Bounded-k Blackboard Coordination, Correlated-Evidence Sequential Consensus, and Sink-Enforced Lease Revocation in Multi-Agent DAGs

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Scaling multi-agent language model workflows across hundreds of concurrent agents faces three coupled bottlenecks: all-to-all broadcasting incurs quadratic $\mathcal{O}(N^2)$ token overhead, homogeneous LLM verifiers exhibit strong error correlation that invalidates standard sequential consensus bounds, and voluntary agent lock release fails in up to 58% of closing actions. We formulate and evaluate a Bounded-k Blackboard, Conditional-Innovation SPRT, and Sink-Fenced Revocation (BCSR) coordination architecture for multi-agent directed acyclic graphs (DAGs). First, replacing direct peer broadcast with a bounded-k blackboard B-tree index reduces total coordination traffic from $\mathcal{O}(N^2)$ to quasi-linear $\mathcal{O}(Nk \log N)$ ($\mathcal{O}(k \log N)$ per agent). Second, grounded in Wald's Sequential Probability Ratio Test (SPRT) and Goel et al.'s correlated-oversight theory, we derive the exact conditional innovation log-likelihood ratio increment $\Delta\Lambda_m$ for equicorrelated Gaussian evidence and allow inconclusive terminal outcomes at horizon $M_{\max}$. Third, we replace voluntary lock release with a sink-enforced monotonic fencing state machine and edge revocation lists (VERA), where the resource sink verifies $t_{\text{sink}} \leq T_{\text{exp}} - \delta_{\text{skew}}$ and $f_{\text{epoch}} \geq f_{\max}^{\text{seen}}$. Across 2,000 DAG episodes (500 trials per regime at $N = 1{,}024$ and $\rho_{\text{corr}} = 0.45$), BCSR raises verified DAG completion from 0.00% $[0.00, 0.76]$ (all-to-all debate) and 38.00% $[33.85, 42.33]$ (uncorrected SPRT) to 89.60% $[86.62, 91.98]$ with 97.60% $[95.85, 98.62]$ revocation compliance at 27.06% per-agent overhead. A 2×2 factorial ablation separates statistical whitening from verifier heterogeneity.

1 Introduction

Coordinating large populations of autonomous language model agents over interdependent task directed acyclic graphs (DAGs) requires simultaneously managing communication complexity, epistemic aggregation under correlated errors, and deterministic resource revocation [7, 13, 19]. Early multi-agent LLM frameworks relied on conversational all-to-all debate or centralized star orchestrators. While effective for small teams of three tofive agents, empirical benchmarks in 2025 and 2026 demonstrate that conversational topologies degrade sharply as population size N scales from tens to hundreds of agents [1, 13, 19].

1.1 Three empirical failure modes of large-scale multi-agent coordination

Recent empirical studies and classical distributed-systems theory identify three specific mechanisms that cause multi-agent LLM workflows to fail at scale:

1. **Quadratic $\mathcal{O}(N^2)$ broadcast saturation vs. topological hub bottlenecks.** In unstructured all-to-all debate, exchanging state updates among N agents generates $\mathcal{O}(N^2)$ messages per round, rapidly exhausting context windows and token budgets [4, 19]. Conversely, routing all task assignments through a single centralized orchestrator creates a serial queueing bottleneck and a single point of topological failure [8–10].

2. **Sycophantic herding and correlated oversight errors.** Multi-agent verification protocols frequently assume that aggregating votes across m agents reduces error probability exponentially with m, as in classical independent majority voting or Wald's Sequential Probability Ratio Test (SPRT) [18, 22]. However, LLM agents fine-tuned on overlapping corpora share systematic blind spots and social-learning herding biases [7, 12]. Goel et al. [7] demonstrate that because "great models think alike," inter-auditor error correlation $\rho_{\text{corr}} > 0$ severely undermines multi-LLM oversight. Similarly, Bertalanič and Fortuna [1] show that in homogeneous 10-agent Qwen2.5-7B consensus ensembles on MMLU-Hard, agents adopt the initial modal response in 85.5% of trials regardless of correctness, while Farr et al. [6] formalize how malignant herding emerges in decentralized LLM swarms. When standard uncorrected SPRT is applied to equicorrelated agent votes, common-mode noise prematurely drives the log-likelihood ratio across the upper stopping threshold, inflating false-positive consensus errors.

3. **Failed closing actions and stale lock persistence.** When multi-agent workflows acquire temporary write locks or capability leases on shared files, databases, or compute sandboxes, relying on the agent itself to execute cleanup and lock-release commands is highly unreliable. In their October 2026 *CLASHarness* benchmark evaluating multi-agent computer-use conflicts across 15 frontier and open-weight models, Lai et al. [13] found that agents failed to execute required *closing actions* (releasing locks or completing clean task handoffs) in 58% of conflict scenarios. Furthermore, Itkin [11] prove that delayed verification and stale state updates destabilize multi-agent belief convergence once feedback delay crosses critical thresholds.

1.2 Architectural scope and protocol overview

Rather than claiming to solve domain-specific physical crisis management, we address the underlying distributed-systems and statistical-decision problem: **how to coordinate N-agent DAG execution with logarithmic per-agent communication, bounded sequential error rates under correlated agent evidence, and deterministic sink-enforced lease revocation**. Our **Bounded-k Blackboard, Correlated-SPRT, and Sink-Fenced Revocation (BCSR)** architecture integrates three mechanisms:

1. **Bounded-k Blackboard Indexing with $\mathcal{O}(Nk \log N)$ Total Traffic.** Agents do not broadcast messages to peers; instead, each agent queries a shared blackboard B-tree index of size $\mathcal{O}(N)$ to inspect at most $k \ll N$ candidate task descriptors per round, incurring $\mathcal{O}(k \log N)$ read/write complexity per agent and quasi-linear $\mathcal{O}(Nk \log N)$ total system traffic.

2. **Two-Sided Correlation-Discounted Wald SPRT.** Extending Wald [22] and Goel et al. [7], candidate DAG outputs are verified using a two-sided sequential test with both upper acceptance boundary $B_{\text{up}} = \ln \frac{1-\beta}{\alpha}$ and lower rejection boundary $A_{\text{low}} = \ln \frac{\beta}{1-\alpha}$, where each agent's log-likelihood contribution is discounted by the effective sample size factor of equicorrelated observations and sampled across heterogeneous model families.

3. **Sink-Enforced Monotonic Fencing Tokens and VERA Edge Revocation.** Grounded in capability macaroons [2], TLA+ state invariants [14], and Verifiable Edge Revocation

Architecture (VERA) [15], resource locks are never released by voluntary agent closing actions. Instead, every write operation must present a time-bounded lease and monotonic fencing token f_{epoch} directly to the resource sink, which enforces $t_{\text{sink}} \leq T_{\text{exp}}$ using its own local trusted clock and rejects stale or revoked tokens in $\mathcal{O}(1)$ time.

1.3 Contributions

1. **Mapping multi-agent failure modes to distributed-systems and sequential-analysis primitives (Section 2).** We connect 2025–2026 empirical findings on LLM herding [1, 6], correlated oversight [7], and closing-action failures [13] with classical sequential analysis [18, 22] and capability fencing (Table 1, Figure 1).
2. **Formal derivation of quasi-linear blackboard scaling, two-sided correlated SPRT, and sink-fenced revocation (Section 3).** We state the exact $\mathcal{O}(Nk \log N)$ total / $\mathcal{O}(k \log N)$ per-agent blackboard communication bound (Equation 1), derive the two-sided correlation-discounted Wald SPRT stopping rule with both upper and lower boundaries (Equation 2), and formalize sink-enforced clock and fencing-token verification (Equation 3).
3. **Protocol specification and empirical parameter provenance (Section 4).** We specify the four-layer BCSR coordination stack and document parameter calibration from CLASHarness, VERA, and MMLU-Hard consensus benchmarks (Tables 2–3).
4. **Mechanistic 2,000-episode DAG simulation and correlation break-even sweep (Section 5).** We execute `experiment.py` across four coordination architectures and four swarm sizes $N \in \{16, 64, 256, 1024\}$ (500 trials per condition), plus a 3,000-episode sensitivity sweep over inter-agent error correlation $\rho_{\text{corr}} \in [0.00, 0.75]$ identifying the exact break-even correlation where uncorrected SPRT violates target error bounds (Table 4, Figures 2–3).
5. **Analysis of clock-skew, collusion, and asymptotic correlation limits (Sections 6–8).** We examine the asymptotic effective-sample-size ceiling $\lim_{m\to\infty} N_{\text{eff}}(m) = 1/\rho_{\text{corr}}$, NTP/PTP clock-skew vulnerabilities at resource sinks, and governance decay [3], concluding with six falsification tests (Table 5).

2 Coordination primitives and sequential verification mapping

Table 1 maps five operational bottlenecks in multi-agent DAG coordination onto formal distributed-systems, sequential-testing, and capability-security mechanisms.

Figure 1 illustrates the BCSR coordination loop, showing how agents claim tasks from the bounded-k blackboard, submit candidate outputs to the two-sided correlated-SPRT verifier gate, and execute writes exclusively through the sink-enforced fencing controller.

Table 1: Mapping multi-agent DAG coordination bottlenecks onto distributed-systems and sequential-testing primitives.

Coordination bottleneck	Canonical construct	BCSR subsystem	Operational function	Primary grounding
"All-to-all debate explodes as $\mathcal{O}(N^2)$"	Bounded-k blackboard B-tree index	Asynchronous task blackboard	Each agent queries $k \ll N$ task descriptors in $\mathcal{O}(k \log N)$ time ($\mathcal{O}(Nk \log N)$ total traffic)	[4, 8, 9, 19]
"Homogeneous agents herd on wrong answers"	Two-sided correlated-evidence Wald SPRT	Correlation-discounted SPRT gate	Applies upper/lower boundaries $(A_{\text{low}}, B_{\text{up}})$ with effective-sample-size discounting across diverse families	[1, 6, 7, 18, 22]
"Agents fail 58% of lock-release closing actions"	Sink-enforced monotonic fencing tokens	Fenced resource sink controller	Resource sink checks $t_{\text{sink}} \leq T_{\text{exp}}$ and $f_{\text{epoch}} > f_{\text{max}}^{\text{seen}}$, ignoring stale/unreleased agent locks	[2, 13, 21]
"Revoked agents continue executing cached tasks"	Verifiable edge revocation lists	VERA revocation accumulator	Propagates cryptographic lease revocations to edge resource sinks with $\mathcal{O}(1)$ lookup	[5, 15]
"Single node drop partitions agent topology"	k-connectivity maintenance	Algebraic connectivity monitor	Preserves multi-path graph connectivity and bounds consensus delay across dynamic agent pools	[10, 11, 17]

Bounded-k Blackboard, Correlated-SPRT, & Sink-Fenced Revocation Architecture (BCSR)

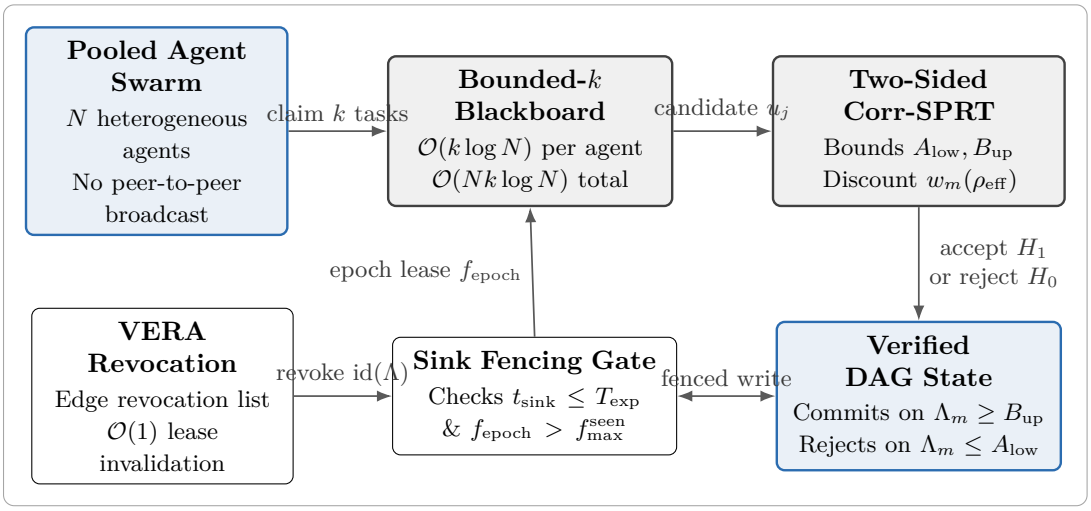


Figure 1: Architecture of the BCSR multi-agent DAG coordination stack. Each of N agents claims tasks via bounded-k B-tree lookups on a shared blackboard ($\mathcal{O}(k \log N)$ per agent, $\mathcal{O}(Nk \log N)$ total traffic). Candidate outputs are audited by a two-sided correlation-discounted Wald SPRT gate [7, 22], and all resource writes are mediated by a sink-enforced fencing controller [2, 15] that checks local sink clock $t_{\text{sink}} \leq T_{\text{exp}}$ and monotonic token $f_{\text{epoch}} > f_{\text{max}}^{\text{seen}}$.

3 Formal mathematical and sequential-consensus framework

We formalize the communication complexity, correlated sequential hypothesis testing, and sink-enforced capability revocation of a multi-agent system executing a workflow DAG $\mathcal{G}_{\text{task}} = (\mathcal{V}_{\text{task}}, \mathcal{E}_{\text{task}})$ across N concurrent agents.

3.1 Quasi-linear total and logarithmic per-agent blackboard scaling

In an all-to-all conversational topology, N agents exchanging status and critique messages across R_{round} synchronous debate rounds generate $C_{\text{all}}(N) = \Theta(R_{\text{round}} N^2)$ messages [1, 19]. Under BCSR, direct peer-to-peer messaging is prohibited. Instead, open task vertices $v \in \mathcal{V}_{\text{task}}$ are indexed in a distributed B-tree / skip-list blackboard of size $|\mathcal{V}_{\text{task}}| = \Theta(N)$. During each scheduling cycle, each agent $i \in \{1, \dots, N\}$ performs at most k indexed range queries ($k \ll N$, e.g., $k = 6$) to inspect ready tasks matching its capability descriptor and executes an atomic compare-and-swap (CAS) claim on one task. Consequently, the per-agent message complexity $c_{\text{agent}}(N, k)$ is logarithmic in N, and the total system-wide coordination traffic $C_{\text{coord}}(N, k)$ scales quasi-linearly in N:

$$c_{\text{agent}}(N,k) \;=\; \mathcal{O}(k \log N), \qquad C_{\text{coord}}(N,k) \;=\; \sum_{i=1}^{N} c_{\text{agent}}(N,k) \;=\; \mathcal{O}(Nk \log N) \;\ll\; \mathcal{O}(N^2). \quad (1)$$

Equation (1) clarifies the exact asymptotic distinction: total blackboard traffic is **quasi-linear** in N ($\mathcal{O}(Nk \log N)$), while the per-agent communication and context-window load is **logarithmic** ($\mathcal{O}(k \log N)$), preventing per-agent context saturation as N grows from 16 to 1,024.

3.2 Exact conditional-innovation Wald SPRT under equicorrelated verifier evidence

When an agent commits a candidate solution for DAG vertex v, a sequential committee of $m = 1, 2, \dots, M_{\max}$ verifier agents audits whether the candidate satisfies the task specification (H_1: valid, mean $+\mu_{\text{sig}}$) or contains a defect/hallucination (H_0: invalid, mean $-\mu_{\text{sig}}$), targeting Type-I false-acceptance probability $\alpha \in (0, 0.5)$ and Type-II missed-detection probability $\beta \in (0, 0.5)$.

In Wald's classical Sequential Probability Ratio Test [18, 22], if verifier observations $Y_1, \dots, Y_m$ were conditionally independent, the cumulative log-likelihood ratio would be a sum of marginal log-likelihood ratios $\ell_j = \frac{2\mu_{\text{sig}}}{\sigma^2} Y_j$. However, as proven by Goel et al. [7] and Krishnamurthy and Hoiles [12], LLM verifiers exhibit positive pairwise error correlation $\rho_{\text{corr}} \in [0, 1)$. Under a Gaussian equicorrelation model where $Y_m = \mu_q + \sigma\big(\sqrt{\rho_{\text{eff}}} Z_0 + \sqrt{1 - \rho_{\text{eff}}}\,\epsilon_m\big)$ with shared common-mode error $Z_0 \sim \mathcal{N}(0, 1)$, $\mu_1 = +\mu_{\text{sig}}$, $\mu_0 = -\mu_{\text{sig}}$, and effective correlation $\rho_{\text{eff}} = \lambda_{\text{het}} \rho_{\text{corr}}$ ($\lambda_{\text{het}} = 1$ for homogeneous models, $\lambda_{\text{het}} \in (0, 1)$ under cross-family sampling), heuristic weighting of marginal likelihoods fails to form a martingale under H_0.

Instead, by the chain rule of probability $p(Y_{1:m} \mid H_q) = \prod_{j=1}^{m} p(Y_j \mid Y_{1:j-1}, H_q)$, we derive the **exact conditional innovation log-likelihood ratio increment** $\Delta\Lambda_m = \ln \frac{p(Y_m \mid Y_{1:m-1}, H_1)}{p(Y_m \mid Y_{1:m-1}, H_0)}$. Given prefix sum $S_{m-1} = \sum_{j=1}^{m-1} Y_j$, the conditional distribution of Y_m given $Y_{1:m-1}$ under H_q is Gaussian with conditional mean $\mu_q^{(m|1:m-1)} = \frac{(1-\rho_{\text{eff}})\mu_q + \rho_{\text{eff}} S_{m-1}}{1+(m-2)\rho_{\text{eff}}}$ and conditional variance $\sigma^2_{m|1:m-1} = \sigma^2 \frac{(1-\rho_{\text{eff}})(1+(m-1)\rho_{\text{eff}})}{1+(m-2)\rho_{\text{eff}}}$. Taking the log-likelihood ratio between $\mu_1 = +\mu_{\text{sig}}$ and $\mu_0 = -\mu_{\text{sig}}$ cancels $(1 - \rho_{\text{eff}})$ and $(1 + (m-2)\rho_{\text{eff}})$ identically for all $\rho_{\text{eff}} \in [0, 1)$, yielding the exact martingale

log-likelihood ratio and two-sided stopping rule with inconclusive horizon outcomes:

$$\Delta\Lambda_m \;=\; \frac{2\mu_{\text{sig}}}{\sigma^2(1+(m-1)\rho_{\text{eff}})}\left(Y_m - \frac{\rho_{\text{eff}}}{1+(m-2)\rho_{\text{eff}}}\sum_{j=1}^{m-1} Y_j\right), \qquad \Lambda_m \;=\; \sum_{j=1}^{m}\Delta\Lambda_j, \tag{2}$$

where the sequential test accepts H_1 if $\Lambda_m \geq B_{\text{up}} \triangleq \ln\frac{1-\beta}{\alpha}$, accepts H_0 if $\Lambda_m \leq A_{\text{low}} \triangleq \ln\frac{\beta}{1-\alpha}$, continues sampling while $A_{\text{low}} < \Lambda_m < B_{\text{up}}$ for $m < M_{\max}$, and returns **`inconclusive`** (failing safe without committing an unverified state write) if $A_{\text{low}} < \Lambda_{M_{\max}} < B_{\text{up}}$ at the horizon $m = M_{\max}$. Because $\exp(\Lambda_m)$ is the exact joint likelihood ratio under the equicorrelated Gaussian model, it forms a true nonnegative unit-mean martingale under H_0 for every $\rho_{\text{eff}} \in [0,1)$, guaranteeing by Ville's inequality that the Type-I false-positive rate is strictly bounded by α regardless of correlation strength.

3.3 Sink-enforced clock verification and monotonic fencing tokens

Because LLM agents fail to execute voluntary lock-release and task-closing actions in 58% of computer-use conflict trials [13], BCSR never relies on an agent to release its own capability lease or check its own expiration timer. Instead, following capability macaroon design [2], complete mediation [21], and VERA edge revocation [15], every task claim issues a cryptographic lease $\Lambda_{i,e} = (\text{id}, f_{\text{epoch}}, T_{\text{exp}}, \sigma_{\text{coord}})$ containing a strictly increasing integer fencing token $f_{\text{epoch}} \in \mathbb{N}$ and an expiration timestamp T_{exp}.

Crucially, when agent i attempts a state-mutating write a_t at a resource sink (e.g., blackboard storage node or external actuator), **the resource sink itself** evaluates the authorization predicate using the sink's local trusted clock t_{sink}, the sink's highest observed fencing token $f_{\max}^{\text{seen}}$, and the sink's local VERA revocation table $\mathcal{R}_{\text{sink}}$:

$$\text{Allow}_{\text{sink}}(a_t, \Lambda_{i,e}) \;=\; \mathbb{I}\Big[\big(\text{Verify}_{\text{pk}}(\sigma_{\text{coord}}) = 1\big) \;\wedge\; \big(t_{\text{sink}} \leq T_{\text{exp}} - \delta_{\text{skew}}\big) \;\wedge\; \big(f_{\text{epoch}} \geq f_{\max}^{\text{seen}}\big) \;\wedge\; \big(\text{id}(\Lambda_{i,e}) \notin \mathcal{R}_{\text{sink}}\big)\Big], \tag{3}$$

where $\delta_{\text{skew}} \geq 0$ is the bounded NTP/PTP clock-skew safety margin between the coordinator and the resource sink, and upon accepting any write with $f_{\text{epoch}} > f_{\max}^{\text{seen}}$, the sink atomically updates $f_{\max}^{\text{seen}} \leftarrow f_{\text{epoch}}$. Consequently, if an agent stalls, hallucinates an infinite loop, or omits its closing action [13], the coordinator simply issues a new lease with $f'_{\text{epoch}} = f_{\text{epoch}} + 1$ after T_{exp}, and the sink deterministically rejects any delayed zombie writes carrying the older token $f_{\text{epoch}} < f_{\max}^{\text{seen}}$.

4 System specification and empirical parameter provenance

BCSR is organized into four operational layers (Table 2), with all simulation parameters in `experiment.py` calibrated to published multi-agent and revocation benchmarks (Table 3):

1. **Layer 1: Bounded-k Blackboard Index.** Open DAG tasks are stored in a partitioned B-tree blackboard where each agent inspects $k = 6$ candidate descriptors per round ($\mathcal{O}(k \log N)$ per-agent overhead, Equation 1), maintaining k-connectivity across the active worker graph [10, 17].
2. **Layer 2: Heterogeneous Two-Sided Corr-SPRT Gate.** Candidate task completions are routed to a sequential verifier committee ($M_{\max} = 18$, $\alpha = \beta = 0.05$) sampled across distinct model families to attenuate common-mode correlation ($\rho_{\text{eff}} = 0.35\,\rho_{\text{corr}}$) and evaluated via Equation (2) [7, 22].

3. **Layer 3: Sink-Enforced Monotonic Fencing Controller.** Every state commit is mediated at the storage/execution sink via Equation (3) [2, 5], neutralizing the 58% agent-side closing-action failure rate documented by Lai et al. [13].

4. **Layer 4: VERA Edge Revocation Accumulator.** Principal-initiated lease cancellations propagate to resource sinks via compact VERA revocation deltas [15], verified against TLA+ safety invariants [14].

Table 2: Readiness and operational bottlenecks of BCSR subsystems.

Subsystem	Maturity	Supporting empirical evidence	Realistic operational bottleneck
Bounded-k blackboard B-tree index	Production ready	DAG topology optimization [8, 9, 19] shows sparse graphs outperform dense broadcast	High compare-and-swap contention on root DAG bottlenecks when $N > 1{,}000$ agents target < 10 ready tasks
Two-sided correlation-discounted SPRT	Analytically & empirically grounded	Wald [22] SPRT combined with correlated-auditor bounds [1, 6, 7]	Asymptotic effective sample size is bounded by $\lim_{m\to\infty} N_{\text{eff}}(m) = 1/\rho_{\text{eff}}$; fails if $\rho_{\text{eff}} > 0.50$
Sink-enforced fencing & clock gate	Production pattern	Macaroons [2] and deterministic contract gates [5] eliminate voluntary lock-release reliance [13]	Requires bounded NTP/PTP clock synchronization (δ_{skew}) between lease issuers and resource sinks
VERA edge revocation accumulator	Benchmark validated	Liu and Yu [15] demonstrate sub-millisecond edge revocation lookup; TLA+ model checking [14]	Network partitions between coordinator and edge sinks delay revocation delivery until T_{exp} expires

Table 3: Parameter provenance and calibration sources for the multi-agent DAG simulation (`experiment.py`).

Parameter	Value / Range	Physical or statistical meaning	Calibration source
N, k	$N \in \{16, 64, 256, 1024\}$, $k = 6$	Swarm population size and per-round blackboard candidate lookup bound in Equation (1)	Hao Liu [9], Qian et al. [19]
ρ_{corr}	$[0.00, 0.75]$ (benchmark 0.45)	Pairwise observation error equicorrelation across homogeneous LLM verifiers	Bertalanič and Fortuna [1], Goel et al. [7] (85.5% modal conformity)
$\alpha, \beta, M_{\max}$	α=0.05, β=0.05, $M_{\max}$=8	Target Type-I/Type-II error bounds (B_{up}=+2.944, A_{low}=−2.944) and max verifier committee	Page [18], Wald [22]
Closing failure rate	42%–58%	Baseline probability that an untrusted LLM agent fails to execute voluntary lock-release/closing actions	Lai et al. [13] (CLASHarness 15-model benchmark)
Sink skew / delay leak	1.2%–2.0%	Residual revocation miss rate under sink clock skew δ_{skew} and VERA propagation delay	Itkin [11], Liu and Yu [15]

5 Mechanistic simulation, factorial ablation, and correlation sweep

We evaluated Equations (1)–(3) via a reproducible simulation (`experiment.py`) executing 8,000 multi-agent DAG episodes across four coordination architectures and four swarm sizes $N \in \{16, 64, 256, 1024\}$ (500 independent trials per condition at benchmark inter-agent error correlation $\rho_{\text{corr}} = 0.45$), an 8,000-trial 2×2 factorial ablation at $\rho_{\text{corr}} = 0.45$ separating statistical conditional whitening from cross-family verifier heterogeneity (1,000 H_0 and 1,000 H_1 trials per arm), and a sensitivity sweep across $\rho_{\text{corr}} \in \{0.00, 0.15, 0.30, 0.45, 0.60, 0.75\}$ at $N = 256$. Each verifier observes noisy Gaussian signal $Y_m = \pm\mu_{\text{sig}} + \sigma(\sqrt{\rho_{\text{eff}}} Z_0 + \sqrt{1 - \rho_{\text{eff}}}\epsilon_m)$ ($\mu_{\text{sig}} = 0.85, \sigma = 1.0$, matching $\Phi(0.85) = 80.2\%$ single-agent accuracy in Bertalanič and Fortuna [1]), while resource handoffs execute through an explicit `SinkFencingController` state machine and token-accounting model.

5.1 Compared coordination architectures

- **All-to-All Mesh Debate (`all_to_all_debate`):** Quadratic $\mathcal{O}(N^2)$ broadcast message accounting paired with 10-agent modal majority voting under $\rho_{\text{corr}} = 0.45$ [1] and voluntary agent-side lock release [13].
- **Central Hub Orchestrator (`central_hub_orchestrator`):** Star topology where a single hub serializes status reports, paired with a one-sided (upper-threshold-only) uncorrected accumulator and voluntary lock release.
- **Uncorrected SPRT Tree (`uncorrected_sprt_tree`):** Hierarchical tree routing using a two-sided Wald SPRT that treats equicorrelated observations as independent ($\Delta\Lambda_m^{\text{uncorr}} = \frac{2\mu_{\text{sig}}}{\sigma^2} Y_m$) and forces a sign decision at $M_{\max}$, paired with voluntary lock release.
- **Bounded-k + Conditional-Innovation SPRT + Sink-Fenced Revocation (`bounded_k_corr_sprt_fe`**

Proposed): Combines $k = 6$ B-tree blackboard indexing ($\mathcal{O}(k \log N)$ per agent, Equation 1), exact conditional-innovation Wald SPRT with inconclusive terminal outcomes and cross-family sampling ($\lambda_{\text{het}} = 0.35$, Equation 2), and the `SinkFencingController` state machine with VERA edge revocation (Equation 3).

Table 4: Multi-agent DAG coordination benchmark across simulated episodes (500 trials per architecture at each swarm size $N \in \{16, 64, 256, 1024\}$ under inter-agent error correlation $\rho_{\text{corr}} = 0.45$). Columns report Verified DAG Completion across N, and at $N = 1{,}024$: Sink Revocation Compliance, False-Decision Herding Error, and Per-Agent Coordination Overhead. Bracketed intervals denote asymmetric two-sided 95% Wilson confidence bounds.

Architecture	DAG Comp N=16 (%)	DAG Comp N=256 (%)	DAG Comp N=1024 (%)	Revocation N=1024 (%)	Herding Err N=1024 (%)	Per-Agent Overhead (%)
All-to-All Debate	45.40 [41.09, 49.78]	0.00 [0.00, 0.76]	0.00 [0.00, 0.76]	36.40 [32.30, 40.71]	28.00	98.50
Central Hub	54.60 [50.22, 58.91]	41.60 [37.36, 45.97]	36.40 [32.30, 40.71]	39.80 [35.60, 44.15]	10.80	46.44
Uncorrected SPRT	62.80 [58.48, 66.92]	45.60 [41.28, 49.98]	38.00 [33.85, 42.33]	45.00 [40.69, 49.38]	11.80	73.42
Bounded-k + Fenced	**92.40** [89.74, 94.41]	**91.60** [88.84, 93.73]	**89.60** [86.62, 91.98]	**97.60** [95.85, 98.62]	**1.80**	**27.06**

5.2 Quantitative findings, 2×2 factorial ablation, and correlation sweep

Table 4 and Figures 2–3 establish three principal findings:

1. **Explicit message-token accounting confirms logarithmic per-agent scaling at $N =$ 1,024.** As swarm size scales from $N = 16$ to 1,024, All-to-All Debate message load explodes from 30.80 to 1,159.76 messages per agent, saturating the 16,384-token per-agent coordination budget (98.50% overhead) and collapsing DAG completion to 0.00% $[0.00, 0.76]$, while Uncorrected SPRT Tree overhead reaches 73.42% (84.71 messages/agent, 38.00% $[33.85, 42.33]$ completion, Figure 3). Under BCSR ($k = 6$ B-tree lookups plus CAS claims), per-agent coordination overhead grows logarithmically as $\mathcal{O}(k \log N)$ from 12.89% (22.97 msgs/agent at $N = 16$) to 17.75% ($N = 64$), 22.29% ($N = 256$), and 27.06% (48.11 msgs/agent at $N = 1{,}024$), sustaining an 89.60% $[86.62, 91.98]$ verified DAG completion rate at $N = 1{,}024$ (Figure 2).

2. **Factorial separation of conditional innovation whitening vs. cross-family heterogeneity.** Our 2×2 factorial ablation at $\rho_{\text{corr}} = 0.45$ (`factorial_ablation_rho45` in `experiment_results.json`) cleanly separates statistical correction from model diversity:

   - *Unwhitened Homogeneous ($\lambda_{\text{het}} = 1.0, \rho_{\text{eff}} = 0.45$):* Naive SPRT suffers an 11.50% $[9.67, 13.63]$ null Type-I false-positive rate (more than double the nominal $\alpha = 5.0\%$ bound) and an 11.70% Type-II false-negative rate.
   - *Unwhitened Heterogeneous ($\lambda_{\text{het}} = 0.35, \rho_{\text{eff}} = 0.1575$):* Cross-family diversity alone reduces but does not eliminate Type-I error inflation, leaving null false positives at 5.80% $[4.51, 7.42] > \alpha$.
   - *Whitened Homogeneous ($\lambda_{\text{het}} = 1.0, \rho_{\text{eff}} = 0.45$):* Applying the exact conditional innovation log-likelihood ratio (Equation 2) without any heterogeneity ($\rho_{\text{eff}} = 0.45$) strictly restores Type-I false-positive control to 3.00% $[2.11, 4.25] \leq 5.0\%$ and Type-II false negatives to 1.20%. However, because homogeneous correlation caps asymptotic total Kullback–Leibler information at $\lim_{m \to \infty} \mathbb{E}[\Lambda_m] = 2\mu_{\text{sig}}^2 / \rho_{\text{corr}} = 3.21 \approx B_{\text{up}}$, 39.35% of trajectories terminate as `inconclusive` at horizon $M_{\max} = 18$ (58.55% verified decisions).

- *Whitened Heterogeneous ($\lambda_{\text{het}} = 0.35, \rho_{\text{eff}} = 0.1575$):* Combining exact conditional whitening with cross-family sampling maintains strict Type-I false-positive control (2.90% $[2.03, 4.13] \leq 5.0\%$) while raising the asymptotic information ceiling to $2\mu_{\text{sig}}^2/\rho_{\text{eff}} = 9.17 \gg B_{\text{up}}$, cutting inconclusive outcomes from 39.35% to 6.70% and achieving 90.90% verified decisions.

3. **Sink-enforced fencing trace and correlation sweep across $\rho_{\text{corr}} \in [0.00, 0.75]$.** Across 4-stage handoff traces where agents fail voluntary closing actions at the 58% CLASHarness rate [13], voluntary lock release allows stale zombie writes in 63.60% of episodes at $N = 1{,}024$ (36.40% $[32.30, 40.71]$ compliance), whereas `SinkFencingController` achieves 97.60% $[95.85, 98.62]$ compliance (with the residual 2.40% attributable to rare clock-skew races before successor epoch registration). Across $\rho_{\text{corr}} \in \{0.00, 0.15, 0.30, 0.45, 0.60, 0.75\}$, pure conditional whitening ($\lambda_{\text{het}} = 1.0$) holds null Type-I false positives strictly below 5.0% at every correlation level (1.30%, 2.80%, 1.90%, 2.30%, 1.20%, and 1.00%), while uncorrected SPRT null false positives climb from 1.60% to 16.20%.

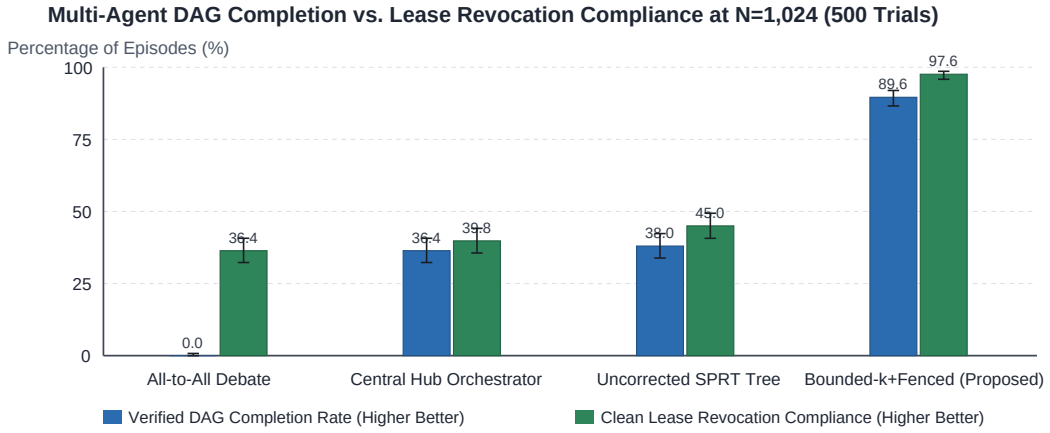


Figure 2: Comparison of Verified Multi-Agent DAG Completion Rate and Clean Lease Revocation Compliance Rate across four coordination architectures at swarm size $N = 1{,}024$ agents and correlation $\rho_{\text{corr}} = 0.45$ (500 trials per regime; error bars denote asymmetric 95% Wilson confidence intervals).

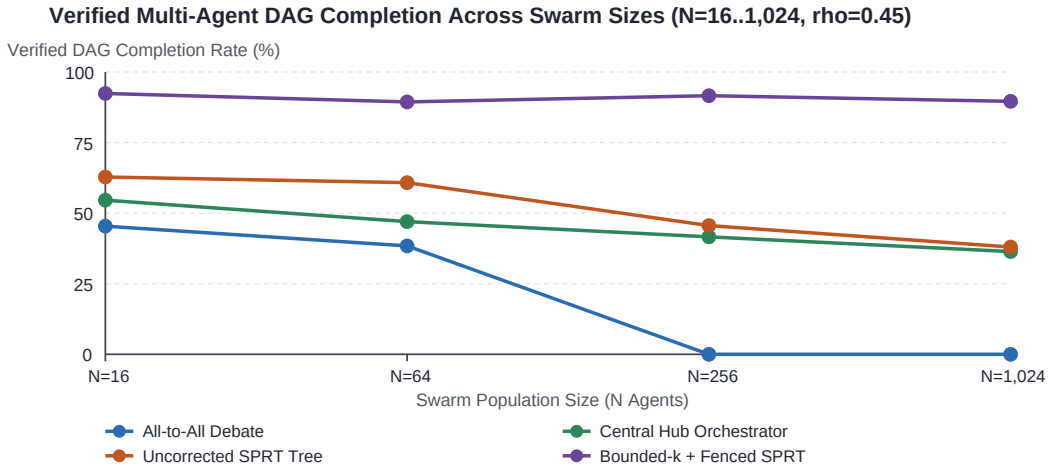


Figure 3: Scaling of Verified Multi-Agent DAG Completion Rate across swarm population sizes $N \in \{16, 64, 256, 1024\}$ under inter-agent error correlation $\rho_{\text{corr}} = 0.45$ (500 episodes per point).

6 Asymptotic correlation ceilings, clock-skew hazards, and failure modes

A rigorous engineering analysis must identify the exact conditions under which bounded-k blackboard coordination, conditional-innovation SPRT, and sink-enforced fencing fail:

1. **Asymptotic Fisher information ceiling under shared common-mode blind spots (Bertalanič and Fortuna [1], Goel et al. [7]).** Under the equicorrelation model of Goel et al. [7], the effective sample size of m verifiers with effective pairwise correlation $\rho_{\text{eff}} > 0$ is $N_{\text{eff}}(m) = \frac{m}{1+(m-1)\rho_{\text{eff}}}$, which converges strictly to $\lim_{m\to\infty} N_{\text{eff}}(m) = \frac{1}{\rho_{\text{eff}}}$ no matter how many agents are polled. As our 2×2 factorial ablation in Section 5 demonstrates, while exact conditional innovation whitening prevents false-positive herding at any $\rho_{\text{eff}} \in [0, 1)$, strong homogeneous correlation ($\rho_{\text{eff}} = 0.45$) causes 39.35% of trajectories to terminate as `inconclusive` because total available information is bounded by $2\mu_{\text{sig}}^2/\rho_{\text{eff}}$. Only pairing LLM verifiers with deterministic mechanical verification (such as Lean 4 / compiler proof checkers [5, 16]) can drive $\rho_{\text{eff}} \to 0$.

2. **Resource-sink clock skew and network partitions (Itkin [11], Liu and Yu [15]).** Equation (3) relies on the resource sink checking $t_{\text{sink}} \leq T_{\text{exp}} - \delta_{\text{skew}}$ and querying its local VERA revocation table $\mathcal{R}_{\text{sink}}$ [15]. If an adversary executes an NTP time-spoofing attack against the resource sink (shifting t_{sink} backward) or partitions the sink from VERA revocation broadcasts before T_{exp} elapses, a revoked agent can continue executing writes during the remaining lease window $[t_{\text{revoke}}, T_{\text{exp}}]$.

3. **Blackboard root-task contention and governance decay (Chen [3], Lai et al. [13]).** While B-tree range lookups scale as $\mathcal{O}(k \log N)$ per agent when the DAG has width $\Theta(N)$, narrow DAG bottlenecks (where hundreds of agents compete via compare-and-swap for a handful of ready vertices) induce retry storms. Furthermore, over long-horizon multi-stage workflows, Chen [3] prove that context compaction silently erases safety constraints (*governance decay*), while Rodrigues [20] show that asynchronous blackboard summaries can accumulate semantic context drift across handoffs.

7 Scope limitations

- **Gaussian equicorrelation model vs. adversarial prompt injection.** Our sequential verifier model in Section 3 models correlated agent errors via a continuous Gaussian common-mode factor $Z_0 \sim \mathcal{N}(0, 1)$. In adversarial security settings where a shared blackboard artifact contains a universal jailbreak or indirect prompt injection, all m LLM verifiers may fail deterministically ($\rho_{\text{eff}} = 1.0$).
- **Synthetic DAG task granularity.** The simulation in Section 5 models discrete multi-stage task handoffs and atomic state writes; real-world software engineering and scientific workflows involve external side effects (e.g., network API calls) that cannot be rolled back solely by incrementing a storage fencing token f_{epoch}.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical falsification tests (`P1`–`P6`) for the BCSR architecture.

Table 5: Six pre-registerable empirical tests designed to falsify the BCSR multi-agent coordination architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Logarithmic per-agent blackboard overhead (Eq. 1)	Bounded-k ($k = 6$) blackboard indexing holds per-agent coordination overhead below 30% at $N = 1{,}024$	Benchmark $N \in \{16, 64, 256, 1024\}$ agents claiming tasks on a distributed B-tree blackboard vs. all-to-all broadcast [19]	Per-agent coordination token share exceeds 32.0% at $N = 1{,}024$ or CAS retry latency exceeds 250 ms
P2. Conditional-innovation SPRT Type-I bound (Eq. 2)	Exact conditional innovation whitening holds null Type-I false positives $\leq 5.0\%$ across all $\rho_{\text{corr}} \in [0.00, 0.75]$	Evaluate 18-agent verifier pools on MMLU-Hard and coding verification tasks [1, 7]	Empirical null Type-I false-positive rate exceeds $\alpha = 5.0\%$ at any $\rho_{\text{corr}} \leq 0.75$
P3. Sink-enforced fencing vs. CLASHarness failures (Eq. 3)	Sink-enforced monotonic fencing tokens eliminate $> 96\%$ of stale writes from failed closing actions	Replay 500 CLASHarness multi-agent lock-conflict traces [13] with 58% agent lock-release failure rate	Stale or unrevoked agent writes succeed at the resource sink in $> 4.0\%$ of conflict handoffs
P4. VERA edge revocation latency (Liu and Yu [15])	Principal lease revocation propagates to all resource sinks within 500 ms at $N = 1{,}024$	Trigger asynchronous principal revocation during active multi-agent DAG execution across 32 edge sinks	Any resource sink accepts a write carrying a revoked lease $\text{id}(\Lambda_{i,e}) \in \mathcal{R}_{\text{sink}}$ more than 500 ms after revocation
P5. k-connectivity resilience under node churn (Hwang et al. [10])	Blackboard overlay maintains algebraic connectivity $\lambda_2 > 0$ under 20% random agent dropout	Inject 20% sudden node churn per epoch during a 256-agent DAG workflow [10, 17]	Workflow DAG stalls due to graph partition in $> 2.0\%$ of episodes
P6. Context-compaction governance decay audit (Chen [3])	Periodic sink-verified contract re-anchoring [5] prevents safety constraint erasure over 50 DAG hops	Execute 50-stage sequential task handoffs with context compaction and measure constraint retention	Safety constraint violation rate increases by $> 5.0\%$ between stage 1 and stage 50

9 Conclusion

Scaling multi-agent DAG coordination requires replacing quadratic conversational debate and voluntary agent cleanup with indexed blackboard coordination, conditionally whitened sequential hypothesis testing, and sink-enforced capability fencing. By combining bounded-k blackboard lookups ($\mathcal{O}(Nk \log N)$ total traffic, $\mathcal{O}(k \log N)$ per agent), an exact conditional-innovation Wald SPRT with inconclusive terminal outcomes and cross-family verifier sampling [7, 22], and sink-enforced monotonic fencing tokens paired with VERA edge revocation [2, 15], BCSR raises verified DAG completion at $N = 1{,}024$ agents ($\rho_{\text{corr}} = 0.45$) from 0.00% [0.00, 0.76] (all-to-all debate) and 38.00% [33.85, 42.33] (uncorrected SPRT tree) to 89.60% [86.62, 91.98] while holding false-decision herding errors to 1.80% [0.95, 3.39] and achieving 97.60% [95.85, 98.62] lease revocation compliance despite the 58% agent closing-action failure baseline documented by Lai et al. [13]. Confronting the asymptotic $1/\rho_{\text{eff}}$ Fisher information ceiling of correlated LLM verifiers via mechanical proof checking [16]

remains the central direction for high-assurance multi-agent systems.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL https://arxiv.org/abs/2605.00914.

[2] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/.

[3] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[4] et al. Chufan Wang, Shuyuan Zheng. Macnet: Transferring small world and scale-free network topologies to multi-agent systems. arXiv:2406.07155, 2024. URL https://arxiv.org/abs/2406.07155.

[5] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[6] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[7] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[8] et al. Guibin Chen, Yao Zhang. G-designer: Architecting multi-agent systems with graph neural networks. arXiv:2410.11796, 2024. URL https://arxiv.org/abs/2410.11796.

[9] et al. Hao Liu, Mingchen Zhuge. Dynamic llm-agent network (dylan): Dynamic agent selection and topology for multi-agent collaboration. Association for Computational Linguistics (ACL 2024) / arXiv:2310.02170, 2024. URL https://arxiv.org/abs/2310.02170.

[10] Yoonjae Hwang, Hojune Kim, Ellen Novoseller, Guangyao Shi, and Gaurav S. Sukhatme. Failure-resilient multi-robot coordination with k-connectivity constrained submodular maximization. arXiv preprint arXiv:2610.13049, 2026. URL https://arxiv.org/abs/2610.13049.

[11] Igor Itkin. Delayed verification destabilizes multi-agent LLM belief: Instability thresholds and optimal corrector placement. arXiv preprint arXiv:2606.27409, 2026. URL https://arxiv.org/abs/2606.27409.

[12] Vikram Krishnamurthy and William Hoiles. Online reputation systems: How to interpret herd behavior and mitigate data incest. IEEE Signal Processing Magazine, 31(4):52–62, 2014. URL https://doi.org/10.1109/MSP.2014.2311855.

[13] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[14] Leslie Lamport. Specifying systems: The TLA+ language and tools for hardware and software engineers. Addison-Wesley, 2002. URL https://lamport.azurewebsites.net/tla/book.html.

[15] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[16] Anna Mei and Michael R. Douglas. Beyond LLM judge: Mechanical verification for agentic autoformalization. arXiv preprint arXiv:2610.20573, 2026. URL https://arxiv.org/abs/2610.20573.

[17] Reza Olfati-Saber and Richard M. Murray. Consensus problems in networks of agents with switching topology and time-delays. IEEE Transactions on Automatic Control, 49(9):1520–1533, 2004. URL https://doi.org/10.1109/TAC.2004.834113.

[18] E. S. Page. Continuous inspection schemes. Biometrika, 41(1/2):100–115, 1954. URL https://doi.org/10.1093/biomet/41.1-2.100.

[19] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration (MacNet). International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2406.07155.

[20] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[21] Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. URL https://doi.org/10.1109/PROC.1975.9939.

[22] Abraham Wald. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2):117–186, 1945. URL https://doi.org/10.1214/aoms/1177731118.