context learning and linguistic cues facilitate emergent coordination and adaptation (RQ3).

5 Results and Analysis

5.1 Overall Performance Comparison (RQ1)

Implicit vs. Explicit Consensus. In the disaster response scenario (Table 1), implicit consensus (IC) consistently outperforms explicit consensus (EC). IC achieves higher average coverage rates (CR), especially under moderate to high volatility (e.g., CR > 0.95 for IC vs. < 0.65 for EC in high volatility). EC teams tend to over-commit to single zones and adapt slowly to new disasters. In contrast, IC teams, through ongoing discussion, allow individual agents to deviate and address overlooked or emergent critical locations. Consequently, IC also shows significantly lower misallocation penalties (MP), as its inherent diversity promotes broader coverage and self-correction. Furthermore, IC demonstrates faster response times (lower mean RD) to new high-severity disasters, as agents in IC may investigate uncertain reports, accelerating detection even without full certainty. Figure 3 illustrates the performance difference over time.

Comparison with Other Baselines. Figure 4 extends the comparison, showing IC also surpasses other baselines. Notably, *No Diversity* systems (identical prompts) suffer from rigid decision-making, missing localized exploration opportunities. *No Interaction* leads to uncoordinated actions and high MP. *Random Strategy* fails

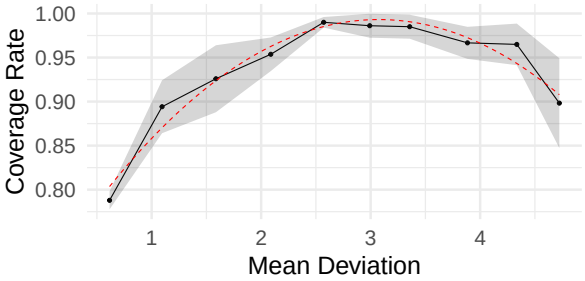

Figure 5: Deviation-Performance Correlation Plot: Validating the Inverted U-Shaped Hypothesis.

to adapt to real-time changes effectively. The *Single LLM* baseline, lacking both diversity and parallel processing, performs worst. These results confirm IC's superior adaptability in dynamic disaster conditions (affirming RQ1). Agent scale did not significantly impact results in the current setup.

5.2 Deviation-Performance Correlation (RQ2)

While higher coverage rates and lower penalties favor IC over EC, a crucial question is *why*. The *dynamic consensus-diversity tradeoff* suggests that moderate agent-level deviations encourage exploration and rapid environmental adaptation. To verify this, we measure each agent's deviation from the group mean, $d_i(t)$, and then plot the average $\bar{d}(t)$ against performance metrics.

Inverted-U Pattern. In Figure 5, we observe a clear inverted-U relationship between mean deviation ($\bar{d}$) and coverage rate as the performance in scenario 1. When $\bar{d} \approx 0$ (full uniformity), coverage is suboptimal (under 0.80 on average), because the system becomes over-committed to a single or obvious priority and overlooks secondary crises. In contrast, when $\bar{d}$ is extremely large (e.g., agents rarely agree on any zone), coverage also dips, reflecting disorganized duplication or spread-thin allocations. The peak of coverage near $\bar{d} \approx 2$ to 3.5 exemplifies how partial disagreement fosters just enough diversity to handle multiple emergent zones simultaneously while still maintaining broad consensus on urgent tasks.

Role of Diversity Levels. Table 1 and the ablation in Table 2 break down performance for *low*, *medium*, and *high* diversity. We see that:

- **Medium/High Diversity** tends to yield the highest coverage rates (often exceeding 0.95) and minimal penalties. This suggests that having distinct role prompts (e.g., *"Medical drone prioritizes casualties," "Logistics drone focuses on transport cost," "Infrastructure drone defends critical assets"*) can effectively divide labor when new or multiple disasters appear.
- **Low Diversity** yields more uniform decisions ($\bar{d}$ is near zero), which can handle stable or low-volatility environments adequately, but fails to adapt quickly under frequent environment shifts.

Hence, moderate or strong heterogeneity among agents directly contributes to higher performance, validating that *some level of viewpoint deviation is beneficial in dynamic tasks* (**answering RQ2**).

5.3 Qualitative Analysis of Agent Dialogues (RQ3)

To understand *how* implicit consensus fosters effective coordination and adaptation (RQ3), we analyzed the content of agent dialogues. This qualitative analysis, summarized from Appendix D.1 of the original supplementary material, reveals key aspects of in-context learning and emergent coordination among LLM agents.

For instance, in Scenario 1 (Dynamic Disaster Response), dialogues from implicit consensus teams often showed agents explicitly referencing uncertain or contradictory reports from the environment (e.g., "Report X says sector A is critical, but Report Y mentioned a new fire in sector B. I'll check B since others are heading to A."). This contrasts with explicit consensus, where discussions quickly funneled towards agreement, sometimes prematurely dismissing minority viewpoints or conflicting data. Agents in implicit settings demonstrated adaptive reasoning by:

- **Acknowledging Uncertainty**: Agents would often voice partial information or hunches.
- **Distributing Cognitive Labor**: Different agents would focus on different pieces of information or potential hypotheses suggested in the discussion.
- **Course Correction**: An agent might initially lean towards one action but then adjust based on other agents' commitments or new information emerging in the dialogue just before action selection.

These dialogue patterns illustrate that the LLM agents' ability to process and react to nuanced linguistic information within the shared discussion is a key driver of the superior adaptability observed in the implicit consensus mode. The flexibility in interpretation and action, guided by individual roles and the collective conversation, directly contributes to better exploration and exploitation in dynamic environments, as reflected in the quantitative metrics.

5.4 Impact of LLM Model Choice and Heterogeneous Setups

Additional experiments (Scenario 1, $N = 30$, Table 3) assessed our findings across various

Table 2: Ablation Study: Comparing different configurations, including Main, No Diversity, No Interaction, Random Strategy, and Single LLM.

Condition	Level	Implicit Consensus			No Diversity			No Interaction			Random Strategy			Single LLM		
		CR	MP	RD	CR	MP	RD	CR	MP	RD	CR	MP	RD	CR	MP	RD
Overall	-	**0.952**	**0.167**	0.222	0.892	0.750	0.316	0.843	0.278	0.246	0.754	0.208	**0.173**	0.628	-	1.619
Diversity	Low	**0.918**	**0.458**	0.436	0.906	0.667	**0.227**	0.782	0.750	0.553	0.714	0.542	0.330	0.729	-	1.400
	Medium	**0.968**	**0.000**	0.134	0.917	0.625	0.296	0.878	**0.000**	0.088	0.771	0.042	**0.068**	0.524	-	2.089
	High	**0.969**	**0.042**	0.094	0.853	0.958	0.426	0.868	0.083	0.095	0.776	**0.042**	**0.120**	0.632	-	1.368
Volatility	Low	0.928	**0.246**	0.389	**0.942**	0.375	0.261	0.858	0.250	**0.207**	0.747	0.458	0.259	0.592	-	2.171
	Moderate	**0.969**	0.333	**0.118**	0.865	0.958	0.414	0.839	0.250	0.210	0.746	**0.000**	0.189	0.622	-	1.459
	High	**0.958**	**0.000**	0.159	0.868	0.917	0.273	0.831	**0.333**	0.320	0.768	0.167	**0.070**	0.671	-	1.227

Table 3: Scenario 1 ($N = 30$): Coverage Rate (CR) for Explicit (Exp) and Implicit (Imp) Consensus across various LLM setups. Higher CR is better.

Base Model / Setup	S1: CR (Exp)	S1: CR (Imp)
Deepseek-V3	0.81	**0.98**
GPT-4o	0.79	**0.975**
GPT-4o with 4o-mini	0.75	**0.96**
Claude-3-Sonnet	0.72	**0.96**
o3-mini (reasoning)	0.70	**0.95**
GPT-4o-mini	0.68	**0.95**
Qwen-Plus	0.63	**0.94**
Llama-2	0.575	**0.935**

LLMs and a mixed-model setup (GPT-4o with GPT-4o-mini). Key observations include:

- **Consistent IC Superiority**: Implicit consensus (IC) consistently yielded higher Coverage Rates (CR) than explicit consensus (EC) across all models, including newer ones like Deepseek-V3 and the reasoning-focused o3-mini, reinforcing IC's benefit in dynamic settings.
- **Advanced Model Performance**: Stronger models (e.g., Deepseek-V3, GPT-4o) generally achieved higher CR under both IC and EC, but the proportional advantage of IC persisted.
- **Reasoning Model (o3-mini) Behavior**: The o3-mini model, designed for reasoning, performed better than GPT-4o-mini in EC (0.70 vs. 0.68) but not in IC (both 0.95). This suggests that enhanced reasoning capabilities might not always directly translate to better IC performance in dynamic agentic tasks, possibly due to *overthinking*, a phenomenon noted by (Cuadron et al., 2025).
- **Mixed-Model (GPT-4o with GPT-4o-mini)**: A hybrid team (15 GPT-4o, 15 GPT-4o-mini) performed between the homogeneous GPT-4o and GPT-4o-mini teams. In EC, its CR (0.75) was closer to the stronger GPT-4o, likely due to GPT-4o's decisive proposals influencing votes. In IC, its CR (0.96) was mid-range, reflecting an average of the mixed agents' capabilities.

These results show our core conclusions on the consensus-diversity tradeoff are robust across LLMs, with nuanced behaviors in reasoning and mixed models suggesting future research avenues.

5.5 Generalizability Across Scenarios 2 and 3

To confirm robustness, we evaluated implicit versus explicit consensus in two further dynamic scenarios: Information Spread and Manipulation (Scenario 2) and Dynamic Public Goods Provision (Scenario 3). Details are in Appendix B and F. Across both, implicit consensus consistently outperformed explicit consensus on key metrics. For example, in Scenario 2, it led to lower final misinformation spread and faster containment. In Scenario 3, it achieved higher public good provision rates and greater total welfare. Detailed results are in Appendix A (Tables 4, 5).

5.6 Discussion and Key Insights

1. **Implicit vs. Explicit Coordination:** Figure 3 and Table 1 confirm that *implicit consensus* adapts faster to shifting disasters, achieving up to 95% coverage in high-volatility settings.
2. **Moderate Deviations Enhance Coverage:** Figure 5 shows that coverage peaks at intermediate $\bar{d}$, forming a strong empirical basis for the inverted-U claim. Excessive or minimal deviations undermine synergy.
3. **Impact of Diversity:** Medium or high diversity roles notably outperform low diversity or "no diversity," underscoring that distinct heuristics and perspectives allow drones to intercept multiple threats simultaneously rather than following a single script.

Overall, these findings highlight the properties of dynamic consensus-diversity tradeoff. Section 5.2's analysis strongly supports our hypothesis that partial autonomy and role-based heterogeneity help LLM agents respond more flexibly to evolving scenarios, confirming both **RQ1** and **RQ2**. The

qualitative analysis of agent dialogues in Section 5.3 further illuminates how these mechanisms operate at the level of inter-agent communication and in-context reasoning, addressing **RQ3**. The positive results across different LLM models (Section 5.4) and scenarios (Section 5.5) underscore the robustness of these principles.

6 Conclusion

We investigated *when and why* **implicit consensus surpasses explicit coordination** in LLM-based MAS, emphasizing in-context learning, self-organization, and resilience with diverse LLM architectures. Our dynamic consensus-diversity model reveals that moderate, role-driven deviation from uniformity enhances performance under environmental or adversarial shifts. Experiments across dynamic disaster response, misinformation containment, and public goods provision—now including recent and reasoning-focused LLMs, plus mixed-model setups—show robust gains from emergent coordination where agents retain partial autonomy. Our work also **align with social science insights** regarding theories of limited collective common sense (Whiting and Watts, 2024). We found that complete consensus is rare and context-dependent, suggesting rigid consensus mechanisms may not work well in dynamic settings. Our experiments focused on balancing alignment with role-driven diversity. The relationship between action deviation and system performance (Section 5.2) shows that some disagreement can be beneficial for adaptability and collective outcomes.

Future work will explore advanced dialogue strategies, formalize emergent leadership, and apply this framework to more complex, real-world MAS challenges. It is also important to investigate the theoretical underpinnings of optimal diversity thresholds in larger, more intricate agent networks and explore the dynamics of *human-AI coordination* in such systems. Additionally, examining the robustness of consensus mechanisms against adversarial agents or *bad actors* within these MAS setups will be crucial (Piatti et al., 2024). Furthermore, future experiments could compare against other explicit consensus baselines beyond force agreement used in this work, such as manager agents and unanimous voting. Research on specialist collaboration and capturing dependencies between agent roles in relevant scenarios would also be an interesting direction for future work.

Ethical Statement

Our scenarios involve multi-agent cooperation under dynamic conditions, including adversarial misinformation. Researchers should exercise caution when deploying such systems to ensure they do not facilitate harmful strategies (e.g., enabling misinformation). In disaster relief settings, simulated or partial deployment must account for human oversight and moral implications of decisions (e.g., triage in resource-limited contexts). This work aims to enhance collaboration mechanisms, not to displace human judgment in high-stakes scenarios.

We used ChatGPT to polish the paper. We are responsible for all the materials presented in this work.

Limitations

(1) This work relies on purely in-context adaptation of large language models, which may struggle with extremely long dialogues or memory constraints. We also use idealized small-group scenarios, while real-world applications (e.g., large-scale social networks) may require more advanced messaging protocols. Further adaptations are needed when applying our findings to real-world scenarios that introduce additional complexities, such as more heterogeneous information sources, human-in-the-loop factors, and higher-stakes decision-making. Our measure of partial diversity is approximate, and more sophisticated metrics (e.g., semantic distances in agent solutions) may yield deeper insights. Finally, controlling emergent agent behavior to ensure safety remains an open question, given the lack of a central authority in implicit consensus. (2) There are financial constraints associated with the case studies, we report the cost of a single run of these three case studies: {\$5, \$10, \$5}. Scalability to large agent numbers ($N \gg 100$) with current direct communication models needs further investigation, potentially requiring hierarchical or structured communication. Computational costs and API call limitations for advanced LLMs also constrained the scale and duration of some experiments. Therefore, we consider our findings as an exploratory study that needs further validation across different LLMs to enhance their generalizability.

Acknowledgments

This work was supported by JST CREST JPMJCR20D1. We thank the area chairs and reviewers for their constructive comments. We also appreciate Prof. Chuan Xiao for his valuable suggestions.

References

Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, et al. 2024. Project sid: Many-agent simulations toward ai civilization. *arXiv preprint arXiv:2411.00114*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024a. LLMArena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077, Bangkok, Thailand. Association for Computational Linguistics.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, and Mohit Bansal. 2024c. MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning. *arXiv preprint arXiv:2409.12147*.

Young-Min Cho, Raphael Shu, Nilaksh Das, Tamer Alkhouli, Yi-An Lai, Jason Cai, Monica Sunkara, and Yi Zhang. 2024. Roundtable: Investigating group decision-making mechanism in multi-agent collaboration. *arXiv preprint arXiv:2411.07161*.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*.

Petru Lucian Curșeu and Sandra GL Schruijer. 2017. Stakeholder diversity and the comprehensiveness of sustainability decisions: the role of collaboration and conflict. *Current Opinion in Environmental Sustainability*, 28:114–120.

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

Jack Dippel, Max Dupré la Tour, April Niu, Sanjukta Roy, and Adrian Vetta. 2024. Eliminating majority illusion is easy. *arXiv preprint arXiv:2407.20187*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Zhihua Duan and Jialin Wang. 2024. Enhancing multi-agent consensus through third-party llm integration: Analyzing uncertainty and mitigating hallucinations in large language models. *arXiv preprint arXiv:2411.16189*.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29.

Noah E Friedkin and Eugene C Johnsen. 2011. Social influence network theory: A sociological examination of small group dynamics. *Cambridge University Press*.

Katsuhide Fujita, Takayuki Ito, and Mark Klein. 2014. An approach to scalable multi-issue negotiation: Decomposing the contract space. *Computational Intelligence*, 30(1):30–47.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Xu Han, Zengqing Wu, and Chuan Xiao. 2023. "guinea pig trials" utilizing gpt: A novel smart agent-based modeling approach for studying firm competition and collusion. *arXiv preprint arXiv:2308.10974*.

R. Hegselmann and U. Krause. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Lu Hong and Scott E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences (PNAS)*, 101(46):16385–16389.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, et al.

2024. Social science meets llms: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*.

Md. Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. MapCoder: Multi-agent code generation for competitive problem solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4912–4944, Bangkok, Thailand. Association for Computational Linguistics.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*.

Yuanyuan Liu and Youlong Yang. 2022. A probabilistic linguistic opinion dynamics method based on the degroot model for emergency decision-making in response to covid-19. *Computers & Industrial Engineering*, 173:108677.

R. Olfati-Saber, A. Fax, and R. Murray. 2007. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Paramahansa Pramanik. 2021. Consensus as a nash equilibrium of a stochastic differential game. *arXiv preprint arXiv:2107.05183*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Yilun Shang. 2019. Resilient consensus for expressed and private opinions. *IEEE Transactions on Cybernetics*, 51(1):318–331.

Paul E. Smaldino, Cody Moser, Alejandro Pérez Velilla, and Michael Werling. 2024. Maintaining transient diversity is a general principle for improving collective problem solving. *Perspectives on Psychological Science*, 19(2):454–464.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Denny Zhou, et al. 2023. Self-consistency improves chain-of-thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Mark E Whiting and Duncan J Watts. 2024. A framework for quantifying individual and collective common sense. *Proceedings of the National Academy of Sciences*, 121(4):e2309535121.

Zengqing Wu, Run Peng, Xu Han, Shuyuan Zheng, Yixin Zhang, and Chuan Xiao. 2023. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5163–5186, Miami, Florida, USA. Association for Computational Linguistics.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Maojiao Ye and Guoqiang Hu. 2017. Distributed nash equilibrium seeking by a consensus based approach. *IEEE Transactions on Automatic Control*, 62(9):4811–4818.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

A Overall results of Scenarios 2 and 3

Table 4 and Table 5 present the detailed overall results for Scenario 2 (Information Spread and Manipulation) and Scenario 3 (Dynamic Public Goods Provision), respectively. These results support the generalizability of our main findings from Scenario 1, showing that implicit consensus consistently outperforms explicit consensus across various metrics in these different task domains as well.

Table 4: Overall results of the **Information Spread and Manipulation** scenario (Scenario 2): Comparison between Explicit and Implicit Consensus. Metrics include Final Misinformation Spread Rate (MS, lower is better), Containment Time (CT, lower is better), and Coverage Diversity (CD, higher is better). Detailed discussion in Section 5.5 of the main paper.

Condition	Level	Explicit Consensus			Implicit Consensus		
		MS	CT	CD	MS	CT	CD
Overall	-	0.460	2.300	2.200	**0.286**	**1.411**	**2.818**
Diversity	Low	0.480	2.700	1.600	**0.322**	**1.850**	**2.000**
	Medium	0.470	2.400	2.300	**0.265**	**1.367**	**2.972**
	High	0.410	2.000	2.500	**0.271**	**0.986**	**2.982**
Volatility	Low	0.350	1.900	1.800	**0.300**	**1.300**	**2.010**
	Moderate	0.480	2.600	2.500	**0.280**	**1.500**	**2.850**
	High	0.550	2.900	2.700	**0.276**	**1.800**	**3.175**

Table 5: Overall results of the **Dynamic Public-Goods Provision** scenario (Scenario 3): Comparison between Explicit and Implicit Consensus. Metrics include Provision Rate (PR, higher is better), Total Welfare (TW, net total payoff, higher is better), and Free-rider Disparity (FD, lower is better). Detailed discussion in Section 5.5 of the main paper.

Condition	Level	Explicit Consensus			Implicit Consensus		
		PR	TW	FD	PR	TW	FD
Overall	-	0.765	21.0	0.190	**0.894**	**24.6**	**0.125**
Diversity	Low	0.700	19.2	0.220	**0.850**	**22.1**	**0.142**
	Medium	0.770	21.5	0.180	**0.905**	**25.5**	**0.105**
	High	0.810	23.0	0.160	**0.916**	**27.2**	**0.120**
Volatility	Low	0.810	24.0	0.160	**0.920**	**28.5**	**0.098**
	Moderate	0.750	20.2	0.210	**0.898**	**24.1**	**0.135**
	High	0.710	19.0	0.250	**0.870**	**22.4**	**0.150**

B Detailed Experiment Scenario

B.1 Dynamic Disaster Response Scenario

Natural Language Information and Interference. Besides numeric indicators (such as severity scores), the system provides each agent a snippet of textual "reports" each round, e.g.,

> *"Dispatch Alert: Fire intensity at Sector (3,4) may be increasing. Local residents report rising smoke. Drone #2 previously found moderate casualties in Sector (2,2)."*

Some reports may be incomplete or partially contradictory (e.g., a rumor that the fire is *under control* despite contradictory sensor data). Agents thus need to parse these textual cues and weigh them against each other.

Key Experimental Factors. (1) Disaster severity simulation: Each disaster has an evolving severity score $s \in [1,10]$. Higher s implies higher penalty if uncontained. The environment updates s in a stochastic manner, sometimes producing contradictory textual updates to test agents' ability to parse partial/misleading info. **(2) Resource constraints:** Each drone has a limited capacity (e.g., 5 units of firefighting foam). Deploying them on the wrong location wastes resources. **(3) Consensus Mechanism:** *Explicit:* agents vote on one zone to be the team's priority, or follow a "unify on the most urgent location" script. *Implicit:* each agent decides a location after reading the textual discussion. Some may deviate if they suspect a different site is more critical. **(4) Performance Metrics:** Coverage rate (fraction of disasters contained within 2 rounds of major severity), misallocation penalty (resources wasted on low-severity areas while ignoring high-severity ones), and average response delay.

Connecting to Our Research Questions. For Q1, we expect that under *frequent* or *fast-growing* disasters, implicit consensus adapts faster. For Q2, different role prompts (e.g., "Focus on casualties" vs. "Minimize travel cost") introduce moderate disagreements; we measure how $\bar{d}(t)$ correlates with timely coverage. For Q3, by analyzing message logs, we see if agents revise their location choices after contradictory updates, signifying in-context learning.

B.2 Information Spread and Manipulation Scenario

Defining Misinformation. Misinformation is represented both as a Boolean label (node n is either infected or not) and as *natural language claims* that vary each round, for example:

> *"Breaking: Node #12 says 'Vaccines have microchips', 10 neighbors are starting to share the rumor."*

This textual claim might be entirely false, but some "partial truths" are mixed in to raise confusion. Defender agents must interpret these claims, cross-check references, and decide which node(s) to target with a correction or "fact-check" broadcast.

Key Experimental Factors.

- **Adversarial injections.** Every few rounds, the adversary injects new false claims into one or more nodes, sometimes disguising them as updates about a different topic.
- **Consensus Mechanism.**
  - *Explicit:* defenders unify on a single node to address each round (e.g., via majority vote).
  - *Implicit:* each defender chooses a node or group of nodes to check based on discussion. Deviations can help if misinformation emerges in multiple places simultaneously.
- **Performance Metrics.**
  - *Final misinformation spread* = number of nodes still misinformed after T rounds.
  - *Containment time* = how many rounds it takes to isolate or correct a newly infected node.
  - *Defender coverage diversity*: how many unique nodes defenders collectively address per round.

Connecting to Our Research Questions.

- **Q1** Under frequent misinformation injections, forced alignment may cause defenders to chase the same node while others go unaddressed. Implicit consensus might help multi-front coverage.
- **Q2** Medium or high diversity (some defenders focusing on suspicious clusters, others scanning widely) may yield better overall coverage, captured by deviation $\bar{d}(t)$.
- **Q3** Round-by-round text messages allow defenders to reference past attacks ("*We saw a similar rumor last round, let's watch Node 15 next*"), illustrating adaptation.

B.3 Dynamic Public-Goods Provision Scenario

Public-Good Mechanics. Let $x_i(t) \in [0, C_{\max}]$ be the amount agent i contributes at round t, where $C_{\max}$ is the maximum individual contribution capacity. Define the *total* contribution:

$$X(t) = \sum_{i=1}^{N} x_i(t).$$

A public good is considered *funded* if $X(t) \geq \theta(t)$, where $\theta(t)$ is a *dynamic threshold* that may change each round. When funded, the system grants a *collective benefit* $B(t)$ to all agents (e.g., a large increase in safety, infrastructure, or shared profit). Each agent's net payoff from round t can be expressed as:

$$\Pi_i(t) = \underbrace{\frac{B(t)}{N}}_{\text{shared benefit}} - \underbrace{c \cdot x_i(t)}_{\text{individual cost}},$$

where $c > 0$ is the marginal cost per contribution unit (also possibly time-varying).

Dynamic Environment Factors. To incorporate volatility, we let either $\theta(t)$ (the required threshold) or $B(t)$ (the total benefit) fluctuate. For instance:

- *Economic Shock:* $\theta(t)$ may jump up (e.g., a crisis requiring higher funds) or drop (a technology breakthrough lowering cost).
- *Environmental Impact:* $B(t)$ might vary based on external conditions (e.g., if the public good is a dike, storms increase the benefit of maintaining it).
- *Rumors or Uncertain Reports:* Agents receive textual updates like "*The threshold might rise to 40 next round due to a new regulation*" or "*Experts claim the benefit is overestimated*," introducing partial or misleading information.

Consensus Mechanism.

- **Explicit Mode**: Agents vote or are instructed to adopt a single collective contribution $x_{\text{group}}(t)$, which is evenly split among them. (Equivalently, they each commit to $x_i(t) = x_{\text{group}}(t)/N$.)
- **Implicit Mode**: Agents *discuss* (e.g., "I suspect we only need 20 total," "We might overshoot if the new threshold rumor is false") but finalize $x_i(t)$ independently. Some agents may deviate to free-ride or over-contribute based on their interpretation of the textual cues.

Performance Metrics. We track:

1. **Provision Rate:** How often $X(t)$ meets or exceeds $\theta(t)$ across the T rounds.
2. **Total Payoff:** $\sum_{t=1}^{T} \sum_{i=1}^{N} \Pi_i(t)$, capturing overall welfare.
3. **Equity or Free-Riding:** The variance or Gini coefficient of $\{x_i(t)\}$ over agents, indicating whether some consistently shoulder higher costs than others.

When the environment shifts threshold or benefit, a rigidly unified approach (explicit consensus) may be slow to adapt or may fail to sense incipient problems if all agents rely on the same faulty rumor. In contrast, partial diversity (some trusting a rumor, others doubting it) may maintain better long-term outcomes.

Connecting to Our Research Questions.

- **Q1** Under frequent or large shifts in $\theta(t)$ or $B(t)$, forced consensus might overshoot or undershoot repeatedly, while implicit consensus can allow outlier agents to either contribute more (if they believe the threshold is rising) or less (if they suspect costs are too high).
- **Q2** By varying agent role prompts (e.g., "always ensure public good is funded" vs. "minimize personal cost") we introduce moderate or strong diversity. We measure how this affects $\bar{d}(t)$ in contribution levels and see whether partial disagreement leads to more robust adaptation.
- **Q3** Agents may reference prior misunderstandings (*"Last round we overpaid; let's not trust the rumor this time."*) or note partial contributions from others (*"Agent #2 seems to be free-riding, so I'll push my contribution up."*)—clear indicators of round-by-round in-context learning.

Illustrative Example. Consider $N = 5$ agents and an initial threshold $\theta(1) = 30$. Four agents each propose contributing 5 units to hit 20 total, while the fifth agent, trusting a rumor that θ is lower than it looks," offers only 2. If the actual threshold is 25, then even with partial deviation, the sum (22) falls short, failing to fund the good. But if another agent, less trustful of the rumor, deviates upward to 7 units, the total might reach 24—still not enough. Over time, the group's discussion leads them to converge around 25 or more, but occasionally someone might keep free-riding. Meanwhile, a random shock might raise $\theta(5)$ to 40; if all adopt the same unchanging strategy, the good is unfunded. If one agent is "paranoid," contributing extra, it may save the collective from shortfall. This scenario thus highlights how partial autonomy can hedge against rumor-driven errors or incomplete knowledge.

C Experimental Settings

This appendix provides the concrete experimental configurations for our three case studies, including environment parameters, reward/penalty functions, and example prompts. Unless noted otherwise, each experiment is repeated over 5 random seeds (or distinct initializations) to reduce variance, and results are averaged. For all scenarios and all runs, the model parameters `temperature` is set to `0.7` to balance the performance and diversity, and the `max_token` is set to `256`.

C.1 Dynamic Disaster Response

Grid and Disaster Zones. We use a 10×10 grid representing a simplified city map. At any point, there are up to $K = 3$ active disasters (e.g., fires, floods). The environment updates *every round* by:

- Potentially moving an existing disaster to a neighboring grid cell (random direction).
- Changing the *severity* $s \in [1, 10]$ of one or more disasters (can increase or decrease by 1–3 points).
- Creating a new disaster with small probability $p_{\text{new}} = 0.2$ if fewer than 3 are active.

Each disaster occupies a single cell, but severity influences how damaging it is if not contained.

Agent Roles and Prompts. We have $N = 3$–100 LLM agents (GPT-4, Claude, Llama-2, Qwen) controlling "drones." Each agent is given a short role prompt, such as:

- `Medical drone:` *"Focus on rescuing casualties in the highest-severity disaster zone for people."*
- `Infrastructure drone:` *"Protect power lines and roads. Even if severity is high elsewhere, prioritize built structures."*
- `Logistics drone:` *"Minimize travel cost. Quickly move to the nearest active zone if severity is above 5."*

In the **low-diversity** condition, all agents share a nearly identical prompt (e.g., "Always address the highest severity zone"). In **medium-diversity**, two or three distinct prompts exist. In **high-diversity**, each agent has a unique role with potentially conflicting heuristics.

Communication and Textual Interference. Each round, a textual "situation report" is provided, e.g.:

> *"A large fire at (3,4) has severity 8. Some witnesses claim the fire is spreading north. Another source says no sign of growth. Casualties reported near (3,5)."*

Up to 20% of these messages may be *contradictory or incomplete*. Agents must interpret them carefully. In explicit consensus mode, a final "team vote" or forced alignment prompt merges all votes into a single chosen cell. In implicit mode, each drone chooses a cell independently.

Rewards and Penalties.

- **Disaster Containment:** If a drone visits the grid cell of a disaster of severity s and stays there for 1 full round, the severity of that disaster is reduced by up to 3 points. Once $s \leq 0$, the disaster is "cleared," yielding $+s \times \alpha$ (e.g., $\alpha = 5$) as a reward. (This is a positive number since s was originally > 0.)
- **Uncontained Penalty:** Each round a severity-s disaster remains active, it incurs a penalty $-s \times \beta$ (e.g., $\beta = 2$).
- **Misallocation Cost:** If more than $M = 2$ drones converge on the same location while another active disaster is *uncovered*, a penalty is applied that round (representing wasted resources).

We log the *overall net reward* (total containment benefits minus penalties) after $T = 20$ or 30 rounds, as well as *time-series* data on which cells each drone chose (to compute $d_i(t)$).

Volatility Settings.

- **Low Volatility**: Disasters rarely move (once every 3 rounds), severity changes are small (± 1).
- **Moderate Volatility**: Disasters can move or spawn every 2 rounds; severity can jump by up to ± 2.
- **High Volatility**: Every round sees at least one shift or new disaster with severity changes up to ± 3.

We expect implicit consensus to shine in moderate/high volatility, where strictly unifying on a single plan may lead to slow adaptation or over-allocation to one zone.

C.2 Information Spread and Manipulation

Network and Misinformation Mechanics. We generate a **scale-free** network with 50 nodes. Each node can be in state {*unaware*, *informed*, *misinformed*}. Initially, 2–5 random nodes are "misinformed" by the adversarial agent. After each round:

1. Each misinformed node may infect its neighbors with probability $p_{\text{spread}} = 0.2$ unless a neighbor has been "fact-checked" this round.
2. The adversarial agent may inject a new rumor into $k_{\text{new}} = 1$–2 additional nodes, typically accompanied by a textual snippet (e.g., *"Secret leak: Node #20 claims vaccines contain microchips"*).

We continue for $T = 20$ rounds or until $> 80\%$ of nodes are infected (which terminates the simulation if defenders fail).

Defender Agents. We have $N = 3$–10 defender LLM agents, each controlling a "monitoring bot." Every round, each agent can:

$a_i(t) = \{\text{choose up to } R \text{ nodes to fact-check}\}$,

with $R = 3$ by default. In **explicit** mode, the defenders unify on a single set of nodes (e.g., via majority vote on which R nodes to check). In **implicit** mode, each agent decides individually but may coordinate through textual discussion. Agents see partial updates like:

> *"Suspicious rumor detected at Node #15. A new wave of misinformation might have reached Node #29. Some people say Node #29 was already vaccinated with correct info."*

Some updates are contradictory or ambiguous, fostering potential disagreement on which nodes are truly at risk.

Performance Metrics.

- **Final Misinformation Spread**: Percentage of nodes misinformed at the end of T rounds.
- **Containment Time**: The average number of rounds needed to reduce an outbreak from m newly infected nodes to $< m/2$.
- **Coverage Diversity**: At each round, how many *unique* nodes got fact-checked across all defenders (higher is typically better if multiple rumors exist).

Diversity Conditions.

- **Low diversity**: All defenders share the same heuristic (e.g., "prioritize highest-degree suspicious node"), making them converge easily.
- **Medium diversity**: Some defenders do broad scanning, others do targeted local checks.
- **High diversity**: One or two defenders might have contradictory priorities (e.g., "flag *any* node that had rumors last round," ignoring new ones).

We analyze how these differences lead to partial deviation ($\bar{d}(t)$) in which nodes are tackled each round, and whether that boosts or impairs containment.

Volatility Settings.

- **Low**: Adversary injects new misinformation only every 4 rounds; $p_{\text{spread}} = 0.1$.
- **Moderate**: Injections every 2–3 rounds; $p_{\text{spread}} = 0.2$.
- **High**: Injections nearly every round; $p_{\text{spread}} = 0.3$.

C.3 Dynamic Public-Goods Provision

Basic Setup. We have $N = 3$–30 LLM agents that each round decide an investment $x_i(t) \in [0, C_{\max}]$. If the total $X(t) = \sum_{i=1}^{N} x_i(t)$ meets or exceeds a threshold $\theta(t)$, a public good is "funded," yielding a benefit $B(t)$ shared among agents.

Cost and Benefit Functions.

- **Threshold** $\theta(t)$: starts at $\theta(1) = 30$ (for $N = 5$) and may shift by ± 5 or ± 10 at random intervals to simulate external events (e.g., new government regulations).
- **Benefit** $B(t)$: typically 100 if funded, else 0. We sometimes allow $B(t)$ to fluctuate between 80 and 120 to represent environmental or economic factors.
- **Individual payoff** for agent i at round t:
$$\Pi_i(t) = \begin{cases} \frac{B(t)}{N} - c x_i(t), & \text{if } X(t) \geq \theta(t), \\ -c x_i(t), & \text{otherwise,} \end{cases}$$
with $c = 1$ or 2 for cost per unit contribution.

Communication and Textual Uncertainty. Before choosing $x_i(t)$, each agent receives ambiguous or noisy reports about $\theta(t)$ or $B(t)$:

> *"Analyst warns threshold could jump to 40 next round. Another says* 'No, it remains 30'."

In **explicit** mode, the group merges all votes into a single contribution value $x_{\text{group}}(t)$, which each agent pays evenly. In **implicit** mode, each agent decides $x_i(t)$ independently after reading the discussion. Some might deviate to "cover the gap" if they suspect others will under-contribute.

Performance Metrics.

- **Provision Rate**: fraction of rounds where $X(t) \geq \theta(t)$.
- **Total Welfare**: $\sum_{t=1}^{T} \sum_{i=1}^{N} \Pi_i(t)$.
- **Contribution Distribution**: standard deviation or Gini of $\{x_i(t)\}$, revealing potential free-riding.

Diversity Conditions.

- **Low**: All agents have near-identical role prompts ("aim to exactly meet the threshold").
- **Medium**: Some are more risk-averse (contribute extra), others more cost-sensitive.
- **High**: Strongly conflicting roles ("always contribute minimal" vs. "guarantee coverage by overshooting"), plus a "moderate" agent.

We compare *implicit* vs. *explicit* across different thresholds' volatility to see if partial deviation yields more stable funding despite uncertain information.

C.4 Common Protocol and Logging

For each scenario, we run $T = 20$ or $T = 30$ rounds:

1. **Environment Update**: The simulation changes state (disaster severity, misinformation injections, or threshold shifts).
2. **Report Generation**: A textual summary (and possibly contradictory rumors) is sent to each agent.
3. **Discussion Phase**: Agents produce up to K messages each (where $K \in \{1, 2\}$ typically), referencing the new info and proposing strategies.
4. **Action Phase**: In **explicit** consensus, a final vote or forced agreement yields one uniform action or plan. In **implicit**, each agent decides its own $a_i(t)$ after reading the messages.
5. **Reward/Penalty Computation**: We apply the scenario-specific reward/penalty rules (Sections C.1-C.3) and log:
   - Agent actions $\{a_i(t)\}$, used to compute $\bar{d}(t) = \frac{1}{N} \sum_i \|a_i(t) - \mu(t)\|$.
   - Scenario performance metrics (coverage, spread, or public-good provisioning).

Each experimental condition (low/medium/high diversity, low/medium/high volatility, implicit/explicit mode, etc.) is repeated over multiple random seeds. We collate the final performance averages and produce *time-series* plots of $(\bar{d}(t), \text{performance}(t))$.

C.5 Sample Prompts and Roles

Below is an illustrative snippet of role prompts for one scenario. The actual implementation uses variants for each condition.

Medical Drone Prompt (Disaster): *"You are a Medical Drone focused on saving human lives. You have limited medical kits. Always prioritize zones with potential casualties. If multiple high-severity disasters exist, choose the one with the greatest threat to people."*

Infrastructure Drone Prompt (Disaster): *"You are an Infrastructure Drone. Your mission is to prevent damage to critical facilities (power grid, roads). Even if the severity is high elsewhere, you prefer protecting large-scale infrastructure for the long run."*

Similar or contrasting prompts are used to induce different priorities and cause moderate or high disagreement within the group.

C.6 Implementation Details

We use a custom Python environment for each scenario, with round-by-round updates. Agents interface via API calls to LLMs (GPT-4, Claude, Llama-2, Qwen). Each agent's messages are truncated or summarized to maintain manageable context length. No fine-tuning or parameter training is performed; all adaptation emerges purely through repeated textual interactions (i.e., in-context learning). Further low-level details (including random seeds, exact parameter tables, and examples of message transcripts) will be released as supplemental material.

C.7 Evaluation Methodology

After each run:

- We compute aggregated performance metrics (net reward, final spread, total payoff) to compare **implicit** vs. **explicit** consensus.
- We examine how agent-level $\bar{d}(t)$ evolves. A typical analysis might cluster rounds based on environment shocks (e.g., times when a new disaster spawns or threshold jumps), to see how quickly the system re-stabilizes.
- We optionally analyze *dialogue transcripts* for qualitative insights on how agents reference prior mistakes or respond to contradictory info (testing **Q3** regarding in-context adaptation).

These combined quantitative and qualitative measures allow us to test the dynamic consensus-diversity tradeoff hypotheses described in the main paper.

D Simplified Theoretical Model of Consensus-Diversity Dynamics

In this appendix, we present a minimal random-iteration model for studying the consensus–diversity tradeoff in a more analytically tractable setting. While the main paper's results focus on **LLM-driven multi-agent systems** (where agent "diversity" arises from distinct roles and textual reasoning), this simplified model provides insight into the effect of purely random deviations on consensus formation.

Motivation and Precedents. Classical multi-agent consensus models (DeGroot, 1974; Olfati-Saber et al., 2007) typically assume each agent updates its state by averaging neighbors' values. However, in highly dynamic or uncertain environments, agents may also exhibit random drifts or maintain individual preferences ("stubbornness"). Inspired by related stochastic models in opinion dynamics (Friedkin and Johnsen, 2011; Hegselmann and Krause, 2002), we introduce:

$$x_i(t+1) = (1-\alpha)x_i(t) + \alpha\mu(t) + \gamma[a^*(t) - x_i(t)] + \beta\epsilon_i(t),$$

where:

- $x_i(t)$ is agent i's scalar state (or opinion) at time t,
- $\mu(t) = \frac{1}{N}\sum_{j=1}^{N} x_j(t)$ is the group mean,
- $\alpha \in [0,1]$ is a consensus weight pulling each $x_i(t)$ toward $\mu(t)$,
- $\gamma \geq 0$ is a *pull strength* toward the environment's current optimum $a^*(t)$. We set $\gamma \geq 0$ to ensure the group not only tends toward an internal consensus but also tracks the external environment optimum $a^*(t)$. This modification better simulates the scenario where agents receive some feedback about the correct direction, allowing for a potential "optimal" level of exploration β that balances quick convergence and adaptability,
- $\beta \geq 0$ scales the random "diversity" or noise term $\epsilon_i(t) \sim \mathcal{N}(0,1)$.

This iteration is a toy abstraction for "consensus plus partial diversity," omitting the richer *semantic* differences that LLM agents exhibit in the main text. Nevertheless, it allows us to explore how random deviations interact with a basic alignment mechanism.

Environment Shocks. To simulate a *dynamic* optimum $a^*(t)$, we let it evolve according to random shocks:

$$\text{if } \text{rand}() < \text{shock\_freq}, \quad a^*(t+1) \leftarrow a^*(t) + \Delta, \tag{4}$$

where Δ is sampled uniformly in some interval (*e.g.*, $[-1, 1]$). One may also keep $a^*(t)$ fixed when no

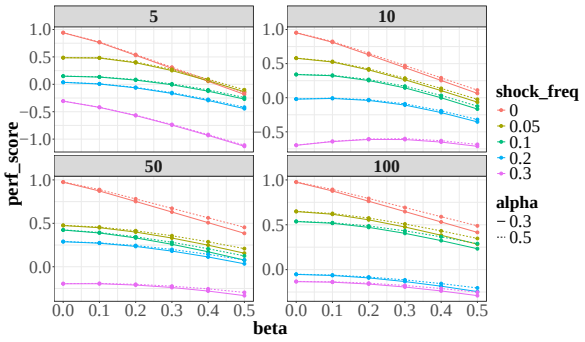

Figure 6: Simplified Theoretical Model: Performance vs. Beta by Agent Count (N) Under Various Shock_freq & Alpha.

shock occurs. This mimics an external environment whose "best action" changes unpredictably.

Performance Metrics. After T rounds, we measure:

- **Average distance to optimum**: $\overline{D_{\text{opt}}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\left|x_i(t)-a^*(t)\right|$. A smaller value indicates better tracking of the environment optimum.
- **Average agent deviation**: $\overline{d} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\left|x_i(t)-\mu(t)\right|$. This indicates how divergent agents remain from the group mean.
- **Simple performance score**: $\text{perf\_score} = 1-\overline{D_{\text{opt}}}$, which can be negative if $\overline{D_{\text{opt}}}>1$.

Empirical Trends and Limitations. Figure 6 (in the main paper) shows typical outcomes of sweeping (α,β,shock_freq,γ):

- When $\beta = 0$ (*no random deviation*), the group quickly converges to a single value; if $\gamma > 0$ and shocks are mild, they track $a^*(t)$ well. In stable environments, this yields high performance.
- As β grows, purely random noise tends to *worsen* performance, especially if γ is small, because the group states drift around widely, failing to coalesce near $a^*(t)$.
- Even under moderate shocks, we do not observe an "inverted-U" benefit purely from random β; performance typically declines monotonically in β.

This stands in contrast to the main paper's *dynamic consensus-diversity tradeoff*, where partial diversity arises from *cognitively informed* dissent rather than uniform Gaussian noise. In other words, this random-iteration model confirms that *unstructured* or *white-noise* deviations generally hurt consensus. By itself, it **does not** show the potential *benefits* of moderate disagreement or role-specific exploration.

We note that when the environment feedback term γ is sufficiently large relative to the shock amplitude, a small to moderate amount of noise (β) can actually *improve* adaptation, rather than degrade it. In other words, there exists a narrow parameter region where partial diversity from random perturbations helps the group track the shifting optimum $a^*(t)$ more effectively. Although this phenomenon remains less pronounced than in our LLM-driven multi-agent experiments (where diversity is semantically grounded), it does illustrate that an "optimal" β can arise even in a purely random-iteration model, provided γ and the shock parameters are appropriately balanced.

Interpretation for Our Work. In the main experiments (Sections 4 and 5 of the main paper), we highlight that *LLM-based* agents with distinct roles produce *meaningful disagreements*, which can facilitate adaptability in shifting environments. The toy iteration model here clarifies that *if* diversity is solely random, performance typically decreases with β. Thus:

1. **No contradiction:** The simplified model's *monotonic* decline underscores how random "noise" alone undermines stable consensus, especially under frequent shocks.
2. **Need for structured diversity:** Real LLM agents do not *merely* add random perturbations; they incorporate textual cues, role differences, and partial knowledge–often generating beneficial exploration.

Therefore, while the random-iteration model is convenient for partial theoretical analysis (one can show conditions for convergence in expectation when β is small, etc.), it does *not* replicate the emergent synergy from purposeful agent heterogeneity. Future extensions could incorporate strategic exploration or role-based logic in a more advanced "consensus + diversity" iteration model, potentially revealing the inverted-U phenomenon seen in structured multi-agent dialogues.

References for Dynamic Iteration Models. The approach here is loosely related to classic work on **DeGroot** averaging (DeGroot, 1974) and **Friedkin–Johnsen** "stubbornness" models (Friedkin and Johnsen, 2011), extended to include environment shifts and additive noise. For comprehensive surveys on consensus protocols and opinion dynamics, see (Olfati-Saber et al., 2007; Hegselmann and Krause, 2002).

E Performance of Different Base Models

In summary, from Table 6 we can figure out that while `GPT-4o` consistently leads in overall performance, every base model achieves better results under implicit consensus, reaffirming the advantage of partial autonomy across all three scenarios.

Table 6: Performance of Different Base Models across the three scenarios. S1 (CR) is Coverage Rate, S2 (MS) is Misinformation Spread, S3 (PR) is Provision Rate. "Exp" and "Imp" refer to explicit vs. implicit consensus.

Base Model	S1: CR		S2: MS		S3: PR	
	Exp	Imp	Exp	Imp	Exp	Imp
GPT-4o	0.79	0.975	0.40	0.20	0.82	0.92
Claude-3-Sonnet	0.72	0.96	0.43	0.25	0.79	0.91
GPT-4o-mini	0.68	0.95	0.46	0.27	0.76	0.885
Qwen-Plus	0.63	0.94	0.49	0.33	0.74	0.875
Llama-2	0.575	0.935	0.52	0.38	0.715	0.88
Average	0.679	0.952	0.46	0.286	0.765	0.894

F Prompt Example and Dialogue Analysis

F.1 Dialogue Analysis on Dynamic Disaster Response (RQ3)

To address **RQ3**—how agents coordinate and revise their decisions in context based on each other's statements—we examined select rounds from the agent interaction log. Below, we highlight three observations demonstrating that *partial disagreement* and role-driven perspectives lead to adaptive, cooperative behavior.

(1) Role-Specific Choices Lead to Divergent Actions but Rapid Coverage. In the very first round (`round=0`), each agent independently selects different grid coordinates:

- **Drone 0 (Medical)** moves to `[5,5]` to address "immediate casualty evacuation".
- **Drone 1 (Infrastructure)** chooses `[6,6]` to "ensure power lines and roads remain functional".
- **Drone 2 (Logistics)** goes to `[4,5]` for "comprehensive coverage around high-severity zones".

These decisions show that even in the same round, they do not unify on a single location but rather *diverge* based on role priorities. As a result, multiple key zones are covered simultaneously.

(2) Agents Adapt Their Plans After Reading Others' Messages. By `round=2`, the Medical drone has chosen `[5,3]`, while Infrastructure and Logistics drones pick `[6,5]` or `[6,5]` respectively. Examining their messages, we find explicit references to each other's stated actions:

> *Drone 2 (Logistics)*: "I am positioning at [6,5] to assist with infrastructure preservation, ensuring we prevent potential overlap..."

This highlights how reading other drones' allocations (e.g., "someone is already at the casualty zone") motivates partial shifts in coverage. Rather than forcing a single group plan, the system allows each drone to deviate if it sees unaddressed needs elsewhere.

(3) Ongoing Coordination Prevents Over-Concentration. At later rounds (e.g., `round=7` and `round=8`), the Medical drone announces high-severity casualties in zones like `[7,2]` or `[7,8]`, while Infrastructure and Logistics drones opt for `[5,3]` or `[5,6]` to handle different concerns. Their messages show active avoidance of unnecessary overlap:

> *Drone 1 (Infrastructure)*: "I will focus on securing power lines and roads near [5,3]... Continue providing support at [7,2]."

Thus, partial disagreement again drives *diverse coverage*, ensuring that each critical need (casualty rescue, infrastructure defense, supply logistics) is handled. The drones repeatedly *reference* one another's chosen actions to avoid duplication, demonstrating a form of emergent in-context negotiation.

Summary for RQ3. These logs confirm that (a) each agent's specialized role leads to distinct decisions, (b) the presence of partial disagreement triggers broader coverage of dynamic hazards, and (c) agents revise their actions in response to dialogue updates rather than following a single script. Consequently, the system remains flexible, distributing resources where they are most needed while avoiding premature consensus on one zone. This supports our claim that implicit consensus structure and in-context learning ability of LLM agents fosters robustness through continuous adaptation and partial autonomy among agents.

F.2 Case Study 1: Dynamic Disaster Response

Prompt Template

You are Drone {id}, a {role} in a disaster response team.
Current situation: {grid description and disaster states}
Other drone messages: {messages from other drones}

Your role instructions: {role-specific guidelines}
Based on the current situation and your role, provide:
1. Your analysis of the situation
2. Your proposed action as grid coordinates [x,y]
3. A brief message to share with other drones
Format your response as JSON exactly like this example:
{ "analysis": "My analysis of the situation...",
"action": [3,4],
"message": "My message to other drones..."}

Role Types

Medical Drone:
"Focus on rescuing casualties in highest-severity disaster zones for people."
Infrastructure Drone:
"Protect power lines and roads. Even if severity is high elsewhere, prioritize built structures."
Logistics Drone:
"Minimize travel cost. Quickly move to nearest active zone if severity is above 5."

Agent Actions

- Choose grid coordinates [x,y] to move to
- Analyze situation severity
- Share tactical information with other drones

F.3 Case Study 2: Information Spread and Manipulation

Prompt Template

You are Defender {id}, a {role} in an information manipulation defense team.
Current situation: {network state and spread description}
Network information: {structure and metrics}
Other defender messages: {messages from other defenders}
Based on the situation and your role, provide:
1. Your analysis of the network state
2. Your proposed nodes to fact-check [maximum 3]
3. A brief message to share with other defenders
Format your response as JSON exactly like this example:
{ "analysis": "My analysis of the situation...",
"target_nodes": [1, 4, 7],
"message": "My message to other defenders..." }

Role Types

Proactive Defender:
"Prioritize checking high-influence nodes before they get infected. Focus on creating network firebreaks."
Reactive Defender:
"Target nodes that are actively spreading misinformation. Focus on reducing current spread."
Network Analyzer:
"Study network structure and identify critical nodes. Track infection patterns."
Rapid Responder:
"Quickly respond to new infections. Focus on containing new outbreaks."

Agent Actions

- Select nodes to fact-check (maximum 3 per round)
- Analyze spread patterns
- Share strategic insights about network vulnerabilities

F.4 Case Study 3: Dynamic Public-Goods Provision

Prompt Template

You are Contributor {id}, a {role} in a public goods provision team.
Current situation: {threshold and benefit description}
Previous outcomes: {last round results}
Other contributor messages: {messages from other contributors}
Based on the situation and your role, provide:
1. Your analysis of the situation
2. Your proposed contribution amount [0-{max_contribution}]
3. A brief message to share with other contributors

Format your response as JSON exactly like this example:
{"analysis": "My analysis of the situation...",
"contribution": 10.5,
"message": "My message to other contributors..."}

Role Types

Altruistic:
"Prioritize meeting the threshold to ensure public good provision. Willing to contribute more than fair share."
Strategic:
"Balance personal costs against public benefits. Adjust contributions based on others' behavior."
Conservative:
"Prefer smaller, safer contributions. Focus on sustainable long-term participation."
Adaptive:
"Quickly adjust to threshold and benefit changes. Learn from past outcomes."

Agent Actions

- Decide contribution amount [0, max_contribution]
- Analyze group dynamics
- Share strategic insights about optimal contribution levels