BEYOND CENTRALITY: POSITION-DECOUPLED VERIFICATION FOR FINITE-HORIZON ATTACKS IN LLM MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model-based multi-agent systems (LLM-MAS) are vulnerable to misinformation that propagates through inter-agent dependencies. Topology-aware defenses often prioritize structurally central sources, implicitly assuming that lower-centrality sources pose lower security risk. We challenge this assumption with matched Hub/Nonhub attacks that differ only in source position, and find that nonhub sources can pose risks comparable to hubs. We further show formally that dominant asymptotic Perron source rankings need not order finite-horizon task-relevant adversarial influence; when such uncalibrated rankings drive selective verification, residual risk can shift toward structurally deprioritized sources. To address this risk–protection mismatch, we propose **CORA-GG**, a position-decoupled assertion-governance framework that separates verification allocation from source centrality, prioritizes risky assertions within each source, tracks semantic lineages across downstream reuse, and gates factual influence on verifier-supported evidence. This prevents unverified or repeatedly inherited assertions from gaining factual authority through propagation. Across Quant, Rigid, and MMLU on MetaGPT, LangGraph, and AutoGen, CORA-GG records no observed targeted attack successes and achieves lower worst-source residual risk than evaluated baselines.

1 INTRODUCTION

Large Language Model-based Multi-Agent Systems (LLM-MAS) increasingly rely on structured communication to coordinate planning, reasoning, retrieval, tool use, and decision making (Guo et al., 2024; Wu et al., 2024; Hong et al., 2024; Li et al., 2023). This collaboration improves task capability but also creates a system-level attack surface: false information produced by one compromised agent can propagate through benign agents and influence downstream decisions (Cemri et al., 2025; He et al., 2025). Securing LLM-MAS therefore requires not only identifying malicious information, but also deciding how limited defensive resources should be allocated.

Existing defenses often use structural information to identify security-critical agents, edges, or communication regions (Wang et al., 2025; Dong et al., 2026; Huang et al., 2026; Wang et al., 2026b; Li et al., 2026; Wang et al., 2026a). Such approaches are well motivated when structural importance is calibrated to adversarial risk. However, using static source centrality directly as a verification priority implicitly assumes that lower-centrality sources are less consequential.

We test this assumption through controlled multi-round attacks in which Hub and Nonhub conditions use the same malicious content, injection schedule, attack opportunities, and role-consistent packaging, differing only in source position. In undefended pooled evaluation, Hub and Nonhub attacks achieve ASRs of 83.3% and 81.1%, respectively, while risk is non-monotone across centrality strata. These results show that lower source centrality does not reliably imply lower attack risk. Rather, adversarial impact depends on whether malicious content is adopted by downstream agents as a semantic commitment or functional premise and subsequently reused during collaboration. Following this propagation view, we represent each message as a set of independently verifiable atomic assertions and track how these assertions are restated, relied upon, or inherited by downstream outputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This observation exposes a risk–protection mismatch in topology-aware selective verification. Although assertions originating from nonhub sources can be as consequential as those from hubs, centrality-based allocation may give them fewer or later verification opportunities. Once an unverified assertion is adopted and reused, its semantic descendants can continue to influence downstream reasoning, allowing insufficient protection at the source to accumulate into greater residual risk along the propagation lineage. We refer to this position-induced imbalance as *coverage skew*. Controlled centrality-weight interventions confirm the mechanism: changing only the centrality preference shifts verification coverage and correspondingly shifts residual risk toward the less-protected source position.

To address this problem, we propose **CORA-GG** (**C**ollaborative **O**utlier-**R**obust **A**ssertion Governance on **G**enealogy **G**raphs), an architecture-preserving message-layer defense that governs factual assertions and their semantic lineages. CORA-GG follows three principles. First, it decouples each source's base verification fraction from static centrality while retaining risk-aware prioritization within each source quota. Second, it tracks inherited, revised, and semantically equivalent assertions through a genealogy graph and promotes factual content to trusted status only after evidence-grounded verification. Third, it allows only verifier-supported lineages to carry factual weight in downstream aggregation, preventing repetition or reuse from creating factual authority. The defense does not modify the underlying agent models, roles, or communication topology.

CORA-GG records no observed targeted attack successes on spanning Quant, Rigid, and MMLU across MetaGPT, LangGraph, and AutoGen. Under separate token-envelope-, false-positive-rate-, and clean-utility-matched operating-point comparisons, it also achieves lower observed worst-source residual risk than the evaluated baselines. Complementary source-position, routing, budget, collusion, and adaptive-attack analyses characterize both the mechanism.

- We provide the first formal analysis in LLM-MAS showing that asymptotic structural centrality does not order finite-horizon task-relevant adversarial influence. We further prove that using uncalibrated rankings to allocate selective verification can reverse the defended source-risk ordering, concentrating residual risk on structurally deprioritized sources.
- We formulate this risk–protection mismatch as a worst-source verification-allocation problem and propose **CORA-GG** to eliminate structurally induced verification blind spots. CORA-GG decouples base verification coverage from uncalibrated source centrality while retaining risk-aware within-source prioritization and semantic-lineage governance; under the stated uncalibrated-hazard model, its position-decoupled base allocation is minimax-optimal for the worst-source residual-risk upper bound.
- We introduce *verification-gated downstream influence control*, which extends verification from detection to propagation control: only verifier-supported, lineage-normalized assertions may contribute factual weight to downstream decisions, while unresolved assertions and inherited repetitions cannot gain factual authority through reuse. This design prevents unverified information from being amplified after propagation and is validated across adaptive-attack stress tests.

2 Related Work

Adversarial propagation in LLM agents and multi-agent systems. LLM agents are vulnerable to prompt injection and tool- or context-mediated attacks that alter downstream reasoning, tool use, or actions (Greshake et al., 2023; Debenedetti et al., 2024). In multi-agent systems, compromised information can be repeatedly consumed, transformed, and propagated across collaborators. Recent work studies topology-dependent security bottlenecks (Yu et al., 2025), security-sensitive planner roles (Dong et al., 2026), multi-hop upstream contamination (Liang et al., 2026), communication-level propagation (He et al., 2025), and sequential error accumulation under repeated reformulation (Singh & Pawar, 2026). Together, these studies establish propagation as a system-level security problem, but leave open how source position should affect defensive allocation under a finite interaction horizon.

Topology-aware defense and finite-horizon influence. Topology-aware defenses use communication structure to identify security-critical agents, edges, or regions. Representative approaches

include topological anomaly intervention (Wang et al., 2025), spatio-temporal topology adaptation (Huang et al., 2026), influential-agent prioritization (Wang et al., 2026b), security-critical edge ranking (Li et al., 2026), and budget-constrained auditing on dynamic dependency graphs (Wang et al., 2026a). More broadly, classical network analysis distinguishes structural prominence from process-dependent influence: eigenvector-style centrality captures recursive structural importance (Bonacich, 1972), whereas influence-spread models depend on the propagation process (Kempe et al., 2003), and temporal-network studies show that time ordering can substantially alter realized connectivity (Grindrod et al., 2011; Holme & Saramäki, 2012). For bounded-round LLM-MAS, this raises a key question: whether static source centrality is sufficiently calibrated to finite-horizon adversarial influence to justify unequal verification coverage.

Assertion-level factuality, verification, and provenance. Fine-grained factuality research evaluates generated content at the claim level. FActScore decomposes responses into atomic facts (Min et al., 2023); FacTool performs tool-augmented factuality detection (Chern et al., 2023); SAFE uses search-augmented claim verification (Wei et al., 2024); and VeriScore distinguishes verifiable from unverifiable claims (Song et al., 2024). Related work also emphasizes evidence attribution and provenance through retrieval-based revision (Gao et al., 2023) and claim-evidence tracking across transformations (Xu et al., 2026; Wang et al., 2026c). In multi-agent workflows, SagaLLM combines validation with state management (Chang & Geng, 2025), while Xie et al. (2026) tracks propagated information using genealogy graphs. These studies motivate claim-level verification and lineage tracking; CORA-GG further addresses how verification capacity should be allocated across source positions and how verification status should govern the downstream influence of inherited assertions.

3 METHOD

3.1 PROBLEM FORMULATION AND MATCHED THREAT MODEL

We model an LLM-based multi-agent system as a directed information-dependency graph $G = (V, E)$ with weighted dependency matrix $A \geq 0$, where $A_{ji} > 0$ indicates that agent j functionally consumes information from agent i. Structural source centrality is the normalized positive left Perron vector

$$\mathbf{c}^\top A = \lambda_{\max}(A)\mathbf{c}^\top. \tag{3.1}$$

We use c_i to define the matched source-position intervention, rather than as a surrogate for realized adversarial impact.

To distinguish dependency support from realized semantic transmission, define

$$M_{ji} = A_{ji}\tau_{ji}, \qquad \tau_{ji} \geq 0, \quad i \neq j, \tag{3.2}$$

with $M_{ii} = d_i \geq 0$. Here τ_{ji} absorbs semantic transmission, retention, filtering, and correction, and d_i captures local persistence. Hence $\mathrm{supp}(M_{\mathrm{off}}) \subseteq \mathrm{supp}(A)$, but in general $M \neq A$.

Let $X_i(t) \in \{0, 1\}$ indicate whether agent i at round t semantically entails, materially depends on, or changes its conclusion because of a target false assertion $a_\star$, and let $p_i(t) = \Pr[X_i(t) = 1]$. For non-negative task readout $\mathbf{r}$, task-relevant contamination is $\Phi(t) = \mathbf{r}^\top \mathbf{p}(t)$.

Matched source-position attack. The attacker controls the outgoing information of one source for a bounded T-round interaction but cannot modify uncompromised model parameters, roles, topology, terminal decision rules, or hidden states. Hub and Nonhub conditions use the same semantic assertion, injection schedule, horizon, attack opportunities, and role-consistent payload construction; only source position changes. The Hub is the highest-c_i functionally relevant coordination source, whereas the Nonhub is a relevant lower-c_i source that is not the terminal coordinator. Source assignments are fixed before attack execution. Full construction details are in Appendix C.1.1.

Running example. Consider a four-role workflow with a central Planner, a lower-centrality Retriever, a Solver, and a final Aggregator. The matched attack inserts the same false factual assertion at either the Planner or the Retriever while keeping timing and interaction opportunities fixed. If the Solver later paraphrases or relies on that claim without obtaining independent evidence, the descendant remains in the same evidential lineage. This example will illustrate the three questions

below: whether structural rank orders finite-horizon impact, how a rank-based verification policy can under-cover one source, and how CORA-GG separates base entitlement from within-source risk prioritization.

Our objective is not to remove structural information from defense. We instead ask whether static source centrality should directly determine base verification entitlement when source-specific finite-horizon hazards have not been independently calibrated.

3.2 Finite-Horizon Influence and Coverage Skew

We proceed from finite-horizon non-orderability to coverage-skew-induced residual-risk reversal, and finally to minimax position-decoupled base allocation under uncalibrated hazards.

Step 1: finite-horizon centrality–risk mismatch. We approximate local semantic propagation by

$$\mathbf{p}(t+1) \approx M\mathbf{p}(t) + \mathbf{u}(t). \tag{3.3}$$

For an injection sequence originating from source i,

$$\Phi_i(T) = \mathbf{r}^\top \sum_{t=0}^{T-1} M^{T-1-t} u_i(t)\mathbf{e}_i. \tag{3.4}$$

Although the dominant left Perron coordinate of primitive M is an asymptotic source prior, Eq. equation 3.4 also depends on non-dominant modes, task readout, injection timing, semantic transmission, persistence, and correction.

Theorem 1 (finite-horizon non-orderability). There exist primitive non-negative matrices M, non-negative readouts $\mathbf{r}$, bounded horizons T, and sources i, j such that the dominant M-Perron source score satisfies $v_i > v_j$ while $\Phi_i(T) < \Phi_j(T)$. Thus the dominant asymptotic source mode is not sufficient to order finite-horizon task-relevant adversarial influence.

The theorem formally isolates the centrality–risk mismatch used in this paper: it is a non-sufficiency result, not a claim that centrality fails on every task. Appendix A.2 gives the constructive proof and spectral interpretation.

Step 2: coverage skew under selective verification. Let $f_i \in [0,1]$ be the routed fraction of eligible source-i assertions, and let $\nu_i \in [0,1]$ be the *effective pre-terminal coverage*: the probability that a relevant false assertion is selected, resolved, and its verdict becomes available before consequential downstream use. Thus f_i is directly controlled by routing, whereas ν_i additionally depends on verifier and workflow timing.

For baseline source hazard q_i^0 and conditional interception efficacy η, we use the mechanism abstraction

$$q_i^D = q_i^0(1 - \nu_i\eta). \tag{3.5}$$

Proposition 1 (residual-risk reversal under asymmetric verification). For common T, Nonhub residual risk exceeds Hub residual risk exactly when $q_\ell^D > q_h^D$; if $q_h^0 > 0$ and $1 - \nu_\ell\eta > 0$, this is equivalent to

$$\frac{q_\ell^0}{q_h^0} > \frac{1 - \nu_h\eta}{1 - \nu_\ell\eta}. \tag{3.6}$$

Hence unequal effective coverage can reverse defended source-risk ordering even when undefended Nonhub hazard is not larger. Appendix A.3 gives the derivation and time-varying extension.

Step 3: robust base allocation under uncalibrated hazards. For candidate volumes $m_i > 0$, controllable verification fractions $f_i \in [0,1]$, budget $\sum_i m_i f_i \le B$, common efficacy $0 < \eta \le 1$, and uncalibrated hazards $0 \le q_i^0 \le \bar{q}$, consider

$$\min_{\boldsymbol{f}} \sup_{q_i^0 \in [0,\bar{q}]} \max_i q_i^0(1 - \eta f_i) \quad \text{s.t.} \quad \sum_i m_i f_i \le B. \tag{3.7}$$

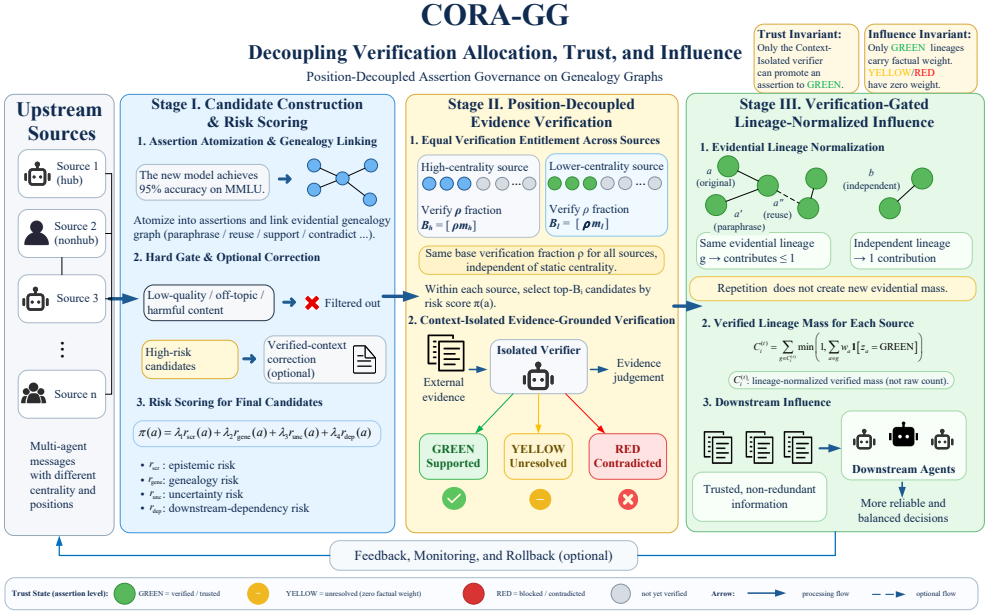


Figure 1: CORA-GG claim lifecycle: equal base verification entitlement across sources, risk-aware selection within each source, evidence-grounded trust promotion, and verified-lineage-only downstream factual influence.

Theorem 2 (minimax position-decoupled base allocation). For $0 < B < \sum_i m_i$, the minimizer of Eq. equation 3.7 is $f_i = \rho = B/(\sum_j m_j)$ for every source, with value $\bar{q}(1 - \eta\rho)$; any feasible allocation with $\min_i f_i < \rho$ has a strictly larger worst-source upper bound.

The theorem concerns controllable routing f_i, not realized coverage ν_i. It motivates source-independent *base* allocation only when no source-specific hazard ordering has been independently validated; calibrated hazard information can justify non-uniform allocation. The proof is in Appendix A.4.

3.3 CORA-GG: Position-Decoupled Assertion Governance

CORA-GG separates *entitlement*, *priority*, *trust*, and *influence*. Static source position does not determine base verification coverage; assertion-level risk orders candidates within each source, evidence-grounded verification determines trusted state, and only trusted content may influence downstream aggregation. Figure 1 summarizes the pipeline.

Assertion state and semantic genealogy. Messages are decomposed into atomic factual assertions with source, round, semantic parents, evidence references, and state $z_a \in \{\textsc{Green}, \textsc{Yellow}, \textsc{Red}\}$, denoting supported, unresolved, and rejected/contradicted assertions, respectively. CORA-GG maintains a genealogy graph $\mathcal{G}^{(t)} = (\mathcal{A}^{(t)}, \mathcal{E}^{(t)}_{\text{gene}})$ encoding inheritance, revision, support, contradiction, and material dependence. Paraphrases and downstream reuse inherit their evidential lineage rather than creating independent support. Independently acquired evidence is represented as a new evidential unit. Appendix B.1 formalizes the distinction between semantic agreement and independent evidence.

Stage I: candidate construction and epistemic risk. A narrow deterministic hard gate first removes low-ambiguity prohibited patterns. Suspicious surviving assertions may then be corrected using source-local verified history and, if needed, verified global lineage context. Screening and correction may reject or revise candidates but cannot assign GREEN; every corrected candidate remains YELLOW until factual verification.

After optional correction, each final unresolved candidate receives:

$$\pi(a) = \lambda_1 r_{\mathrm{scr}}(a) + \lambda_2 r_{\mathrm{gene}}(a) + \lambda_3 r_{\mathrm{unc}}(a) + \lambda_4 r_{\mathrm{dep}}(a), \tag{3.8}$$

where r_{scr} measures screening/local–global discrepancy risk, r_{gene} lineage-associated risk, r_{unc} epistemic uncertainty, and r_{dep} estimated downstream consequence. The dependency term may use current task structure, so position decoupling is not topology-agnostic.

Stage II: position-decoupled evidence verification. For unresolved candidates $Q_i^{(t)}$ from source i, with $m_i = |Q_i^{(t)}|$, CORA-GG assigns

$$B_i = \min(m_i, \lceil \rho m_i \rceil), \qquad Q_{\rho,i}^{(t)} = \mathrm{Top}\text{-}B_i\left(Q_i^{(t)}, \pi\right). \tag{3.9}$$

Thus static source centrality does not determine *how much* base verification a source receives, whereas assertion risk determines *which* candidates consume that quota. Position-decoupled verification is therefore neither random verification nor globally uniform assertion selection.

A context-isolated verifier checks selected assertions against trusted external evidence and verified genealogy. A SUPPORTED verdict promotes an assertion to GREEN; CONTRADICTED assigns RED; insufficient evidence and unselected assertions remain YELLOW. No preprocessing operation can assign GREEN, giving the verified-only invariant

$$z_a = \textsc{Green} \implies \mathrm{Verify}(a, E(a)) = \textsc{Supported}. \tag{3.10}$$

The primary implementation uses fractional source budgets; hard global-call caps are evaluated separately in Appendix G.

Stage III: verification-gated, lineage-normalized influence. Let $\mathcal{C}^{(t)}$ be the global partition of evidential-lineage units, and let $o(g)$ be the unique provenance owner of unit g, i.e., the source that introduced its independently grounded evidential root into the governed workflow. Define

$$\mathcal{C}_i^{(t)} = \{g \in \mathcal{C}^{(t)} : o(g) = i\}. \tag{3.11}$$

The partition and single-valued owner prevent one inherited lineage from being credited again to every downstream agent that repeats it. If a downstream agent obtains genuinely independent evidence, that evidence forms a new unit.

The verified contribution mass of source i is

$$C_i^{(t)} = \sum_{g \in \mathcal{C}_i^{(t)}} \min\left(1, \sum_{a \in g} w_a \mathbf{1}[z_a = \textsc{Green}]\right), \tag{3.12}$$

where $0 < w_a \leq 1$. The outer cap prevents repetition, paraphrasing, or assertion splitting within one evidential lineage from accumulating unbounded factual weight. Aggregation uses

$$\alpha_i^{(t)} = \frac{\varepsilon + C_i^{(t)}}{\sum_j (\varepsilon + C_j^{(t)})}. \tag{3.13}$$

The governed factual message $\widetilde{M}_i^{(t)}$ contains only GREEN factual content. RED assertions are excluded; YELLOW assertions may be surfaced only as explicit uncertainty metadata with zero factual weight. Hence the smoothing constant ε cannot restore factual influence to unresolved content.

CORA-GG therefore separates allocation, trust, and influence: source centrality does not directly determine base coverage, non-verification heuristics cannot create trusted state, and inherited repetition cannot create independent factual authority. Detailed guarantees, complexity, and runtime pseudocode are given in Appendix B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate MetaGPT (Hong et al., 2024), LangGraph (LangChain Team, 2024), and AutoGen (Wu et al., 2024) on Quant, Rigid, and MMLU-Pro STEM using 200 held-out base-task IDs per family,

Table 1: Held-out security, attacked-task utility, false positives, and efficiency. ASR, Attack Acc, and FPR are in %; latency is in seconds. Mixed ASR uses a 1:1 Hub/Nonhub interleaving.

Method	Hub ASR	Nonhub ASR	Worst ASR	Mixed ASR	Attack Acc	FPR	Tokens	Latency
None	83.3	81.1	83.3	82.2	32.8	0.0	9.2k	71.3
Speed	12.2	17.8	17.8	15.0	33.3	0.8	22.4k	192.5
Balanced	3.3	10.0	10.0	6.7	41.1	2.1	33.7k	204.1
Strict	3.3	7.8	7.8	5.6	35.6	5.8	50.1k	223.7
Reflection	70.0	71.1	71.1	70.6	30.6	1.2	16.8k	157.3
G-Safeguard	2.8	9.7	9.7	6.3	36.4	3.1	40.8k	237.2
BlindGuard	4.2	11.3	11.3	7.8	34.9	2.7	37.5k	229.6
ShieldAgent	3.6	10.8	10.8	7.2	35.2	3.4	45.2k	251.8
CORA-GG	**0.0**	**0.0**	**0.0**	**0.0**	**41.3**	2.3	**29.8k**	**171.6**

five repetitions,and matched Hub/Nonhub conditions following the $T = 3$ construction in Section 3. Baselines are no defense, Speed, Balanced, Strict, Reflection, G-Safeguard (Wang et al., 2025), BlindGuard (Miao et al., 2026), and ShieldAgent (Chen et al., 2025). Balanced is the primary source-position-prioritized reference. Each method contributes 18,000 executions, for 162,000 total; inference treats the 600 base tasks as the highest-level sampling clusters. Full accounting is in Appendix C.1.3.

All agent backbones use Qwen2.5-7B-Instruct at temperature zero. CORA-GG uses DeBERTa-v3-small for screening and a context-isolated Qwen2.5-7B-Instruct verifier; methods using an LLM verifier share the same verifier backbone and prompt template. CORA-GG fixes $\rho = 0.30$, selected on a disjoint validation partition as the smallest measured ratio with zero validation ASR (Appendix C.2). The primary metric is targeted ASR, with $ASR_{\text{worst}} = \max(ASR_{\text{Hub}}, ASR_{\text{Nonhub}})$. *Verification Share* is the realized routed fraction of eligible assertions (f_i), while *Effective Pre-Terminal Coverage* additionally requires a verdict before consequential downstream use (ν_i). Mixed ASR uses a prespecified 1:1 Hub/Nonhub interleaving from the same held-out pool and is not an independent attack sample. Full procedures are in Appendix C.1.

4.2 Overall Security and Cost-Matched Comparison

We compare CORA-GG at the frozen operating point $\rho = 0.30$ with all baselines on the same 600 held-out base tasks across three frameworks and task families. To isolate governance effects from operating-point advantages, Balanced, G-Safeguard, and Strict are separately validation-tuned to match the token envelope, false-positive rate, or clean-task utility, while CORA-GG remains fixed. Because these constraints cannot be matched simultaneously at a single operating point, they are evaluated separately.

Across the 600 unique held-out base tasks, CORA-GG records no observed targeted success in either source-position condition (Table 1). Because the 18,000 executions reuse base-task clusters across framework, source position, and repetition, they are not treated as independent Bernoulli samples; inference instead uses base-task ID as the conservative sampling unit, giving a one-sided 95% upper bound of 0.50% for the zero-event result (Appendix F.2.2).

Table 2: Strongest validation-optimized baseline under each matched operating-point constraint on the prespecified constrained-comparison slice. ASR is in %.

Constraint	Best Baseline	Baseline Mixed	Baseline Worst	CORA Mixed	CORA Worst
Iso-token	G-Safeguard	10.2	13.5	**0.0**	**0.0**
Iso-FPR	Strict	7.0	9.8	**0.0**	**0.0**
Iso-utility	G-Safeguard	9.2	13.2	**0.0**	**0.0**

The strongest matched baseline retains nonzero mixed and worst-source ASR under all three operating-point controls (Table 2), so the observed security gain is not explained solely by higher token use, a more aggressive false-positive regime, or lower clean-task utility. The result also

replicates on 600 fresh base-task IDs disjoint from validation and the primary grid: across 18,000 matched executions, CORA-GG again records 0.0% Hub and Nonhub ASR, compared with 3.5%/10.5% for Balanced and 3.1%/9.8% for G-Safeguard (Appendix Table 21). The full constrained baseline grid is reported in Appendix E.1.

4.3 Does Static Centrality Order Finite-Horizon Risk?

The central premise behind source-prioritized defense is that a structurally more central source should also be more consequential when compromised. We test this premise first with the matched undefended Hub/Nonhub comparison and then with a broader analysis over four centrality strata and 45 attacked source positions. For each position, we compare three source scores—dependency Perron centrality, a fitted asymptotic propagation-Perron score, and estimated finite-horizon influence—against empirical ASR. This design tests whether static structural rank orders the adversarial impact that is actually realized within the task horizon, rather than whether centrality carries any structural information at all.

Table 3: Descriptive association between source scores and empirical ASR over 45 attacked source positions. Pearson r is reported together with R^2 from a one-score linear prediction of ASR; no source-level significance test is used for the headline claim.

Score	Pearson r	R^2
Dependency Perron (c_A)	0.12	0.014
Fitted propagation Perron ($c_{\hat{M}}$)	0.18	0.032
Finite-horizon influence ($\hat{\Phi}(T)$)	**0.67**	**0.448**

Undefended Hub and Nonhub attacks reach pooled ASRs of 83.3% and 81.1%, and source risk is non-monotone across the four centrality quantiles (83.3%, 78.5%, 82.1%, and 81.1%). Over the 45 attacked source positions, finite-horizon influence is substantially more predictive of empirical ASR ($r = 0.67$, $R^2 = 0.448$) than dependency Perron centrality ($r = 0.12$, $R^2 = 0.014$) or fitted asymptotic Perron centrality ($r = 0.18$, $R^2 = 0.032$) (Table 3). These results align with Theorem 1: static centrality captures dependency opportunity, but it is not sufficient to order bounded-horizon task-relevant adversarial impact. The point is therefore not that topology is useless, but that uncalibrated static source rank is an unreliable basis for assigning unequal verification entitlement.

4.4 Mechanism: Allocation Bias Moves Residual Risk

The previous section establishes a centrality–risk mismatch; here we test whether that mismatch becomes a causal defense failure when structural rank directly controls verification allocation. Balanced provides the motivating example: Hub/Nonhub effective pre-terminal coverage is 91.4%/28.7%, while residual ASR is 3.3%/10.0%, yielding a 62.7-point coverage gap and a 3.03 Nonhub/Hub residual-risk ratio (Appendix Table 10). We then perform two controlled interventions while holding the remaining governance components fixed. First, we vary only the centrality-preference coefficient

$$B_i(\gamma) = B_{\text{tot}} \frac{m_i \exp(\gamma \tilde{c}_i)}{\sum_j m_j \exp(\gamma \tilde{c}_j)}, \tag{4.1}$$

where $\gamma = 0$ gives equal source fractions, $\gamma > 0$ favors more-central sources, and $\gamma < 0$ reverses the preference. Second, at the same governance operating point, we compare Hub-preferential, uniform, global risk-only, and position-decoupled per-source routing. Together, these interventions test whether changing only the allocation rule changes which source is under-covered and where residual attack risk remains.

The centrality dose–response produces the predicted directional reversal (Table 4). When Hub is deprioritized ($\gamma < 0$), residual risk appears primarily on the Hub side; when Nonhub is deprioritized ($\gamma > 0$), residual risk shifts toward Nonhub. Among the tested doses, $\gamma = 0$ is the only point with both zero observed targeted ASR and zero verification-share gap.

The routing-only control gives the same qualitative conclusion from a different intervention (Table 5): Hub-preferential, uniform, and global risk-only routing retain nonzero residual ASR, whereas

Table 4: Centrality dose–response under a fixed total verification budget. Verif. Share is the routed fraction of eligible assertions; Verif. Gap is Hub minus Nonhub share.

γ	Hub ASR	Nonhub ASR	Worst ASR	Mixed ASR	Hub Verif.	Nonhub Verif.	Verif. Gap
-2.0	4.5	0.1	4.5	2.3	15.2	42.8	-27.6
-1.0	1.8	0.0	1.8	0.9	22.5	38.5	-16.0
-0.5	0.6	0.0	0.6	0.3	26.8	34.2	-7.4
0.0	**0.0**	**0.0**	**0.0**	**0.0**	30.1	30.1	0.0
$+0.5$	0.0	0.2	0.2	0.1	34.5	25.8	$+8.7$
$+1.0$	0.0	0.6	0.6	0.3	41.2	18.5	$+22.7$
$+2.0$	0.0	2.2	2.2	1.1	55.8	8.2	$+47.6$

Table 5: Routing-only controlled intervention under a fixed governance pipeline. Eff. Cov. Gap is Hub minus Nonhub effective pre-terminal coverage.

Routing Strategy	Hub ASR	Nonhub ASR	Worst ASR	Mixed ASR	Eff. Cov. Gap	$\mathrm{RR}_{\mathrm{N/H}}$
Hub-preferential	3.3	10.0	10.0	6.7	+62.7	3.03
Uniform	5.8	6.2	6.2	6.0	+4.1	1.07
Risk-only	4.5	5.1	5.1	4.8	+2.3	1.13
Position-Decoupled	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–

position-decoupled per-source quotas eliminate both the observed coverage gap and targeted successes. Additional controls show that this is not merely an artifact of the full CORA-GG stack: in a minimal router-plus-verifier system, position decoupling reduces mixed ASR from 81.9% to 17.4%, and in a routing $\times$ hard-gating factorial it reduces mixed ASR from 6.7% to 0.0% with the gate enabled and from 13.4% to 0.6% with it disabled. The sign reversal is therefore stronger evidence than a cross-setting correlation alone: changing only the allocation bias changes which source is under-served and where residual risk remains. Event counts and the remaining controls are reported in Appendix Table 14 and Appendix D.1.

4.5 Robustness and Failure Modes

We further probe operating boundaries beyond the matched single-source setting by varying attack horizon, backbone/verifier pairing, simultaneous compromise, verifier-call capacity, paraphrases, obfuscation, genealogy dilution, assertion flooding, and trusted-evidence poisoning.

Across $T \in 1, 3, 5, 7$, CORA-GG records source-specific ASR of 0.0–0.4%, with a Hub–Nonhub effective-coverage gap of at most 0.1 percentage points; backbone/verifier transfer yields 0.1–0.4% residual ASR. Under simultaneous compromise, mixed ASR is 0.3–0.4% for $K = 2$ and at most 1.2% for tested $K = 3$ settings, with source-specific ASR up to 1.5%. The clearest degradation arises when resource or trust assumptions are stressed: under a 180-call cap, increasing assertion volume from $1\times$ to $8\times$ reduces verified share from 30.1% to 4.2% and raises mixed ASR from 0.0% to 3.0%, whereas a 360-call cap at $8\times$ lowers it to 0.5%. Semantically equivalent paraphrases raise ASR to 1.1%; gate-aware obfuscation, genealogy dilution, and assertion flooding yield 0.5%, 0.2%, and 0.7%, respectively, while retrieval poisoning is the strongest tested boundary at 3.5%. Overall, robustness remains stable across tested horizons and transfer settings but degrades under verifier-capacity saturation, larger collusion, and corrupted trusted evidence. Full stress-test results are in Appendix G.

5 Conclusion

Static source centrality captures dependency opportunity but is not a reliable proxy for finite-horizon verification priority. CORA-GG addresses this mismatch through position-decoupled verification, assertion-level risk ranking, semantic genealogy, evidence-grounded trust, and verified-only downstream influence.Its remaining limitations include trusted-evidence poisoning, verifier-capacity saturation, larger-scale collusion, longer horizons, dynamic topology, and verifier compromise.

REFERENCES

Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120, 1972. doi: 10.1080/0022250X.1972.9989806.

M. Cemri, M. Z. Pan, S. Yang, L. A. Agrawal, B. Chopra, R. Tiwari, K. Keutzer, A. Parameswaran, D. Klein, K. Ramchandran, M. A. Zaharia, J. E. Gonzalez, and I. Stoica. Why do multi-agent LLM systems fail? *Advances in Neural Information Processing Systems*, 38, 2025. Datasets and Benchmarks Track.

Edward Y. Chang and Longling Geng. SagaLLM: Context management, validation, and transaction guarantees for multi-agent LLM planning. *Proceedings of the VLDB Endowment*, 18(12):4874–4886, 2025. doi: 10.14778/3750601.3750611.

Z. Chen, M. Kang, and B. Li. ShieldAgent: Shielding agents via verifiable safety policy reasoning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8313–8344, 2025.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.

E. Debenedetti, J. Zhang, M. Balunovi'c, L. Beurer-Kellner, M. Fischer, and F. Tram'er. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems*, 37, 2024. Datasets and Benchmarks Track.

Shen Dong, Mingxuan Zhang, Pengfei He, Li Ma, Bhavani Thuraisingham, Hui Liu, and Yue Xing. PEAR: Planner-executor agent robustness benchmark. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4547–4567, 2026. doi: 10.18653/v1/2026.findings-eacl.237.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 16477–16508, 2023. doi: 10.18653/v1/2023.acl-long.910.

K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, pp. 79–90, 2023.

Peter Grindrod, Mark C. Parsons, Desmond J. Higham, and Ernesto Estrada. Communicability across evolving networks. *Physical Review E*, 83(4):046120, 2011. doi: 10.1103/PhysRevE.83.046120.

T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pp. 8048–8057, 2024. doi: 10.24963/ijcai.2024/890.

P. He, Y. Lin, S. Dong, H. Xu, Y. Xing, and H. Liu. Red-teaming LLM multi-agent systems via communication attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6726–6747, 2025. doi: 10.18653/v1/2025.findings-acl.349.

Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012. doi: 10.1016/j.physrep.2012.03.001.

S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

R. Huang, C. Wang, T. Zhang, F. Liu, J. Sun, H. Wang, and Y. Wu. TopoSHIELD: Reshaping the flow of malice via spatio-temporal risk-aware topological evolution in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8740–8768, 2026. doi: 10.18653/v1/2026.findings-acl.426.

David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003. doi: 10.1145/956750.956769.

LangChain Team. Langgraph. LangChain Blog, January 2024. Published January 17, 2024.

G. Li, H. Hammoud, D. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

Kunyang Li, Kyle Domico, Jonathan Gregory, and Patrick McDaniel. MESA: Prioritizing vulnerable communication channels for securing multi-agent systems. *arXiv preprint arXiv:2606.30602*, 2026. doi: 10.48550/arXiv.2606.30602.

Ruichao Liang, Le Yin, Jing Chen, Yebo Feng, Cong Wu, Xiaoyu Zhang, Huangpeng Gu, Zijian Zhang, and Yang Liu. Don't trust your upstream: Exploiting LLM multi-agent system via topology-guided adversarial propagation. *arXiv preprint arXiv:2512.04129*, 2026. doi: 10.48550/arXiv.2512.04129.

R. Miao, Y. Liu, Y. Wang, X. Shen, Y. Tan, Y. Dai, S. Pan, and X. Wang. BlindGuard: Safeguarding LLM-based multi-agent systems under unknown attacks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39215–39234, 2026. doi: 10.18653/v1/2026.acl-long.1819.

S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Prabhjot Singh and Bhushan Pawar. The hallucination snowball: Modeling error propagation as state transitions in multi-agent llm pipelines. In *FAGEN Workshop: Failure Modes in Agentic AI at ICML*, 2026.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, 2024. doi: 10.18653/v1/2024.findings-emnlp.552.

Kaixiang Wang, Yidan Lin, Jiong Lou, and Jie Li. BRA-Audit: Budgeted runtime auditing for LLM multi-agent systems via cumulative-exposure audit-point placement. *arXiv preprint arXiv:2608.14668*, 2026a. doi: 10.48550/arXiv.2608.14668.

Kaixiang Wang, Zhaojiacheng Zhou, Bunyod Suvonov, Jiong Lou, and Jie Li. MAS-Shield: A defense framework for secure and efficient LLM MAS. *arXiv preprint arXiv:2511.22924*, 2026b. doi: 10.48550/arXiv.2511.22924.

S. Wang, G. Zhang, M. Yu, G. Wan, F. Meng, C. Guo, K. Wang, and Y. Wang. G-Safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7261–7276, 2025. doi: 10.18653/v1/2025.acl-long.359.

Yiqi Wang, Jiaqi Zhang, Taotao Cai, Zirui Liu, Qingqiang Sun, Zequn Sun, Zhangkai Wu, Manqing Dong, Mingkai Zheng, Xuefei Yin, and Yanming Zhu. From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents. *arXiv preprint arXiv:2606.04990*, 2026c. doi: 10.48550/arXiv.2606.04990.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in LLM-based multi-agent collaboration. *arXiv preprint arXiv:2603.04474*, 2026. doi: 10.48550/arXiv.2603.04474.

Ruoyu Xu, Gaoxiang Li, and Victor S. Sheng. GAVEL: Evidence-contract debate with mechanized scrutiny for provenance-grounded fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35907–35920, 2026. doi: 10.18653/v1/2026.findings-acl.1789.

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*, pp. 2905–2938, 2025. doi: 10.18653/v1/2025.findings-acl.150.

A Proofs and Supporting Theory

This section gives the assumptions and proofs supporting the finite-horizon source-risk analysis in Section 3.2. Definitions already introduced in the main text are not repeated.

A.1 Assumptions

Assumption A1 (Local propagation approximation). In a low-contamination neighborhood,

$$\mathbf{p}(t+1) \approx M\mathbf{p}(t) + \mathbf{u}(t),$$

where M summarizes semantic transmission, persistence, filtering, and correction.

Assumption A2 (Fixed-horizon structural model). The spectral argument treats M as fixed over the analyzed horizon. A time-varying workflow can instead be represented by products of round-specific propagation operators.

Assumption A3 (Verification attenuation). For source i, baseline conditional hazard q_i^0, effective pre-terminal coverage ν_i, and interception efficacy η satisfy

$$q_i^D = q_i^0(1 - \nu_i\eta).$$

Assumption A4 (No-new-evidence inheritance). For downstream realizations $R_1, \ldots, R_m$ derived from evidential lineage L without independent factual evidence,

$$Y \to L \to (R_1, \ldots, R_m).$$

Assumption A5 (Verifier reliability). Whenever a probabilistic false-trust bound is invoked,

$$\Pr[\mathrm{Verify}(a) = \textsc{Supported} \mid a \text{ is false}] \le \varepsilon_{\mathrm{fs}}.$$

Assumption A5 is not required for the architectural verified-only invariant.

A.2 Proof of Theorem 1

Let M be primitive. With dominant eigenvalue λ_1 and biorthogonally normalized left/right eigenvectors $\mathbf{v}_1, \mathbf{u}_1$, the asymptotic decomposition is

$$M^k = \lambda_1^k \mathbf{u}_1\mathbf{v}_1^\top + R_k, \qquad \lambda_1^{-k}R_k \to 0.$$

Thus the Perron source coordinate $v_{1,i}$ is an asymptotic source prior once the dominant mode controls the response. At finite horizon, however, a diagonalizable M gives

$$\Phi_i(T) = \sum_k (\mathbf{r}^\top \mathbf{u}_k) v_{k,i} \sum_{t=0}^{T-1} \lambda_k^{T-1-t} u_i(t),$$

so the source ordering also depends on non-Perron modes, the task readout, and injection timing.

A constructive counterexample suffices. Consider

$$M = \begin{pmatrix} 0.8 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}, \qquad \mathbf{r} = (0,1)^\top.$$

The matrix is positive and hence primitive. Its dominant eigenvalue is $\lambda_1 \approx 0.8303$, with positive left Perron vector proportional to $(1, 0.3028)^\top$, so $v_1 > v_2$. For $T = 2$ and a single unit injection at $t = 0$, Eq. equation 3.4 gives

$$\Phi_1(2) = \mathbf{r}^\top M\mathbf{e}_1 = 0.1 < 0.5 = \mathbf{r}^\top M\mathbf{e}_2 = \Phi_2(2).$$

Hence the dominant Perron source ranking need not order finite-horizon, task-relevant influence. The conclusion does not require diagonalizability: in the Jordan form, finite powers also contain non-dominant generalized-eigenspace terms that need not be negligible at bounded T.

Because this counterexample already grants the ranking access to the realized propagation operator M, an A-based structural ranking requires an additional calibration argument linking dependency support to semantic transmission, timing, persistence, correction, and task readout.

A.3 Proof of Proposition 1

For common horizon T, define $f_T(q) = 1 - (1-q)^T$. Since $f_T'(q) = T(1-q)^{T-1} > 0$ for $q \in (0,1)$,

$$P_\ell^D(T) > P_h^D(T) \iff q_\ell^D > q_h^D.$$

Using Assumption A3,

$$q_\ell^0(1-\nu_\ell\eta) > q_h^0(1-\nu_h\eta).$$

If $q_h^0 > 0$ and $1 - \nu_\ell\eta > 0$, division yields Eq. equation 3.6. In particular, equal effective coverage $\nu_h = \nu_\ell$ preserves the baseline hazard ratio and introduces no additional coverage-induced distortion.

For time-varying hazards, coverage, and efficacy, $q_i^D(t) = q_i^0(t)(1-\nu_i(t)\eta(t))$ and

$$P_i^D(T) = 1 - \prod_{t=1}^{T}(1 - q_i^D(t)).$$

The same comparison therefore reduces to the corresponding cumulative-survival ordering; Proposition 1 is the constant-hazard special case.

A.4 Proof of Theorem 2

The optimization variable in Eq. equation 3.7 is the controllable routing fraction f_i, not realized effective coverage ν_i. For any feasible $\boldsymbol{f}$, the inner supremum can place $q_i^0 = \bar{q}$ on a least-covered source, giving

$$\sup_{q_i^0 \in [0,\bar{q}]} \max_i q_i^0(1-\eta f_i) = \bar{q}\left(1 - \eta \min_i f_i\right).$$

The problem is therefore equivalent to maximizing $\min_i f_i$ under $\sum_i m_i f_i \le B$. Let

$$\rho = \frac{B}{\sum_j m_j}.$$

Then

$$\min_i f_i \le \frac{\sum_i m_i f_i}{\sum_i m_i} \le \rho.$$

The common allocation $f_i = \rho$ is feasible and attains the upper bound, with objective $\bar{q}(1-\eta\rho)$. Any feasible allocation with $\min_i f_i < \rho$ has a strictly larger worst-source upper bound. The conclusion is conditional on the absence of a validated source-specific hazard ordering.

Corollary A2 (integer implementation). For $m_i > 0$ and $B_i = \min(m_i, \lceil \rho m_i \rceil)$,

$$\rho \le \frac{B_i}{m_i} \le 1, \qquad B_i < \rho m_i + 1,$$

and therefore

$$\sum_i B_i < \rho \sum_i m_i + |\{i : m_i > 0\}|.$$

Thus the implemented scheduler realizes the common source-independent target up to less than one rounding call per nonempty source. Experiments with an additional binding global call cap are reported in Appendix G.4.

B CORA-GG Guarantees and Runtime

This section proves the two governance properties not established by the allocation theorem—evidential non-duplication and verified-only factual influence—and gives the runtime and complexity.

B.1 Genealogy and Evidential Independence

Let Y denote the truth state of a proposition, L its evidential lineage, and $R_1, \ldots, R_m$ downstream realizations that introduce no independent evidence.

Lemma A1 (No-new-evidence inheritance). Under Assumption A4,

$$I(Y; R_1, \ldots, R_m \mid L) = 0, \qquad p(Y \mid R_1, \ldots, R_m, L) = p(Y \mid L).$$

Proof. The Markov condition gives $p(R_1, \ldots, R_m \mid Y, L) = p(R_1, \ldots, R_m \mid L)$. Bayes' rule then yields the posterior equality, and the conditional mutual information is zero.

The lemma concerns descendants that reuse the same evidential lineage. Independently retrieved or derived evidence forms a new evidential unit; the guarantee is therefore conditional on correct genealogy construction.

B.2 Verified-Only Trust and Lineage-Gated Influence

Proposition A3 (Verified-only promotion). Under Algorithm 1,

$$z_a = \textsc{Green} \Longrightarrow \mathrm{Verify}(a, E(a)) = \textsc{Supported}.$$

Proof. Atomization, genealogy linking, correction, risk scoring, and routing do not assign GREEN; hard gating can only reject. The only transition into GREEN is a SUPPORTED verifier verdict. Under Assumption A5, a false routed assertion is therefore promoted with probability at most $\varepsilon_{\mathrm{fs}}$.

Proposition A4 (Bounded lineage amplification). For any correctly linked evidential unit g, replacing an assertion by finitely many semantic descendants that introduce no independent evidence cannot make the contribution of g in Eq. equation 3.12 exceed one.

Proof. The outer cap in Eq. equation 3.12 is one, independent of the number of descendants. This is a bounded-amplification result, not exact splitting invariance; the latter would additionally require mass-conserving descendant weights.

Proposition A5 (Unverified-content non-influence). At a fixed aggregation stage, hold the GREEN assertions, lineage partition, ownership, and Green weights fixed. Arbitrary changes to YELLOW or RED assertions that do not promote them to GREEN leave the factual aggregation output unchanged.

Proof. By construction, $\widetilde{M}_i^{(t)}$ contains only GREEN factual assertions, and $C_i^{(t)}$ depends only on Green indicators. Hence $\widetilde{M}_i^{(t)}$, $C_i^{(t)}$, and Eq. equation 3.13 are unchanged. The statement is stage-local: re-executing earlier workflow stages may alter future candidate sets or verifier decisions.

B.3 Runtime Algorithm

Candidate correction is preprocessing only: it may revise an assertion using verified local history and, if needed, verified global lineage context, but the revised candidate remains YELLOW. Within each source quota, the operational score $\pi(a)$ from Eq. equation 3.8 orders candidates; it does not affect factual status.

B.4 Computational Complexity

Let N_t be the number of current atomic assertions and G_{t-1} the stored genealogy size. State bookkeeping is $O(N_t)$. Naive genealogy matching is $O(N_t G_{t-1})$; indexed approximate retrieval reduces candidate lookup to roughly $O(N_t \log G_{t-1})$ plus bounded reranking. Source-wise exact ranking costs $\sum_i O(m_i \log m_i)$. By Corollary A2, verifier calls are below $\rho \sum_i m_i + |\{i : m_i > 0\}|$ absent an additional global cap. Genealogy storage is $O(G_t)$; bounded-history retention or lineage compression can control long-run growth. Resource saturation is evaluated in Appendix G.4.

Algorithm 1 CORA-GG Runtime Governance

Require: messages $\{M_i^{(t)}\}$, genealogy $\mathcal{G}^{(t)}$, verified genealogy $\mathcal{G}_+^{(t)}$, trusted evidence $\mathcal{K}_{\text{trust}}$, verification ratio ρ, parameters $\Theta = (\tau_g, \boldsymbol{\lambda}, \theta_c, H)$
Ensure: governed messages, assertion states, updated genealogies, output

initialize $Q_i^{(t)} \leftarrow \emptyset$ and $\widehat{\mathcal{A}}_i^{(t)} \leftarrow \emptyset$ for all i
▷ Stage I: candidate construction and risk
for each source i **do**
  for each $a \in \text{Atomize}(M_i^{(t)})$ **do**
    $z_a \leftarrow$ YELLOW; record source and round
    if $\text{HardGate}(a, H) = 1$ **then**
      $z_a \leftarrow$ RED; $\text{AddNode}(a, \mathcal{G}^{(t)})$; **continue**
    end if
    $\text{LinkGenealogy}(a, \mathcal{G}^{(t)}, \tau_g)$
    if $\text{NeedsCorrection}(a, \theta_c)$ **then**
      $a \leftarrow \text{CorrectWithVerifiedContext}(a, \mathcal{G}_+^{(t)})$ ▷ remains YELLOW
    end if
    $\text{AddNode}(a, \mathcal{G}^{(t)})$; $\widehat{\mathcal{A}}_i^{(t)} \leftarrow \widehat{\mathcal{A}}_i^{(t)} \cup \{a\}$
    if $\text{RequiresFactualResolution}(a)$ **then**
      $\pi(a) \leftarrow \text{RiskScore}(a; \boldsymbol{\lambda})$; $Q_i^{(t)} \leftarrow Q_i^{(t)} \cup \{a\}$
    end if
  end for
end for
▷ Stage II: position-decoupled verification
for each source i **do**
  $B_i \leftarrow \min(|Q_i^{(t)}|, \lceil \rho |Q_i^{(t)}| \rceil)$; $S_i \leftarrow \text{Top-}B_i(Q_i^{(t)}, \pi)$
  for each $a \in S_i$ **do**
    $E(a) \leftarrow \text{RetrieveTrustedEvidence}(a, \mathcal{K}_{\text{trust}}, \mathcal{G}_+^{(t)})$
    $v_a \leftarrow \text{ContextIsolatedVerify}(a, E(a))$
    if $v_a =$ SUPPORTED **then**
      $z_a \leftarrow$ GREEN; $\text{AddVerifiedNode}(a, \mathcal{G}_+^{(t)})$
    else if $v_a =$ CONTRADICTED **then**
      $z_a \leftarrow$ RED; $\text{QueueDependentDescendants}(a)$
    end if
  end for
end for
▷ Stage III: verification-gated lineage-normalized influence
$\mathcal{C}^{(t)} \leftarrow \text{BuildOwnedEvidentialLineages}(\mathcal{G}^{(t)})$
for each source i **do**
  $\widetilde{M}_i^{(t)} \leftarrow \text{AssembleGreenFactualContent}(\widehat{\mathcal{A}}_i^{(t)})$
  $\mathcal{C}_i^{(t)} \leftarrow \{g \in \mathcal{C}^{(t)} : o(g) = i\}$
  $C_i^{(t)} \leftarrow \sum_{g \in \mathcal{C}_i^{(t)}} \min\Big(1, \sum_{a \in g} w_a \mathbf{1}[z_a = \text{GREEN}]\Big)$
end for
$\alpha_i^{(t)} \leftarrow (\varepsilon + C_i^{(t)}) / \sum_j (\varepsilon + C_j^{(t)})$ for all i
$O^{(t)} \leftarrow \text{Aggregate}(\{\widetilde{M}_i^{(t)}\}, \{\alpha_i^{(t)}\})$
return $\{\widetilde{M}_i^{(t)}\}, \{z_a\}, \mathcal{G}^{(t)}, \mathcal{G}_+^{(t)}, O^{(t)}$

C Evaluation Protocol and Operating-Point Selection

This section records details needed to reproduce the evaluation: task-specific attack criteria, run accounting, statistical analysis, and validation-only selection of ρ. Evaluations sharing a base-task ID are treated as paired repeated measurements.

C.1 Evaluation Protocol and Run Accounting

C.1.1 Task Construction and Matched-Attack Specification

The Hub/Nonhub intervention follows Section 3.1; only the task-specific attack target and success criterion vary across families. **Quant** contains 200 held-out custom grade-school mathematics tasks

with adversarial distractors; the injected target is a false algebraic identity or incorrect numerical premise, and attack success requires the final answer to deviate by more than 5% from ground truth because of that injected premise. **Rigid** contains 200 held-out custom rule-based symbolic-reasoning tasks; the attack injects a violated constraint or reversed logical order, and success requires the final plan to violate the injected rule when executed. **MMLU** contains 200 held-out tasks from the MMLU-Pro STEM subset using the benchmark-provided validation split, which is distinct from the method-selection validation pool and held out from tuning; the injected target is a factually incorrect scientific claim, and success requires the selected answer to reflect the injected false claim rather than ground truth.

C.1.2 Targeted-success attribution

Targeted ASR is separated from ordinary task error. Task correctness is first computed using the family-specific ground-truth criteria stated above. A run is counted as a targeted attack success only when the terminal failure satisfies that task criterion and the terminal decision is semantically attributable to the injected false assertion $a_\star$ or to a tracked semantic descendant that materially depends on it. Errors that are not attributable to the attack lineage are therefore not counted as targeted successes.

The automated semantic-attribution component uses the DeBERTa screening/judging pipeline described in the implementation. Its attribution labels are audited on a 5% random sample in Table 18 against adjudicated human judgments and a GPT-4o chain-of-thought judge. The audit concerns semantic contamination attribution only; task correctness itself is determined from the task-specific success criteria rather than from DeBERTa alone.

C.1.3 Run Accounting and Data Provenance

The main held-out comparison uses the same 600 unique base-task IDs for every method. A base task appearing under multiple frameworks, source positions, methods, or repetitions is a paired repeated observation rather than a new independent task. Table 6 summarizes the execution counts and evidence-layer provenance used in the evaluation. Exp-M is the primary held-out comparison,

Table 6: Run accounting for the principal evidence layers. Counts refer to the primary comparison, model-selection sweep, primary-grid diagnostics, paraphrase study, and independent replication. Additional diagnostic interventions, validation-stability checks, and robustness stress tests are not enumerated here; their local denominators are reported in the corresponding table captions.

ID	Evidence Layer	Executions
Exp-M	Common-grid main comparison	162,000
Exp-B	Verification-ratio selection	6,750
Exp-D	Primary-grid CORA-GG diagnostics (reuse Exp-M)	18,000
Exp-F	Paraphrase robustness	3,600
Exp-G	Independent matched replication	18,000

Exp-B is validation-only selection, Exp-D reuses Exp-M for diagnostics, Exp-F is the paraphrase study, and Exp-G is the disjoint replication. Execution counts in Table 6 summarizes the principal evidence layers used for the main evaluation and model selection. Exp-D denotes diagnostics derived directly from the Exp-M primary grid. Additional intervention experiments and robustness studies use experiment-specific denominators, which are reported in their respective table captions.

C.1.4 Statistical Analysis

For base task u, framework f, source-position condition a, method d, and repetition $r \in \{1, \ldots, 5\}$, let $Y_{u,f,a,d,r}$ denote the targeted attack-success indicator. We first aggregate repetitions within each matched condition:

$$\bar{Y}_{u,f,a,d} = \frac{1}{5} \sum_{r=1}^{5} Y_{u,f,a,d,r}. \tag{C.1}$$

The common base-task ID u is retained as the highest-level resampling cluster. Framework, source-position, method, and repetition observations derived from the same u are therefore paired repeated

measurements rather than independent samples. The primary inferential unit is the base task; run-level percentages are descriptive summaries of the execution grid.

Pairwise method comparisons use base-task-clustered permutation or resampling procedures with 10,000 draws, preserving all observations belonging to the same u within a resampled cluster. Risk ratios and continuous-cost differences use clustered bootstrap intervals over base-task IDs. Mixed-effects sensitivity analyses include task-family and framework effects. Benjamini–Hochberg correction is applied within predefined comparison families.

Zero-event uncertainty is computed at the base-task cluster level rather than from the repeated execution count; Appendix F.2.2 gives the estimand and exact bound.

C.2 Verification-Ratio Selection

CORA-GG selects the verification ratio exclusively on a disjoint validation partition containing 45 unique base-task IDs, with five IDs in each of the nine framework–task-family strata. Each validation base task is evaluated under both matched Hub and Nonhub source-position conditions. Across 15 candidate verification ratios and five repetitions per condition, the selection sweep contains $45 \times 2 \times 15 \times 5 = 6{,}750$ executions. Table 7 reports the 15 measured validation ratios used for model selection. Validation mixed ASR first reaches zero at $\rho = 0.30$. Larger measured ratios do not reduce observed validation ASR further but require additional verification. We therefore select $\rho = 0.30$ and freeze it before the primary held-out evaluation.A five-fold nested scan on a separate 60-ID validation pool selected $\rho = 0.30$ in all folds, providing an additional selection-stability check without reusing the primary test set. This supplementary stability scan is not included in the principal run accounting of Table 6.

Table 7: Validation sweep used to select the verification ratio. Each source-position condition contains 225 executions (45 base-task IDs $\times$ 5 repetitions); Mixed ASR pools the 450 Hub/Nonhub executions. ASR is in %.

ρ	Hub ASR	Nonhub ASR	Mixed ASR
0.05	10.7	17.8	14.2
0.10	6.7	14.2	10.4
0.15	3.6	10.2	6.9
0.20	1.3	4.9	3.1
0.25	0.4	2.7	1.6
0.30	0.0	0.0	0.0
0.35	0.0	0.0	0.0
0.40	0.0	0.0	0.0
0.45	0.0	0.0	0.0
0.50	0.0	0.0	0.0
0.60	0.0	0.0	0.0
0.70	0.0	0.0	0.0
0.80	0.0	0.0	0.0
0.90	0.0	0.0	0.0
1.00	0.0	0.0	0.0

D Source-Position and Allocation Diagnostics

This section reports the controls needed to separate source-position allocation from role identity, topology, hard gating, and assertion-level risk ranking.

D.1 Matched Source-Position Controls

D.1.1 Topology and Role-Position Controls

Tables 8 and 9 separate topology and source position from framework and role identity. Table 9 reports the corresponding role-position swap while holding the AutoGen mesh fixed. The topology control shows that undefended source ordering can change when the collaboration structure changes.

Table 8: Within-AutoGen topology control. The same agent set is rewired while prompts, roles, tools, and budgets remain fixed. ASR is in %.

Topology	Hub ASR	Nonhub ASR	$RR_{N/H}$	Pattern
Chain	78.3	65.0	0.83	Hub-dominant
Star	85.0	80.0	0.94	Similar
Mesh	76.7	88.3	1.15	Nonhub-dominant

Table 9: Undefended role-position swap in a fixed AutoGen mesh. The same agent template is placed at Hub or Nonhub positions while the topology, prompts, tools, and terminal rule remain fixed.

Condition	ASR (%)	Pattern
Hub (coordinator position)	81.7	–
Nonhub (peripheral position)	79.2	Comparable risk

The role-position control holds the agent template fixed while changing structural position. Together these results argue against attributing the Hub/Nonhub pattern solely to framework identity or role semantics.

D.1.2 Effective-Coverage Asymmetry in Structural Baselines

Table 10 reports the effective pre-terminal coverage asymmetry of the structural baselines. Balanced

Table 10: Residual attack risk and realized effective pre-terminal verification coverage. Gap is Hub minus Nonhub effective coverage. ASR and coverage are in %.

Defense	Hub ASR	Hub Eff. Cov.	Nonhub ASR	Nonhub Eff. Cov.	$RR_{N/H}$	Gap
None	83.3	–	81.1	–	0.97	–
Speed	12.2	68.2	17.8	12.5	1.46	55.7
Balanced	3.3	91.4	10.0	28.7	3.03	62.7
Strict	3.3	97.6	7.8	45.2	2.36	52.4

is the clearest source-position-prioritized baseline because it explicitly favors Hub-origin assertions. Speed and Strict are intensity controls. Their effective coverage can remain asymmetric even when their nominal routing policies are broader, because the verification result must become available before consequential downstream use.

D.1.3 Source-Centrality Quantiles

Table 11 broadens the endpoint Hub/Nonhub comparison to four centrality strata. Undefended ASR

Table 11: Source-position analysis over four centrality quantiles. Verif. Share is the fraction of eligible assertions routed to verification. ASR and verification share are in %.

Centrality Quantile	No-Defense ASR	CORA-GG ASR	CORA-GG Verif. Share
Q1 (Hub)	83.3	0.0	30.2
Q2	78.5	0.2	30.1
Q3	82.1	0.1	30.0
Q4 (Nonhub)	81.1	0.0	30.1

is non-monotone across the four source-centrality groups, whereas CORA-GG maintains an approximately constant verification share of 30.0–30.2%. This broadens the source-position analysis beyond the binary Hub/Nonhub endpoint comparison.

D.1.4 MINIMAL-SYSTEM ISOLATION

Table 12 isolates routing from the remaining CORA-GG governance stack. To isolate the allocation rule from the complete governance stack, we disable atomization, semantic genealogy, hard gating, and verification-gated aggregation and retain only the router plus a basic verifier. The effect is

Table 12: Minimal-system routing isolation. ASR, accuracy, and FPR are in %.

Configuration	Hub ASR	Nonhub ASR	Mixed ASR	Attack Acc	FPR
No defense	83.3	81.1	82.2	32.8	0.0
Minimal Hub-preferential	72.5	91.2	81.9	30.2	1.2
Minimal Uniform	68.4	69.8	69.1	31.5	1.5
Minimal Position-Decoupled	18.5	16.2	17.4	38.6	3.1

weaker than in the complete pipeline but remains large, showing that the routing result does not require every other CORA-GG component to be present.

D.1.5 ROUTING × HARD-GATING FACTORIAL

Table 13 tests whether the routing effect survives removal of hard gating. The factorial is used

Table 13: Routing × hard-gating factorial. Eff. Cov. Gap is Hub minus Nonhub effective pre-terminal coverage. ASR and coverage gap are in %.

Routing	Gate	Hub ASR	Nonhub ASR	Mixed ASR	Eff. Cov. Gap	Interpretation
Hub-preferential	On	3.3	10.0	6.7	+62.7	allocation bias
Hub-preferential	Off	8.5	18.2	13.4	+58.3	bias persists
Position-decoupled	On	0.0	0.0	0.0	0.0	lowest observed risk
Position-decoupled	Off	0.5	0.6	0.6	+0.1	routing effect persists

as a descriptive mechanism control rather than as a separate inferential experiment. With the gate enabled, position decoupling changes mixed ASR from 6.7% to 0.0%; with the gate disabled, it changes mixed ASR from 13.4% to 0.6%. Because all four cells reuse the same base-task pool, we do not attach an additional factorial F test to these descriptive contrasts. The persistence of the routing contrast with the gate disabled supports a role for allocation beyond hard gating alone.

D.2 CENTRALITY DOSE–RESPONSE RUN ACCOUNTING

Table 14 gives the full centrality dose–response together with event-count accounting.

Table 14: Full centrality dose–response with event counts. Verif. Share denotes realized verification share; $k_H/k_N/k_M$ denotes Hub/Nonhub/Mixed targeted-success counts. Percentages are reported without the percent sign inside cells.

γ	Hub ASR	Nonhub ASR	Mixed ASR	Hub Verif.	Nonhub Verif.	Verif. Gap	Base Tasks	Reps	$k_H/k_N/k_M$
-2.0	4.5	0.1	2.3	15.2	42.8	-27.6	200	5	45/1/23
-1.0	1.8	0.0	0.9	22.5	38.5	-16.0	200	5	18/0/9
-0.5	0.6	0.0	0.3	26.8	34.2	-7.4	200	5	6/0/3
0.0	0.0	0.0	0.0	30.1	30.1	0.0	200	5	0/0/0
$+0.5$	0.0	0.2	0.1	34.5	25.8	+8.7	200	5	0/2/1
$+1.0$	0.0	0.6	0.3	41.2	18.5	+22.7	200	5	0/6/3
$+2.0$	0.0	2.2	1.1	55.8	8.2	+47.6	200	5	0/22/11

E EXTENDED BASELINE AND UTILITY RESULTS

This section gives the constrained operating-point comparison, clean-task utility, and the semantic-attribution audit used by the primary evaluation.

E.1 VALIDATION-OPTIMIZED CONSTRAINED BASELINES

Each baseline is tuned on the same validation partition under three pre-specified operational constraints and frozen before evaluation on a common prespecified 200-base-task operating-point slice, stratified across framework–task-family cells. Every base task is evaluated under both matched Hub and Nonhub source-position conditions with five repetitions, yielding 1,000 executions per source-position condition and method, or 2,000 executions per method operating point. Iso-token matches CORA-GG's approximately 29.8k-token envelope, iso-FPR matches its 2.3% clean false-positive rate, and iso-utility matches its 72.8% clean-task accuracy. This constrained slice is an operating-point comparison and is not counted as an independent replication of Exp-M. Tables 15 and 16 separate the security outcomes from the matched operating-point diagnostics to improve readability.

Table 15: Validation-optimized constrained comparison on the common 200-base-task operating-point slice: ASR metrics. Each method point contains 1,000 executions per source-position condition (2,000 total across Hub and Nonhub). Worst ASR is $\max(\mathrm{ASR}_H, \mathrm{ASR}_N)$. All values are in %.

Method	Constraint	Hub ASR	Nonhub ASR	Mixed ASR	Worst ASR
CORA-GG	Fixed point	0.0	0.0	0.0	0.0
Balanced	Iso-token	7.2	15.1	11.2	15.1
G-Safeguard	Iso-token	6.8	13.5	10.2	13.5
Strict	Iso-token	10.2	19.5	14.9	19.5
Balanced	Iso-FPR	6.5	14.2	10.4	14.2
G-Safeguard	Iso-FPR	3.5	11.5	7.5	11.5
Strict	Iso-FPR	4.2	9.8	7.0	9.8
Balanced	Iso-utility	5.8	13.5	9.7	13.5
G-Safeguard	Iso-utility	5.2	13.2	9.2	13.2
Strict	Iso-utility	13.5	26.2	19.9	26.2

Table 16: Matched operating-point diagnostics for the same constrained comparison. FPR and Clean Acc are in %; Tokens are in thousands. Events H/N/M gives the observed event counts for Hub, Nonhub, and Mixed conditions.

Method	Constraint	FPR	Clean Acc	Tokens (k)	Events H/N/M
CORA-GG	Fixed point	2.3	72.8	29.8	0/0/0
Balanced	Iso-token	2.8	70.1	30.1	72/151/112
G-Safeguard	Iso-token	2.6	69.2	29.5	68/135/102
Strict	Iso-token	2.1	58.5	29.8	102/195/149
Balanced	Iso-FPR	2.3	69.5	32.5	65/142/104
G-Safeguard	Iso-FPR	2.3	67.8	41.5	35/115/75
Strict	Iso-FPR	2.3	65.4	48.1	42/98/70
Balanced	Iso-utility	3.1	72.8	35.2	58/135/97
G-Safeguard	Iso-utility	1.5	72.8	38.5	52/132/92
Strict	Iso-utility	0.5	72.8	28.5	135/262/199

At the validation-selected operating points, each tested baseline retains nonzero mixed and worst-source ASR under all three matched constraints (Table 15), whereas CORA-GG records zero observed ASR. Table 16 reports the corresponding false-positive, clean-utility, token, and event-count diagnostics. The comparison is intended to control the operating point rather than establish uniform Pareto dominance; individual baselines can be preferable on individual secondary metrics.

E.2 CLEAN-TASK UTILITY

Table 17 reports clean-task utility and the safe-and-correct rate for all methods. The primary table already reports token use and latency; no second cost table is repeated here.

Table 17: Clean-task utility and safe-and-correct rate (SCR). All reported rates are in %.

Method	Clean Acc	Block	FPR	SCR	Atk. Acc.
None	74.2	0.0	0.0	74.2	32.8
Speed	71.5	1.2	0.8	70.7	33.3
Balanced	68.3	3.5	2.1	65.8	41.1
Strict	62.1	8.7	5.8	56.4	35.6
Reflection	69.8	2.1	1.2	68.3	30.6
G-Safeguard	65.4	5.2	3.1	62.1	36.4
BlindGuard	66.7	4.5	2.7	63.7	34.9
ShieldAgent	64.2	6.1	3.4	60.3	35.2
CORA-GG	72.8	2.6	2.3	70.9	41.3

E.3 Semantic Attribution Audit

Table 18 reports the semantic-attribution audit against human adjudication and a GPT-4o judge. This audit evaluates semantic contamination attribution. Task correctness itself is determined using the task-specific success criteria rather than by DeBERTa alone.

Table 18: Judge audit on a 5% random sample ($n = 900$ runs, 180 task units).

Judge	Agreement w/ DeBERTa	Precision	Recall
Human (adjudicated)	0.982	1.000	0.972
GPT-4o (CoT)	0.976	0.947	0.986

F Disaggregated Results and Replication

This section separates subgroup reporting from independent replication and finite-sample zero-event accounting.

F.1 Disaggregated Held-Out Results

F.1.1 Framework-Level Results

Table 19 reports the selected framework-level held-out results. The subgroup results emphasize heterogeneity. The three frameworks do not exhibit a common undefended Hub/Nonhub ordering,

Table 19: Selected framework-level ASR results for entries whose subgroup aggregation is directly reported from the common-grid logs. Each source-position aggregate within a framework contains 3,000 executions (3 task families $\times$ 200 base tasks $\times$ 5 repetitions); Hub and Nonhub reuse the same underlying base-task IDs. ASR is in %.

Framework	Defense	Hub ASR	Nonhub ASR
MetaGPT	None	66.7	70.0
MetaGPT	Strict	6.7	3.3
MetaGPT	BlindGuard	5.0	8.3
MetaGPT	CORA-GG	0.0	0.0
LangGraph	None	100.0	73.3
LangGraph	Strict	3.3	16.7
LangGraph	BlindGuard	5.0	13.3
LangGraph	CORA-GG	0.0	0.0
AutoGen	None	83.3	100.0
AutoGen	Strict	0.0	3.3
AutoGen	BlindGuard	2.5	12.5
AutoGen	CORA-GG	0.0	0.0

which is why the pooled 83.3%/81.1% result is not interpreted as evidence of universal source-position equivalence.

F.1.2 TASK-FAMILY RESULTS

Table 20 reports the selected task-family slices. The magnitude and direction of source-position effects depend on task family. The supported claim is therefore insufficient calibration of static source centrality, rather than a constant Hub/Nonhub effect magnitude across all tasks.

Table 20: Selected task-family slices. ASR and attacked-task accuracy are in %.

Task	Defense	H ASR	N ASR	H Acc	N Acc
Quant	None	83.3	66.7	10.0	3.3
Quant	Balanced	3.3	16.7	26.7	13.3
Quant	Strict	0.0	13.3	20.0	16.7
Rigid	None	73.3	80.0	3.3	0.0
Rigid	Balanced	0.0	0.0	16.7	16.7
Rigid	Strict	0.0	3.3	6.7	6.7
MMLU	None	93.3	96.7	86.7	93.3
MMLU	Balanced	6.7	13.3	86.7	86.7
MMLU	Strict	10.0	6.7	86.7	76.7

F.2 INDEPENDENT REPLICATION AND ZERO-EVENT ACCOUNTING

F.2.1 INDEPENDENT REPLICATION

Table 21 reports the matched disjoint 600-ID replication. We independently evaluate CORA-GG, Balanced, and G-Safeguard on a disjoint replication pool of 600 fresh base-task IDs that overlap with neither the method-selection validation pool nor the primary Exp-M task set. Each fresh base task is evaluated under both matched Hub and Nonhub source-position conditions, with five repetitions for each of the three replicated methods. Thus Exp-G contains $600 \times 2 \times 5 \times 3 = 18{,}000$ executions in total, corresponding to 3,000 executions per source-position condition and method. The repeated Hub/Nonhub executions remain paired through the shared base-task ID, which is retained as the inferential cluster. Replication outcomes are reported separately and are not pooled with Exp-M for inference. Table 21 reports the independent matched replication on the disjoint 600-ID test set. The independent replication reproduces the qualitative ordering of the primary comparison and is reported separately rather than pooled with Exp-M.

Table 21: Independent matched replication on a disjoint 600-ID test set. Each method is evaluated on both source positions with five repetitions, yielding 3,000 executions per source-position condition and 6,000 per method. ASR, attacked-task accuracy, and FPR are in %.

Method	Hub ASR	Nonhub ASR	Mixed ASR	Attack Acc	FPR
Balanced	3.5	10.5	7.0	40.8	2.2
G-Safeguard	3.1	9.8	6.5	36.1	2.9
CORA-GG	0.0	0.0	0.0	41.0	2.1

F.2.2 ZERO-EVENT CONFIDENCE BOUNDS

CORA-GG records no observed targeted successes in its 18,000-execution Exp-M grid. Because these executions reuse 600 underlying base-task IDs across frameworks, source positions, and repetitions, two different estimands must be distinguished. The run-level calculation treats executions as independent Bernoulli trials and is shown only descriptively. The inferential calculation treats the base-task cluster event $Z_{u,a}$ (or Z_u for the complete grid) defined in Appendix C.1.4 as the Bernoulli observation, conditional on the base-task sampling model stated there. This cluster event

asks whether a base task exhibits at least one targeted failure anywhere in the prespecified repeated-measures grid covered by the corresponding row. For zero events among N clusters, the one-sided Clopper–Pearson 95% upper bound is

$$U_{0.95}(N) = 1 - 0.05^{1/N}. \tag{F.1}$$

For $N = 600$, $U_{0.95} = 0.498\%$, reported as 0.50%. Table 22 reports the exact one-sided upper bounds for the zero-event result at the stated analysis levels. The run-level values are descriptive

Table 22: One-sided exact 95% upper bounds after zero observed targeted successes. The base-task column concerns the cluster-level any-failure event defined in Appendix C.1.4; the run-level column is descriptive only.

Scope	Base-task N	Run n	Cluster any-failure upper	Descriptive run upper
Fine-grained cell	200	1,000	1.49%	0.30%
Per source-position condition (Exp-M)	600	9,000	0.50%	0.03%
Complete Hub+Nonhub grid	600	18,000	0.50%	0.02%

because run-level independence is incompatible with the repeated-measures design. The 0.50% cluster value instead bounds, under the stated base-task sampling model, the probability that a newly sampled base task exhibits at least one targeted failure anywhere in the prespecified grid represented by the row. It is not a 0.50% upper bound on arbitrary per-run ASR in a new deployment. The cluster-level statement is therefore the conservative headline uncertainty summary for the zero-event result.

G Robustness Stress Tests

This section reports the stress tests summarized in Section 4.5.

G.1 Attack Horizon

Table 23 evaluates the finite-horizon result over $T \in \{1, 3, 5, 7\}$. Containment weakens slightly as

Table 23: Finite-horizon stress test. ASR and effective-coverage gap are in %.

T	No-Def Hub	No-Def Nonhub	Balanced Nonhub	CORA Hub	CORA Nonhub	CORA Eff. Cov. Gap
1	45.2	38.5	8.2	0.1	0.2	+0.1
3	83.3	81.1	10.0	0.0	0.0	0.0
5	91.5	89.2	12.5	0.3	0.2	+0.1
7	95.1	93.8	14.2	0.4	0.3	+0.1

the interaction horizon increases, but the Hub–Nonhub effective-coverage gap does not reopen over the tested $T \leq 7$ range.

G.2 Backbone and Verifier Transfer

Table 24 reports targeted agent- and verifier-backbone transfer checks. These are targeted sensitivity

Table 24: Targeted backbone and verifier transfer for CORA-GG. ASR, FPR, and attacked-task accuracy are in %.

Agent Backbone	Verifier Backbone	Hub ASR	Nonhub ASR	Mixed ASR	FPR	Attack Acc
Qwen2.5-7B	Qwen2.5-7B	0.0	0.0	0.0	2.3	41.3
Llama-3.1-8B	Llama-3.1-8B	0.2	0.3	0.3	3.1	39.8
Qwen2.5-7B	Mistral-7B-v0.3	0.1	0.4	0.3	2.8	40.5

checks rather than a complete agent-backbone $\times$ verifier-backbone factorial. They therefore support robustness over the measured configurations but not model-family invariance.

G.3 Simultaneous Multi-Source Compromise

Table 25 extends the primary single-source threat model to $K \leq 3$ simultaneous compromises. Ob-

Table 25: Simultaneous multi-source compromise. Cov. Floor is the minimum effective pre-terminal coverage among compromised sources. n/source cond. is the execution denominator for each reported Hub and Nonhub rate; Mixed is the prespecified 1:1 derived interleaving, not a third independent sample. ASR and coverage are in %.

K	Sources	Routing	Hub ASR	Nonhub ASR	Mixed ASR	Cov. Floor	n/source cond.	Events H/N/M
1	–	Hub-pref.	3.3	10.0	6.7	28.7	1000	33/100/67
1	–	Pos-dec.	0.0	0.0	0.0	30.1	1000	0/0/0
2	H+L	Hub-pref.	8.5	18.2	13.4	22.5	1000	85/182/134
2	H+L	Pos-dec.	0.3	0.5	0.4	29.8	1000	3/5/4
2	L+L	Hub-pref.	2.1	22.5	12.3	15.2	1000	21/225/123
2	L+L	Pos-dec.	0.2	0.4	0.3	30.0	1000	2/4/3
3	H+L+L	Hub-pref.	12.5	25.8	19.2	12.8	1000	125/258/192
3	H+L+L	Pos-dec.	0.8	1.5	1.2	28.5	1000	8/15/12
3	random	Hub-pref.	9.2	20.1	14.7	18.5	1000	92/201/147
3	random	Pos-dec.	0.5	1.0	0.8	29.2	1000	5/10/8

served residual risk increases as the number of compromised sources grows in these tested settings. We therefore describe the result as robustness over the measured $K \leq 3$ regimes rather than as collusion invariance or a monotone law beyond the tested configurations.

G.4 Hard Global Budget under Assertion Flooding

Table 26 measures degradation when a fixed verifier-call cap becomes binding. A binding global

Table 26: Hard verifier-call caps crossed with assertion-volume multipliers. n/source cond. is the denominator for each Hub and Nonhub rate; Mixed is the derived 1:1 interleaving. ASR and verified share are in %.

Budget	Volume	Routing	Hub ASR	Nonhub ASR	Mixed ASR	Verified Share	n/source cond.	Events H/N/M
100	1×	Pos-dec.	0.5	0.6	0.6	28.5	1000	5/6/6
180	1×	Pos-dec.	0.0	0.0	0.0	30.1	1000	0/0/0
180	2×	Pos-dec.	0.2	0.3	0.3	15.2	1000	2/3/3
180	4×	Pos-dec.	0.8	1.2	1.0	7.8	1000	8/12/10
180	8×	Pos-dec.	2.5	3.5	3.0	4.2	1000	25/35/30
360	8×	Pos-dec.	0.4	0.6	0.5	8.5	1000	4/6/5
180	4×	Hub-pref.	2.1	18.5	10.3	5.2	1000	21/185/103

budget creates residual risk when adversarial assertion volume reduces the fraction that can be verified. Increasing the available call budget partially recovers containment. This stress test characterizes degradation over the measured resource-contention range and does not establish denial-of-service robustness under unbounded load.

G.5 Paraphrase Robustness

We separately evaluate lexical robustness by replacing the standard Nonhub realization with semantically equivalent payloads having less than 30% lexical overlap and no explicit wrapper markers. The semantic target and the $T = 3$ attack schedule remain unchanged. This evaluation uses 900 runs and 180 matched condition-task observations per method. Table 27 reports robustness to semantically equivalent low-overlap paraphrases. CORA-GG exhibits a 1.1-point increase under the paraphrased payloads. This experiment tests robustness to substantial lexical variation and is reported separately from the unified defense-aware adaptive evaluation below.

Table 27: Robustness to semantic paraphrasing. ASR is in %.

Method	Standard	Paraphrased	ΔASR
Balanced	10.0	22.2	+12.2
G-Safeguard	9.7	18.5	+8.8
BlindGuard	11.3	15.6	+4.3
CORA-GG	0.0	**1.1**	+1.1

G.6 Defense-Aware Adaptive Attacks

Table 28 evaluates defense-aware payloads on a common AutoGen/MMLU/Nonhub slice. Gate-aware obfuscation, genealogy dilution, and assertion flooding yield 0.5%, 0.2%, and 0.7% ASR for CORA-GG, respectively. Retrieval poisoning produces the largest degradation, reaching 3.5%, because it directly attacks the trusted-evidence channel on which verification depends. Evidence integrity is therefore an explicit security assumption.

Table 28: Defense-aware adaptive attacks on a common AutoGen/MMLU/Nonhub evaluation slice. ASR is in %. Each attack uses 200 base tasks, five repetitions, and 1,000 runs per defense cell. Retrieval poisoning directly targets the trusted-evidence assumption.

Defense	Standard	Gate-Aware	Genealogy Dilution	Retrieval Poisoning	Flooding
Balanced	10.0	18.5	14.2	16.7	31.5
G-Safeguard	9.7	15.2	12.8	14.4	28.3
CORA-GG	**0.0**	**0.5**	**0.2**	**3.5**	**0.7**

H Reproducibility and Implementation

H.1 Reproducibility

The artifact contains held-out and validation task identifiers, matched payloads and source assignments, run-level outcomes and cost logs, analysis scripts, generation and verification templates, the CORA-GG runtime configuration, and a frozen execution environment. Exact baseline commits, checkpoints, prompts, evidence manifests, and retrieval settings are recorded there.

H.2 Practical Limitations

The stress tests identify three operational boundaries: corrupted trusted evidence, binding verifier capacity, and larger coordinated compromise. The experiments also do not establish security for arbitrary-duration workflows, dynamically changing topologies or roles, broader model families, verifier compromise, collusion beyond the tested range, or unbounded resource-exhaustion attacks. Zero observed targeted successes are finite-sample results; Appendix F.2.2 gives the corresponding cluster-level uncertainty.

H.3 Baseline Implementation Checkpoints

Speed uses centrality-oriented routing with a non-generative check; **Balanced** is the explicitly Hub-preferential governance reference; **Strict** nominally routes all unresolved assertions to comprehensive verification, subject to pre-terminal timing; and **Reflection** performs agent self-review without an external verifier. **G-Safeguard**, **BlindGuard**, and **ShieldAgent** use released implementations where available, with task-interface adaptations documented in the artifact. All methods that invoke an LLM verifier use the same Qwen2.5-7B-Instruct verifier backbone and matched prompt template in the primary comparison. CORA-GG uses assertion routing, genealogy governance, and a context-isolated verifier.