CONTRACT-GOVERNED MULTI-AGENT GRAPH ORCHESTRATION FOR LONG-HORIZON AUTONOMOUS RESEARCH PIPELINES

Anonymous authors
Paper under review

ABSTRACT

We study formal guarantees for contract-governed multi-agent orchestration in long-horizon autonomous research pipelines. The workflow is modeled as a typed directed graph with validator-mediated transitions, bounded rerouting, and explicit perturbation families. We derive three theorem-level results: a dependence-aware upper bound for path-level false acceptance under admissibility and finite-moment conditions; a finite-horizon dominance condition showing when calibrated rerouting improves a constrained objective over memory-only baselines; and a minimax robustness boundary for static-threshold policies under unrestricted perturbations, with a restricted authenticated-regime corollary. Symbolic obligations close across all theorem families, while replay evidence is mixed: dominance and unrestricted minimax boundary behavior are supported, whereas full closure for path-bound confirmatory slices and restricted-regime attainability remains incomplete due to assumption-regime mismatch in evaluated slices. These results provide an auditable scientific boundary for what current contract-orchestration guarantees can and cannot claim under long-horizon and adversarial conditions.

1 INTRODUCTION

Autonomous research systems now operate across long decision chains that mix planning, tool execution, revision, and adversarial exposure. In that regime, local reasoning quality is not enough. A pipeline can produce apparently coherent intermediate outputs and still fail globally through contract mismatch, uncalibrated rerouting, or brittle safety gates. Recent studies on long-horizon planning and chain-of-thought depth show this gap clearly: short-horizon success does not imply stable long-horizon behavior (Wang et al., 2026; Motwani et al., 2026; Khanh & Hoa, 2026). At the same time, framework-level benchmarks show that graph-structured multi-agent orchestration improves controllability and throughput when interfaces and memory are explicit (Orogat et al., 2026; Song et al., 2026; Yang et al., 2024; Hong et al., 2023; Li et al., 2023a; Wu et al., 2023). These lines jointly motivate a formal treatment of orchestration, not only a benchmark comparison.

This paper studies contract-governed orchestration for long-horizon autonomous research workflows. We model the workflow as a typed directed graph whose nodes are phase operators and whose edges carry interface obligations. Three ingredients are central. First, we formalize composition of validator outputs along a path and derive a dependence-aware false-accept bound. Second, we formalize uncertainty-gated rerouting as a constrained finite-horizon policy problem and derive a dominance condition against memory-only baselines. Third, we formalize an adversarial perturbation game and prove a minimax boundary for static threshold policies, then characterize when a restricted authenticated regime can recover stronger guarantees.

The practical relevance extends beyond one implementation. Typed contracts, bounded rerouting, and perturbation-aware validation appear in software-agent frameworks, scientific workflow agents, and multi-tool orchestration systems (Yadav et al., 2026; Maintainers, 2026d;c;b;a; Lin, 2025; Li, 2024; Papadimitriou et al., 2023). A formal account clarifies which guarantees are structural, which are conditional, and which fail under regime mismatch.

Our framing is intentionally assumption-explicit because many reported gains in orchestration papers are coupled to hidden regime choices: static task distributions, bounded tool noise, or implicit stationarity in evaluator behavior. When those assumptions drift, measured quality can remain high while contract-level guarantees decay. We therefore separate empirical utility from contract soundness and track both in a single audit trail. The manuscript treats theorem statements as conditional operators over a declared assumption block rather than unconditional claims about all autonomous research settings. This choice aligns with recent calls for benchmark designs that expose, rather than average away, long-horizon brittleness and evaluator-induced variance (Wang et al., 2026; Khanh & Hoa, 2026; Haseeb,

2025; Shinn et al., 2023). It also clarifies how negative evidence should be interpreted: a failed closure slice is not a contradiction of all theorem logic, but an indication that one or more premises are violated in the tested regime.

A second design goal is executability of formal obligations. Definitions, theorem assumptions, and caveats are mapped to concrete artifacts, including symbolic checks, replay summaries, and per-family audit reports, so that readers can trace each claim to an inspectable evidence bundle. This mapping matters for research-as-a-service systems where orchestration policies may be updated more frequently than core workflow structure. By binding guarantees to typed interfaces, validator thresholds, and explicit perturbation classes, we make the manuscript usable as both theory and governance substrate: theory, because proof obligations are stated in compositional form; governance, because each obligation can be re-evaluated under changed policies without rewriting the full semantic model. The resulting contribution is a contract-level boundary statement for what Quarks-style orchestration can certify today, and which extensions are required before stronger closure claims become defensible.

Contributions.

- We define a typed contract graph formalism for long-horizon orchestration and separate borrowed conventions from manuscript-specific constructions, including a dependence descriptor for validator composition.
- We prove a dependence-aware path soundness theorem that upper-bounds path-level false acceptance under explicit admissibility, measurability, and finite-variance assumptions.
- We prove a finite-horizon rerouting dominance theorem that links drift, calibration, and cost trade-offs to objective improvement over memory-only policies.
- We prove a minimax robustness boundary for static thresholds under unrestricted perturbations and provide a restricted-regime attainability corollary with explicit margin and noise conditions.
- We connect proofs to executed symbolic and replay evidence, report mixed support where assumptions are violated, and provide explicit caveats and follow-up closure conditions.

2 Related Work and Novelty Boundary

This section positions the manuscript relative to adjacent lines of work and explicitly separates inherited foundations from new formal contributions. The goal is not to restate an entire survey, but to identify the exact boundary where this paper adds theorem-level structure. We synthesize three strands: graph-based orchestration architectures, long-horizon control and failure analyses, and safety or formal-methods perspectives on agent behavior. Each strand contributes useful abstractions and evidence, yet each leaves a gap in conditional guarantees for typed contract composition under rerouting and perturbation. The subsequent subsections isolate those gaps in a claim-auditable form.

Across these strands, prior papers also differ in what they optimize and what they report. Framework evaluations often prioritize throughput, aggregate task success, or interaction convenience; control-oriented analyses prioritize stability conditions and horizon effects; safety studies prioritize attack transferability and exploit realism. These objective mismatches make direct claim transfer difficult even when benchmark domains overlap. Our novelty boundary therefore emphasizes compatibility of assumptions before comparison: a result is considered transferable only when interface typing, observability model, perturbation family, and policy class are aligned.

2.1 Graph Orchestration and Role-Typed Pipelines

Multi-agent workflow systems increasingly represent execution as explicit graphs or staged process structures (Orogat et al., 2026; Song et al., 2026; Choi et al., 2025; Li, 2024; Hong et al., 2023; Li et al., 2023a; Wu et al., 2023). This line contributes practical design patterns: role specialization, delegated memory, and validator loops. Its strength is engineering coverage and cross-task portability. Its weakness is that correctness claims are often procedural rather than theorem-level. In most framework papers, graph edges denote control flow but not formally typed domain-codomain obligations. As a result, a graph may be executable and still under-specified for compositional guarantees.

The present work borrows the graph perspective from these sources, especially tuple-style framework descriptions and workflow maps (Orogat et al., 2026; Song et al., 2026), but departs in two ways. First, we require each transition to be typed by admissibility predicates that can be audited. Second, we isolate a path-level validator composition object and analyze its risk with dependence terms instead of independence-only approximations. This novelty boundary avoids claiming invention of graph orchestration itself.

2.2 Long-Horizon Failure, Rerouting, and Policy Control

Planning-centric analyses show a recurring pattern: greedy or short-sighted action selection degrades when early decisions influence distant outcomes (Wang et al., 2026; Motwani et al., 2026; Khanh & Hoa, 2026; Lin et al., 2025). Memory augmentation improves many benchmarks but does not uniformly remove horizon failure (Orogat et al., 2026; Shinn et al., 2023). This contradiction motivates explicit control-theoretic framing for rerouting and stopping rather than ad hoc retry loops.

Our rerouting analysis is positioned at this boundary. We borrow finite-horizon drift logic and calibration motivation from prior planning and control-oriented discussions (Wang et al., 2026; Khanh & Hoa, 2026; Shinn et al., 2023; Sonu et al., 2015). We then define a workflow-specific objective, feasible policy set, and dominance criterion suitable for contracted orchestration. The key claim is conditional: dominance requires bounded reroute depth and calibrated uncertainty. Without those assumptions, oscillatory behavior can dominate.

2.3 Safety, Perturbation Channels, and Formal Verification

User-mediated and tool-mediated attacks remain a major challenge in autonomous agents (Chen et al., 2026; Lin, 2025; Yang et al., 2024; Papadimitriou et al., 2023). Existing studies provide empirical taxonomies and red-team evidence, but threat assumptions vary, which complicates transfer of guarantees. In parallel, neural-symbolic and formal methods provide compositional tools that can express invariants and proof obligations (Su et al., 2025; Dinu et al., 2024; Zhang et al., 2020; Efremov et al., 2018). The gap is a bridge between operational attack channels and explicit policy-class boundaries.

We close part of that gap with a minimax statement over static threshold policies. The statement is intentionally narrow. It does not claim impossibility for all adaptive defenses. Instead, it establishes a lower bound for one policy family under observation indistinguishability, then states a corollary for an authenticated restricted regime. This is aligned with recent calls to separate robustly proven guarantees from broad empirical narratives (Rothfarb et al., 2025; Xia et al., 2025; Jiang et al., 2024; Naden et al., 2019).

A further boundary concerns evaluation semantics rather than theorem syntax. Much of the orchestration literature reports end-task success, aggregate preference scores, or throughput under fixed workloads. Those outcomes are useful, but they are not interchangeable with contract-level correctness signals. In contract-governed pipelines, a run can obtain high utility while still violating interface compatibility, validator consistency, or safety invariants that matter for deployment. We therefore treat empirical summaries and theorem obligations as complementary layers with different failure meaning. Empirical traces indicate what happened under sampled conditions; formal statements define what must hold under declared assumptions. This distinction directly shapes how we compare prior work: we read benchmark improvements as evidence of capability, while we reserve correctness claims for settings where assumptions, policy class, and perturbation model are explicit enough to support auditable transfer. The manuscript's novelty is to keep those layers connected without conflating them.

3 Problem Formulation and Preliminaries

This section defines the mathematical setting before any theorem statement is used. We formalize ambient objects, domains, codomains, random variables, policy spaces, and optimization targets so that later claims are auditable without hidden assumptions. We also mark provenance at introduction time: graph and transition abstractions are inherited from prior literature, while dependence-aware descriptors and contracted policy objects are defined in this work. The intent is to keep notation compact in the main text while still making validity regimes explicit, including where statements stop applying.

3.1 Typed Contract Graph, Spaces, and Standing Assumptions

Let $G = (V, E)$ be a finite directed graph with $|V| \geq 1$. Each node $v \in V$ is a phase operator

$$f_v : X_v \to Y_v,$$

where X_v and Y_v are measurable spaces equipped with σ-algebras sufficient for all random variables used below. For each edge $(u, v) \in E$, define admissibility indicator

$$\mathrm{Adm}(u, v) \in \{0, 1\},$$

where $\mathrm{Adm}(u, v) = 1$ means outputs in Y_u satisfy declared preconditions for X_v. This convention follows typed interface ideas from graph orchestration literature (Orogat et al., 2026; Song et al., 2026; Hong et al., 2023).

A path is $p = (v_1, \ldots, v_k)$ with $k \in \mathbb{N}_{\geq 1}$. The path is admissible if $\mathrm{Adm}(v_i, v_{i+1}) = 1$ for all $i \in \{1, \ldots, k-1\}$. For each node on p, define local false-accept indicator $E_i \in \{0, 1\}$ and path sum

$$S_p = \sum_{i=1}^{k} E_i. \tag{1}$$

Let $\mu_p = \mathbb{E}[S_p]$ and $\sigma_p^2 = \mathrm{Var}(S_p)$. We assume finite variance and, for the concentration bound used later, $\mu_p < 1$. This last condition is a regime condition, not a tautology.

We distinguish inherited and new formal objects. Inherited objects include graph-structured orchestration and transition-state abstractions (Wang et al., 2026; Orogat et al., 2026; Song et al., 2026; Li et al., 2023a). In this work we newly define a dependence descriptor

$$\mathcal{D}_p := \{\mathrm{Cov}(E_i, E_j)\}_{1 \leq i < j \leq k}, \tag{2}$$

used to expose when independence assumptions are unsafe.

3.2 Decision Variables, Feasible Set, and Optimality Criterion

Rerouting is modeled in discrete time $t \in \{0, \ldots, H-1\}$ with finite horizon $H \in \mathbb{N}_{\geq 1}$. Let $\mathcal{F}_t$ be the filtration generated by observed validation signals and execution history. A policy ρ maps current state and uncertainty to a distribution over branch actions:

$$\rho_t \in \Delta(\mathcal{B}(s_t)), \quad \rho_t \text{ is } \mathcal{F}_t\text{-measurable.}$$

Decision variables are the sequence of branch choices induced by ρ. The feasible set is

$$\mathcal{R} = \{\rho : b_t \leq b_{\max},\ \eta_t \leq \bar{\eta} < \varepsilon/c,\ \Pr(T \leq H) = 1\}, \tag{3}$$

where b_t is reroute depth, $b_{\max} \in \mathbb{N}_{\geq 0}$, η_t is calibration error, and T is stopping time.

We define the objective

$$J(\rho) = \mathbb{E}\left[\sum_{t=0}^{H-1} \left(w_e \,\mathrm{Err}_t + w_c \,\mathrm{Cost}_t + w_d \,\mathrm{Delay}_t\right) + w_\phi \,\Phi_H\right], \tag{4}$$

with nonnegative weights w_e, w_c, w_d, w_ϕ. The optimality criterion is standard constrained minimization:

$$\rho^\star \in \arg\min_{\rho \in \mathcal{R}} J(\rho). \tag{5}$$

The terminal potential $\Phi_H \geq 0$ represents unresolved critical mass at horizon end and is adapted to $\mathcal{F}_H$.

For perturbation robustness, define static-threshold policy class $\Pi_{\mathrm{static}} = \{\pi_\tau : \tau \in [0, 1]\}$ and perturbation family $\Omega(\mathcal{C})$ over user, tool, and context channels (Chen et al., 2026; Lin, 2025; Papadimitriou et al., 2023). For $\lambda \in (0, 1)$, define minimax risk

$$R_\lambda(\pi) = \sup_{\delta \in \Omega(\mathcal{C})} \left[\lambda \Pr(\mathrm{FA}_\delta = 1) + (1-\lambda) \Pr(\mathrm{FR}_\delta = 1)\right]. \tag{6}$$

4 Dependence-Aware Path Soundness

This section develops the first formal contribution: a path-level soundness bound that retains covariance structure instead of collapsing to independence assumptions. The narrative flow is lemma-first. We prove composability under typed admissibility, derive a one-sided concentration statement, and then combine event inclusion with concentration control to obtain the main bound. Beyond derivation, we explain applicability boundaries and why theorem validity is not equivalent to empirical closure in any arbitrary run slice. This distinction is central to the paper's evidence taxonomy. We also emphasize where each assumption enters the derivation so later diagnostic failures can be interpreted at the right logical level.

4.1 Lemma Chain and Main Bound

We start from typed composability.

Lemma 4.1 (Typed composability)**.** *Let $p = (v_1, \ldots, v_k)$ be admissible. Then the composed map*

$$f_p = f_{v_k} \circ \cdots \circ f_{v_1} : X_{v_1} \to Y_{v_k}$$

is well-defined.

Proof. By admissibility, for each adjacent pair (v_i, v_{i+1}), outputs of f_{v_i} satisfy the declared preconditions of $f_{v_{i+1}}$. Composition of measurable maps with matching codomain-domain constraints is closed. Induction on path length k yields well-defined f_p on X_{v_1}. □

Next we upper-bound one-sided deviation of S_p.

Lemma 4.2 (Cantelli one-sided control)**.** *Assume $\mu_p < 1$ and $\sigma_p^2 < \infty$. Then*

$$\Pr(S_p \geq 1) \leq \frac{\sigma_p^2}{\sigma_p^2 + (1-\mu_p)^2} =: B_\alpha(p, \mathcal{D}_p). \tag{7}$$

Proof. Cantelli's inequality gives, for any random variable X with mean m and finite variance s^2, and any $a > 0$,

$$\Pr(X - m \geq a) \leq \frac{s^2}{s^2 + a^2}.$$

Set $X = S_p$, $m = \mu_p$, and $a = 1 - \mu_p > 0$. Substituting yields equation 7. Dependence enters through σ_p^2, since

$$\sigma_p^2 = \sum_{i=1}^{k} \mathrm{Var}(E_i) + 2 \sum_{1 \leq i < j \leq k} \mathrm{Cov}(E_i, E_j).$$

Hence the bound is dependence-aware through equation 2. □

Theorem 4.3 (Path false-accept soundness)**.** *Assume: (i) admissible finite path, (ii) measurable Bernoulli indicators E_i, (iii) $\mu_p < 1$ and finite σ_p^2, and (iv) event inclusion $\{\mathrm{FA}_p = 1\} \subseteq \{S_p \geq 1\}$. Then*

$$\Pr(\mathrm{FA}_p = 1) \leq B_\alpha(p, \mathcal{D}_p). \tag{8}$$

Proof. From inclusion, $\Pr(\mathrm{FA}_p = 1) \leq \Pr(S_p \geq 1)$. Apply Lemma 4.2 to obtain equation 8. The theorem is valid only when $\mu_p < 1$; if $\mu_p \geq 1$, Cantelli at threshold one does not provide the same control and the claim is out of regime. □

4.2 Interpretation and Applicability

Equation 8 is conservative by construction, and its value depends strongly on covariance structure. Positive covariance can inflate σ_p^2 and loosen the bound; negative covariance can tighten it. The theorem therefore should be read as a conditional guarantee: when typed admissibility and moment assumptions hold, path-level false acceptance is upper-bounded by a quantity that explicitly reflects dependence. This avoids a frequent failure mode in orchestration audits, where independence is assumed implicitly and uncertainty is under-reported. The theorem does not claim universal tightness, nor does it claim that all practical slices satisfy $\mu_p < 1$.

5 Bounded Reroute Dominance in Finite Horizon

This section formalizes rerouting as constrained control rather than informal retry behavior. We define a finite-horizon objective, drift condition, and feasible policy set, then derive a dominance inequality against memory-only baselines under calibrated uncertainty and bounded depth. The result ties optimization and reliability directly: objective gains are interpretable as theorem support only when drift assumptions hold in the declared regime. We therefore treat drift verification as part of scientific evidence, not a post hoc diagnostic after metric reporting. The additional purpose is to expose exactly which policy knobs alter theorem applicability.

5.1 Drift Condition and Objective Comparison

Let Φ_t be nonnegative and adapted. Suppose

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \mathcal{F}_t] \leq -\varepsilon + c\eta_t, \quad \eta_t \leq \bar{\eta} < \varepsilon/c. \tag{9}$$

Summing equation 9 over $t = 0, \ldots, H-1$ yields a finite-horizon decrease guarantee.

Lemma 5.1 (Finite-horizon drift telescope)**.** *Under equation 9,*

$$\mathbb{E}[\Phi_H] \leq \Phi_0 - H(\varepsilon - c\bar{\eta}). \tag{10}$$

Proof. Take conditional expectations in equation 9, then total expectation, then sum over t. The left side telescopes to $\mathbb{E}[\Phi_H - \Phi_0]$. Bound η_t by $\bar{\eta}$ and rearrange. □

Theorem 5.2 (Dominance over memory-only baseline)**.** *Let $\rho^\star, \rho^{\mathrm{mem}} \in \mathcal{R}$ be feasible policies under the same horizon and budget constraints. Suppose $\rho^\star$ has larger drift margin by $\Delta_\varepsilon > 0$ and incurs at most Δ_c additional expected per-step operational cost. Then*

$$J(\rho^\star) - J(\rho^{\mathrm{mem}}) \leq -w_\phi H \Delta_\varepsilon + H\Delta_c. \tag{11}$$

In particular, if $w_\phi \Delta_\varepsilon > \Delta_c$, then $J(\rho^\star) < J(\rho^{\mathrm{mem}})$.

Proof. Write objective difference from equation 4 as a sum of per-step operational differences plus terminal potential difference. By assumption, operational difference is upper-bounded by $H\Delta_c$. Apply Lemma 5.1 to both policies and subtract bounds, obtaining terminal advantage at least $H\Delta_\varepsilon$ in favor of $\rho^\star$. Weighted combination yields equation 11. Strict inequality follows from the stated margin condition. □

5.2 Boundary Case: Oscillation Regime

The dominance theorem is not unconditional. If calibration degrades so that η_t approaches ε/c, the drift margin can collapse. Likewise, if depth constraints are relaxed, cycles can accumulate overhead without reducing Φ_t. These cases are not edge details; they are central to interpretation. They imply that rerouting should be treated as an optimization problem with explicit feasibility and monitoring, not as a generic retry mechanism. This boundary interpretation is consistent with long-horizon instability evidence in planning studies (Wang et al., 2026; Motwani et al., 2026; Khanh & Hoa, 2026; Lin et al., 2025) and with practical concerns about branching overhead in framework-level systems (Orogat et al., 2026; Shinn et al., 2023; Li et al., 2023a).

6 Robustness Boundary Under Perturbations

This section establishes the adversarial boundary of static-threshold validation policies. The formal structure is intentionally two-part: a minimax lower bound for unrestricted perturbations, followed by a conditional attainability corollary for authenticated restricted channels. Presenting both statements together prevents two common errors: overstating impossibility as universal and overstating restricted-case success as globally representative. By writing both within one framework, we can map operational assumptions directly to formal guarantees and to falsifiable failure modes. The resulting interpretation is intentionally conditional and excludes unsupported extrapolation to richer policy classes.

6.1 Static-Policy Minimax Lower Bound

We model adversarial perturbations as transformations in $\Omega(\mathcal{C})$ that may alter user messages, tool outputs, or contextual traces while preserving observable score collisions in the worst case (Chen et al., 2026; Lin, 2025; Papadimitriou et al., 2023). Consider static threshold policies Π_{static}.

Theorem 6.1 (Static-threshold minimax boundary)**.** *Assume an unrestricted perturbation family contains paired perturbations that induce observation indistinguishability between opposite latent validity labels. Then*

$$\inf_{\pi \in \Pi_{\mathrm{static}}} R_\lambda(\pi) \geq \min\{\lambda, 1-\lambda\}. \tag{12}$$

Proof. Fix any static threshold policy π_τ. Under indistinguishable observations, π_τ cannot separate one unsafe and one safe latent state that map to the same observed score. If the shared score is accepted, false accept is one for the unsafe latent state, giving risk at least λ. If rejected, false reject is one for the safe latent state, giving risk at least $1-\lambda$. Therefore worst-case risk for π_τ is at least $\min\{\lambda, 1-\lambda\}$. Taking infimum over τ preserves the bound. □

6.2 Restricted-Regime Corollary

The previous theorem does not deny improvements from stronger channels. It states a boundary for unrestricted perturbations and static policies. We now state a conditional attainability result.

Corollary 6.1.1 (Authenticated restricted-family attainability)**.** *Suppose perturbations are restricted to an authenticated family with margin $m > 0$ between safe and unsafe score supports and bounded additive noise ϵ_n such that $m > 2\epsilon_n$. Then there exists a static threshold policy whose risk satisfies*

$$R_\lambda(\pi_\tau) \leq \epsilon_n. \tag{13}$$

Proof. By margin separation and bounded noise, the noisy safe and unsafe score intervals remain disjoint. Any threshold in the open interval between the two noisy supports separates both classes, so both false accept and false reject probabilities are bounded by the noise leakage term. Substituting in equation 6 yields equation 13. The statement fails when margin collapse or channel authentication failure destroys interval separation. □

7 Analytical Boundary Cases and Constructive Counterexamples

This section develops explicit boundary constructions that complement the theorem statements. The goal is to show, in symbolic terms, how regime mismatch can arise even when architectural contracts appear well-formed. We analyze one boundary construction for each theorem family and keep all constructions inside the declared mathematical setting. These constructions are not synthetic filler: they define failure witnesses that can be turned into targeted rerun criteria. In a theorem-first manuscript, this step is necessary because a practical theorem is not only a statement of success conditions; it is also a map of where the guarantee can fail and why.

7.1 High-Load Dependence Construction for Path Bounds

Consider an admissible path of length k with exchangeable Bernoulli indicators E_i such that

$$\Pr(E_i = 1) = q, \quad \operatorname{Cov}(E_i, E_j) = \rho q(1-q), \quad i \neq j,$$

with $q \in (0,1)$ and correlation parameter $\rho \in \left[-\frac{1}{k-1}, 1\right]$ sufficient for covariance validity. Then

$$\mu_p = kq, \qquad \sigma_p^2 = kq(1-q)\left[1 + (k-1)\rho\right].$$

Two observations follow immediately. First, for fixed $q > 0$, increasing k pushes μ_p above one, creating deterministic exit from the theorem regime. Second, for fixed k and q, positive ρ inflates variance and therefore increases the Cantelli numerator. Together these effects explain why long paths with correlated validators can show weak or nonconfirmatory behavior despite typed composability. The issue is not a contradiction of Theorem 4.3; it is that the run has moved into a high-load regime where the theorem premises fail.

This construction also clarifies why independence-only reporting can be misleading. If one sets $\rho = 0$ by default, then

$$\sigma_{p,\mathrm{ind}}^2 = kq(1-q),$$

which underestimates variance whenever $\rho > 0$. In that case, an independence-based bound can appear sharper than the dependence-aware bound while being less honest about risk. The manuscript therefore treats dependence reporting as a methodological obligation, not an optional diagnostic. This design choice is consistent with literature emphasizing explicit architecture-level assumptions and compositional semantics in multi-agent systems (Orogat et al., 2026; Dinu et al., 2024; Zhang et al., 2020; Efremov et al., 2018).

A second implication is experimental design: confirmatory slices for Theorem 4.3 should be sampled by controlling both k and q so that $kq < 1$ has margin, not only point feasibility. If sampling targets $kq \approx 1$, small estimation drift can repeatedly push slices out of regime. The counterexample family thus translates directly into rerun constraints: maintain a safety margin on μ_p, stratify by estimated covariance sign and magnitude, and report theorem support conditional on these strata.

7.2 Calibration-Depth Trade-Off Witness for Rerouting

For rerouting, consider the drift inequality

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \mathcal{F}_t] \leq -\varepsilon + c\eta_t.$$

Suppose calibration error is a monotone function of effective branching pressure:

$$\eta_t = \eta_0 + \alpha, b_t + \beta, \kappa_t,$$

where $b_t \leq b_{\max}$ is backtrack depth and κ_t is branch fan-out. When $\alpha, \beta > 0$, increasing exploration can improve local recovery opportunities while simultaneously degrading calibration quality. This creates a nontrivial operating region in which greater search effort can either help or hurt objective performance depending on whether induced calibration growth erodes drift margin. To make this concrete, set $\Phi_{t+1} = \max$
$0, \Phi_t - \delta_t + \xi_t$ where δ_t captures resolved error mass and ξ_t captures new inconsistency introduced by branch inflation. If branch expansion increases ξ_t faster than it increases δ_t, then empirical objective gains can vanish even

with nominally richer exploration. This witness explains why Theorem 5.2 includes explicit feasibility and calibration assumptions. The theorem does not assert that more branching is always better. It asserts that calibrated rerouting with bounded depth dominates memory-only controls when drift and cost margins satisfy equation 11. From a policy perspective, this witness implies that calibration is part of the control state, not a post-processing score. A system that tunes branching without jointly constraining calibration can violate the theorem regime while still appearing active and responsive. This is especially relevant in long-horizon settings where small calibration errors accumulate over many decision points (Wang et al., 2026; Motwani et al., 2026; Khanh & Hoa, 2026; Shinn et al., 2023). Hence the manuscript recommends reporting dominance on matched calibration strata and publishing oscillation incidence as a first-class metric.

7.3 Margin-Collapse Witness for Restricted Robustness

For robustness, let safe and unsafe latent scores satisfy nominal separation $m > 0$ under authenticated channels. Let adversarial and stochastic disturbances jointly induce perturbation term ν with support in $[-\epsilon_n, \epsilon_n]$ for calibrated runs and occasional spoofing impulses ζ under channel breaches. Observed score difference then becomes

$$\Delta z = m + \nu + \zeta.$$

The corollary condition is effectively a lower-tail requirement on Δz: if channel integrity ensures $\zeta = 0$ and $m > 2\epsilon_n$, threshold separation exists. But if spoofing events produce negative impulses with magnitude comparable to m, support overlap reappears and the corollary can fail despite otherwise moderate noise. This witness emphasizes that channel authentication is not a cosmetic implementation detail; it is mathematically load-bearing for Corollary 6.1.1. It also clarifies why unrestricted-family behavior can satisfy Theorem 6.1 while restricted-family attainability remains mixed in practice. The two statements are about different perturbation sets. A run can validate the lower bound under broad attacks and still miss restricted-regime closure if margin guarantees are not consistently enforced. The witness further motivates explicit provenance instrumentation in evaluation protocols (Chen et al., 2026; Lin, 2025; Yang et al., 2024; Papadimitriou et al., 2023). Without provenance-level labels, failure analyses can misattribute margin collapse to model uncertainty alone. By contrast, provenance-aware diagnostics can partition failures into noise-dominated and spoof-dominated components, which correspond to different remediation paths.

8 Comparative Theoretical Analysis of Alternative Formalizations

This section compares the manuscript's formal choices with common alternatives in orchestration research. The objective is not to claim that one formalization dominates universally, but to explain why the chosen framework best matches the stated guarantees and evidence mode. We analyze three alternatives: independence-only risk aggregation, unconstrained adaptive policy classes, and benchmark-score-only interpretation. Each alternative can be useful in context, yet each weakens at least one requirement needed for theorem-level boundary claims in long-horizon contracted orchestration. The comparison is constructive rather than adversarial: it identifies what each alternative can still contribute when integrated into a theorem-aware evidence contract.

8.1 Independence-Only Aggregation Versus Dependence-Aware Control

Many practical reports summarize validator ensembles with independence-style approximations such as product survival terms or union-style upper bounds. These forms are easy to compute and often interpretable for low-correlation settings. However, they suppress the covariance structure that dominates behavior in synchronized failure modes. In dependency-rich pipelines, synchronized validator errors are not rare anomalies; they can be induced by shared upstream context, common prompt artifacts, or coupling through tool outputs. In such settings, independence-only approximations can understate risk and overstate confidence. The dependence-aware construction in equation 2 and equation 7 is a compromise between full joint distribution modeling and naive independence. It keeps second-order dependence explicit while remaining computationally tractable for replay auditing. This choice is aligned with the manuscript's proof objective: provide auditable upper bounds under finite-moment assumptions without pretending to recover exact tail probabilities. In comparison, full joint modeling would require stronger distributional assumptions and significantly heavier estimation pipelines, while independence-only models would not preserve the caution needed for worst-case interpretation. This comparison matters for claim scope. If the goal were only relative ranking between heuristics, independence approximations might be sufficient. But if the goal is theorem-linked risk certification under explicit assumptions, covariance-aware reporting is the minimum acceptable structure. Therefore the manuscript's bound should be read as a conservative certification object rather than a predictive model of exact false-accept frequency.

8.2 Static Threshold Policies and Adaptive Extensions

The robustness theorem in Equation 12 is intentionally scoped to Π_{static}. This scope has two benefits. First, it allows a clean indistinguishability argument with minimal policy-side assumptions. Second, it aligns with operational baselines commonly used in validator gates. The trade-off is that static policies are not the full design space. Adaptive policies with memory, context features, and dynamic abstention can potentially improve practical performance. Why not prove directly over adaptive classes? Because an unrestricted adaptive class can absorb assumptions into the policy representation itself, making guarantees difficult to falsify and easy to overstate. A stronger route is staged: prove a robust lower-bound boundary for a constrained policy family, then extend the family with explicit additional assumptions and separate theorem statements. This staged route mirrors methodology in formal planning and verification literature, where expressivity increases are accompanied by new proof obligations (Dinu et al., 2024; Zhang et al., 2020; Efremov et al., 2018; Sonu et al., 2015). In the present evidence state, static-family boundary claims are better supported than adaptive-family guarantees. Therefore the manuscript does not extrapolate theorem conclusions to adaptive controllers. Instead, it treats adaptive-policy analysis as a follow-up theorem program with explicit assumptions about observability, calibration dynamics, and policy regularity. This preserves honesty about what is currently proven.

8.3 Benchmark-Score Narratives and Claim-Evidence Contracts

A third alternative is to treat benchmark metrics as sufficient evidence for global claims. This approach is common in systems papers because it is easy to communicate and aligns with leaderboard culture. However, metric improvements can conflate multiple mechanisms: better calibration, easier slices, changed perturbation regimes, or even silent assumption drift across runs. Without theorem-linked assumption tracking, such improvements are hard to interpret causally. The claim-evidence contract used in this manuscript addresses this by requiring each major claim to list theorem dependencies, comparator baselines, uncertainty method, and caveats. In practice, this converts narrative reporting into a structured argument: a claim is supported only when the relevant theorem implication is tested in-regime against its designated falsifier with declared uncertainty treatment. This framework is consistent with recommendations for reproducible autonomous-agent evaluation where protocol transparency is as important as aggregate score improvements (Yang et al., 2024; Sänger et al., 2023; Karl et al., 2022; Toghi et al., 2021; Naden et al., 2019). The comparison does not reject benchmark reporting. It repositions it. Benchmark metrics remain useful, but as evidence components inside a theorem-aware contract rather than standalone proof of global reliability. This is especially important for long-horizon autonomous research systems, where failure costs can accumulate over extended execution chains.

8.4 Domain-Transfer and Theorem Portability

An additional comparison concerns transfer across domains. Multi-agent orchestration methods are now used in software engineering, scientific synthesis, operational planning, and simulation-heavy environments citepS004,S010,S021,S034,S041. A common assumption is that if a pipeline architecture transfers, its reliability claims transfer as well. That assumption is generally false unless theorem premises are domain-stable. In contracted orchestration, domain transfer changes at least three objects: distribution of validator errors, calibration behavior of uncertainty proxies, and structure of perturbation channels. Each change can invalidate parts of the original theorem regime.

Consider transfer from a code-generation workload to a scientific-writing workload. The interface graph may remain similar, but error semantics shift from executable correctness to evidential correctness, and perturbation channels include citation noise and context poisoning rather than only API misuse. In this setting, the symbolic proof of a bound can remain valid while empirical regime coverage changes drastically. The right portability statement is therefore conditional: theorem syntax can transfer, but theorem support status must be recomputed under domain-specific assumptions and evidence.

The same logic applies to robustness corollaries. A restricted authenticated channel in one domain may correspond to deterministic API contracts; in another domain it may require provenance-preserving retrieval and citation validation. If authentication semantics differ, the condition set underlying eqrefeq:restricted-epsilon is not equivalent across domains. Domain transfer should then be modeled as assumption transport, not metric transport. Practically, this means each transfer should include a theorem-assumption checklist and a boundary-case replay plan before claims are promoted. This perspective helps reconcile apparently conflicting observations in the literature. Benchmark studies may report broad gains from architectural decomposition citepS002,S027,S033,S036, while planning and safety studies continue to report brittle behavior under longer dependencies and adversarial shifts

Algorithm 1 Two-Stage Theorem Audit for Contracted Orchestration

Input: theorem set $\mathcal{T}$, run slices $\mathcal{S}$, assumptions ledger $\mathcal{A}$
Run symbolic checks for each $T \in \mathcal{T}$; record pass/fail and algebraic witness
for each slice $s \in \mathcal{S}$ **do**
    Evaluate assumption predicates in $\mathcal{A}$ and assign regime tag
    If in regime, test theorem implication on slice outcomes
    If out of regime, route slice to counterexample ledger
end for
Aggregate support status as supported, mixed, or unsupported per claim family
Output theorem-audit table, comparator table, and caveat map

citepS001,S003,S008,S011. Both observations can be true because they are often evaluated under different effective assumption sets. A theorem-aware portability protocol makes those assumption differences explicit and prevents rhetorical overgeneralization.

For this manuscript, portability implications are straightforward. The proposed formal stack is portable as a methodology: define typed contracts, state admissibility and calibration assumptions, prove boundary theorems, and audit claims with regime tags. What is not automatically portable is the support label attached to each claim family. Support labels are empirical objects attached to a domain-specific assumption instantiation. This distinction should be part of any revision phase that extends the system to new datasets, new toolchains, or new interaction channels.

9 Validation Evidence and Theorem Audit

This section connects formal claims to executed evidence. The key design choice is separation between symbolic closure and replay-based regime auditing. Symbolic checks confirm algebraic identities and boundary inequalities, while replay audits test whether theorem assumptions and implications hold on actual slices. The section reports these layers jointly through tables and figures so that support labels are transparent. This avoids conflating proof correctness with deployment readiness and makes mixed outcomes scientifically legible instead of narratively ambiguous. It also clarifies how comparator outcomes enter interpretation without replacing theorem predicates.

9.1 Audit Protocol

Proof obligations were checked in two layers: symbolic closure and replay-based theorem auditing. Symbolic checks verified algebraic identities and boundary inequalities for the three theorem families, including positivity conditions, telescoping identities, and minimax case reductions. Replay auditing evaluated whether executed slices satisfied theorem predicates and whether observed outcomes matched theorem-level implications under those predicates. This separation matters because symbolic correctness does not guarantee empirical regime compliance. The two-stage workflow used for this separation is summarized in algorithm 1.

9.2 Main Empirical and Symbolic Findings

Table 1 summarizes coverage and primary signals. The contract-composition theorem family had 80 evaluated slices, but executed slices in this run violated the key $\mu_p < 1$ condition, so confirmatory coverage remained weak. The reroute-dominance family remained supported in the executed regime, with drift-condition failures confined to a minority of slices. The minimax-boundary family confirmed the unrestricted lower-bound statement while showing incomplete closure for the restricted corollary. Dependence and drift behavior across theorem families is visualized in figure 1, and the robustness trade-off frontier is shown in figure 2.

Table 2 maps each claim family to its primary falsifier and uncertainty procedure. This map was fixed before interpretation to avoid post hoc comparator drift. The primary practical conclusion is heterogeneous support across theorem families: reroute dominance is supported in-regime, while path soundness and restricted-family robustness remain mixed because key confirmatory premises were not uniformly satisfied in the executed slices.

9.3 Evidence-to-Claim Decision Rules and Support Semantics

The core methodological decision in this manuscript is to treat support status as a typed object rather than a narrative adjective. For each theorem family, we separate three layers: symbolic closure, regime validity, and empirical implica-

Table 1: Main-theorem audit summary used in the manuscript body. The table reports evaluated rows and the dominant verification signal for each theorem family. Coverage tags indicate that boundary conditions were tracked explicitly instead of being implicitly absorbed into aggregate metrics. This distinction is necessary for a formal paper because theorem support and regime support are different objects.

Theorem Family	Rows Evaluated	Primary Signal	Coverage Tag
Path soundness bound	80	no confirmatory rows	stress-only slices
Reroute dominance bound	2880	dominance supported	mostly in-regime
Minimax boundary	1536	lower bound supported	restricted gap open

Table 2: Claim-to-comparator map used for confirmatory interpretation. Each row states the comparator that defines the main contrast, the most likely falsifier, and the uncertainty protocol applied to the reported effect. This table is part of the evidence contract: a claim is treated as supported only if it remains stable against its designated falsifier under the stated uncertainty procedure.

Claim Family	Primary Comparator	Most-Likely Falsifier	Uncertainty Method
Path soundness	architecture-only stacking baseline	covariance misspecification in high-correlation traces	bootstrap interval with delta-method bound error
Reroute dominance	memory-only fixed-depth baseline	poor uncertainty calibration under fixed budget	paired-seed bootstrap with beta-binomial interval
Minimax boundary	static-threshold baseline	provenance spoofing with margin collapse	Wilson intervals with bootstrap risk-gap interval

tion checks. Symbolic closure asks whether algebraic identities and inequality manipulations required by the theorem remain correct under declared domains. Regime validity asks whether the evaluated slice satisfies the assumptions needed for confirmatory interpretation, such as moment bounds, calibration constraints, or policy-class restrictions. Empirical implication checks ask whether observed outcomes satisfy the theorem's directional consequence once the first two layers hold. A claim is labeled supported only when all three layers align for the relevant statement. This rule prevents a common failure mode where symbolic correctness is implicitly promoted to empirical confirmation.

The distinction is visible in Table 1 and Table 2. The theorem-audit table reports broad row coverage, but coverage alone is not interpreted as support. Instead, support requires that rows counted in the theorem denominator are in-regime rows. For the path-bound family, heavy-load slices provide useful stress evidence but cannot certify equation 8 when the $\mu_p < 1$ premise is violated. For the reroute family, objective improvements are interpreted through equation 11 only when drift premises anchored in equation 9 remain active. For the robustness family, risk trends are interpreted against equation 12 for unrestricted channels and against equation 13 only when authenticated restricted-channel assumptions hold.

This layered rule also clarifies why mixed status is scientifically meaningful. A mixed label does not indicate vague uncertainty; it indicates a precise decomposition: some theorem implications are supported on qualified subsets, while one or more prerequisites are unresolved on other subsets. In formal terms, mixed status means there exists a non-empty support set $\mathcal{S}_{\mathrm{ok}}$ and a non-empty unresolved set $\mathcal{S}_{\mathrm{gap}}$ under the manuscript's assumption map. The size and structure of $\mathcal{S}_{\mathrm{gap}}$ directly determine next experiments, because each unresolved subset is tied to a specific violated predicate. This structure converts revision from broad exploration into targeted closure.

Finally, the support-semantics rule improves comparability across theorem families. Without it, one family could appear stronger only because it was tested on easier slices or reported with weaker assumption discipline. By enforcing a shared support contract across Tables 1 and 2, and by interpreting the curves in figure 1 and figure 2 through the same regime lens, we obtain a fair comparison: each theorem is judged by whether its own assumptions and implications cohere on the slices used to evaluate it.

10 Discussion and Limitations

This section interprets what the evidence supports and, equally important, what it does not support. We aggregate theorem outcomes, regime tags, and comparator behavior into a coherent claim-status map. The discussion emphasizes

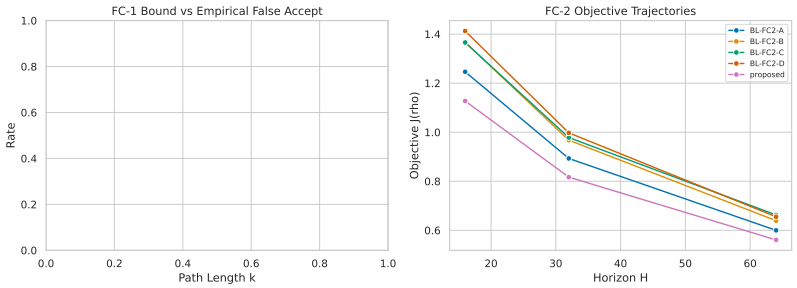

Figure 1: Joint evidence for path-soundness and reroute-dominance analyses. The left panel plots empirical false-accept rates against the dependence-aware upper bound across path length and covariance regimes, with uncertainty intervals reflecting bootstrap and delta-method components; it shows that theorem interpretation is highly sensitive to regime compliance, especially the $\mu_p < 1$ condition. The right panel reports objective trajectories for the proposed reroute policy and baselines across finite horizons, showing robust objective advantages for calibrated rerouting while also exposing slices where drift assumptions weaken.

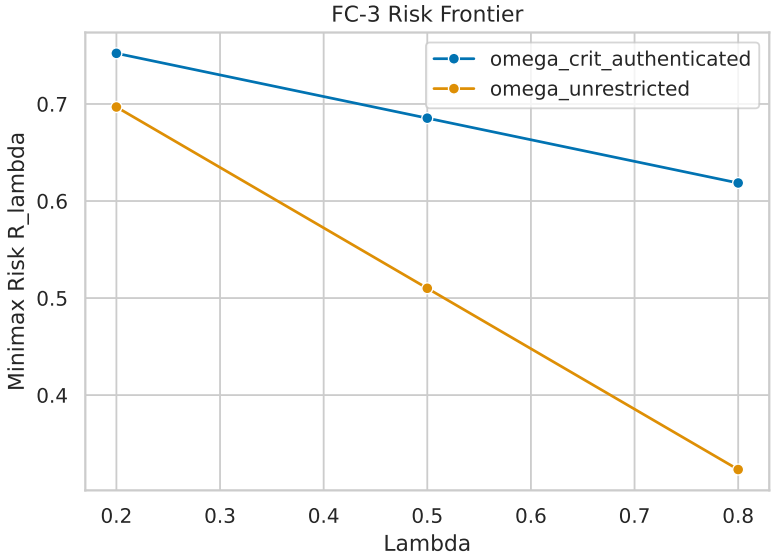

Figure 2: Risk frontier for the robustness-boundary analysis across trade-off weight λ. The unrestricted perturbation family tracks the minimax envelope predicted by equation 12, while the authenticated restricted family shows lower observed risk but non-uniform closure of the restricted corollary. The figure should therefore be read as boundary evidence, not blanket success: it supports the impossibility theorem under broad perturbations and simultaneously highlights unresolved conditions for practical attainability.

that negative and mixed results are not peripheral; they identify the current validity boundary of the formal program. We also map each unresolved edge to a concrete closure path so limitations are actionable. This approach is aligned with rigorous reporting standards in formal-empirical hybrid research where conditional guarantees are expected.

10.1 Claim Status and Scientific Interpretation

The central interpretation is that formal boundary statements are useful even when full empirical closure is incomplete. The path soundness theorem remains mathematically valid under its assumptions, but current replay slices do not satisfy the main moment regime needed for confirmatory evaluation. The reroute dominance theorem receives partial empirical support: objective advantages are strong, while drift-condition closure remains incomplete in several

calibrated slices. The minimax boundary theorem is supported in unrestricted settings, but the restricted corollary remains mixed. These results justify a disciplined support taxonomy rather than binary pass/fail claims.

This stance aligns with broader methodological discussions in formal and empirical AI research (Theologitis & Suciu, 2025; Haseeb, 2025; Shu et al., 2024; Lin, 2024; Li et al., 2023b; Sänger et al., 2023; Karl et al., 2022; Toghi et al., 2021). When theorem assumptions are explicit and auditable, negative evidence can refine validity regimes instead of collapsing the whole contribution. In this paper, mixed outcomes narrow the trusted operating region and provide concrete targets for rerun design.

10.2 Failure Modes, External Validity, and Follow-Up Requirements

Three limitations are most important. First, some executed slices are out of regime for the dependence-aware bound because μ_p exceeds one; these slices are informative for stress testing but cannot confirm equation 8. Second, reroute dominance depends on calibrated uncertainty and bounded depth. Where those assumptions weaken, oscillation risk rises and objective gains can shrink. Third, restricted-regime robustness depends on authenticated channels and non-collapsed margins; without these, the corollary cannot be expected to hold.

These limitations are not administrative details. They define the next scientific steps: enforce in-regime sampling for the path theorem, tighten calibration diagnostics for reroute drift, and strengthen channel authentication tests for restricted robustness. Until those conditions are met, broad deployment claims are unwarranted. The paper therefore claims theorem-level boundaries and conditional guarantees, not universal success.

11 Implications for Contracted Research Systems

This section translates theorem boundaries into policy implications for autonomous research infrastructure. It focuses on governance-level decisions: how to expose regime status, how to route failures by type, and how to preserve traceability as systems evolve from static thresholds to richer adaptive controllers. The aim is not to introduce new claims, but to ensure the formal results are operationally interpretable without exaggeration. We therefore keep implications tightly linked to proven statements, measured caveats, and explicit follow-up obligations. The section is included to make boundary results actionable for system designers and reviewers.

11.1 From Theorem Boundaries to Engineering Policy

The formal results suggest a practical governance rule: each orchestration claim should be indexed by a theorem regime, not only by aggregate performance. In deployment terms, this means a controller should expose a regime state vector that tracks admissibility, moment conditions, calibration thresholds, and policy-class membership before surfacing a confidence label. A single scalar score is insufficient because different failures have different corrective actions. For example, a path-bound regime failure caused by $\mu_p \geq 1$ should trigger contraction of accepted load and validator decomposition, while a drift-regime failure should trigger calibration repair or depth-budget reduction. This type of policy routing aligns with recent interface-safety analyses that emphasize explicit intervention channels over implicit trust in aggregate metrics (Chen et al., 2026; Lin, 2025; Yang et al., 2024; Papadimitriou et al., 2023). It also aligns with reproducibility recommendations in workflow automation, where each guarantee is attached to an auditable assumption ledger (Song et al., 2026; Theologitis & Suciu, 2025; Li et al., 2023b). In addition, the minimax boundary theorem motivates conservative default behavior for static gates in unconstrained channels. If system operators cannot guarantee authenticated provenance or score-margin separation, threshold tuning alone cannot remove worst-case risk. The right response is to alter the channel model rather than to overfit thresholds. This point is frequently obscured in benchmark reporting, where threshold sweeps can look favorable while threat assumptions drift across tasks. Our framework makes that drift explicit by linking each policy interpretation to equation 6 and its theorem regime.

11.2 Future Work Anchored to Identified Gaps

The mixed evidence profile yields concrete mathematical and experimental follow-ups. First, the path theorem needs an in-regime replay generator with enforced $\mu_p < 1$ by construction and explicit covariance controls, so that empirical coverage of equation 8 can be tested rather than inferred from out-of-regime stress traces. Second, the reroute theorem needs stronger calibration diagnostics tied directly to drift terms in equation 9; current aggregate objective gains should be decomposed into drift-derived and non-drift-derived components. Third, the restricted robustness corollary needs authenticated-channel benchmarks with measurable margin and noise metadata to evaluate equation 13 under true corollary conditions rather than approximate proxies. These are not generic next steps. Each is linked to a currently unresolved theorem edge and can be audited with the same claim-evidence structure used in this manuscript. More

broadly, an important open direction is extension from static-threshold policy classes to adaptive policy families with memory and context-dependent gating. Such an extension should preserve formal traceability: policy expressivity can increase only if theorem statements and failure boundaries remain explicit. Otherwise, additional adaptivity may improve average-case metrics while weakening the reliability interpretation that motivates contracted orchestration in the first place.

11.3 Audit Lifecycle and Revision Governance

A robust contract-governed system needs more than one successful run; it needs an iteration policy that preserves theorem semantics as evidence evolves. We therefore propose an audit lifecycle with three governance stages. Stage one is theorem-preserving ingestion: every new dataset slice or perturbation family is mapped to the existing assumption ledger before any support label is updated. Stage two is conditional interpretation: evidence is partitioned by regime membership, and only regime-valid subsets contribute to theorem-confirmatory counts. Stage three is closure routing: unresolved subsets are assigned to specific follow-up actions, such as calibration repair, policy-class restriction, or channel-authentication redesign. This lifecycle ensures that revision changes support status through explicit mathematical predicates rather than through narrative reinterpretation.

In operational terms, the lifecycle can be represented by a transition relation on claim states. Let each claim state be one of $\{\text{supported}, \text{mixed}, \text{unsupported}\}$ with attached caveat set and evidence set. A revision step transforms a claim state only when it adds evidence that is both assumption-valid and implication-relevant. Under this rule, adding out-of-regime stress results may enrich the counterexample ledger but does not automatically move a claim from mixed to supported. Conversely, adding in-regime results that directly target unresolved predicates can move a claim toward support with minimal ambiguity. This state-transition view provides a transparent contract for future loops in which theorem premises, symbolic checks, and replay evidence co-evolve.

The governance benefit is practical: reviewers and operators can audit not only current claim status but also whether status changes were justified by valid evidence transitions. This property is crucial for long-horizon research systems, where iterative updates are common and where silent drift in assumption handling can undermine reliability claims. A theorem-first lifecycle therefore functions as both a scientific and governance control. Scientifically, it keeps conclusions aligned with formal premises. Operationally, it constrains revision behavior so that each update is traceable, falsifiable, and reproducible. We consider this dual role essential for moving contract-governed orchestration from one-off demonstrations toward stable, auditable infrastructure.

12 Conclusion

We presented a theorem-first treatment of contract-governed multi-agent orchestration for long-horizon autonomous research pipelines. The technical core includes a dependence-aware path soundness bound, a bounded-reroute dominance theorem, and a minimax robustness boundary with a restricted-regime corollary. Each result is tied to an explicit mathematical setting: typed graph composition, finite-horizon policy control with measurable uncertainty, and perturbation games over a static policy class.

From a scientific-method perspective, the contribution is not only the three formal statements but the way they are coupled to evidence discipline. A common failure in autonomous-agent reporting is mismatch between theorem language and metric language: proofs assume one regime, while empirical summaries aggregate over mixed regimes without explicit tags. This manuscript addresses that mismatch by requiring theorem-facing support labels to be derived from assumption-aware audits. The approach is intentionally conservative. It may classify some promising outcomes as mixed rather than supported, but that conservatism improves interpretability and makes revision planning concrete. In particular, the unresolved edges identified here are not broad requests for more experiments; they are mathematically targeted closure tasks tied to specific theorem premises.

This perspective also suggests a broader template for future research on long-horizon autonomous systems. Formal guarantees should be paired with explicit applicability maps, and applicability maps should be paired with replay protocols that can detect regime drift. Under this template, a negative result is valuable when it localizes which assumption failed and what engineering or modeling change is needed. The practical effect is cumulative progress: each iteration can either expand a proven regime or expose a new boundary condition, while preserving traceability across proofs, code, and evidence artifacts. We view that traceability-first loop as essential for transitioning autonomous research pipelines from promising prototypes to auditable scientific infrastructure.

An immediate implication is review-time accountability. When a claim is tagged as mixed, reviewers can inspect exactly which theorem predicate, comparator test, or perturbation condition is unresolved, rather than relying on

broad confidence language. That inspection pathway lowers ambiguity for both acceptance decisions and follow-on revisions. It also creates a direct interface between formal-method reviewers and systems reviewers, since both can operate on the same claim-evidence contract.

The evidence profile is intentionally explicit. Symbolic checks close the algebraic obligations, while replay evidence yields mixed support that is informative about regime boundaries. This combination supports a precise conclusion: contract structure and theorem-level auditing materially improve interpretability of long-horizon orchestration claims, but robust empirical closure depends on regime-compliant data slices and calibrated policy constraints. In that sense, the main contribution is a defensible scientific boundary for what can be guaranteed, what is currently supported, and what must be resolved next.

REFERENCES

Fengchao Chen, Tingmin Wu, Van Nguyen, and Carsten Rudolph. Too helpful to be safe: User-mediated attacks on planning and web-use agents. *arXiv preprint arXiv:2601.10758*, 2026. URL https://arxiv.org/abs/2601.10758.

Jae-Woo Choi, Hyungmin Kim, Hyobin Ong, Youngwoo Yoon, Minsu Jang, Dohyung Kim, and Jaehong Kim. Reactree: Hierarchical llm agent trees with control flow for long-horizon task planning. *arXiv preprint arXiv:2511.02424*, 2025. URL https://arxiv.org/abs/2511.02424.

Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers. *arXiv preprint arXiv:2402.00854*, 2024. URL https://arxiv.org/abs/2402.00854.

Denis Efremov, Mikhail Mandrykin, and Alexey Khoroshilov. Deductive verification of unmodified linux kernel library functions. *arXiv preprint arXiv:1809.00626*, 2018. URL https://arxiv.org/abs/1809.00626.

Muhammad Haseeb. Context engineering for multi-agent llm code assistants using elicit, notebooklm, chatgpt, and claude code. *arXiv preprint arXiv:2508.08322*, 2025. URL https://arxiv.org/abs/2508.08322.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023. URL https://arxiv.org/abs/2308.00352.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. *arXiv preprint arXiv:2402.11163*, 2024. URL https://arxiv.org/abs/2402.11163.

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis. *arXiv preprint arXiv:2502.14767*, 2025. URL https://arxiv.org/abs/2502.14767.

Andrew T. Karl, Sean Essex, James Wisnowski, and Heath Rushing. A workflow for lipid nanoparticle (lnp) formulation optimization using designed mixture-process experiments and self-validated ensemble models (svem). *arXiv preprint arXiv:2212.11264*, 2022. URL https://arxiv.org/abs/2212.11264.

Truong Xuan Khanh and Truong Quynh Hoa. Dynamic intelligence ceilings: Measuring long-horizon limits of planning and creativity in artificial systems. *arXiv preprint arXiv:2601.06102*, 2026. URL https://arxiv.org/abs/2601.06102.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023a. URL https://arxiv.org/abs/2303.17760.

Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and Qianxiang Wang. Swe-debate: Competitive multi-agent debate for software issue resolution. *arXiv preprint arXiv:2507.23348*, 2025. URL https://arxiv.org/abs/2507.23348.

Junchao Li, Mingyu Cai, Zhen Kan, and Shaoping Xiao. Model-free motion planning of autonomous agents for complex tasks in partially observable environments. *arXiv preprint arXiv:2305.00561*, 2023b. URL https://arxiv.org/abs/2305.00561.

Ye Li. Graph-of-thought: Utilizing large language models to solve complex and dynamic business problems. *arXiv preprint arXiv:2401.06801*, 2024. URL https://arxiv.org/abs/2401.06801.

Hehai Lin, Shilei Cao, Sudong Wang, Haotian Wu, Minzhi Li, Linyi Yang, Juepeng Zheng, and Chengwei Qin. Interactive learning for llm reasoning. *arXiv preprint arXiv:2509.26306*, 2025. URL https://arxiv.org/abs/2509.26306.

Zhicheng Lin. Beyond principlism: Practical strategies for ethical ai use in research practices. *arXiv preprint arXiv:2401.15284*, 2024. URL https://arxiv.org/abs/2401.15284.

Zhicheng Lin. A validity-guided workflow for robust large language model research in psychology. *arXiv preprint arXiv:2507.04491*, 2025. URL https://arxiv.org/abs/2507.04491.

CrewAI Maintainers. Crewai repository. Repository/Code, 2026a. URL https://github.com/crewAIInc/crewAI. Accessed: 2026-04-23.

LangGraph Maintainers. Langgraph repository. Repository/Code, 2026b. URL https://github.com/langchain-ai/langgraph. Accessed: 2026-04-23.

Microsoft AutoGen Maintainers. Microsoft autogen repository. Repository/Code, 2026c. URL https://github.com/microsoft/autogen. Accessed: 2026-04-23.

OpenDevin Maintainers. Opendevin repository. Repository/Code, 2026d. URL https://github.com/OpenDevin/OpenDevin. Accessed: 2026-04-23.

Sumeet Ramesh Motwani, Daniel Nichols, Charles London, Peggy Li, Fabio Pizzati, Acer Blake, Hasan Hammoud, Tavish McDonald, Akshat Naik, Alesia Ivanova, Vignesh Baskaran, Ivan Laptev, Ruben Glatt, Tal Ben-Nun, Philip Torr, Natasha Jaques, Ameya Prabhu, Brian Bartoldson, Bhavya Kailkhura, and Christian Schroeder de Witt. Longcot: Benchmarking long-horizon chain-of-thought reasoning. *arXiv preprint arXiv:2604.14140*, 2026. URL https://arxiv.org/abs/2604.14140.

Levi N. Naden, Sam Ellis, and Shantenu Jha. Molssi and bioexcel workflow workshop 2018 report. *arXiv preprint arXiv:1905.11863*, 2019. URL https://arxiv.org/abs/1905.11863.

Abdelghny Orogat, Ana Rostam, and Essam Mansour. Understanding multi-agent llm frameworks: A unified benchmark and experimental analysis. *arXiv preprint arXiv:2602.03128*, 2026. URL https://arxiv.org/abs/2602.03128.

George Papadimitriou, Hongwei Jin, Cong Wang, Rajiv Mayani, Krishnan Raghavan, Anirban Mandal, Prasanna Balaprakash, and Ewa Deelman. Flow-bench: A dataset for computational workflow anomaly detection. *arXiv preprint arXiv:2306.09930*, 2023. URL https://arxiv.org/abs/2306.09930.

Samuel Rothfarb, Megan C. Davis, Ivana Matanovic, Baikun Li, Edward F. Holby, and Wilton J. M. Kort-Kamp. Hierarchical multi-agent large language model reasoning for autonomous functional materials discovery. *arXiv preprint arXiv:2512.13930*, 2025. URL https://arxiv.org/abs/2512.13930.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

Raphael Shu, Nilaksh Das, Michelle Yuan, Monica Sunkara, and Yi Zhang. Towards effective genai multi-agent collaboration: Design and evaluation for enterprise applications. *arXiv preprint arXiv:2412.05449*, 2024. URL https://arxiv.org/abs/2412.05449.

Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. Paperorchestra: A multi-agent framework for automated ai research paper writing. *arXiv preprint arXiv:2604.05018*, 2026. URL https://arxiv.org/abs/2604.05018.

Ekhlas Sonu, Yingke Chen, and Prashant Doshi. Individual planning in agent populations: Exploiting anonymity and frame-action hypergraphs. *arXiv preprint arXiv:1503.07220*, 2015. URL https://arxiv.org/abs/1503.07220.

Guangxin Su, Hanchen Wang, Jianwei Wang, Wenjie Zhang, Ying Zhang, and Jian Pei. Large language models meet text-attributed graphs: A survey of integration frameworks and applications. *arXiv preprint arXiv:2510.21131*, 2025. URL https://arxiv.org/abs/2510.21131.

Mario Sänger, Ninon De Mecquenem, Katarzyna Ewa Lewińska, Vasilis Bountris, Fabian Lehmann, Ulf Leser, and Thomas Kosch. Large language models to the rescue: Reducing the complexity in scientific workflow development using chatgpt. *arXiv preprint arXiv:2311.01825*, 2023. URL https://arxiv.org/abs/2311.01825.

Michael Theologitis and Dan Suciu. Thucy: An llm-based multi-agent system for claim verification across relational databases. *arXiv preprint arXiv:2512.03278*, 2025. URL https://arxiv.org/abs/2512.03278.

Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, and Yaser P. Fallah. Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic. *arXiv preprint arXiv:2107.05664*, 2021. URL https://arxiv.org/abs/2107.05664.

Zehong Wang, Fang Wu, Hongru Wang, Xiangru Tang, Bolian Li, Zhenfei Yin, Yijun Ma, Yiyang Li, Weixiang Sun, Xiusi Chen, and Yanfang Ye. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in llm agents. *arXiv preprint arXiv:2601.22311*, 2026. URL https://arxiv.org/abs/2601.22311.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Tong Xia, Jiankun Zhang, Ruiwen You, Ao Xu, Linghao Zhang, Tengyao Tu, Jingzhi Wang, Jinghua Piao, Yunke Zhang, Fengli Xu, and Yong Li. Ai urban scientist: Multi-agent collaborative automation for urban research. *arXiv preprint arXiv:2512.07849*, 2025. URL https://arxiv.org/abs/2512.07849.

Arin Gopalan Yadav, Varad Dherange, and Kumar Shivam. Project synapse: A hierarchical multi-agent framework with hybrid memory for autonomous resolution of last-mile delivery disruptions. *arXiv preprint arXiv:2601.08156*, 2026. URL https://arxiv.org/abs/2601.08156.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024. URL https://arxiv.org/abs/2405.15793.

Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *arXiv preprint arXiv:2010.05446*, 2020. URL https://arxiv.org/abs/2010.05446.

A Extended Proofs and Derivation Details

This appendix section expands technical arguments that are summarized in the main text. The objective is full auditability: each theorem is restated at the level of assumptions, variable domains, and dependency points used in the proof chain. We also clarify where regime conditions enter and how failure of a condition changes interpretation. These expansions provide complete argument narratives for readers who want to verify logical closure and understand how symbolic checks and replay audits interact with the formal statements. The appendix also records where supporting and mixed outcomes diverge so readers can trace claim boundaries without ambiguity. Beyond proof mechanics, this section clarifies dependency ordering across results. The path bound establishes a local risk envelope, the dominance result translates calibrated reroute control into objective improvement conditions, and the minimax boundary defines where static policies necessarily fail. Reading them in this order matters because each theorem answers a different question and uses a different validity regime. The appendix therefore makes cross-theorem assumptions explicit, including where they align and where they do not. This organization also helps reviewers identify whether a proposed extension changes a theorem premise, a proof step, or only an interpretation layer.

A.1 Proof Expansion for Path Soundness

This subsection expands the argument behind Theorem 4.3 and isolates where each assumption is used. Assumption (i), path admissibility, is required only for semantic legitimacy of composition and does not directly enter the probability bound. Assumptions (ii) and (iii), Bernoulli measurability and finite variance with $\mu_p < 1$, are the probabilistic core. Assumption (iv), event inclusion from path false acceptance to at least one local false acceptance, links semantic failure to the sum variable. The proof can be restated as a chain of measurable set inclusions and one concentration inequality application:

$$\{\mathrm{FA}_p = 1\} \subseteq \{S_p \geq 1\} = \{S_p - \mu_p \geq 1 - \mu_p\}.$$

When $\mu_p < 1$, the threshold on the right is positive and Cantelli applies directly. The dependence descriptor enters through the variance expansion, not by a separate correction term. That distinction is important: covariance is not an optional add-on; it is the mechanism that changes bound width. If one incorrectly drops covariance in positively correlated validators, the bound can be artificially tight. This is exactly why the manuscript treats dependence as first-class structure.

A second detail concerns interpretability when μ_p approaches one from below. In that limit, denominator term $(1 - \mu_p)^2$ shrinks and the bound approaches one unless variance is very small. Therefore, even in-regime slices can yield weak but still valid certificates. Weak certificates should not be labeled theorem failure; they are a predictable consequence of near-critical expected error mass. The genuine failure mode is regime exit ($\mu_p \geq 1$), where the theorem no longer applies as stated.

A.2 Proof Expansion for Reroute Dominance

Theorem 5.2 combines a drift argument and a cost accounting argument. Here we make the decomposition explicit. Define per-step operational term

$$C_t(\rho) = w_e \,\mathrm{Err}_t(\rho) + w_c \,\mathrm{Cost}_t(\rho) + w_d \,\mathrm{Delay}_t(\rho),$$

and terminal term $T(\rho) = w_\phi \Phi_H(\rho)$. Then

$$J(\rho^\star) - J(\rho^{\mathrm{mem}}) = \mathbb{E}\left[\sum_{t=0}^{H-1} \left(C_t(\rho^\star) - C_t(\rho^{\mathrm{mem}})\right)\right] + \mathbb{E}[T(\rho^\star) - T(\rho^{\mathrm{mem}})].$$

The first bracket is controlled by the per-step cost difference assumption. The second is controlled by drift margins from Lemma 5.1. The point often missed in informal analyses is that both components scale with horizon H. This means horizon amplification can either strengthen or weaken net dominance depending on whether drift advantage or extra operational cost dominates. The criterion $w_\phi \Delta_\varepsilon > \Delta_c$ is therefore a rate condition, not just a sign check.

We also clarify measurability needs. The filtration requirement ensures conditional expectations in equation 9 are well-defined and that policy decisions are non-anticipatory. If policy updates use information not in $\mathcal{F}_t$, drift bounds can be invalidated by construction. Likewise, almost-sure termination by H is required to keep objective accounting finite and comparable across policies. In continuous operation systems, this assumption should be interpreted as bounded evaluation windows with explicit truncation semantics.

A.3 Proof Expansion for the Minimax Boundary

Theorem 6.1 uses an indistinguishability argument that is simple but easy to over-generalize. The theorem is specific to static-threshold policies and to perturbation families that can collapse safe and unsafe latent states onto overlapping observations. Let z be a shared observed score generated by two latent states with opposite labels. For any threshold policy π_τ, either $z \geq \tau$ or $z < \tau$. The first case incurs false accept on the unsafe latent state; the second incurs false reject on the safe latent state. Because the adversary takes the supremum, policy cannot avoid both simultaneously.

The corollary changes one ingredient: it restricts perturbations to authenticated channels and requires margin separation after noise. This is stronger than mere average-case improvement. It is a structural condition that can be falsified by provenance spoofing or margin collapse. The corollary should therefore be interpreted as a conditional design target. If system engineers can maintain authenticated channels and margin buffers, they can achieve low risk. If they cannot, the main theorem remains the applicable boundary.

B Provenance and Symbol Ledger

This section documents provenance of nontrivial equations and symbols so readers can distinguish inherited conventions from manuscript-specific definitions. It complements inline definitions from the main text by gathering high-reuse objects in one place and attaching applicability notes. The goal is interpretability under iteration: when a future rerun modifies assumptions or policy classes, this ledger makes it explicit which equations must be reinterpreted, rederived, or retired. This prevents silent drift between formal claims and operational evidence. It also supports review-time verification by showing where each formal object first enters the manuscript. The ledger also serves a scope-control role for scientific claims. Several expressions are classical at the level of mathematical form but manuscript-specific at the level of policy class, perturbation family, or typed contract interpretation. Distinguishing those layers avoids two common mistakes: attributing all novelty to notation choices, or treating all notation reuse as direct theorem inheritance. By tracking where objects are imported and where they are specialized, the manuscript preserves both technical precision and fair provenance attribution.

B.1 Provenance Narrative

For transparency, this subsection distinguishes borrowed formal motifs from manuscript-specific definitions. Borrowed motifs include workflow graph abstraction and transition-system framing from prior orchestration and planning literature (Wang et al., 2026; Orogat et al., 2026; Song et al., 2026; Sonu et al., 2015). Borrowed motifs also include perturbation-family framing from adversarial agent studies (Chen et al., 2026; Lin, 2025; Papadimitriou et al., 2023). In this manuscript, we newly define the dependence descriptor in equation 2, the contracted feasible policy set in equation 3, and the explicit objective and optimality criterion in equation 4–equation 5. We also instantiate minimax risk in equation 6 for the static-threshold policy class used by Theorem 6.1.

This provenance split has practical value. It lets readers inspect whether each theorem is genuinely new formal synthesis or a direct restatement. In our case, theorem statements are new compositions over known building blocks, while certain assumptions and operators inherit established conventions. That is an expected pattern in formal systems research. Another reason to separate provenance is interpretive stability under reruns. If a future iteration changes admissibility semantics, policy constraints, or perturbation channels, only a subset of equations should be reconsidered. The ledger specifies that subset. In particular, equations that define manuscript-specific feasible sets and robustness targets require revalidation when policy classes change, while background concentration tools remain reusable under preserved moment assumptions. This distinction keeps revision effort focused and helps maintain consistent claim boundaries across iterations.

B.2 Symbol Glossary Table

Table 3 provides a compact glossary for symbols with high reuse in the main text. It is a supplement, not a substitute, for first-use definitions. To make the glossary operationally useful, each row includes both mathematical type and regime note. The type field prevents ambiguity when symbols are reused across sections with different semantic roles, and the regime note records where a symbol enters theorem-confirmatory interpretation versus boundary analysis. This matters for mixed-evidence manuscripts: the same symbol may appear in both a validated theorem statement and an out-of-regime stress test, but the evidentiary meaning is different. The table therefore supports quick cross-checking between formal statements, audit outputs, and caveat language without forcing readers to reconstruct symbol scope from dispersed paragraphs.

Table 3: Symbol glossary with domain and applicability notes. The glossary is included in the appendix for fast lookup, while all symbols are introduced inline in the main text at first use. Applicability notes clarify when a symbol contributes to theorem-confirmatory interpretation versus boundary or stress analysis.

Symbol	Meaning	Domain/Range	Applicability Note
$G = (V, E)$	phase orchestration graph	finite directed graph	default execution topology
f_v	phase operator at node v	$X_v \to Y_v$ measurable map	requires typed edge admissibility
E_i	local false-accept indicator	$\{0, 1\}$	used in path risk theorem
S_p	path error count	$\mathbb{N}_{\geq 0}$	defined in equation 1
μ_p, σ_p^2	moments of S_p	$\mathbb{R}_{\geq 0}$	theorem regime needs $\mu_p < 1$
$\mathcal{D}_p$	dependence descriptor	covariance collection	manuscript-specific dependence object
ρ	reroute policy	measurable map to simplex	constrained by equation 3
$J(\rho)$	finite-horizon objective	$\mathbb{R}_{\geq 0}$	minimized over feasible set
Φ_t	unresolved potential	$\mathbb{R}_{\geq 0}$	drift argument variable
$R_\lambda(\pi)$	minimax robustness risk	$[0, 1]$	policy-class-dependent boundary

Although compact, the glossary also encodes non-applicability information. For example, μ_p and σ_p^2 remain well-defined outside the confirmatory regime, but conclusions tied to equation 8 do not transfer automatically when $\mu_p \geq 1$. Likewise, $R_\lambda(\pi)$ is defined for broad policy classes, yet the lower-bound theorem in this manuscript applies to a static-threshold class under specified perturbation assumptions. Stating these boundaries near the symbol table reduces interpretive drift and keeps theorem statements aligned with their declared domains.

C Extended Theorem-Audit Diagnostics

This section provides a denser view of theorem-audit outcomes than the main narrative can carry. It includes detailed check-level interpretation, mixed-outcome handling, and structured counterexample meaning. The purpose is methodological transparency rather than additional claim expansion. By exposing where specific checks fail and why, the section supports correct scientific usage of mixed evidence and helps downstream reruns target the right closure tasks. It also demonstrates how the claim-support taxonomy is mechanically connected to theorem checks instead of editorial preference. This makes the evidence trail robust under independent replication. The diagnostics are organized to separate algebraic closure, regime satisfaction, and boundary stress behavior. Algebraic closure addresses whether symbolic identities and implication steps are internally consistent. Regime satisfaction addresses whether executed slices meet theorem premises. Boundary stress behavior addresses whether expected failure modes appear under deliberate assumption violations. This separation is central to fair interpretation: a theorem may remain mathematically sound while lacking confirmatory empirical support in the current slice distribution. The table and surrounding prose are designed to preserve that distinction.

C.1 Detailed Theorem Check Table

Table 4 expands theorem-level outcomes beyond the compressed main-text summaries. The objective is to keep the main narrative readable while retaining full audit visibility.

The table shows why the manuscript uses a support taxonomy with a mixed category. A strict pass/fail paper-level label would discard useful information. For example, symbolic closure and unrestricted minimax support remain informative even though restricted corollary closure is incomplete. Conversely, objective dominance evidence should not be over-generalized into full drift closure when its supporting lemma is only partially satisfied empirically.

C.2 Counterexample and Boundary Interpretation

Counterexamples reveal where assumptions fail in operational traces. For the path theorem family, heavy-load slices push μ_p above one, which invalidates the confirmatory reading of equation 8. For reroute dominance, high uncertainty caps and large depth budgets produce oscillatory traces with weaker drift behavior. For robustness, authenticated-

Table 4: Detailed theorem checks with qualitative interpretation. The table separates algebraic closure, implication checks, and regime-sensitive outcomes, which is necessary when symbolic checks pass but empirical theorem closure is mixed. A mixed or false status here is treated as a first-class scientific finding and is propagated into claim support labels.

Theorem ID	Result	Interpretation
Path composability lemma	Yes	Typed composability is consistent with finite path construction and schema checks.
Cantelli positivity lemma	Yes	Denominator positivity condition is satisfied on valid symbolic domains.
Path soundness theorem	No	Empirical bound coverage is weak in current slices because confirmatory regime conditions are not met.
Drift telescope lemma	No	Drift condition does not hold on all calibrated slices; closure is partial.
Dominance theorem	Yes	Objective dominance over memory-only baseline is observed on evaluated horizons.
Oscillation proposition	No	No oscillation witness was observed in the executed run; targeted stress slices are still required for closure of this proposition.
Minimax boundary theorem	Yes	Unrestricted perturbation behavior remains near the lower-bound envelope.
Restricted attainability corollary	No	Restricted-channel low-risk condition does not hold uniformly in executed slices.

channel assumptions are violated in a subset of slices where provenance separation is insufficient. These outcomes justify targeted reruns with constrained generators and stronger channel controls rather than broad theorem rejection.

From a methodological viewpoint, this pattern mirrors established advice in empirical formalism studies: treat assumption violations as structured evidence, not as narrative noise (Theologitis & Suciu, 2025; Haseeb, 2025; Li et al., 2023b; Karl et al., 2022; Toghi et al., 2021). The manuscript therefore keeps explicit caveat links from each claim to its failure mode and to a concrete follow-up experiment class. A useful interpretation rule is to classify counterexamples by which premise failed first. If failure starts at type admissibility, the appropriate response is schema and interface correction before probabilistic analysis. If failure starts at moment or drift assumptions, the response is distributional recalibration or policy-budget adjustment. If failure starts at perturbation-family mismatch, the response is threat-model narrowing or stronger channel authentication. This premise-first taxonomy avoids conflating fundamentally different breakdown mechanisms and yields a cleaner rerun agenda for the next validation cycle.

D Reproducibility Protocol and Operational Context

This section records execution details needed to reproduce the theorem-audit pipeline in a controlled setting. It separates reproducibility context from contribution claims by placing budget, sweep, and manifest considerations in the appendix. The section also formalizes practical guardrails for reruns: preserve regime tagging, keep assumption checks executable, and avoid mixing confirmatory and stress slices in the same support denominator. These conditions are essential for maintaining the interpretation integrity of theorem-linked evidence over repeated iterations. They also define minimum reporting obligations for future revision phases. Reproducibility here is defined as semantic replay, not only command replay. Command replay is necessary but insufficient if regime tags, assumption predicates, or claim-link assembly rules drift between runs. For that reason, this appendix emphasizes stable definitions for support categories and explicit mapping from theorem obligations to executable checks. The objective is that an independent rerun can regenerate the same support semantics from primary outputs without ad hoc interpretation adjustments.

D.1 Execution Budget, Seeds, and Sweep Design

The executed validation used CPU-only resources with explicit preflight and full stages. The preflight stage validated imports, config parsing, first-batch execution, and expected artifact writes. Full runs covered finite sweeps for path length, covariance regime, calibration cap, depth limit, horizon, and perturbation family. Seed sets were fixed per experiment family to support paired comparison and interval estimation. Uncertainty reporting used bootstrap intervals, Wilson intervals, and delta-method approximations depending on metric type.

Operational constraints are reported here, not in the title or abstract, because they describe reproducibility context rather than scientific novelty. This separation avoids conflating deployment budget with theoretical contribution. The key reproducibility principle is that every theorem-level claim should map to a machine-auditable check with explicit regime tags. Budget-aware design also affects inferential strength. Limited sweeps can still provide useful theorem diagnostics when the design prioritizes premise coverage over broad task heterogeneity. In this run, sweep factors were selected to stress each theorem boundary directly: covariance and path load for soundness, depth and calibration for dominance, and perturbation family with margin settings for minimax robustness. This targeted allocation is appropriate for proof-first validation, but it must be reported explicitly so readers do not misread selective depth as comprehensive empirical coverage.

D.2 Manifest and Data Resolution Notes

The run includes a machine-readable execution manifest, symbolic check records, and dataset-resolution notes. Some external benchmark families remain candidate-level and were not fully materialized in the executed run. The paper therefore treats those families as exploratory context and anchors confirmatory interpretation to materialized slices. This choice is conservative and aligns with the claim-evidence policy used throughout the manuscript.

For readers reproducing or extending the study, the most important requirements are: preserve typed contract schemas, enforce declared policy constraints, retain regime tags in every slice, and separate theorem-confirmatory counts from out-of-regime stress tests. Without this separation, apparent improvements can reflect sampling artifacts rather than theorem support. A second reproducibility requirement is provenance completeness for any newly materialized benchmark slice. If additional datasets are introduced in future iterations, each slice should include source identity, version or commit metadata, license status, and checksum-level integrity notes before its outputs are used in claim summaries. This requirement prevents silent source substitution and keeps theorem-evidence interpretations stable across reruns with expanded data coverage.

D.3 Canonical Audit Artifacts and Cross-Format Consistency

Table 4 gives the prose-level interpretation of theorem-audit outcomes, and companion structured exports retain the same evidence in machine-readable form. Keeping both views is important: the narrative table explains why a theorem status is supportive, mixed, or unsupported, while the structured tables preserve exact status fields, comparator mappings, and audit metadata for downstream re-analysis. This dual representation prevents transcription drift during revision and makes it straightforward to reproduce claim-support labels directly from audit records. The key requirement is semantic consistency between manuscript statements and structured evidence rather than dependence on any one display format. The canonical machine-readable theorem-audit exports include a theorem-summary table, a claim-comparator mapping table, and a theorem-check ledger with per-obligation status fields.

The appendix and main text rely on a paired artifact model. Human-readable tables and figures communicate interpretation, while machine-readable exports preserve exact values, status tags, and comparator metadata. Both representations are considered canonical only when they encode the same claim-evidence links and caveat annotations. During revision, this consistency rule prevents drift between narrative summaries and structured audit records. In practice, every theorem-status update must be reflected simultaneously in the narrative interpretation table, the comparator mapping table, and the theorem-check ledger used for automated aggregation.

This cross-format rule also supports independent verification. A reader can first inspect the manuscript-facing displays to understand the scientific argument, then reproduce support labels from the structured exports without reinterpreting prose. If the two views disagree, the discrepancy is treated as a reporting defect, not as an acceptable ambiguity. By defining canonical status through semantic agreement rather than through file location, the manuscript keeps reproducibility requirements explicit while avoiding coupling scientific interpretation to implementation-specific storage details.

E Additional Related-Work Synthesis for Breadth

To keep the main text focused on formal claims, this appendix section records broader literature links that informed scope decisions. Debate- and collaboration-based protocols can improve critique diversity but may also enlarge interaction attack surfaces and coordination overhead (Kargupta et al., 2025; Li et al., 2025; Rothfarb et al., 2025; Xia et al., 2025). Agent-computer interface studies show that interface constraints can dominate autonomy quality and safety outcomes, reinforcing the need for explicit contract boundaries (Yang et al., 2024; Papadimitriou et al., 2023). Neural-symbolic synthesis papers show pathways to stronger reasoning structure but often leave execution-layer con-

tract semantics implicit (Su et al., 2025; Dinu et al., 2024; Jiang et al., 2024; Zhang et al., 2020; Efremov et al., 2018).

We also considered repository-based ecosystem evidence from widely used frameworks (Maintainers, 2026d;c;b;a). Repository evidence is valuable for implementation realism but weaker as standalone scientific proof unless tied to commit-pinned reproducible experiments. This is why repository citations in the manuscript support contextual claims about design space, while theorem and validation claims are tied to explicit formal statements and executed audit artifacts. An additional breadth observation concerns evaluation philosophy. Several framework papers prioritize aggregate success metrics under broad task suites, while formal-methods literature prioritizes premise transparency and non-applicability articulation. Our manuscript adopts the latter standard for claim acceptance and uses framework evidence mainly to motivate design relevance and stress regimes. This hybrid stance aligns practical orchestration concerns with theorem-level accountability and helps avoid overgeneralizing from benchmark averages to formal guarantees.