SolidWall: Achieving Security and Efficiency in Decentralized LLM Based Multi-Agent System

Anonymous Authors

Abstract

The transition to decentralized LLM-based Multi-Agent Systems (LLM-MAS) is driven by the imperative for data sovereignty and model heterogeneity. However, this architectural shift introduces a critical tension between security and efficiency: the absence of central oversight and the inherent distrust among agents expose systems to heightened semantic attacks, which can trigger cascading failures. Meanwhile, existing defenses relying on exhaustive agent-to-agent consensus impose communication overhead ($O(n^2)$) and high latency. To reconcile this conflict, we propose **SolidWall**, a framework designed to ensure robust alignment with minimal overhead. Architecturally, SolidWall decouples the probabilistic *LLM Reasoning Core* from a deterministic *Mandatory Safety Module (MSM)*, establishing a verifiable local root of trust that reduces communication complexity to a linear scale ($O(n)$). Algorithmically, we introduce *Random BFT Auditing* to sample dynamic, unbiased committees for lightweight consensus. Furthermore, by exploiting the *Generative-Discriminative Gap*, we implement *Light Auditing*, which replaces costly generative reconstruction with efficient discriminative verification. Empirical evaluations demonstrate that SolidWall restores over 80% of system performance under severe adversarial attacks while reducing inference latency by approximately 48%, effectively validating the feasibility of secure and scalable decentralized intelligence. Our work is available on https://anonymous.4open.science/r/MAS-security-2BB3/

1 Introduction

The rapid evolution of Large Language Models (LLMs) has catalyzed the development of LLM-based Multi-Agent Systems (LLM-MAS) [Brown *et al.*, 2020; Li *et al.*, 2024]. By leveraging inter-agent communication to facilitate collective reasoning, these systems exhibit emergent intelligence, enabling them to decompose and tackle complex, multi-step tasks that surpass the inherent capabilities of individual models [Guo *et al.*, 2024; Wang *et al.*, 2024].

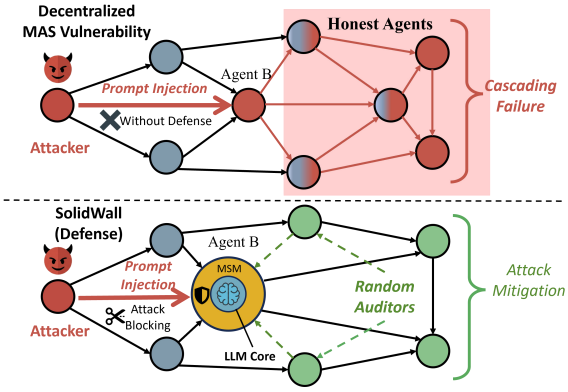

Figure 1: The Vulnerability of Decentralized MAS and our defense

In the current landscape, MAS are increasingly evolving towards decentralized architectures. While centralized orchestration facilitates unified control and simplifies agent coordination, it encounters inherent limitations in practical deployment. Specifically, reliance on a single central authority introduces scalability bottlenecks and critical single-point-of-failure risks [Jin *et al.*, 2024; Zhao *et al.*, 2025]. More critically, this transition is compelled by stringent privacy constraints and data sovereignty requirements. In the context of cross-organizational agent collaboration, to safeguard proprietary knowledge, core LLMs are strictly confined within organizational boundaries; they can neither be deployed locally by third parties nor can their behaviors be directly manipulated by external central nodes. Therefore, constructing a *trustless* MAS capable of collaboration without a central authority has emerged as a pivotal research topic [Yang *et al.*, 2025; Ding *et al.*, 2025].

The transition to decentralized LLM-MAS significantly expands the attack surface by fundamentally shifting the trust architecture [Yu *et al.*, 2025a; Kushwaha *et al.*, 2025]. In the absence of central oversight, the inherent distrust among agents renders systems vulnerable to semantic attacks—such as hallucinations and adversarial prompt injections—where a single compromised node can trigger cascading failures that corrupt the global consensus (as illustrated in the top panel of Figure 1) [Lee and Tiwari, 2024; Yu *et al.*, 2025b; Zhang *et al.*, 2024b]. Current defenses are structurally ill-suited for

this context. Centralized strategies relying on global auditors introduce critical single points of failure and deployment hurdles [Wang *et al.*, 2025; Shen *et al.*, 2025]. Meanwhile, due to the absence of a trusted central authority, existing decentralized mechanisms (e.g., debate or voting) necessitate that all agents broadcast their results globally to perform exhaustive verification for each interaction [Chen *et al.*, 2024; Zhao *et al.*, 2024]. This requirement imposes a quadratic communication overhead ($\mathcal{O}(n^2)$) and high latency, severely bottlenecking real-time efficiency. *Consequently, there is an urgent demand for a framework that ensures robust alignment in decentralized LLM MAS settings while still maintaining its high efficiency*.

To address these challenges, we propose **SolidWall**, a decentralized framework designed to reconcile safety with efficiency (as shown in the bottom panel of Figure 1). Structurally, we establish a reliable local trust anchor by decoupling the agent into a probabilistic *LLM Reasoning Core* and a deterministic *Mandatory Safety Module (MSM)*. To mitigate the overhead of exhaustive consensus, we implement a *Random BFT Auditing* mechanism that stochastically samples a minimal committee of unbiased auditors, thereby avoiding the redundancy of global broadcasting. This synergistic combination of architectural decoupling and committee-based auditing reduces the communication complexity from quadratic to linear ($O(n)$). Furthermore, we exploit the *Generative-Discriminative Gap* to introduce *Light Auditing*, enabling auditors to validate outputs via a single forward pass rather than computationally expensive autoregressive reconstruction, thereby ensuring real-time responsiveness.

Extensive theoretical analysis and experiments show that SolidWall can achieve a recovery of 80% under attacks in decentralized LLM-MAS, while reducing auditing latency by about 48% compared to existing defense methods. Our main contributions are as follows:

Decentralized-Native Architecture: We introduce a framework that reconciles the tension between alignment rigor and efficiency. By architecturally decoupling probabilistic reasoning from deterministic Mandatory Safety Modules, we establish a verifiable local trust root, reducing communication complexity from quadratic $\mathcal{O}(n^2)$ to linear $\mathcal{O}(n)$.

Efficient Random Auditing: We devise a Random BFT Auditing mechanism that exploits the *Generative-Discriminative Gap*. By replacing costly autoregressive reconstruction with lightweight discriminative verification within minimal committees, we minimize computational overhead without compromising probabilistic safety.

Rigorous Validation: Extensive evaluations demonstrate that SolidWall provides topology-agnostic resilience, restoring over 80% of system performance under severe attacks while reducing inference latency by approximately 48% compared to existing methods.

2 Related Work

Decentralized MAS research has evolved from independent learning to sophisticated coordination. While IPPO serves as a robust baseline [Yu *et al.*, 2022], it often struggles with non-stationarity. To address this, architectures like MAT [Wen *et al.*, 2022] and HAPPO [Kuba *et al.*, 2021] explicitly model agent interdependencies. Addressing security in open networks, recent frameworks utilize blockchain for decentralized identity [Ding *et al.*, 2025] and hybrid consensus to mitigate Byzantine failures. Furthermore, LLMs now enable cognitive synergy through semantic planning and natural language debate [Yang *et al.*, 2025].

MAS Attacks. Security threats against MAS are garnering increasing attention [Yu *et al.*, 2025a]. Agent autonomy and communication capabilities introduce vulnerabilities beyond traditional LLMs [Luo *et al.*, 2025; Dong *et al.*, 2024]. Compromised agents can corrupt shared information [Yu *et al.*, 2025b], tamper with memory [Dong *et al.*, 2025], or propagate misinformation via falsified consensus [Wang *et al.*, 2025; Chern *et al.*, 2024]. Unlike isolated failures, these attacks spread through networked dependencies, causing systemic cascades that complicate defense [Zheng *et al.*, 2025; Gu *et al.*, 2024].

MAS Defenses. For multi-agent systems, defense strategies often focus on ensuring robust interactions and communications. One category of approaches employs dedicated supervisory agents to monitor information exchange and verify message correctness within the system. These methods have demonstrated dynamic and efficient detection capabilities [Wu *et al.*, 2025; Lin *et al.*, 2025], or leverage techniques like Graph Neural Networks (GNNs) to detect the propagation of malicious information [Wang *et al.*, 2025; Shen *et al.*, 2025]. Another prominent category relies on decentralized consensus processes among multiple agents to eliminate single points of failure. These methods ensure output safety through multi-round inter-agent debates or voting mechanisms [Chen *et al.*, 2024; Zhao *et al.*, 2024; Huang *et al.*, 2024].

We propose **SolidWall**, a novel framework that introduces a dual-layer agent architecture paradigm synergized with a Random BFT Auditing protocol. This design ensures security in decentralized environments while maintaining lightweight communication and minimal latency overhead.

3 Preliminary

This section defines the graph-theoretical models for multi-agent systems and elaborates on the complex adversarial threats the systems face.

3.1 Multi-Agent System

We abstract the Multi-Agent System (MAS) as a graph

$$G = (V, E)$$

$V = \{A_1, A_2, \ldots, A_n\}, E \subseteq V \times V$, where V is the set of n agents and E represents the communication channels. A directed edge $(A_i, A_j) \in E$ signifies that agent A_i can transmit messages or intermediate results to agent A_j. This abstraction supports an arbitrary network topology, allowing for flexible collaboration patterns such as chain, tree, ring, or fully connected structures, without imposing specific architectural constraints.

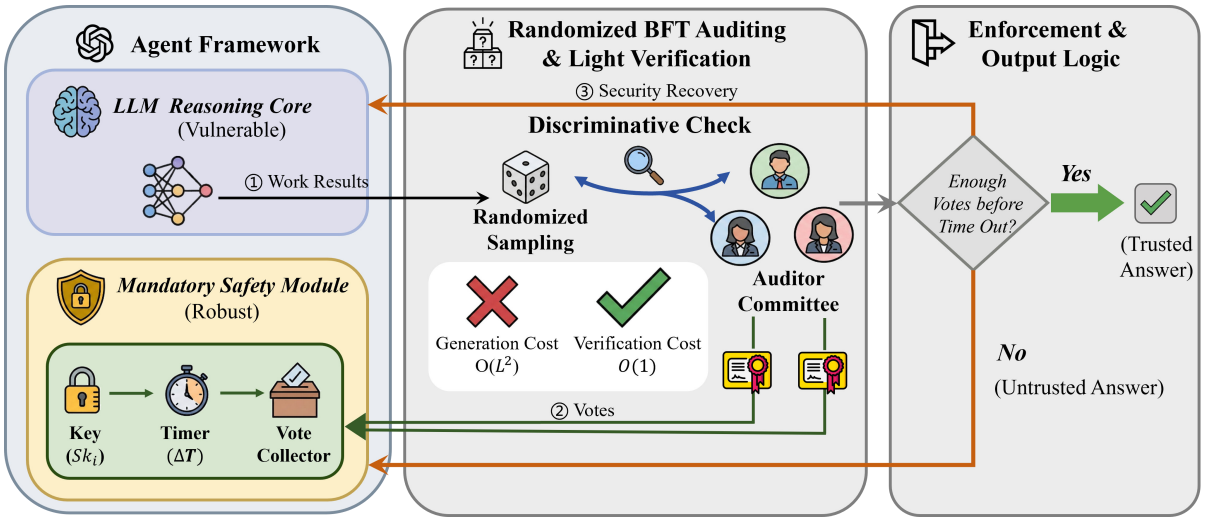

Figure 2: Overview of **SolidWall**

3.2 Attack Strategy

We consider a system of n agents facing a *Semantically Adaptive Adversary* who controls f malicious nodes. This adversary specifically targets the inherent vulnerabilities of the LLM reasoning core. The attack strategy is twofold:

- **Exploiting Internal Stochasticity (Hallucinations):** The adversary leverages the probabilistic nature of LLMs to amplify stochastic hallucinations. By subtly manipulating context, compromised agents can induce "confabulations"—plausible but factually incorrect outputs—that erode the semantic integrity of the consensus process without triggering syntactic error detectors.
- **External Semantic Injection (Prompt & Backdoor Attacks):** The adversary actively deploys sophisticated external attacks, including Misinformation Injection, Bias Injection and Role Hijacking [Lee and Tiwari, 2024; Zhang *et al.*, 2024a]. Furthermore, the adversary may utilize latent backdoors activated by specific triggers, forcing the LLM to deviate from its alignment protocols and generate toxic or deceptive plans.

Ultimately, these attacks aim to precipitate systemic cascading failures by corrupting the global consensus.

4 Method

We propose **SolidWall**, a decentralized alignment framework that bridges the expressive power of probabilistic generative models with the rigor of deterministic verification. The core philosophy of SolidWall is to decouple the *stochastic reasoning process* (handled by the LLM as a *Generator*) from the deterministic Protocol Enforcement (enforced by the Mandatory Safety Module, MSM). To address the inherent epistemic uncertainty and potential adversarial deviations in MAS, SolidWall introduces a Random BFT Auditing mechanism. This approach stochastically selects a subset of agents to function as unbiased Discriminators, estimating the global semantic consistency with minimal communication overhead.

4.1 Framework Architecture

To ensure robust coordination under uncertainty, SolidWall decouples the agent architecture into a probabilistic generator and a deterministic constraint enforcer (see Fig. 2). This separation of concerns allows us to isolate high-entropy reasoning risks from low-entropy protocol execution.

LLM Reasoning Core (Probabilistic Generator). This module handles *semantic processing tasks* and functions as the agent's cognitive engine. Acting as the generative policy π_θ, it synthesizes high-dimensional plans and responses based on the current context. Crucially, in our framework, the LLM is versatile: it acts as a generator when proposing solutions and shifts roles to function as a semantic discriminator when designated as an auditor. However, due to the inherent stochasticity and hallucination risks of LLMs, all outputs from this core are treated strictly as *tentative proposals*. They remain uncommitted and invisible to the global consensus until validated by the safety layer.

Mandatory Safety Module (Deterministic Protocol Enforcement). This module executes *protocol enforcement tasks* and serves as a verifiable local trust anchor. Distinct from the neural network, the Mandatory Safety Module (MSM) implements rigid, deterministic logic to govern the consensus lifecycle. It acts as a strict gatekeeper, asynchronously aggregating votes from peer auditors to filter out stochastic errors. A proposal is formally committed only if the accumulated approval exceeds a predefined Byzantine threshold τ within a specific time window ΔT. To optimize system efficiency, the MSM enforces a mathematically rigorous *early rejection* policy: if the count of negative votes reaches a point where the consensus threshold τ becomes theoretically unreachable (i.e., exceeding $m - \tau$), the proposal is discarded immediately. This mechanism ensures that individual reasoning failures are intercepted locally before they can propagate as cascading errors.

4.2 Algorithmic Process

To ensure robust security with minimal latency, we introduce *Randomized BFT Auditing*. This algorithm decouples the verification complexity from the total network scale, enabling the system to maintain high throughput even as the number of agents grows.

(1) Probabilistic Auditor Sampling. To mitigate the overhead of global consensus, we adopt a lightweight auditing paradigm via a compact, dynamic committee $\mathcal{A} \subset \mathcal{V}$ ($|\mathcal{A}| = m \ll n$). We employ BLS threshold cryptography to construct a verifiable distributed randomness beacon [Syta *et al.*, 2017]. The sampling seed is derived from a unique BLS signature aggregated from context-specific shares. Unlike simple hash-based methods, this ensures: (i) **Unbiasability**, preventing adversaries from manipulating the random output without controlling a threshold of shares; and (ii) **Unpredictability**, concealing the committee composition until signature reconstruction. These properties prevent adaptive adversaries from launching targeted corruption attacks in advance, ensuring the committee serves as a robust and unbiased estimator of the global consensus.

(2) Semantic Voting and Certification. The core consensus mechanism relies on semantic verification rather than mere protocol adherence. Each selected auditor $A_j \in \mathcal{A}$ acts as a discriminator, evaluating the semantic consistency and logical validity of the proposal $\mathcal{O}$ against the shared context C. Unlike the high-temperature generation used for reasoning, auditors employ low-temperature discriminative prompting to output a binary verification signal. A vote v_j is cast only if the discriminative confidence exceeds a rigid safety threshold ρ:

$$v_j = \mathbb{I}\left[P_\theta(\text{Valid} \mid \mathcal{O}, C) \geq \rho\right]$$

While individual auditors remain vulnerable to sophisticated adversarial attacks, Random BFT neutralizes this threat as the collective deception probability P_{succ} decays exponentially with the committee size. Subsequently, the MSM asynchronously collects these digital signatures. Upon reaching the threshold τ, the proposal transitions from a tentative state to a committed state. If the proposal fails to garner sufficient support or triggers the early rejection condition, the MSM initiates a security recovery protocol, prompting the reasoning core to regenerate a compliant response.

4.3 Optimization: Light Auditing

To mitigate the overhead inherent in multi-agent coordination, we propose *Light Auditing*, a paradigm that exploits the fundamental computational asymmetry in LLM reasoning.

The Generative-Discriminative Gap. SolidWall significantly reduces computational costs by shifting the consensus burden from generative reconstruction to discriminative verification. Traditional consensus mechanisms often require peer agents to re-generate reasoning paths or engage in verbose debates, incurring an autoregressive cost of $O(L \cdot (N + L))$, where L is the output length. In contrast, we treat the proposal as a fixed context and compute the discriminative probability $P(\text{Valid}|\mathcal{O}_i, \text{Context})$. This operation requires only a single forward pass, reducing the complexity to $O(1 \cdot (N + L))$. This yields an efficiency gain $\eta \approx L$ that scales linearly with reasoning depth, allowing the system to support a larger, more robust auditor committee m without the prohibitive latency penalty associated with generative ensembles, thereby preserving real-time system responsiveness.

Reliability via Verification Asymmetry. This optimization also enhances system robustness against stochastic hallucinations. We leverage the principle that verification is inherently more reliable than generation ($P_{\text{verify}} > P_{\text{generate}}$), as the search space for binary discrimination is vastly smaller than that of open-ended generation. By applying the *Condorcet Jury Theorem*, we observe that the probability of collective error decays exponentially as the committee size m increases, provided the auditors are sampled independently. This effectively filters out high-entropy hallucinations, ensuring that the system's final output reflects a robust semantic consensus rather than correlated stochastic noise.

5 Analysis

We analyze SolidWall through a game-theoretic lens, modeling the interaction between the defense protocol and a rational adversary as a resource-constrained game. We demonstrate that our mechanism imposes an asymmetric cost structure, where the cost of successful subversion grows exponentially while the verification overhead remains linear.

5.1 Security: Adversarial Utility Decay

Consider a system of n agents where an adversary controls a fraction f/n. The adversary's goal is to corrupt the consensus output, obtaining a utility U_{adv}. In our Random BFT framework, a successful attack requires the adversary to dominate the dynamically sampled committee $\mathcal{A}$ of size m. Since the Mandatory Safety Module (MSM) acts as a deterministic commitment device enforcing a strict threshold τ, the probability of adversarial success, P_{succ}, is governed by the tail of the hypergeometric distribution $H(k; n, f, m)$:

$$P_{succ} = P(X \geq \tau) = \sum_{k=\tau}^{m} \frac{\binom{f}{k}\binom{n-f}{m-k}}{\binom{n}{m}}$$

Using the Chvátal-Hoeffding bound, this probability decays exponentially with respect to the committee size m:

$$P_{succ} \leq \exp\left(-2m\left(\frac{\tau}{m} - \frac{f}{n}\right)^2\right)$$

For a rational adversary, the expected payoff is $\mathbb{E}[U] = G \cdot P_{succ} - C_{attack}$, where G is the gain from corruption and C_{attack} is the resource cost. As $P_{succ} \to 0$ exponentially with m, the marginal cost to maintain a non-negligible attack probability becomes prohibitive. Thus, honesty (or non-participation) becomes the dominant strategy for resource-bounded adversaries.

5.2 Efficiency: Multi-Dimensional Cost Reduction

We define system burden $\mathcal{J}$ as the product of communication complexity and verification cost. Traditional defenses incur a quadratic penalty $\mathcal{J}_{base} \propto n^2 \cdot \mathcal{C}_{gen}$ via autoregressive

Table 1: Defense performance of SolidWall across different attack types. Baseline and Attack represent the scenarios without attack and under attack (no defense), respectively. SolidWall denotes our proposed defense. Bold values indicate the best performance under attack.

		Misinformation injection			Role hijacking			Bias injection		
		Baseline	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**
CSQA	**Chain**	90.74	66.34	87.69	90.74	74.07	86.73	90.74	76.93	88.12
	Circle	89.89	65.96	87.04	89.89	72.24	85.79	89.89	74.15	86.49
	Complete	88.76	68.52	85.19	88.76	72.37	86.04	88.76	73.87	85.24
	Star	91.59	72.22	88.31	91.59	74.67	88.89	91.59	74.62	88.40
	Tree	90.30	70.37	87.13	90.30	73.76	87.94	90.30	75.03	87.72
		Misinformation injection			**Role hijacking**			**Bias injection**		
		Baseline	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**
GSM8K	**Chain**	80.1	51.67	72.71	80.1	60.54	73.56	80.1	63.37	73.81
	Circle	76.67	50.78	70.97	76.67	58.33	72.07	76.67	65.03	71.73
	Complete	75.81	48.34	68.33	75.81	57.75	71.38	75.81	63.71	69.14
	Star	81.67	55.10	74.32	81.67	61.76	75.35	81.67	66.71	75.52
	Tree	78.33	52.36	69.75	78.33	60.19	71.06	78.33	65.35	72.09
		Misinformation injection			**Role hijacking**			**Bias injection**		
		Baseline	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**
MMLU	**Chain**	90.04	68.37	86.44	90.04	73.65	87.12	90.04	80.04	88.10
	Circle	89.97	71.32	85.03	89.97	73.33	85.37	89.97	81.15	86.24
	Complete	91.63	66.75	87.16	91.63	70.04	86.87	91.63	76.92	89.23
	Star	90.13	70.30	86.69	90.13	71.83	85.98	90.13	78.33	85.79
	Tree	92.14	75.13	88.12	92.14	75.27	87.67	92.14	76.19	88.62

Table 2: Performance of SolidWall against centralized defenses (cent) and Blockagent

	Baseline	**Attack**	**Cent**	**Blockagent**	**SolidWall**
chain	90.74	66.34	81.76	88.17	**87.69**
cycle	89.89	65.96	78.65	88.54	**87.04**
complete	88.76	68.52	80.38	86.75	**85.19**
Star	91.59	72.22	84.62	89.36	**88.31**
tree	90.30	70.37	83.79	87.95	**87.13**

debates. SolidWall optimizes both dimensions: restricting consensus to committee $\mathcal{A}$ reduces communication to $O(n)$, while Light Auditing replaces generation with efficient discriminative verification ($\mathcal{C}_{disc} \ll \mathcal{C}_{gen}$). The total cost $\mathcal{J}_{SW}$ is derived as:

$$\mathcal{J}_{SW} \approx \underbrace{O(n)}_{\text{Topology}} \cdot \underbrace{\mathcal{C}_{disc}}_{\text{Computation}} \ll \underbrace{O(n^2)}_{\text{Topology}} \cdot \underbrace{\mathcal{C}_{gen}}_{\text{Computation}}$$

With output length L, the efficiency gain scales as $\Gamma = \frac{\mathcal{J}_{base}}{\mathcal{J}_{SW}} \propto n \cdot L$. This confirms SolidWall decouples verification from network scale (n) and reasoning depth (L), ensuring scalable high-throughput MAS.

6 Experiment

In this section, we evaluate the performance of **SolidWall** under various adversarial conditions and multi-agent topologies. Our experiments aim to comprehensively assess the framework's *security*, *scalability*, and *efficiency* when deployed in different collaborative environments.

Through these experiments, we address the following research questions:

RQ1: Can *SolidWall* effectively defend against various types of attacks in MAS and maintain correct cooperative outcomes?

RQ2: Does *SolidWall* demonstrate strong scalability and portability, enabling seamless integration across MAS of different sizes and model backbones?

RQ3: Does *SolidWall* achieve robust protection with low additional computational and communication overhead?

6.1 Setup

Dataset. We evaluate the defense performance of SolidWall against adversarial attacks using three types of datasets: (1) Undisputed Facts, (2) Simple Reasoning, and (3) Complex Reasoning. All datasets are constructed from well-known benchmarks, including *CSQA (CommonsenseQA)* [Talmor *et al.*, 2018], *GSM8K* [Cobbe *et al.*, 2021], and *MMLU* [Hendrycks *et al.*, 2020]. For each dataset, we randomly select 800 samples as the basis of our evaluation.

Experiment Settings. We simulated the agent collaboration environment on the AutoGen framework [Wu *et al.*, 2024]. We comprehensively evaluate the performance of SolidWall under different attack strategies, network topologies, and large language model (LLM) settings. Specifically, the evaluation includes four categories of attacks: Role Hijacking [Zhang *et al.*, 2024a], misinformation and bias injection [Lee and Tiwari, 2024], bias injection [Yu *et al.*, 2025b], and **collusion between malicious workers and auditors**. For the communication structure, five classical multi-agent topologies are considered: Chain, Cycle, Tree, Star, and Complete graphs. We use gemini-2.5-flash as the primary model in our experiments, and further extend the evaluation to MAS with different scales and model backbones. The testing accuracy% is reported.

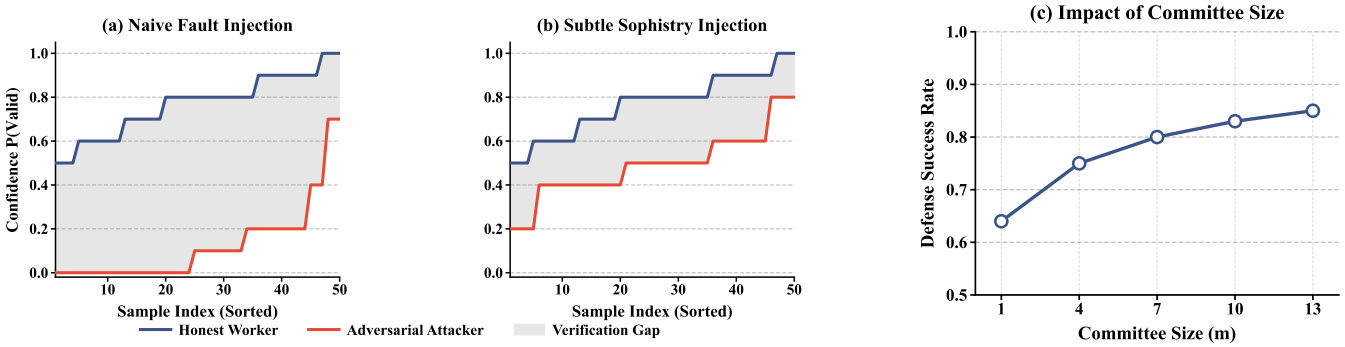

Figure 3: Validation of discriminative auditing and collective auditing. The erosion of individual decision boundaries (a, b) necessitates the Random BFT consensus (c).

Table 3: SolidWall Performance under Different Model

	GPT3.5			deepseek-V3			qwen-3-max			gemini-2.5		
	Baseline	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**	**Baseline**	**Attack**	**SolidWall**
Chain	79.84	55.08	76.75	75.26	54.13	73.84	88.89	65.75	86.81	90.56	65.46	87.61
Cycle	77.14	51.72	73.92	76.10	53.37	73.35	90.34	64.87	85.17	88.79	67.15	86.32
Complete	74.66	48.94	72.21	74.93	57.54	72.81	97.94	65.62	83.56	89.04	63.59	88.10
Star	75.91	47.60	71.42	76.75	53.47	72.13	78.57	66.73	78.29	91.27	63.22	89.71
Tree	73.17	53.45	71.71	77.54	56.92	73.67	80.14	63.58	80.73	90.75	66.13	88.36

6.2 Empirical Validation of Auditing Mechanism

To validate the discriminative auditing paradigm,we sampled 50 instances from MMLU to evaluate auditor confidence ($P(\text{Valid})$) and analyzed the distribution of confidence scores. Under *Naive Fault Injection* (Fig. 3a), a distinct *verification gap* allows honest reasoning to be easily distinguished from overt errors. However, *Subtle Sophistry Injection* (Fig. 3b) employs authoritative mimicry to erode this decision boundary, rendering individual verification unreliable ($P(\text{Valid}) > 0.6$ for adversarial samples).

This contrast motivates our Random BFT mechanism. To demonstrate the efficacy of collective consensus, we evaluate defense robustness under subtle attacks across varying committee sizes $m \in \{1, 4, 7, 10, 13\}$ with corresponding Byzantine thresholds $\tau \in \{1, 3, 5, 7, 9\}$. As illustrated in Fig. 3c, increasing the committee size effectively suppresses the high-entropy tail of sophisticated sophistry. By leveraging the *variance reduction* of collective voting, SolidWall statistically neutralizes false positives that fool individual auditors, confirming the necessity of multi-agent consensus in complex adversarial regimes.

6.3 Defense Performance

Robustness Across Topologies. We evaluate SolidWall on an 8-agent MAS (1 malicious) with auditing parameters $m = 4$ and $\tau = 3$. As summarized in Table 1, while attacks cause severe degradation, SolidWall consistently restores performance to near-baseline levels. Across all 45 configurations, our method achieves over 80% recovery, typically maintaining accuracy within a 3%–5% margin of the benign baseline. Notably, this effectiveness remains topology-agnostic, demonstrating stability even in complex structures like *Complete* and *Tree*.

Furthermore, we benchmark SolidWall against centralized defenses [Zhu *et al.*, 2023] and heavy decentralized defenses BlockAgent [Chen *et al.*, 2024] on CSQA (Table 2). SolidWall outperforms centralized methods and matches BlockAgent's robustness. Crucially, it achieves this defense parity via lightweight auditing, avoiding the computational overhead of multi-round debates and thus optimizing the security-efficiency trade-off.

6.4 Scalability and Portability

Model Agnosticism. We further assess portability across diverse LLM backbones in Table 3. While absolute performance varies by base model capabilities, SolidWall consistently recovers collaborative accuracy across all architectures. This underscores the framework's model-agnostic adaptability, allowing seamless integration into diverse MAS ecosystems without requiring model-specific tuning.

Scalability vs. Adversarial Density. Table 4 evaluates SolidWall on the CSQA benchmark across varying MAS sizes and adversarial ratios. With sparse adversaries, the system maintains near-optimal accuracy indistinguishable from benign baselines. This indicates robust scalability, with SolidWall effectively suppressing error propagation even in high-saturation adversarial environments.

6.5 Efficiency

Finally, we evaluate the efficiency of SolidWall. We compare our framework against the prominent decentralized defense method *BlockAgent* under different attack settings with 6 agents. Figure 4 reports the additional time overhead relative to a system without any defense.

In the no-attack setting, SolidWall introduces only 35.4% average additional time overhead compared to the no-defense

Table 4: Performance of SolidWall across various agent populations (n) and malicious ratios (f). The values represent accuracy%

Topology	$n = 10$			$n = 20$			$n = 50$		
	$f = 0$	$f = 0.1$	$f = 0.3$	$f = 0$	$f = 0.1$	$f = 0.3$	$f = 0$	$f = 0.1$	$f = 0.3$
Chain	93.44	91.57	89.13	84.71	83.19	80.45	82.13	81.57	77.35
Cycle	92.71	89.93	88.34	85.50	85.47	79.83	83.75	82.14	78.19
Complete	88.62	84.88	78.57	79.93	79.64	76.58	82.46	79.95	76.22
Star	89.35	88.51	86.54	89.27	90.04	87.90	84.78	84.16	79.37
Tree	91.76	91.10	85.92	83.75	82.15	81.32	83.31	82.48	78.05

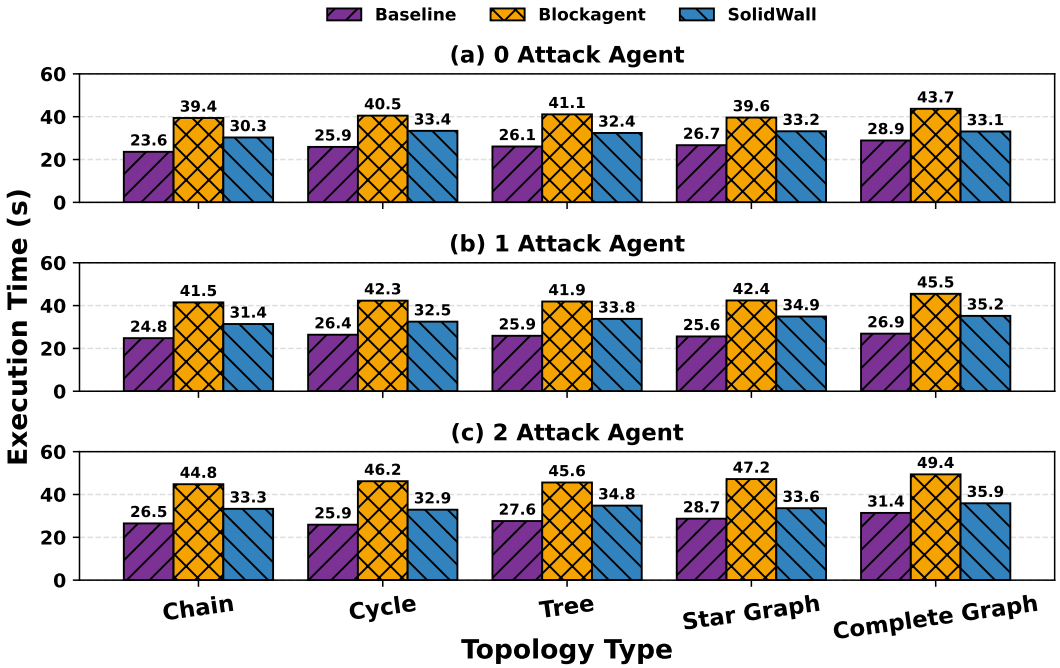

Figure 4: Execution time of different methods with a total of 6 agents under varying numbers of attack agents; Baseline denotes the case without any defense mechanism.

baseline. By contrast, BlockAgent [Chen *et al.*, 2024] incurs 67.8% overhead. This corresponds to a 47.8% reduction in extra cost, demonstrating the low-overhead and high-efficiency nature of our design. Even in the presence of malicious agents, SolidWall maintains an average overhead of only 37.1%, which is still much lower than the 72.3% overhead of BlockAgent. A possible reason for reduced overhead during attacks is that, without security auditing, agents may require longer deliberation to handle malicious information.

7 Conclusion

This paper addresses the vulnerability of decentralized LLM-MAS to cascading reasoning failures. We propose **SolidWall**, a framework that reconciles verification rigor with efficiency by decoupling probabilistic generation from deterministic safety enforcement via a verifiable Mandatory Safety Module (MSM). This architecture enables scalable Random BFT Auditing with linear communication complexity ($\mathcal{O}(n)$) and facilitates feedback-driven policy refinement. Empirical results demonstrate that SolidWall recovers over 80% of system performance under attacks while reducing inference latency by 48% compared to baselines. By establishing a verifiable trust boundary, SolidWall provides a foundational step toward resilient and scalable decentralized intelligence.

8 Discussion

SolidWall primarily addresses semantic-level threats, specifically focusing on mitigating adversarial prompt injections and stochastic hallucinations in LLM-MAS. While our framework establishes a local root of trust through the Mandatory Safety Module (MSM), future work should explore broader system-level defenses. Integrating Trusted Execution Environments (TEE) or advanced cryptographic primitives like Zero-Knowledge Proofs could further harden the decentralized protocol. However, a critical trade-off exists: while these methods enhance security, they introduce significant computational and communication overhead. In bandwidth-constrained environments, balancing rigorous verification with real-time responsiveness remains a pivotal challenge for scalable decentralized intelligence.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2024] Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. Blockagents: Towards byzantine-robust llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, pages 187–192, 2024.

[Chern *et al.*, 2024] Steffi Chern, Zhen Fan, and Andy Liu. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*, 2024.

[Cobbe *et al.*, 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[Ding *et al.*, 2025] Yepeng Ding, Ahmed Twabi, Junwei Yu, Lingfeng Zhang, Tohru Kondo, and Hiroyuki Sato. Decentralized multi-agent system with trust-aware communication. pages 1439–1445, 10 2025.

[Dong *et al.*, 2024] Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024.

[Dong *et al.*, 2025] Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. A practical memory injection attack against llm agents. *arXiv preprint arXiv:2503.03704*, 2025.

[Gu *et al.*, 2024] Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*, 2024.

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

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[Huang *et al.*, 2024] Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael R Lyu, and Maarten Sap. On the resilience of llm-based multi-agent collaboration with faulty agents. *arXiv preprint arXiv:2408.00989*, 2024.

[Jin *et al.*, 2024] Anan Jin, Yuhang Ye, Brian Lee, and Yuansong Qiao. Decoagent: Large language model empowered decentralized autonomous collaboration agents based on smart contracts. *IEEE Access*, 2024.

[Kuba *et al.*, 2021] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

[Kushwaha *et al.*, 2025] Ankita Kushwaha, Kiran Ravish, Preeti Lamba, and Pawan Kumar. A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. *arXiv preprint arXiv:2505.17342*, 2025.

[Lee and Tiwari, 2024] Donghyun Lee and Mo Tiwari. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*, 2024.

[Li *et al.*, 2024] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

[Lin *et al.*, 2025] Fulin Lin, Shaowen Chen, Ruishan Fang, Hongwei Wang, and Tao Lin. Stop wasting your tokens: Towards efficient runtime multi-agent systems. *arXiv preprint arXiv:2510.26585*, 2025.

[Luo *et al.*, 2025] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

[Shen *et al.*, 2025] Xu Shen, Qi Zhang, Song Wang, Zhen Tan, Xinyu Zhao, Laura Yao, Vaishnav Tadiparthi, Hossein Nourkhiz Mahjoub, Ehsan Moradi Pari, Kwonjoon Lee, et al. Metacognitive self-correction for multi-agent system via prototype-guided next-execution reconstruction. *arXiv preprint arXiv:2510.14319*, 2025.

[Syta *et al.*, 2017] Ewa Syta, Philipp Jovanovic, Eleftherios Kokoris Kogias, Nicolas Gailly, Linus Gasser, Ismail Khoffi, Michael J Fischer, and Bryan Ford. Scalable bias-resistant distributed randomness. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 444–460. Ieee, 2017.

[Talmor *et al.*, 2018] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[Wang *et al.*, 2024] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[Wang *et al.*, 2025] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. *arXiv preprint arXiv:2502.11127*, 2025.

[Wen *et al.*, 2022] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

[Wu *et al.*, 2024] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[Wu *et al.*, 2025] Chengcan Wu, Zhixin Zhang, Mingqian Xu, Zeming Wei, and Meng Sun. Monitoring llm-based multi-agent systems against corruptions via node evaluation. *arXiv preprint arXiv:2510.19420*, 2025.

[Yang *et al.*, 2025] Yingxuan Yang, Huacan Chai, Shuai Shao, Yuanyi Song, Siyuan Qi, Renting Rui, and Weinan Zhang. Agentnet: Decentralized evolutionary coordination for LLM-based multi-agent systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

[Yu *et al.*, 2025a] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6216–6226, 2025.

[Yu *et al.*, 2025b] Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Kun Wang, Qingsong Wen, and Yang Wang. Netsafe: Exploring the topological safety of multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2905–2938, 2025.

[Zhang *et al.*, 2024a] Yuyang Zhang, Kangjie Chen, Xudong Jiang, Yuxiang Sun, Run Wang, and Lina Wang. Towards action hijacking of large language model-based agent. *arXiv preprint arXiv:2412.10807*, 2024.

[Zhang *et al.*, 2024b] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. *arXiv preprint arXiv:2401.11880*, 2024.

[Zhao *et al.*, 2024] Xiutian Zhao, Ke Wang, and Wei Peng. An electoral approach to diversify llm-based multi-agent collective decision-making. *arXiv preprint arXiv:2410.15168*, 2024.

[Zhao *et al.*, 2025] Yang Zhao, Hanjiang Luo, Hang Tao, Jinyin Li, Chao Liu, and Jiehan Zhou. Large language model enhanced multi-uav direct cross-boundary maritime data collection scheme. In *2025 34th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–9. IEEE, 2025.

[Zheng *et al.*, 2025] Can Zheng, Yuhan Cao, Xiaoning Dong, and Tianxing He. Demonstrations of integrity attacks in multi-agent systems. *arXiv preprint arXiv:2506.04572*, 2025.

[Zhu *et al.*, 2023] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.