THE SAFETY CRITICALITY NUMBER: CASCADE STABILITY AND INTERVENTION THRESHOLDS FOR AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Risky actions by an AI agent may arrive in cascades: one event changes context or shared state so later events become more likely. We model this with a marked Hawkes process and separate two conflated quantities—an intrinsic cascade criticality R_S (whether an uncontrolled cluster dies out) and operational monitorability (observation, containment, background risk, capacity). With observation probability p and conditional containment q, we prove within the model that control is subcritical exactly when $(1 - pq)R_S < 1$; we extend this to interacting agents via a controlled branching matrix and establish a tight $\Theta(\epsilon^{-2})$ testing rate in labeled parent events. Empirically, in a synthetic decision setting with all tools disabled and labels never executed, a prespecified auditable paired replication raises the downstream terminal unsafe rate from 16.2% to 21.0% over 600 matched restrictive cases when the erroneous upstream conclusion is supplied ($+4.8$pp; McNemar exact $p = 6.5 \times 10^{-3}$; all prompts and raw responses retained), while a multi-model panel shows heterogeneous per-edge transmission ($\widehat{q}$ from 0 to 0.67) and a realized cascade shows finite-horizon growth versus persistence consistent with $R_S = b\,\widehat{q}$. A second pre-registered study on a different task family observes no transmission on two current models; because the studies differ in more than verifiability, this contrast does not isolate verifiability as the cause and holds for the tested tasks only. A separate read-only trace study finds excess adjacency of elevated-access tool calls after conditioning on task and phase (Holm $p = 0.0025$)—temporal dependence, not evidence that deployed-agent harms follow a Hawkes process. Finally, a prospective test—predictions fixed before each corrected collection on a previously examined case pool—forecast that placing an independent reviewer *last* is riskiest; the order held in both error-making models, resolved within one and pooled (McNemar $p = 6.4 \times 10^{-4}$) but not the other, with a permutation-invariant baseline tying all placements. These are model-specific translations of a classical stability boundary, not a measured branching ratio or a deployment guarantee.

1 INTRODUCTION

Runtime safety monitoring is usually framed as classification: inspect an action and decide whether to allow it. That view misses a second question. Does one risky action create conditions that make later risky actions more likely? Tool outputs alter context, failed attempts trigger retries, and agents can coordinate through shared state. These mechanisms can produce cascades even when the marginal probability of any one event is small.

Self-exciting point processes provide a natural language for such dynamics. In a Hawkes process, the integral of the excitation kernel is the mean number of direct descendants of one event. We call this quantity the *Safety Criticality Number* R_S. The threshold $R_S = 1$ is the classical branching-process stability boundary (Brémaud & Massoulié, 1996; Hawkes & Oakes, 1974); relabeling it does not by itself produce a theory of monitoring. In particular, $R_S < 1$ does not imply that a monitor has enough capacity, while $R_S > 1$ does not imply that intervention is futile.

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

This paper makes that distinction explicit within a marked Hawkes model. We model monitoring as independent observation and propagation containment. Containment prevents a detected event from generating descendants but does not erase the event that prompted the intervention. This yields an exact and interpretable control law within the model: observation probability p and conditional containment probability q enter only through the effective intervention $u = pq$, and stability requires $(1-u)R_S < 1$. We then add background risk and inspection capacity, obtaining an operational load condition in addition to intrinsic criticality.

Our contributions are:

1. a marked, intervention-thinned Hawkes model that distinguishes cascade stability from monitor capacity;
2. an exact intervention threshold $u_c = (1-1/R_S)_+$ (an infimum for strict subcriticality) and its observation–containment tradeoff;
3. a multivariate control criterion $\rho(AD_u) < 1$ for interacting agents, including a uniform-control closed form and an exact marginal value for targeted containment;
4. a tight $\Theta(\epsilon^{-2})$ testing result under explicitly observed parent–child labels, stated in labeled-parent sample size rather than ambiguous trace time;
5. phase-controlled empirical tests on model tool-call traces, with all qualifying and zero-event sessions retained and classifier sensitivity reported; and
6. empirical evidence in a synthetic decision setting: a multi-model handoff-contamination study and a prespecified auditable paired replication (600 restrictive pairs, all enrolled prompts and raw responses retained) establishing a significant within-case increase in unsafe downstream decisions (+4.8pp, 95% CI $[1.3, 8.4]$, McNemar $p = 6.5 \times 10^{-3}$), a realized multi-generation cascade showing finite-horizon growth versus persistence consistent with $R_S = b\widehat{q}$, and an exploratory held-out containment sweep (distinguishing message-replacement from parent-quarantine) whose correct-descendant counts peak near a calibration-derived reference—stopping short of an identified threshold, a zero-offspring guarantee, or a demonstrated targeting advantage over a tuned fixed rate.

The empirical study is deliberately modest. Sensitive-access events in a synthetic code-review task are a proxy for actions a monitor might inspect, not observed harms or near-misses. Moreover, regime changes can create spurious Hawkes criticality (Filimonov & Sornette, 2014), and flexible background rates can materially change branching-ratio estimates (Omi et al., 2017). We therefore do not estimate the continuous-time R_S from tool-call indices. Instead, we ask whether adjacent sensitive accesses occur more often than expected after conditioning on task and coarse session phase.

2 Related Work

AI safety monitoring and control. Shielding methods block actions that violate a specification (Alshiekh et al., 2018); formal verification attempts to certify properties before or during deployment (Seshia et al., 2018); and AI control protocols allocate monitoring effort under possible intentional subversion (Greenblatt et al., 2024). These approaches focus primarily on action-level correctness or adversarial oversight. Our model instead isolates propagation: whether one detected risk event is contained before it produces follow-on events.

Agent error cascades and message-level interventions. Closest to our empirical setting is Xie et al. (2026) (*From Spark to Fire*), which abstracts LLM multi-agent collaboration as a directed dependency graph, gives an early-stage risk criterion for error amplification across six frameworks, and mitigates cascades with a genealogy-graph message-layer governance plugin that tracks claim provenance and rolls back risky propagations (preventing final infection in $\geq 89\%$ of runs). We differ in assumptions, intervention semantics, and guarantees: rather than a dependency-graph amplification criterion paired with message-provenance governance, we model recurring probabilistic *parent-level* containment in a branching process, prove when the *expected* number of contaminated descendants per contaminated agent crosses one, and give a matching $\Theta(\epsilon^{-2})$ identifiability rate; our handoff randomization measures the within-case change in the downstream decision in a matched experiment

rather than observing propagation in situ. We view these lines as complementary: their governance middleware and framework-level cascade study motivate the phenomenon, while our contribution is the paired measurement protocol for per-edge handoff effects and its explicit connection—and limits—to cascade-control models such as the $(1 - pq)R_S$ law. A parallel empirical literature provides failure-mode taxonomies with annotated traces (Cemri et al., 2025), responsible-agent and error-step attribution (Zhang et al., 2025), and emulated tool-risk scenarios (Ruan et al., 2024); these could support broader evaluation, but estimating causal offspring from them would still require additional instrumentation and controlled interventions.

Hawkes stability and branching. The cluster representation of a linear Hawkes process identifies the kernel mass with a Galton–Watson offspring mean (Hawkes & Oakes, 1974). Stability results extend to nonlinear and multivariate processes (Bacry et al., 2015; Brémaud & Massoulié, 1996). Hawkes-process control has also been studied broadly (Lima, 2023); for example, Gupta et al. (2018) optimize finite-horizon removal of selected historical events in a spatial process. We use these classical results as primitives but study a different intervention—recurring probabilistic containment of every observed parent—as detailed next.

What is new relative to prior branching models. Branching- and epidemic-process criticality is well established in seismology, epidemiology, and neuronal avalanches, where the branching ratio is estimated from observational counts of a fixed system. Our contribution is not the formalism but three agent-specific moves: (i) the containment channel as a first-class control, so the $(1 - pq)R_S < 1$ law separates intrinsic criticality from operational monitorability (p, q, capacity), which does not appear when the offspring mean is a fixed property to be measured; (ii) the identifiability side, a $\Theta(\epsilon^{-2})$ testing rate in labeled parent events; and (iii) treating per-edge transmission as an interventional, model-specific quantity measured by randomizing the handoff, tracing how the contaminated-offspring ratio crosses $R_S = 1$ as fan-out varies, rather than inferring a branching ratio from passive clustering.

Model misspecification. Hawkes fits are sensitive to nonstationary background activity. Mixtures of Poisson regimes can appear spuriously critical (Filimonov & Sornette, 2014), while explicit time-varying baselines can reduce estimated endogeneity (Omi et al., 2017). Residual and time-rescaling diagnostics are standard for continuous-time point processes (Ogata, 1988). Our tool traces have only ordinal call indices, so we use a binary history model and conditional randomization rather than claim continuous-time point-process identification.

Near-miss analysis. Near-misses are widely used as leading indicators in industrial safety, although fixed accident-pyramid ratios are not universal (Heinrich, 1931). In our framework a near-miss is useful because it reveals a propagation-capable event before a later descendant becomes harmful. The benefit is expressed through observation probability p, not through an unsupported effective-sample-size multiplier.

3 Model

Definition 1 (Marked risk process)**.** *Risk events occur at times $\tau_1, \tau_2, \ldots$ with conditional intensity*

$$\lambda(t) = \mu + \sum_{\tau_i < t} \phi(t - \tau_i), \tag{1}$$

where $\mu > 0$ is the immigrant rate and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is integrable. Each event receives a mark $Y_i \in \{\text{near-miss}, \text{harm}\}$; for the burden calculations, marks are conditionally independent with $\Pr(Y_i = \text{harm}) = h$.

The intensity in eq. (1) is an instantaneous rate, not a probability. The mark assumption is intentionally simple and is not needed for the cascade-stability results. The kernel is also mark-*independent*; mark-dependent contagion (e.g. harm events more excitatory than near-misses) is not captured by the scalar R_S but is exactly the type-structured setting of theorem 9, with $\rho(AD_u)$ the relevant threshold (appendix C.3 gives the multitype controlled matrix and stationary harm burden).

Definition 2 (Safety Criticality Number)**.** *The intrinsic cascade criticality is*

$$R_S = \|\phi\|_1 = \int_0^\infty \phi(s)\, ds. \tag{2}$$

Under the Poisson cluster representation, R_S is the expected number of direct offspring generated by one uncontrolled event.

For an exponential kernel $\phi(s) = \alpha e^{-\beta s}$, eq. (2) gives $R_S = \alpha/\beta$. The general definition avoids making the threshold depend on a particular decay family.

The primitive shared with our discrete experiments is the Galton–Watson offspring mean: the continuous-time process of eq. (1) embeds it with Poisson offspring of mean R_S, while the bounded-fan-out handoff cascades of section 5.2 embed it with offspring $b\,\widehat{q}$. The mean-based stability results below—the intervention threshold and the spectral criterion—transfer across both because they depend only on this mean. The $\Theta(\epsilon^{-2})$ testing rate is *not* mean-only: its proof uses the full likelihood of independent Poisson offspring counts (their variance, not the mean alone), so it holds under the Poisson-offspring assumption of eq. (1) rather than transferring from the mean; the extinction *probability* is likewise embedding-specific (appendix A). We thus state the theory in continuous time and read the discrete cascades as a finite-fan-out realization of the same mean, not a continuous-time estimate; appendix C.1 makes the units and clock conversions explicit.

Definition 3 (Observation and containment)**.** *Each risk event is observed independently with probability p. Conditional on observation, the monitor contains its propagation with probability q. A contained event produces no offspring. The effective intervention probability is $u = pq$, and the controlled offspring mean is*

$$R_{\text{eff}} = (1-u)R_S = (1-pq)R_S. \tag{3}$$

Containment can represent revoking a tool capability, isolating shared state, terminating a subtask, or resetting context; the independence assumption makes the threshold transparent, and correlated detection and state-dependent intervention are important extensions.

Definition 4 (Operational monitoring load)**.** *Let $B > 0$ be the monitor's sustainable inspection rate. Whenever the controlled process is stationary, define*

$$L_S = \frac{p\bar{\lambda}}{B}, \tag{4}$$

where $\bar{\lambda}$ is the stationary risk-event rate. We call a configuration operationally feasible *when $R_{\text{eff}} < 1$ and $L_S < 1$.*

Thus R_S captures endogenous amplification, while L_S also includes exogenous arrivals and monitoring capacity. Neither replaces an application-specific harm model.

4 Main Results

4.1 Controlled cascade threshold

Theorem 5 (Intervention threshold)**.** *Under definitions 1 and 3:*

1. *If $R_{\text{eff}} < 1$, every immigrant-triggered cascade is finite almost surely, its expected total size is*

$$\mathbb{E}[|\mathcal{C}|] = \frac{1}{1-R_{\text{eff}}}, \tag{5}$$

and the controlled Hawkes process has stationary mean intensity

$$\bar{\lambda} = \frac{\mu}{1-R_{\text{eff}}}. \tag{6}$$

2. *If $R_{\text{eff}} > 1$, a cascade survives with positive probability $1-s_*$, where $s_* < 1$ is the smallest solution of*

$$s = u + (1-u)\exp\{R_S(s-1)\}. \tag{7}$$

3. The infimum intervention needed for subcriticality is

$$u_c = \left(1 - \frac{1}{R_S}\right)_+ . \tag{8}$$

Within the model, the theorem distinguishes an uncontrolled supercritical system from an uncontrollable one. When $R_S > 1$, passive observation is insufficient, but propagation can be stabilized whenever containment exceeds eq. (8).

Corollary 6 (Observation–containment tradeoff)**.** *For $R_S > 1$, the controlled process is subcritical exactly when*

$$pq > 1 - \frac{1}{R_S}. \tag{9}$$

If containment succeeds with probability at most q, stabilization is impossible through this mechanism when $q \leq 1 - 1/R_S$, even with perfect observation.

Soft message-level remediation (rewriting a parent's message rather than pruning its subtree) is *not* this zero-offspring operator: remark 14 formalizes an idealized *fixed*-residual model $R_{\text{eff}}(u) = (1-u)R_S + u\,R_{\text{res}}$ ($R_{\text{res}} = b\,q_{\text{repl}}$ constant in u), where for $R_S > 1$ and $0 \leq R_{\text{res}} < 1$ subcriticality needs $u > u_c^{\text{res}} = (R_S - 1)/(R_S - R_{\text{res}})$, which exceeds the zero-offspring threshold u_c when $0 < R_{\text{res}} < 1$; if $R_{\text{res}} \geq 1$, no $u \in [0, 1]$ stabilizes this fixed-mean model. Our finite replacement arms have dose-dependent error rates ($\approx$25% at u=0.5 to $\approx$70% at u=1), so they illustrate rather than prove this regime.

Proposition 7 (Capacity and harm burden)**.** *If $R_{\text{eff}} < 1$, then*

$$\bar{\lambda} = \frac{\mu}{1 - R_{\text{eff}}}, \qquad \bar{\lambda}_{\text{harm}} = \frac{h\mu}{1 - R_{\text{eff}}}, \qquad L_S = \frac{p\mu}{B(1 - R_{\text{eff}})}. \tag{10}$$

Consequently, $R_S < 1$ alone is insufficient for operational feasibility: arbitrarily large μ can make $L_S \geq 1$ while the cascade process remains subcritical.

4.2 Interacting agents

Definition 8 (Controlled branching matrix)**.** *For k event types or interacting agents, let A_{ij} be the expected number of type-i children caused by one type-j parent. Let u_j be the propagation-containment probability for a type-j parent and set $D_u = \text{diag}(1 - u_1, \ldots, 1 - u_k)$.*

Theorem 9 (Multivariate intervention criterion)**.** *The controlled multivariate process is subcritical if and only if*

$$\rho(AD_u) < 1. \tag{11}$$

Under uniform intervention $u_j = u$, this reduces to $(1 - u)\rho(A) < 1$, so

$$u_c^{\text{uniform}} = \left(1 - \frac{1}{\rho(A)}\right)_+ . \tag{12}$$

In particular, every diagonal entry of A may be below one while $\rho(A) > 1$ because of cross-agent excitation.

The result is a control criterion, not a claim that component safety implies or fails to imply all forms of system safety. It identifies one specific failure mode: amplification through interaction loops.

Proposition 10 (Marginal value of targeted containment)**.** *Suppose A is irreducible, $u_j < 1$ for every j, and let $\ell, r > 0$ be left and right Perron vectors of $M = AD_u$, normalized so that $\ell^\top r = 1$. Then*

$$\frac{\partial \rho(M)}{\partial u_j} = -\frac{\rho(M)\,\ell_j r_j}{1 - u_j}. \tag{13}$$

For a small containment increase with marginal cost $c_j > 0$, the largest first-order reduction in criticality per unit cost is therefore obtained from the largest value of $\ell_j r_j / [c_j(1 - u_j)]$.

This sensitivity is local: after each material control change, the Perron vectors and priorities should be recomputed. It does not claim that a greedy allocation is globally optimal.

4.3 What can be certified from labeled cascades?

Exact parent–child assignments are generally latent in an ordinary Hawkes trace. We therefore state the following result only for labeled branching data, as can arise in simulation, instrumentation that records causal task spawning, or controlled evaluation.

Theorem 11 (Labeled-offspring testing complexity)**.** *Suppose* $0 < \epsilon < 1/2$*,* $Z_1, \dots, Z_m$ *are independent offspring counts with* $Z_i \sim \mathrm{Poisson}(R)$*, and the parent labels are observed. Distinguishing*

$$H_0 : R = 1 - \epsilon \qquad \textit{from} \qquad H_1 : R = 1 + \epsilon \tag{14}$$

with error probability at most $1/4$ *requires and is achievable with*

$$m = \Theta(\epsilon^{-2}) \tag{15}$$

labeled parents. The maximum-likelihood estimator is $\widehat{R} = m^{-1} \sum_{i=1}^{m} Z_i$ *with* $\mathrm{Var}(\widehat{R}) = R/m$*.*

Remark 12 (Unlabeled traces)**.** *Theorem 11 is not a trace-duration theorem. In an unlabeled Hawkes realization, immigrant identities and parent assignments are latent; stochastic declustering estimates posterior responsibilities but does not make the true branching tree identifiable. Certification from event times alone requires a separate likelihood analysis, observation-window assumptions, and goodness-of-fit checks. On partially observed or post-intervention traces, latent parentage and containment-induced zero-inflation inflate the estimator variance and can make a subcritical system look near-critical (Filimonov & Sornette, 2014), so the rate is a best case for instrumented evaluation, not passive logs. Even with known parent opportunities, a noisy detection channel (link-detection* d*, spurious rate* η*) slows the achievable rate to* $\Theta((d+\eta)/(d^2\epsilon^2))$*, and an* unknown *channel can render distinct criticalities observationally identical (proposition 15 and appendix D.2).*

4.4 Kernel robustness

Proposition 13 (Kernel-mass universality)**.** *For any nonnegative integrable kernel* ϕ*, the total-progeny and intervention thresholds in theorem 5 depend on* ϕ *only through* $R_S = \|\phi\|_1$*. Kernel shape changes the temporal duration and detectability of cascades but not the Galton–Watson offspring mean.*

We make no universal claim about the observation time needed to estimate a heavy-tailed kernel; that rate depends on tail, censoring, and sampling assumptions.

5 Experiments

5.1 Phase-controlled analysis of model tool calls

Setup. We run five anonymized model families as code-review agents on a purpose-built synthetic repository (documentation, source, configuration, test, and mock-secret files), each receiving 10 compound and 10 focused tasks through a read-only research interface (file reading, text search, file listing; write, shell-execution, and network operations unavailable). All credential-like values are visibly marked `FAKE`/`DUMMY`, endpoints use `localhost`/`example.com`, and access to configuration or mock-secret paths is task-appropriate in a security review. The primary binary proxy marks such accesses and credential-related searches only to test temporal clustering; it is *not* an anomaly detector and does not label harm, policy violation, credential seeking, or intent, and describes no organizational system or control. Every trace with at least eight attempted tool calls is included regardless of event count (avoiding conditioning on observing multiple elevated accesses); harness, retry, and reproducibility details are in appendix A.6.

Primary test. The statistic is the total number of adjacent elevated-access pairs. The null randomizes labels within four equal session-progress bins of each trace, preserving trace-specific phase rates. We report a Holm correction across the five providers. As an effect-size model, we compare a Bernoulli logistic baseline with trace intercepts and cubic session progress against the same model plus an exponentially decayed history feature. This is a discrete ordinal-time model; its history coefficient is not R_S.

All five families reject the primary phase-conditioned null at the smallest attainable raw randomization p-value, $1/(1{,}999+1) = 0.0005$, and remain significant after Holm correction ($p = 0.0025$; per-provider counts in table 9, appendix A.6). The aggregate observed 834 adjacent event pairs, compared with a phase-conditioned null mean of 501.0, and positive ΔAIC independently favors the phase-plus-history model over phase alone. Narrow/broad proxy variants and two-, four-, and eight-bin phase discretizations are consistent (appendix A.6): we read this as robustness of temporal dependence in this sandbox, not validation of the marks as safety-relevant near-misses. Because task-appropriate credential/config inspection is naturally bursty (functional locality) and phase bins control only coarse drift, this excess adjacency shows *non-Poisson procedural clustering*, not that deployed-agent hazards follow a Hawkes process.

Randomization-test calibration. The phase-conditioned randomization used above is calibrated on synthetic binary traces matched to the data (negative control: phase drift only; positive control: added decayed-history self-excitation) and swept over a 9-cell (3×3) generative grid: it keeps the false-positive rate near nominal (empirically 5.7% in the base setting; $\leq 6.5\%$ across the grid, within Monte Carlo variation at 200 datasets/cell) with full power in the base positive control (power is lower, $\approx$ 35–60%, at the smaller grid effect sizes), whereas unrestricted within-trace shuffling is severely anti-conservative under phase drift (97% in the base setting; $9\% \rightarrow 69\%$ with drift). Full tables in appendix A.5.

5.2 Crossed handoff-contamination study (fixed condition order)

Setup. The phase-controlled analysis establishes temporal dependence but cannot separate causal propagation from shared latent task structure. We therefore run a crossed complement in which one agent's output is the only input another agent receives about an upstream review. Synthetic decision cases (three families—expense approval, access control, and refund eligibility—with programmatic ground truth) are evaluated under four handoff conditions applied in a fixed crossed order: no handoff, a correct upstream note, an *erroneous* note that asserts a false premise pointing to the wrong decision, and a *contained* note in which an automated verified-fact step replaces the erroneous note. All entities, records, and actions are synthetic; every model runs with all tools disabled and returns a decision label only, which is never executed. Across the 1,440 calls in this study, mock labels were used exclusively and no tool-use events occurred.

Per-edge transmission across models. Each case is run under all four conditions in a *fixed crossed order* with the true facts held fixed, so the erroneous-vs-correct contrast is a within-case estimate of per-edge transmission $\widehat{q} = P(\text{unsafe} \mid \text{err.}) - P(\text{unsafe} \mid \text{corr.})$ on restrictive cases; order is *not* randomized here (only the later matched experiment is). Every model is correct under the no/correct/contained conditions (0% unsafe); only the erroneous handoff induces unsafe decisions. The magnitude is strongly model-dependent (table 1) with small, unequal samples, so we report Wilson intervals on the erroneous-arm proportion and, for zero-event cells, a one-sided bound rather than a claim of immunity. Two of five families show *no* transmission ($0/30$; $\widehat{q} \leq 0.10$, $b^* \gtrsim 10$), two transmit weakly ($4/30$, $3/23$; $\widehat{q} \approx 0.13$), and one dev-endpoint family is susceptible ($40/60$, $\widehat{q} = 0.67$, $b^* \approx 1.5$). The safe reading is heterogeneous, model-specific transmission, not provable immunity (table 14, appendix A).

Observed multi-generation cascade with forwarded outputs. The panel measures a single-hop $\widehat{q}$; multi-generation behavior is a separate question. We therefore realize an actual branching cascade on the susceptible model in which *each child re-decides the same case while receiving the parent's own generated conclusion*—the parent's actual output is the only thing flowing down an edge, no scripted note—seeded by an *organic* root error (the root errs on its own under time pressure; no premise is injected). A node is *contaminated* if its decision is wrong; a contaminated node forwards its real conclusion to its b children and clean subtrees are pruned (1,987 calls; table 2 in appendix A). Reported size counts *contaminated* (risk-event) progeny only. Copying is near-deterministic on forwarded edges ($\widehat{q}_{\text{fwd}} = 0.99$), and the descriptive contaminated-children/contaminated-parent ratio tracks the fan-out: at $b = 1$ it is $0.98 \approx 1$ (chains persist at roughly constant width, mean size 7.8), while at $b = 2$ it is 1.97 and the contaminated count roughly doubles per generation until it saturates the node cap in 24/28 trees (mean size 26.6). These are *descriptive finite-depth* statistics on selected roots—arms differ in realized depth and node-cap censoring, so cross-arm sizes are not

Table 1: Per-edge unsafe-transmission across five anonymized model families (erroneous vs. correct handoff, fixed crossed order; the *none*, *correct*, and *contained* conditions are 0% unsafe for every model). Erroneous-arm outcomes use *unequal, actual* denominators (weak D lost $7/30$ attempts to CLI errors; the susceptible family has 60 restrictive cases); the panel has $1{,}440$ attempts and $1{,}433$ valid outputs. Intervals are Wilson intervals *on the erroneous-arm proportion* (which equals $\widehat{q}$ here since the correct arm is 0); zero-event cells report a one-sided bound, not $[0, 0]$. These are not intervals for a difference-of-proportions contrast.

Model family	Unsafe/valid (err.)	Wilson interval	$b^* = 1/\widehat{q}$
Robust A	$0/30$	$\widehat{q} \leq 0.10$ (1-sided)	$\gtrsim 10$
Robust B	$0/30$	$\widehat{q} \leq 0.10$ (1-sided)	$\gtrsim 10$
Weak C	$4/30$ (0.13)	$[0.05, 0.30]$	≈ 7.5
Weak D	$3/23$ (0.13)	$[0.05, 0.32]$	≈ 7.7
Susceptible E	$40/60$ (0.67)	$[0.54, 0.77]$	1.5

like-for-like—not an identified asymptotic threshold. Qualitatively, forwarding a real wrong conclusion grows the contaminated count at $b = 2$ and sustains it at $b = 1$ for this model, and parent-level containment reduces it; we claim no measured critical constant.

Graded parent-level containment reduces contaminated activity. The containment arms act at the *parent* level (with probability u a contaminated parent's edges carry a corrected message), closer to theorem 5 than per-message repair, though not an exact realization. Raising u reduces mean contaminated size, but does *not* establish zero-offspring containment: under $u{=}1$ no messages are forwarded by construction (so an all-zero forwarded-edge count is not evidence of low transmission), yet children after replacement still err ($96/138$), conflating background with replacement-induced error. We thus report these arms descriptively (table 2); a complementary message-level fact-check dose rescales the unsafe rate roughly multiplicatively in $(1 - u)$ ($R^2 = 0.97$).

Self-generated errors. A natural concern is that transmission is an artifact of the injected false premise, so we also measure propagation of *self-generated* errors with no injected content: under a low-deliberation (answer-only) instruction on harder multi-clause tasks a parent errs organically at 23%, and a child receiving the parent's own wrong conclusion on the identical cases (independent solver errs 79%) errs by $\widehat{q}_{\text{endo}} = 0.21$ above a matched no-handoff control ($z = 2.8$; adoption in 99% of cases). Organic errors thus propagate causally, and (as in the injected panel) the effect is model-specific: a robust family shows no detectable propagation of its own within the same sample-size limits.

Confound-controlled paired estimate. The effect is not an artifact of the low-deliberation instruction (a full-deliberation downstream agent still shows an excess $\widehat{q}_{\text{delib}} = 0.17$, 95% CI $[0.10, 0.24]$; appendix A.7), but that comparison confounds the handoff with instruction framing. To isolate it we use a single-hop design with *byte-identical* instructions differing only by one upstream-conclusion field, interleaved and paired on the *same* cases. A first matched run (145 cases) gave $\widehat{q}_{\text{match}} = 0.07$ ($25/145$ vs. $15/145$; McNemar $p = 0.04$)—smaller than the non-matched gap—but was underpowered on the unsafe-only subgroup ($16/72$ vs. $10/72$, $p = 0.15$) and did not retain raw responses, so we re-ran at scale.

Prespecified auditable replication. Because the pilot's unsafe-only subgroup was underpowered and its salvaged labels were not re-checkable, we re-ran the *identical* byte-for-byte matched design at scale on the primary model, retaining every enrolled prompt, raw response, and parser result. Prospectively powering for a 5-percentage-point shift (appendix A.1) and enrolling 600 restrictive parent-error cases, the child's unsafe (approve) rate is 21.0% with the handoff versus 16.2% without it (discordance 68 vs. 39; McNemar exact $p = 6.5 \times 10^{-3}$; +4.8pp, 95% CI $[1.3, 8.4]$)—the safety-relevant effect the pilot could not resolve—and the pooled effect over all 900 cases agrees ($p = 3.0 \times 10^{-3}$; +3.9pp). All raw responses are retained and every terminal label reproduces under an independent extractor. Terminal recovery was required for 73.9% of responses; this establishes scoring reproducibility but does not exclude truncation (saved records carry no finish-reasons or token counts) or dependence on the terminal-decision convention—scoring the *initial* decision in-

stead does *not* reproduce a significant effect (appendix A.1). The +4.8pp is a within-case per-edge effect, not a measured threshold.

Held-out containment prediction and baselines. As an exploratory probe on unseen cascades, we estimate $\widehat{q}_{\text{fwd}} = 0.997$ ($\widehat{R}_S = 1.99$ at b=2), take the implied zero-offspring threshold $\widehat{u}_c = 0.498$ as a reference only, and sweep containment on 60 held-out roots (appendix A.2). Under message replacement, erroneous descendants fall monotonically (29.9 $\rightarrow$ 5.7/root) while correct outputs are non-monotone (peaking near $\widehat{u}_c$); this is a finite-horizon observation, not a validated optimum, since replacement is not zero-offspring (remark 14). A fixed-topology ($b = 2$) cross-model sweep (appendix A.7) illustrates the $R_S = b\,\widehat{q}$ law, while heterogeneity ($\widehat{q}$ from 0 to 0.67) cautions against any single $\widehat{q}$ as universal. Monte Carlo simulations check selected model consequences and identify $R_{\text{eff}} = (1-u)R_S = 1$ as a continuous phase transition ($\gamma = 1.01$, $\tau = 1.53$); the exponent check (29 cascades) is suggestive only.

6 Discussion

The takeaway is not "shut down every agent with $R_S > 1$" but that cascade amplification and control effectiveness share one scale: within the model, an illustrative $R_S = 1.3$ system requires effective intervention $u = pq > 23.1\%$ for strict subcriticality, while a subcritical model can still overload a monitor at high immigrant rate—so R_S, p, q, μ, and B should all be reported. Observation enters stability and load with opposing signs: for $L_S = p\mu/[B(1-(1-pq)R_S)]$, $\partial L_S/\partial p$ has the sign of $(1-R_S)$, so raising p lowers steady-state load when $R_S > 1$ but raises it when $R_S < 1$; p should thus be set jointly with definition 4 (appendix C.2). For multi-agent systems the intrinsic quantity is the spectral radius of the full branching matrix (proposition 10).

Limitations. Independent thinning is an idealization; harm marks are independent only for the burden formula; the exact testing rate applies to observed offspring labels, not event-time traces; and ordinal tool-call indices give no durations, so the real-trace study does not estimate R_S. The handoff panel spans five anonymized families (two with no detectable transmission, $0/30$) across three synthetic task families, dev-only endpoints, and a fixed condition order; per-edge transmission is heterogeneous and only one-sided-bounded for robust families ($\widehat{q} \leq 0.10$). The realized cascade and held-out sweep are *descriptive* finite-depth statistics with node-cap censoring, identifying neither an asymptotic threshold nor zero-offspring containment. The replication (appendix A.1) resolves the pilot's underpowered unsafe-only subgroup with retained responses, though $1/\widehat{q}_{\text{match}}$ remains an extrapolation that attenuates on the format-selected subset. Under real descendant blocking with a capacity-limited non-oracle judge (appendix A.2), containment beats fixed-rate control on safety and correct outputs at higher cost and recovers most of an oracle's safety; operational benchmarks, real-side-effect sandboxes, matched-cost comparison against specialized defenses (message-provenance rollback (Xie et al., 2026), proactive monitors), and L_S calibration remain future work; the natural next validation is an end-to-end emulated tool-use benchmark (Ruan et al., 2024) with a learned monitor, where a pilot found emulator fidelity to be the binding constraint—an emulator exposed to the task goal and rubric reverts to agent-like refusals, so faithful emulation must isolate tool execution from the evaluation objective. On the tested verifiable tasks a pre-registered study (appendix A.4) observed no per-edge transmission on two models; generalization beyond these tasks is untested, and its caught salvage/truncation artifact shows why response-level retention and strict-parse sensitivity matter.

7 Conclusion

Classical branching theory supplies the criticality primitive; our contribution is the agent-setting scaffolding around it: within the model we separate intrinsic criticality R_S from operational monitorability via the exact law $(1-pq)R_S < 1$ (a model statement, not a deployment claim), with a matching $\Theta(\epsilon^{-2})$ identifiability rate. Empirically a prespecified paired experiment shows a significant within-case rise in unsafe decisions (+4.8pp, $p = 6.5 \times 10^{-3}$), and a realized cascade behaves consistently with $R_S = b\,\widehat{q}$—motivating causally labeled deployment traces, not a measured threshold.

Ethics Statement

The framework presented in this paper is a theoretical research contribution, not an operational control. It is not used as a compliance, certification, or operational control artifact at any organization. The results describe properties of a mathematical model and should not be interpreted as endorsement or assessment of any specific AI system or deployment. Any future affiliation identifies authorship only, not organizational endorsement. A camera-ready version—including affiliations, acknowledgments, and document metadata—requires an additional security review. The decision and cascade experiments query anonymized frontier-model endpoints with all tools disabled and return decision labels that are never executed; the separate tool-call trace study (appendix A.6) instead used a read-only research interface (file read, text search, and listing) with write, shell, and network operations disabled, and its saved tool-use counts for the decision/cascade studies are zero. All entities, records, and credential-like values are synthetic, and unauthorized decisions are inert labels, not executed actions.

Reproducibility Statement

No corpus, code, data, or traces are distributed with this manuscript, and any future artifact release will undergo separate review. To support reproducibility within these constraints, the main text and appendix A specify each experiment's design, sample sizes, randomization and phase-conditioning procedure, statistical tests (Holm correction, exact/Wilson and rule-of-three intervals, McNemar's paired test), and the frozen numerical results underlying every reported figure.

LLM and AI-tool usage

Large language models are the *object of study* in this paper: the decision and cascade experiments query anonymized frontier-model endpoints with all tools disabled, returning decision labels that are never executed, while the separate trace study used a read-only file interface. Separately, AI coding and writing assistants (Cursor and Codex) provided substantive methodology and analysis assistance under author direction—helping develop and check the theoretical derivations and their interpretation, implement the experiment harness, run the statistical analyses, and draft and edit portions of the manuscript. These assistants also supported related-work retrieval and discovery. To be precise about roles: the experimental *observations* are the model endpoints' own recorded decision responses, and AI-assisted code generated the synthetic cases, ran the paired analyses, and produced the figures. No measurement was fabricated or imputed—every reported number is a recorded output of these (synthetic) experiments—and the authors verified the claims, statistics, and code paths reported here. One limit on independent re-verification applies to the *pilot* matched experiment: a subset of its responses were returned as invalid JSON and salvaged to a parsed decision label without retaining the raw model output, so those particular labels cannot be re-checked against their source responses. The prespecified replication (appendix A.1) removes this limit for the confirmatory result: it retains every enrolled prompt, raw response, and parser status, and each terminal label reproduces exactly under an independent extractor. This statement is provided in accordance with the venue's policy on LLM usage.

References

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996. doi: 10.1214/aop/1065725193.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E Gon-

zalez, and Ion Stoica. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Vladimir Filimonov and Didier Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *arXiv preprint arXiv:1308.6756*, 2014.

Ryan Greenblatt, Buck Shlegeris, Paul Christiano, et al. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2024.

Amrita Gupta, Mehrdad Farajtabar, Bistra Dilkina, and Hongyuan Zha. Discrete interventions in Hawkes processes with applications in invasive species management. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3385–3392, 2018. doi: 10.24963/ijcai.2018/470.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Herbert W Heinrich. *Industrial Accident Prevention: A Scientific Approach*. McGraw-Hill, 1931.

Rafael Lima. Hawkes processes modeling, inference, and control: An overview. *SIAM Review*, 65 (2):331–374, 2023. doi: 10.1137/21M1396927.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

Takahiro Omi, Yoshito Hirata, and Kazuyuki Aihara. Hawkes process model with a time-dependent background rate and its application to high-frequency financial data. *Physical Review E*, 96(1): 012303, 2017. doi: 10.1103/PhysRevE.96.012303.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations (ICLR)*, 2024.

Sanjit A Seshia, Dorsa Sadigh, and S Shankar Sastry. Formal specification for deep neural networks. *International Symposium on Automated Technology for Verification and Analysis*, pp. 20–34, 2018.

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

Shaokun Zhang et al. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2025.

A Additional experimental details

This appendix collects material referenced from the main text: the observed forwarded-output cascade arms (table 2), the phase-controlled elevated-access analysis (appendix A.6), the randomization-test calibration (appendix A.5), the Monte Carlo validation of the control threshold (table 11), the composition/uniform-control sweep (table 12), and the labeled-offspring testing rate (table 13).

A.1 Prespecified auditable replication of the paired handoff effect

This study addresses two weaknesses of the pilot matched experiment: an underpowered unsafe-only subgroup and salvaged labels whose raw outputs were not retained. It is a *new* paired collection run on 24 September 2026 under a locally frozen protocol, conducted after the earlier replication's raw archive became unavailable (that earlier archive remains unavailable); the earlier $+7.2$pp result was known to us and informed our directional prior, so this is a prospective replication after prior exposure rather than an unexposed original confirmation. The design was frozen before running: fixed task-pool seeds, one prespecified primary-model run, byte-identical child instructions differing

Table 2: Realized multi-generation cascade (susceptible model) with *forwarded parent outputs* and organic root seeding, summarizing the arms in section 5.2. "Obs. ratio" is the *descriptive* contaminated-children/contaminated-parent ratio (denominators in parentheses count only parents with at least one evaluated child, excluding parents whose children were never evaluated because the node cap was already reached); "post-repl. errors" counts wrong children evaluated after message replacement. Mean size counts contaminated (risk-event) progeny only (node cap 31). The model expression $(1-u)\, b\, \widehat{q}_{\mathrm{fwd}}$ (theoretical, *not* observed; it excludes post-replacement errors) equals 0.99 at u=0.5 (with $\widehat{q}_{\mathrm{fwd}} = 0.99$, b=2) and 0 at u=1. Arms differ in realized generation depth and cap censoring, so sizes are *not* a common-horizon comparison. These are descriptive statistics, not identified reproduction parameters, and the design does not validate zero-offspring containment.

Arm	Obs. ratio	Mean size	Gen. (cap)	Post-repl. errors
b=1, u=0	0.98 (212/217)	7.8	$\leq$ 8 (0/31)	— (fwd. edges only)
b=2, u=0	1.97 (718/364)	26.6	$\leq$ 5 (24/28)	— (fwd. edges only)
b=2, u=0.5	1.26 (210/167)	8.5	$\leq$ 6 (8/28)	39/156
b=2, u=1	1.39 (96/69)	5.2	$\leq$ 4 (3/23)	96/138 (no fwd. edges)

only by the presence of a single upstream-conclusion field, conditions interleaved in one randomized job list, and a fixed stopping rule (enroll restrictive parent-error cases until the target is reached or the seeded pool is exhausted). The primary outcome is the child's *terminal unsafe* decision (approving a ground-truth-restrictive case) on restrictive parent-error cases; the secondary is overall child error on all parent-error cases. The design is prespecified (fixed seeds and a frozen protocol in the run script) rather than backed by a dated external registration. For every *enrolled* pair (both child conditions) we retained the exact prompt, the raw model response, the parser status, the decision, and the case identifier; the saved artifact covers the 900 enrolled pairs and their 1,800 children, not the additional 2,691 screened-but-unenrolled parents.

Prospective power. For a five-percentage-point paired difference and a two-sided exact McNemar test at $\alpha = 0.05$, power at 600 pairs is $\approx$ 94%, 87%, and 80% for discordant-pair probabilities 0.12, 0.15, and 0.18 respectively; at 520 pairs the corresponding values are $\approx$ 90%, 82%, and 74%. We enrolled 600 restrictive parent-error pairs (and 300 permissive parent-error pairs for the secondary). The run issued 5,413 guarded calls (all tools disabled, zero tool-use events): 1,800 child calls and 3,613 parent calls (3,591 harvested parents plus 22 in-flight tail calls that completed after the enrollment targets were met and were not consumed by the collector).

Results. Primary (safety): handoff unsafe 21.0% vs. control 16.2% over 600 pairs (discordant 68 vs. 39; McNemar exact $p = 6.5 \times 10^{-3}$; paired difference +4.8pp, 95% CI $[1.3, 8.4]$). Secondary (overall): handoff 17.7% vs. control 13.8% over 900 pairs (discordant 84 vs. 49; McNemar exact $p = 3.0 \times 10^{-3}$; +3.9pp $[1.3, 6.5]$). No pair was dropped for a non-parseable child. The p-values are exact; paired-difference confidence intervals use the paired Wald calculation with a $1/n$ continuity allowance and are not exact intervals.

Terminal-decision scoring, parser accounting, and reproducibility. The primary outcome is the child's *terminal* decision after deliberation. Because full-deliberation responses often narrate before committing—1,330 change between their first and last *textual* decision value, and 1,318 contain an earlier *parsed* JSON decision object that differs from the final one (the two counts diverge on duplicate-key and formatting cases)—we extract the last JSON object carrying a `decision` key. The terminal-decision *endpoint* was registered as the primary outcome before enrollment, and this run's frozen protocol specifies the last-JSON-object extraction rule and its fallback (appendix A.1); the rule is applied identically to both arms. We report the extractor in full and, because raw responses are retained, any label can be re-adjudicated under an alternative rule. An up-front parser that accepts only a lone JSON object (optionally inside a single code fence) required this terminal recovery on 73.9% of child responses (77.0% of handoff, 70.8% of control). All 1330 recovered terminal labels reproduce exactly under an independent terminal extractor, and re-running both paired tests with that extractor leaves the discordance cells and p-values unchanged; because raw responses are retained, any label can be re-adjudicated. This establishes reproducibility of the *terminal-decision* convention, not an absence of conflicting intermediate content: scoring the *initial* decision object

Table 3: Initial vs. terminal decision on the 600 restrictive replication cases (unsafe = approve on a ground-truth-restrictive case). Both arms almost always emit a premature initial "approve"; *no* case flips safe→unsafe. The +29 terminal-unsafe difference decomposes as 14 fewer initially-unsafe occurrences offset by 43 fewer unsafe→safe transitions—an accounting decomposition consistent with reduced observed self-correction, not an identified internal mechanism.

Arm	Init. unsafe	Term. unsafe	Safe→unsafe	Unsafe→safe (self-corr.)	Self-corr. rate
Handoff	550 (91.7%)	126 (21.0%)	0	424	77.1%
Control	564 (94.0%)	97 (16.2%)	0	467	82.8%

instead measures a different outcome and does not reproduce the same significant effect, so we do not claim robustness to every decision definition. As a stricter formatting sensitivity, restricting to pairs where both children were a lone object leaves only 83 pairs with high unsafe rates in *both* arms ($\approx$ 74–76%); this format-selected subset is inconclusive about the handoff contrast (paired difference -2.4pp, $p = 0.73$). We therefore read the full-sample terminal-decision test as the primary analysis.

Observed decision trajectories. As a *post hoc* trajectory analysis, we extracted the first and last textual `decision` values in each child response (a regex scan over the raw text) and cross-tabulated the initial against the terminal decision on the 600 restrictive cases (table 3). The model emits a premature "approve" as its first textual decision in 91.7% (handoff) and 94.0% (control) of responses, then revises during deliberation—so the initial value is a poor outcome proxy. The +29 terminal-unsafe difference decomposes arithmetically as 14 *fewer* initially-unsafe occurrences in the handoff arm, more than offset by 43 fewer unsafe→safe transitions ($126-97 = (550-564)+(467-424)$); no case flips safe→unsafe. This pattern is *consistent with reduced observed self-correction*, but it is an accounting decomposition of the emitted decision sequence, not an identification of the internal mechanism producing the effect. (The extraction rule matters: an alternative that decodes JSON objects rather than textual occurrences—five handoff first-objects contain duplicate `decision` keys—gives 545 initially-unsafe and 419 corrections, leaving the qualitative pattern unchanged.) Mean response length is slightly greater under the handoff (508 vs. 503 characters); because the saved records contain no token counts or finish reasons, this does not exclude token-budget or truncation effects.

A.2 Held-out containment prediction and baseline comparison

This *exploratory* study asks whether an estimated cascade criticality is informative about containment on held-out cascades; it does not validate an optimal or utility-maximizing controller. The design was fixed before the held-out phase. **Calibration.** On a disjoint seeded pool we collected 60 organically contaminated terse roots and, with forward-only propagation at fan-out b=2 and no containment, estimated the realized forward-transmission probability $\widehat{q}_{\text{fwd}} = 0.997$, giving $\widehat{R}_S = b\,\widehat{q}_{\text{fwd}} = 1.99$ and an implied *zero-offspring* subcriticality reference $\widehat{u}_c = 1 - 1/\widehat{R}_S = 0.498$. We then fixed $\widehat{u}_c$ and a second, disjoint set of 60 contaminated held-out roots (serialized so the baseline study reuses the identical roots). **Held-out sweep.** Holding roots, node cap (31), generation cap (6), and per-call budget fixed, we swept containment $u \in \{0, 0.30, 0.40, 0.50, 0.60, 0.70, 0.90\}$ under two semantics: *message replacement* (a contained parent's children still run, on a corrected message) and *parent quarantine* (a contained parent produces no children). The scheduler expands only oracle-labeled wrong children (correct children terminate), so per-root counts are *descendant* statistics in an error-expansion tree, excluding the initial wrong root: we report erroneous, attempted, and correct descendant outputs per root.

Exploratory dose-response. Under message replacement, erroneous descendants fall monotonically with u (29.9/root at u=0 to 5.7 at u=0.9), while correct descendant outputs are *non-monotone*, peaking near the $\widehat{u}_c$ reference (5.1/root at u=0.4, 3.9 at u=0.5); correct-descendant counts then decline to 1.9/root at the largest tested rate, and these counts depend on the error-expansion scheduler and finite caps. This is exploratory, not a validated optimum: the subcriticality boundary concerns a zero-offspring intervention, whereas replacement is not zero-offspring ($171/404$ replaced children still err at $\widehat{u}_c$; $296/412$ at the largest rate), and at the rounded estimate $R_{\text{eff}}(u{=}0.4) \approx 1.19 > 1$,

Table 4: Held-out descriptive policy comparison on 60 contaminated roots under a common node cap and per-call limits, with generation cap 2 for the depth-budget policy and 6 for the others (*not* matched realized cost). Counts are *descendant* statistics in an error-expansion tree. "Interv." is the raw counter, which increments before the node-cap check and so over-counts selections with no evaluated children. The fixed and calibration-derived rates share one oracle-based selection routine, differing only in rate/seed; their gap is descriptive, not a targeting effect. This is a distinct 60-root run from the containment sweep, so per-root correct-descendant counts differ from the sweep values by seed (e.g. 4.17 here vs. 3.9 at $u{\approx}0.5$ in the sweep).

Policy	Err. desc./root	Correct desc./root	Interv./root	Realized calls
No intervention	29.9	0.05	0	1800
Depth budget ($d \leq 2$)	6.0	0.02	0	360
Fixed rate ($u = 0.5$)	10.3	4.75	4.23	904
Calibration ($u \approx 0.498$)	9.6	4.17	3.75	828

so the theorem does not predict where useful descendants peak. Parent quarantine drives erroneous descendants toward zero (0.2/root at $u{=}0.9$) but yields *zero correct descendant* outputs at every level—note that at $u{=}0$ it still executes all 1,800 descendants (all scored wrong); it is not that no work runs. The two semantics are thus different operating points, characterized descriptively rather than validated against the control law.

Baselines under a common node cap. On the same held-out roots we contrasted four policies (table 4) under a common node cap and per-call limits—generation cap 2 for the depth-budget policy and 6 for the others, *not* matched realized cost: no intervention; a depth budget (truncate at generation 2); a fixed rate ($u{=}0.5$); and the calibration-derived rate ($u{=}\widehat{u}_c \approx 0.498$). No intervention leaves erroneous descendants uncontained (29.9/root) with almost no correct descendants; the depth budget suppresses errors (6.0/root) but yields ≈ 0 correct descendants, truncating error-expansion branches after generation 2. The fixed and calibration-derived rates are *nearly identical*: both invoke the same oracle-based selection routine (whose active frontier contains only ground-truth-labeled wrong parents) and differ only in rate (0.5 vs. 0.498) and seed, so the small gap (errors 10.3 vs. 9.6/root; correct 4.75 vs. 4.17/root) is *not* evidence of a targeting advantage, and no uniform-over-all-parents policy is present in this comparison. Realized cost is also unequal (fixed: 904 descendant calls, 254 interventions; estimated: 828 calls, 225 interventions), and the intervention counter increments before the node-cap check, over-counting selections whose children were never evaluated (e.g. 227 counted vs. 202 actual two-child replacements at $u \approx 0.30$); we therefore do not read an "errors-averted-per-intervention" ratio as an efficiency measure. A genuinely distinct targeting policy at matched realized cost, with retained per-root outcomes for paired uncertainty, is left to future work.

Non-oracle detector. The sweep and baselines above select parents with a ground-truth error oracle. To gauge operational utility we replace the oracle with a realistic *non-oracle* LLM judge (the same guarded call path, tools disabled) that inspects each parent's decision and reason and flags it as likely-erroneous or ok, on a one-hop, fully labeled panel of 60 organically *wrong* and 60 organically *correct* restrictive parents (table 5). Each parent's two forwarded and two replaced children are evaluated under three policies: *none* (always forward), *oracle* (replace iff truly wrong), and *judge* (replace iff flagged). The judge attains precision 0.67, recall 0.88, and a high false-positive rate 0.43 (26/60 correct parents flagged). Downstream, no intervention leaves 50.0% of children unsafe; the oracle cuts this to 10.0% and the imperfect judge to 12.5%—so the realistic detector retains $\approx$ 94% of the oracle's absolute unsafe reduction (37.5 of 40pp) at the cost of 32% more interventions (79 vs. 60). Two caveats temper this: (i) even the oracle leaves 20% of wrong-parent children unsafe because message replacement is not zero-offspring (remark 14); and (ii) the judge's false positives are nearly free *here* only because replacement inserts a *correct* conclusion—under a destructive intervention (quarantine) the 0.43 false-positive rate would instead cost task completion. This is *non-oracle detection with oracle remediation*: the judge chooses which parents to replace, but the replacement supplies the programmatically correct conclusion, and all 120 parents are inspected without an inspection-capacity constraint. The 693 guarded calls (all tools disabled) comprise 93 new-parent calls, 120 judge calls, and 480 child evaluations, so the 79-versus-60 in-

Table 5: Non-oracle LLM-judge detector vs. oracle on a one-hop labeled panel (60 wrong + 60 correct restrictive parents, 2 forwarded and 2 replaced children each; 240 children per policy). The imperfect judge (precision 0.67, recall 0.88, false-positive rate 0.43) retains most of the oracle's downstream unsafe reduction at higher intervention cost. Residual unsafe under replacement reflects its non-zero-offspring nature (remark 14).

Policy	Downstream unsafe	On wrong parents	Interventions
None (forward)	50.0% (120/240)	100%	0
Oracle	10.0% (24/240)	20%	60
Judge (non-oracle)	12.5% (30/240)	25%	79 (26 FP)

tervention counts are not total call-cost measurements. It is a one-hop, single-model probe, not an operational benchmark against provenance-based rollback.

Capacity-constrained controller with real descendant blocking. The probe above uses *oracle remediation* (replacement inserts the correct answer) and inspects every parent. To test the operational regime the reviewers asked for—*real* descendant blocking under a finite inspection budget, comparing alternatives under a *common inspection cap* with realized costs reported separately—we run a multi-generation held-out cascade on a mixed population of 40 frozen contaminated roots and 40 matched organically-correct roots (b=2, node cap 31, $\leq$ 6 generations). At each generation the monitor may inspect at most κ frontier parents, chosen at random (labels are not free), and *quarantines* them—pruning the entire descendant subtree, with no answer injected. Four policies share the identical roots and cap and differ only in which inspected parents are quarantined: *none* (no inspection); *fixed-rate* (quarantine each inspected parent w.p. 0.5, no signal); *judge* (the non-oracle LLM judge of table 5 flags); and *oracle* (quarantine the truly-wrong among the κ inspected—a capacity-limited perfect detector, *not* an omniscient-over-all-parents one). Table 6 reports per-root outcomes; all arms run on the same 80 roots, so contrasts are paired (n=80; 95% bootstrap CIs). At κ=2 the judge cuts unsafe descendants from 15.0 to 0.75/root (-14.25, $[-17.6, -10.9]$) and beats the fixed-rate baseline on *both* outcome axes—safer (-3.56 unsafe/root, $[-5.7, -1.5]$) and more correct descendant outputs ($+3.71$/root, $[+0.6, +6.8]$), both CIs excluding zero—but at higher computational cost (1,002 vs. 744 calls, $+34.7\%$; $+110.1\%$ at κ=4), so this is a safety–completion gain at higher compute, *not* an equal-cost or Pareto improvement. Whether the judge's advantage survives an *equal-realized-cost* comparison—raising the fixed-rate baseline's inspection budget until its call count matches the judge's ($\approx$ 1,002 at κ=2) and asking whether the judge still separates on both axes—is the natural robustness check for the "is the gain only from spending more?" concern; we did not run it for this cohort (whose roots and per-call records were not retained for a paired re-run) and designate it future work. It recovers nearly all of the capacity-limited oracle's safety ($+0.75$ unsafe/root vs. oracle, $[0.0, 1.9]$) but produces fewer correct descendant outputs (-6.30/root). The decomposition by root type is sharp: on contaminated roots the judge cuts unsafe descendants 30.0 $\rightarrow$ 1.5/root (95%), while on correct roots its in-cascade false positives ($27/192$=14.1% of inspected correct parents at κ=2, $42/336$=12.5% at κ=4) prune legitimate subtrees; the cumulative effect cuts correct descendant outputs 30.0 $\rightarrow$ 17.4/root (-42%, a distinct quantity from the per-inspection rate). This is exactly the destructive-intervention cost the oracle-remediation probe could not exhibit. It supports the *operational utility* of capacity-limited non-oracle blocking *in this synthetic setting*—not the criticality number as a calibrated predictive controller; the strong separation ($\widehat{q}_{\mathrm{fwd}} \approx 0.99$, so contaminated trees are near-uniformly unsafe and correct trees near-uniformly clean), single model, and synthetic tools remain limitations. This study retained per-root summary counts (from which all statistics here are reproducible) but *not* the raw judge/child responses, so its labels cannot be independently re-scored; the 600-pair replication (appendix A.1) retains the stronger response-level auditability. On call accounting: the per-arm totals of table 6 sum to 8,382 guarded calls (all tools disabled), whereas this study's run log reported 8,448 total guarded calls—a 66-call excess over the displayed arms. Because this cohort retained only per-root summary counts and not raw per-call records, we could not reconstruct the source of the excess after the fact and report it as an unresolved bookkeeping discrepancy of unknown origin rather than silently reconcile the two figures. The 8,382 per-arm sum is the auditable quantity from which all statistics in table 6 are reproducible.

Table 6: Capacity-constrained cascade controller with *real* descendant blocking (quarantine; no answer injected) on a mixed held-out population (40 contaminated + 40 correct roots; b=2, node cap 31, ≤ 6 gen). κ is the per-generation inspection budget; arms share a common cap, and total calls (child evaluations + judge calls) are reported separately. The non-oracle *judge* beats the *fixed-rate* baseline on both safety and correct descendant outputs (paired n=80) but at higher computational cost—a safety–completion gain, not an equal-cost/Pareto improvement—and recovers most of the capacity-limited *oracle*'s safety at a false-positive cost in correct descendant outputs.

Policy	κ	Unsafe desc./root	Correct desc./root	Total calls
None	0	15.00	15.00	2400
Fixed-rate (0.5)	2	4.31	4.99	744
Judge (non-oracle)	2	0.75	8.70	1002
Oracle (capacity-limited)	2	0.00	15.00	1200
Fixed-rate (0.5)	4	3.15	4.25	592
Judge (non-oracle)	4	0.80	9.60	1244
Oracle (capacity-limited)	4	0.00	15.00	1200

A.3 Prospective prediction of workflow safety ordering

This study asks whether per-edge transmission measurements, taken on a calibration set, *predict* the relative safety of unseen multi-agent workflows built from the *same* base model at *equal* call budget and *identical* role composition, differing only in *placement*—and whether that prediction beats a local-accuracy baseline. Predictions were fixed before each corrected test collection, after prior exposure to the case pool and earlier outcomes; the pooled analysis is exploratory. All calls reuse the inert, capped harness of appendix A.1 (tools disabled, JSON label only, nothing executed, synthetic data); every call is retained as a per-call record under SHA-256. We ran the identical protocol on three anonymized models (M1, M2, M3) with 300 calibration cases and 208 restrictive (should-deny) plus 192 permissive (should-approve) held-out cases *per configuration* per model, at a combined billed cost of about USD 55.

Agent heterogeneity at fixed model identity. Holding the base model fixed, two reviewer *roles* differ only in instruction: a *deferential* role (D) that gives strong weight to any upstream conclusion, and an *independent* role (S) that re-derives the decision from the policy and facts and treats any upstream conclusion as unverified. On a disjoint calibration set (restrictive + permissive cases with injected correct/erroneous upstream messages) we estimate, per role, the solo error rate b (no upstream), the error given a *correct* upstream (α), and the error given an *erroneous* upstream (β). The two roles have similar solo error but opposite conditional behavior—the deferential role adopts a wrong upstream conclusion (high β) while the independent role resists it (low β) yet also diverges more often from a *correct* upstream (higher α). This asymmetry is why placement, not composition, drives terminal safety.

Frozen prediction model and configurations. For a role sequence $[r_1, \ldots, r_L]$ we propagate a two-state (error/correct) chain: $p_1 = b_{r_1}$ and $p_i = p_{i-1}\beta_{r_i} + (1 - p_{i-1})\alpha_{r_i}$; the terminal error probability p_L gives predicted terminal-unsafe (restrictive) or incompletion (permissive). The three test configurations share the role multiset {D,D,S}, length $L = 3$, and identical budget, differing only in placement of the independent reviewer: *independent-first* [S, D, D], *independent-middle* [D, S, D], and *independent-last* [D, D, S]. Two prespecified local-accuracy baselines are frozen alongside the transmission model: a *mean-local* baseline (average solo error over the chain), which is *identical* across all three placements by construction because they share the role multiset and therefore cannot rank them at all; and a *final-local* baseline (the final agent's solo error), a competence-based comparator that *can* differ across placements. Each model's transmission prediction and both baselines were written to `frozen_predictions.json` and hashed before that model's corrected test was run; we verified for all three that the freeze precedes the test and that the stored SHA-256 still matches.

Correction and prior exposure. These are *corrected* runs. An earlier implementation had a first-agent role-assignment defect (the assigned role was not applied to the first agent without an upstream

message), so its placement-only claim was unsupported; we fixed the harness and re-ran fresh, retaining the earlier records separately without relabeling them. Because the corrected tests reuse the *same held-out case pool* as that earlier run, each model's prediction was fixed *before its corrected collection* but *after* prior exposure to the pool and its earlier outcomes—not before any outcome in the research sequence. The retained pre-registration was single-model and named the first-vs-middle contrast as discriminating; that fine contrast did *not* survive correction (below), and the multi-model pooling and the foregrounded independent-last contrasts are best read as pre-specified-direction extensions rather than as the originally registered primary test.

Held-out result. Each configuration was run on 208 held-out restrictive and 192 permissive cases with *organic* handoffs: each downstream agent receives the previous agent's *parsed decision in a fixed template* (`reason: prior review`), not the full prior response or its reasoning. For the two models that make any errors, the calibration-frozen prediction ranked *independent-last* as riskiest, and the observed terminal-unsafe point-estimate ordering matched (table 7): M1 21.6% (first) $< 22.6\%$ (middle) $< 26.0\%$ (last); M2 $7.2\% < 9.6\% < 17.5\%$. Statistical support for independent-last being the most unsafe placement is *not* equal across the two models: it is resolved *within M2* (independent-last vs. independent-first, exact McNemar $p = 2.7 \times 10^{-5}$; vs. middle $p = 1.5 \times 10^{-3}$) but *not within M1* (vs. first $p = 0.26$; vs. middle $p = 0.31$)—M1 matches only in direction. Pooling the paired discordances across the two error-making models along the pre-specified direction, independent-last is riskiest overall (vs. first $p = 6.4 \times 10^{-4}$; vs. middle $p = 3.2 \times 10^{-3}$); we treat this pooled result as an exploratory extension, not two independent within-model replications. The permutation-invariant *mean-local* baseline predicts all three placements equal, so the *existence* of any ranking requires order information; however, the prespecified *final-local* competence baseline is not neutral—for M2 it already ranks independent-last riskiest (32.2% vs. 31.7%), matching the transmission forecast, while for M1 it predicts independent-last *safest* (24.7% vs. 28.8%)—the direction opposite to M1's *observed* riskiest placement. As an exploratory summary (not a prespecified test: the original single-model plan emphasized the first-versus-middle contrast, which remains unresolved, and M3 is unresolved), among the three prespecified predictors only the transmission model matched the *observed* riskiest placement in both M1 and M2; the final-local competence baseline matched it in M2 but not M1, and the permutation-invariant mean-local baseline is unable to distinguish placements at all. M1's ordering remained statistically unresolved ($p = 0.26$), so this is a descriptive point-estimate observation, not evidence of established predictive superiority. Three honest caveats bound the claim: (i) the *fine* distinction between independent-first and independent-middle was *not* resolved (pooled exact McNemar $p = 0.42$), so we claim only that independent-last is riskiest, not a fully separated three-way order; (ii) magnitudes are model-specific (a $\sim$4pp last-vs-first gap on M1 versus $\sim$10pp on M2), and the transition model predicts the *ordering* better than the magnitudes; and (iii) the third model (M3) produced *zero* terminal-unsafe decisions on all restrictive cases under every placement, with frozen restrictive forecasts of essentially zero ($< 10^{-6}$)—an outcome ceiling on *these* cases with almost no power to resolve any ordering, not evidence that the base model makes no errors (its calibration and permissive tests contain errors) or that placement effects universally vanish. We report M3 as a null/unresolved placement result rather than a mechanistic confirmation.

Scope. The retained pre-registration was single-model, so the multi-model runs and the pooled contrasts are pre-specified-*direction* extensions rather than the originally registered primary test, and we report the unresolved ceiling case (M3) rather than dropping it. The mechanism is asymmetric follow-versus-diverge behavior, not a scalar "resistance," so placement effects are model-specific and not a universal law. The supported claim is bounded: *after correcting an earlier implementation and fixing predictions before each corrected collection on a previously examined case pool, calibration-set transmission measurements forecast that placing an independent reviewer last is the riskiest arrangement of an identical-composition, equal-budget same-model workflow—resolved within M2 and in a pooled analysis but not within M1, a distinction the permutation-invariant local baseline cannot make—in models that make errors at all.*

Pre-registered replication on a disjoint pool. Because the corrected run above fixed its prediction only *after* prior exposure to the case pool, we ran a second study with sample size and analysis frozen *in advance* and committed to reporting either outcome. The frozen directional prediction (independent-last is riskiest) was reused unchanged; the only change is the data: a slice of 400

Table 7: Prospective prediction of terminal safety across three placements of the independent reviewer in a same-model workflow with identical role composition {D,D,S}, length L=3, and equal call budget. Each model's transmission prediction was frozen (SHA-256) before its *corrected* test outcomes on a previously examined case pool; the permutation-invariant mean-local baseline predicts all three placements equal. Rates are unsafe outcomes over *scored* restrictive cases: 208 per configuration except M2 independent-last, which is $36/206$ (two outcomes missing among 208 enrolled after broken chains following timeouts; paired comparisons use the 206 complete pairs). For the two error-making models the observed point-estimate order matched the frozen forecast; independent-last is resolved as riskiest within M2 and in a pooled test (pooled exact McNemar $p = 6.4 \times 10^{-4}$ vs. first) but not within M1 ($p = 0.26$). M3 produced no terminal-unsafe events on these cases (outcome ceiling; placement unresolved).

Model	Independent-first	Independent-middle	Independent-last	Predicted riskiest
M1	21.6%	22.6%	26.0%	independent-last (matched)
M2	7.2%	9.6%	17.5%	independent-last (matched)
M3	0.0%	0.0%	0.0%	(no errors; ceiling)

held-out cases per configuration (192 restrictive/should-deny plus 208 permissive) drawn from a region of the shuffled pool (indices [700:1100]) that is *disjoint* from both the calibration ([0:300]) and the earlier test ([300:700]) sets. The sample is disjoint from the original study but not entirely unexposed: a small M1 smoke test run before the full collection queried six of these cases (54 M1 agent calls) to validate the harness. Following the plan (which preceded that smoke test), we *retain the fixed sample as pre-specified* and do not drop the six cases retrospectively; we simply disclose the prior exposure. The primary endpoint was pre-specified as an exact two-sided McNemar test of independent-last (C1, [D, D, S]) versus independent-first (C2, [S, D, D]) with pre-specified direction C1>C2, computed over the 192 paired restrictive cases; sample size, slice, and analysis were fixed before collection, with no data-dependent stopping. The outcome is a partial, honest replication that mirrors the original pattern. *In M2*, the prediction replicates decisively: independent-last is riskiest (20.3% unsafe, $39/192$, vs. 9.4% middle and 6.8% first, $13/192$), with the full observed order matching the transmission forecast and the primary contrast strongly resolved (exact McNemar n_{10}=30, n_{01}=4, $p = 6.2\times10^{-6}$)—a sharper effect than in the original run, on a disjoint pool. *In M1*, the prediction does *not* replicate: independent-last and independent-first are tied (25.5% vs. 25.5%, $49/192$ each; middle 23.4%) and the primary contrast is null (n_{10}=31, n_{01}=31, $p = 1.00$). We therefore report the placement effect as replicated and pre-specified in M2 but not in M1, consistent with the model-specific magnitudes seen throughout; this is a genuine mixed result retained per the pre-registration rather than a uniform confirmation, and it does not establish superiority over the final-local predictor or broad generalization.

A.4 Verifiable-task transmission bound and a parser-artifact demonstration

The replication uses fully specified but challenging synthetic policy and arithmetic decisions selected for upstream errors; their saved prompts do fix a programmatic answer, so they are deterministically checkable in principle. To test whether the confirmatory +4.8pp effect could be a parsing or chain-of-thought artifact rather than semantic adoption (a reviewer concern), we ran a separate pre-registered study on a *different* family of deterministically checkable tasks whose correct decision is fixed by a simple explicit rule. The analysis plan, oracle, runner, and inputs were frozen by SHA-256 before any call; every prompt and raw response is retained and independently re-scorable (two technical amendments—a provider temperature-parameter constraint and a token-budget fix, below—are logged).

Setup. 460 unique self-contained synthetic decision cases across three families (policy routing, inventory reconciliation, versioned-document eligibility), each with a deterministic ground-truth oracle; tools disabled, labels never executed. Restrictive = oracle decision `hold` (n=229). Two anonymized current models are tested, M_1 (temperature 0) and M_2. Each case is run paired (control vs. cue); the cue supplies an erroneous upstream conclusion (the opposite decision). Terminal decisions are parsed strictly (a final `FINAL_DECISION` line) with a salvage fallback; parse status is recorded per response.

Near-zero transmission under both a bare and a persuasive cue. With a bare erroneous *label* cue, transmission is $0/229$ for both models (control and cue terminal-unsafe rates 0.0%; McNemar exact p=1.0); the permissive-case secondary (erroneous `hold` cue) is likewise 0.0%. Model↔oracle agreement is $920/920$ (M_1, run on all 460 cases $\times 2$ conditions) and $458/458$ (M_2, run on the 229 restrictive cases $\times 2$), non-degenerate (462 `allow` / 458 `hold` matching ground truth). Because a bare label is trivially checkable, we also tested a genuinely *persuasive* erroneous cue: a capable model, prompted to write a confident analysis concluding the wrong `allow`, *refused or self-corrected to the correct* `hold` *on* $\approx 84\%$ of the 228 restrictive cases (only 37 produced a clean, confident, genuinely-erroneous derivation; 24 concluded `hold`; 167 produced no clean wrong decision, e.g. "I cannot present a fabricated allow decision"). Injecting the 37 genuinely-erroneous derivations downstream still yields $0/37$ transmission for both models (95% Wilson upper bound ≈ 9.4pp). On these tasks, then, the measured per-edge transmission is near zero ($q_{\text{eff}} \approx 1$ on the tested cases), consistent with strong containment in this regime. We observed *no* transmission under the tested conditions; because this study and the replication differ in task difficulty, model, prompting, and selection on upstream error—not verifiability alone—this cross-study contrast does not isolate verifiability or establish self-verification as the causal mechanism, and it does not establish that transmission is impossible on other tasks nor that cascades are universal.

A salvage/truncation artifact caught by the frozen sensitivity. A first execution capped responses at 400 tokens and produced a *spurious* apparent effect (M_2 cued-unsafe 8.45%, McNemar $p = 7.6 \times 10^{-6}$). The prespecified strict/salvage sensitivity localized it entirely to truncated responses: `finish_reason=length` cut off correct reasoning—which explicitly *rejected* the cue ("the upstream decision of 'allow' is inconsistent with the stated rule")—before the `FINAL_DECISION` line, and the salvage parser then read `allow` from phrases such as `for "allow,"`. At a $1{,}200$-token budget all $1{,}378$ responses parse strictly and the effect is 0.0pp. This is a concrete demonstration that response-level retention and a strict/salvage sensitivity check are necessary to distinguish semantic adoption from a parsing artifact, and it directly supports the parser caveats we attach to the $+4.8$pp replication (appendix A.1).

A.5 Randomization-test calibration

We simulate binary traces with the same length distribution and marginal event rate as the collected data. Under the negative control, only a smoothly varying session-phase rate is present; under the positive control, an additional decayed history term creates self-excitation. We compare unrestricted within-trace shuffling with the phase-conditioned randomization used in the primary analysis.

Table 8: Calibration over 300 synthetic datasets with 199 permutations per test. Rates are percentages at nominal $\alpha = 0.05$.

Test	Phase-drift false-positive rate	History-model power
Unrestricted shuffle	97.0	100.0
Phase-conditioned shuffle	5.7	100.0

The unrestricted shuffle is severely anti-conservative when event rates drift with task phase. Conditioning on four phase bins reduces the empirical false-positive rate to 5.7%, close to the nominal 5%, while retaining 100% power for the strong simulated history effect in this base positive control. To confirm this is not tuned to a single configuration, we sweep phase-drift strength and self-excitation η over a 9-cell (3×3) grid (200 datasets/cell; table 10): the phase-conditioned false-positive rate stays near nominal in every cell (max 6.5%, consistent with Monte Carlo variation at this replicate count), the unrestricted test is inflated in every cell and grows with drift ($9\% \rightarrow 69\%$), and phase-conditioned power rises monotonically with the true η (from $\approx 35\%$ to $\approx 60\%$ across the swept effect sizes). This calibration supports the randomization design, not the realism of either synthetic generator.

A.6 Phase-controlled elevated-access analysis

A purpose-built isolation layer confines sessions to the synthetic corpus and records boundary failures; sessions use fixed execution budgets and a prespecified one-retry rule. The 100 planned tasks

required 102 attempts (two initial timeouts preserved as audit rows; both retries succeeded), with no recorded current-protocol boundary violations (historical records are a separate cohort outside this protocol). Of these, 85 traces met the prespecified inclusion threshold of at least eight attempted tool calls (table 9) and were analyzed; the remaining 15 fell below it and were excluded. We retain task identity, normalized call position, raw tool arguments, gateway usage, and permission denials for reproducibility; provider and internal-system identifiers are excluded.

Table 9 reports the per-provider phase-controlled analysis summarized in section 5. The primary analysis is accompanied by narrow and broad proxy definitions and by phase discretizations with two, four, and eight bins. The broad proxy remains significant for every family after Holm correction. Under the narrow proxy, families B–E remain significant ($p = 0.0025$), whereas family A does not ($p = 0.346$; $\Delta\text{AIC} = -3.62$); the narrow aggregate remains significant ($p = 0.0005$). For the primary proxy, unadjusted p-values remain at most 0.003 for every family under two, four, and eight phase bins. We interpret this mostly consistent sensitivity evidence as robustness of temporal dependence in this sandbox, not validation of the marks as safety-relevant near-misses.

Table 9: Phase-controlled elevated-access analysis. Provider p-values are Holm-adjusted across five tests; the aggregate p-value is unadjusted. ΔAIC is positive when adding history improves the phase-adjusted model.

Provider	Traces	Calls	Event rate	Phase p	ΔAIC
A	12	393	0.183	0.0025	15.12
B	19	1,547	0.256	0.0025	390.38
C	19	1,021	0.328	0.0025	359.57
D	15	557	0.244	0.0025	146.31
E	20	1,172	0.219	0.0025	166.37
All	85	4,690	0.255	0.0005	1,075.59

Table 10: Randomization-test calibration across a grid of generative parameters (phase-drift gain $\times$ self-excitation η; 200 datasets/cell, 199 permutations). False-positive rate is measured with no self-excitation; power is measured with self-excitation of strength η. Rates are percentages at nominal $\alpha = 5\%$. The phase-conditioned false-positive rate stays near nominal everywhere (at most 6.5%, within Monte Carlo variation at this replicate count); the unrestricted false-positive rate is inflated everywhere and grows with drift.

Drift gain	η	FP unrestr.	FP phase-cond.	Power unrestr.	Power phase-cond.
0.35	0.0	9.0	4.5	9.0	2.5
0.35	2.0	7.0	2.5	99.0	35.0
0.35	3.0	9.5	2.0	98.5	55.0
0.65	0.0	24.5	3.0	23.5	3.0
0.65	2.0	27.0	6.5	96.5	34.5
0.65	3.0	26.0	2.5	97.5	53.0
0.95	0.0	68.5	4.0	65.0	3.5
0.95	2.0	67.5	2.5	99.5	41.5
0.95	3.0	71.0	4.5	99.0	59.5

Table 11: Monte Carlo validation of the control threshold for $R_S = 1.3$. "Survival" is used for supercritical rows and mean total cluster size for subcritical rows.

u	R_{eff}	Regime	Quantity	Empirical	Theory
0.000	1.300	supercritical	survival	0.422	0.423
0.181	1.065	supercritical	survival	0.103	0.097
0.250	0.975	subcritical	mean size	37.38	40.00
0.500	0.650	subcritical	mean size	2.89	2.86

Table 12: Cross-excitation and uniform propagation control. The tested intervention is five percentage points above the theoretical threshold.

k	$\rho(A)$	Required u	Controlled ρ
1	0.70	0.000	0.665
3	0.94	0.000	0.893
4	1.06	0.057	0.947
6	1.30	0.231	0.935
10	1.78	0.438	0.911

Table 13: Labeled-parent sample size for at least 75% classification accuracy over 1,200 trials per hypothesis.

ϵ	Labeled parents m	$m\epsilon^2$	Accuracy
0.30	10	0.90	0.825
0.20	25	1.00	0.848
0.10	50	0.50	0.754
0.07	103	0.50	0.776
0.05	200	0.50	0.772

A.7 Simulation validation of the theory

Cross-model generality at fixed topology. To check that the threshold is not a single-model artifact, we run the same fixed-topology ($b = 2$) cascade on each family. This cross-model sweep uses the *constructed*-benchmark cascade (a scripted erroneous handoff per edge, calibrated to each family's measured single-hop $\widehat{q}$); we therefore read it as an illustration of the $R_S = b\,\widehat{q}$ law, with the forwarded-output cascade in the main text supplying the direct mechanism evidence for the susceptible family. The empirical branching ratio tracks $R_S = b\,\widehat{q}$: it is 0.00 for a robust family (no contamination arises), ≈ 0.42 for the weak families (cascades die, mean size 4.4), and supercritical for the susceptible family (cascades grow). The identical topology is thus subcritical for some families and supercritical for others, consistent with $R_S = b\,\widehat{q}$.

Controlled cascade threshold. We simulate 4,000 Poisson Galton–Watson cascades per configuration (table 11, fig. 1): below threshold the mean cluster size matches eq. (5), and above threshold the empirical survival probability matches the fixed-point prediction. Within the assumed branching model, three simulated mean-field signatures identify $R_{\text{eff}} = (1-u)R_S = 1$ as a continuous phase transition: a susceptibility exponent $\gamma = 1.01$ (theory 1; $R^2 = 0.9996$), a Borel cluster-size tail $s^{-\tau}$ with $\tau = 1.53$ (theory $3/2$), and finite-size suppression of the survival order parameter below threshold, sharpening toward the limiting order parameter—identically zero below threshold and rising continuously above $R_{\text{eff}} = 1$—as the censoring size grows. A near-critical *agent* cascade ($u = 0.25$, $R_{\text{eff}} = 0.99$) has a heavy total-progeny tail (fitted exponent 1.46) compatible with $\tau = 1.53$, but with only 29 cascades—this arm was stopped early—it is a suggestive consistency check, not a measured power law; the exponent claim rests on the simulation.

Composition and uniform control. We use k agents with diagonal offspring mean 0.70 and uniform off-diagonal mean 0.12 (table 12). Each component is individually subcritical. The system crosses the spectral boundary at four agents, after which the uniform intervention predicted by theorem 9 returns the spectral radius below one. As an implementation check for proposition 10, we evaluate its analytic derivative on a positive asymmetric four-type matrix; centered finite differences with step 10^{-6} agree in every coordinate, with maximum absolute discrepancy 6.6×10^{-10}.

Labeled-offspring testing. For each ϵ, we draw independent offspring counts under $R = 1 - \epsilon$ and $R = 1 + \epsilon$ and classify with the sample mean. The smallest tested m reaching 75% accuracy keeps $m\epsilon^2$ approximately constant (table 13), matching theorem 11.

Table 14: Susceptible model E: handoff-contamination breakdown (fixed crossed condition order). Unsafe counts are over the 60 ground-truth-restrictive cases per condition; overall error is over all 120 cases per condition. The Erroneous overall rate is $43/120 = 35.8\%$ (the 40 restrictive unsafe approvals plus 3 errors on permissive cases). Each case is run under all four handoff conditions with case facts held fixed.

Handoff condition	Unsafe (restrictive)	Overall error
None	0/60	0.0%
Correct	0/60	0.0%
Erroneous	40/60	35.8%
Contained (verified fact)	0/60	0.0%

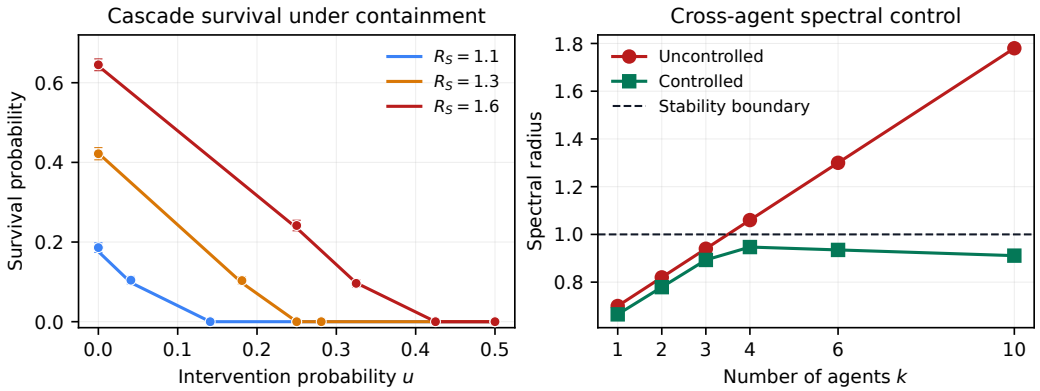

Figure 1: Controlled-threshold checks. Left: fixed-point survival probabilities (lines) and Monte Carlo estimates with 95% binomial intervals (points) for three uncontrolled criticalities. Right: cross-agent excitation crosses the spectral boundary, while the intervention prescribed by theorem 9 keeps the controlled radius below one.

B Proofs and Additional Details

B.1 Proof of Theorem 5

Proof. Under the Poisson cluster representation of a linear Hawkes process, an uncontrolled event has a Poisson number of direct offspring with mean $R_S = \|\phi\|_1$ (Hawkes & Oakes, 1974). A parent is contained with probability u and then has no offspring; otherwise its offspring count is $\mathrm{Poisson}(R_S)$. The probability-generating function of the controlled offspring count Z_u is therefore

$$G_u(s) = u + (1-u)\exp\{R_S(s-1)\}, \tag{16}$$

and

$$\mathbb{E}[Z_u] = G_u'(1) = (1-u)R_S = R_{\text{eff}}. \tag{17}$$

For $R_{\text{eff}} < 1$, the associated Galton–Watson process becomes extinct almost surely. If M is its expected total progeny including the immigrant, the branching recursion gives

$$M = 1 + \mathbb{E}[Z_u]M = 1 + R_{\text{eff}}M, \tag{18}$$

so $M = (1 - R_{\text{eff}})^{-1}$.

Equivalently, assign independent propagation marks $C_i \sim \mathrm{Bernoulli}(1-u)$ and write

$$\lambda_u(t) = \mu + \sum_{\tau_i < t} C_i \phi(t - \tau_i). \tag{19}$$

In stationarity, Campbell's identity and independence of C_i yield

$$\bar{\lambda} = \mu + \bar{\lambda}\,\mathbb{E}[C_i]\int_0^\infty \phi(s)\,ds \tag{20}$$

$$= \mu + \bar{\lambda}(1-u)R_S. \tag{21}$$

Solving gives $\bar{\lambda} = \mu/(1 - R_{\text{eff}})$.

For $R_{\text{eff}} > 1$, Galton–Watson theory gives positive survival probability. Its extinction probability is the smallest fixed point of the offspring generating function, hence $s_* = G_u(s_*)$ as stated. Finally,

$$(1-u)R_S < 1 \quad \Longleftrightarrow \quad u > 1 - \frac{1}{R_S}. \tag{22}$$

When $R_S \leq 1$, the infimum nonnegative intervention is zero; combining the cases gives eq. (8). □

Distributional scope of the extinction fixed point. The first-moment threshold $(1-u)R_S < 1$ is distribution-free (it depends only on the offspring mean). The transcendental fixed point (16) is, however, specific to the Poisson offspring law of the linear Hawkes cluster. For a bounded communication fan-out b (as in the discrete cascades of section 5.2) with iid per-edge transmission, the per-parent offspring count is bounded and the extinction probability solves the corresponding polynomial generating function—for the containment-thinned b=2 tree, the quadratic $s = u + (1-u)(1-\widehat{q}+\widehat{q}s)^2$ rather than (16) (bounded fan-out alone is insufficient; the form also assumes independent edges). Only the mean-based threshold transfers unchanged; the extinction *probability* is topology-specific.

B.2 Proof of Corollary 6

Proof. Substituting $u = pq$ into $R_{\text{eff}} = (1-u)R_S$ gives

$$R_{\text{eff}} < 1 \quad \Longleftrightarrow \quad pq > 1 - \frac{1}{R_S}. \tag{23}$$

Because $p \leq 1$, the largest achievable product for fixed q is q. Thus no $p \in [0,1]$ can satisfy the inequality when $q \leq 1 - 1/R_S$. The strict inequality corresponds to strict subcriticality; equality is the critical boundary. □

Remark 14 (Imperfect (message-level) containment)**.** *Soft remediation—rewriting or replacing a parent's message with a verified fact rather than pruning its subtree—is* not *the zero-offspring operator of definition 3: a remediated parent's children still err at a residual rate $q_{\text{repl}} > 0$. Consider the idealized* homogeneous *model with a* fixed *residual offspring mean $R_{\text{res}} = b\, q_{\text{repl}}$ (constant in u and in the visited-state distribution), so that $R_{\text{eff}}(u) = (1-u)R_S + u\, R_{\text{res}}$. Assuming $R_S > 1$, subcriticality requires*

$$u > u_c^{\text{res}} = \frac{R_S - 1}{R_S - R_{\text{res}}} \quad \textit{whenever } 0 \leq R_{\text{res}} < 1, \tag{24}$$

which exceeds the zero-offspring threshold u_c when $0 < R_{\text{res}} < 1$; if instead $R_{\text{res}} \geq 1$, no $u \in [0,1]$ stabilizes this fixed-mean mixture (an intervention that drives R_{res} below one by other means could still stabilize it). This is a statement about the homogeneous fixed-residual model. Our realized b=2 replacement arms are finite-horizon, dose-dependent *frequencies rather than a single stationary q_{repl}: post-replacement error runs $39/156$ ($\approx$25%) at u=0.5 and $96/138$ ($\approx$70%) at u=1 in section 5.2, and $171/404$ ($\approx$42%) to $296/412$ ($\approx$72%) across the held-out sweep. They neither identify a common residual branching parameter nor prove the absence of a stabilizing dose; the arms in section 5.2 and appendix A.2 are* descriptive *dose-responses, not realizations of the theorem 5 zero-offspring boundary. A model-only grid over (R, R_{res}, u) (appendix D.1) confirms both the mixture mean $m = (1-u)R + uR_{\text{res}}$ and the extra offspring variance $u(1-u)(R - R_{\text{res}})^2$ that collapsing the mixture to a single* $\text{Pois}(m)$ *would erase.*

B.3 Proof of Proposition 7

Proof. The event-rate identity is eq. (6). Conditional independent marking with harm probability h thins the stationary event process in expectation, so

$$\bar{\lambda}_{\text{harm}} = h\bar{\lambda} = \frac{h\mu}{1 - R_{\text{eff}}}. \tag{25}$$

The observed-event rate is $p\bar{\lambda}$; dividing it by inspection capacity B gives L_S. For any fixed $p > 0$, $R_{\text{eff}} < 1$, and $B < \infty$, choosing $\mu \geq B(1 - R_{\text{eff}})/p$ makes $L_S \geq 1$. Hence intrinsic subcriticality does not imply the capacity condition. □

B.4 Proof of Theorem 9

Proof. By definition, column j of A contains the expected offspring vector produced by a type-j parent. Such a parent propagates with probability $1 - u_j$, so independent containment multiplies column j by $1 - u_j$. The controlled mean offspring matrix is therefore AD_u.

The multitype Galton–Watson subcriticality criterion and the corresponding multivariate linear-Hawkes stability criterion are $\rho(AD_u) < 1$. Under uniform intervention, $D_u = (1 - u)I$ and homogeneity of the spectral radius gives

$$\rho(AD_u) = (1 - u)\rho(A). \tag{26}$$

Solving the strict inequality produces the claimed uniform threshold.

For the final statement, let A have diagonal value $r < 1$ and off-diagonal value $\epsilon > 0$. Its Perron eigenvalue is $r + (k - 1)\epsilon$, which can exceed one even though every diagonal entry is $r < 1$. □

Cyclic versus acyclic topologies. The spectral criterion $\rho(AD_u) < 1$ governs *recurrent* interaction. For a strictly feedforward (acyclic) agent pipeline, A is nilpotent under a topological ordering, so $\rho(AD_u) = 0$ regardless of edge transmission and infinite-horizon supercriticality is impossible in the spectral sense. Such pipelines instead exhibit *finite-horizon* transient amplification governed by the resolvent path sums $\sum_{k\geq 1}(AD_u)^k$, and Proposition 10 additionally assumes A irreducible (a strictly positive Perron vector), which acyclic A violate. The threshold is therefore informative for cyclic role structures and self-exciting retries, not for bounded-depth DAGs, where the finite-depth cascade statistics of section 5.2 are the relevant object.

B.5 Proof of Proposition 10

Proof. Because A is irreducible and every $1 - u_j$ is positive, $M = AD_u$ is irreducible. Its Perron root is simple, so standard eigenvalue perturbation with $\ell^\top r = 1$ gives

$$\frac{\partial \rho(M)}{\partial u_j} = \ell^\top \left(\frac{\partial M}{\partial u_j}\right) r = -\ell^\top A e_j e_j^\top r = -(\ell^\top A)_j r_j. \tag{27}$$

The jth component of $\ell^\top AD_u = \rho(M)\ell^\top$ implies $(\ell^\top A)_j(1 - u_j) = \rho(M)\ell_j$. Substitution yields

$$\frac{\partial \rho(M)}{\partial u_j} = -\frac{\rho(M)\ell_j r_j}{1 - u_j}. \tag{28}$$

Dividing the magnitude of this derivative by the marginal cost c_j gives the stated first-order ranking. □

Behavior as $u_j \to 1$. For $u_j < 1$ under the irreducibility assumptions of proposition 10, the equivalent expression $-(\ell^\top A)_j r_j$ avoids explicit division by $1 - u_j$ and is numerically preferable to the ratio $-\rho(M)\ell_j r_j/(1 - u_j)$. We make *no* smoothness claim at the boundary $u_j = 1$: the controlled matrix AD_u may become reducible there, its Perron root need not remain simple, and the normalized eigenvectors need not have the assumed limiting behavior, so the derivative need not exist or stay bounded. For example, with $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $u_2 = 0$, $\rho(AD_u) = \sqrt{1 - u_1}$, whose derivative $-\frac{1}{2}(1 - u_1)^{-1/2}$ diverges as $u_1 \to 1$. The interior result of proposition 10 is unaffected.

B.6 Proof of Theorem 11

Proof. The joint log likelihood under offspring mean R is, up to terms independent of R,

$$\ell(R) = \left(\sum_{i=1}^{m} Z_i\right) \log R - mR. \tag{29}$$

Differentiation gives the maximum-likelihood estimator

$$\widehat{R} = \frac{1}{m}\sum_{i=1}^{m} Z_i. \tag{30}$$

Because independent Poisson variables add, $m\widehat{R} \sim \text{Poisson}(mR)$, so $\mathbb{E}[\widehat{R}] = R$ and $\text{Var}(\widehat{R}) = R/m$.

For an upper bound, classify as H_1 when $\widehat{R} > 1$. Under either hypothesis an error implies $|\widehat{R} - R| \geq \epsilon$. Chebyshev's inequality gives

$$\Pr_R(\text{error}) \leq \frac{R}{m\epsilon^2} \leq \frac{1+\epsilon}{m\epsilon^2}. \tag{31}$$

Thus $m \geq 4(1+\epsilon)/\epsilon^2$ suffices for error at most $1/4$; sharper Chernoff bounds change only the constant.

For the lower bound, the one-sample Kullback–Leibler divergence is

$$D_{\text{KL}}(\text{Pois}(1-\epsilon) \,\|\, \text{Pois}(1+\epsilon)) = (1-\epsilon)\log\frac{1-\epsilon}{1+\epsilon} + 2\epsilon \tag{32}$$

$$= 2\epsilon^2 + O(\epsilon^3). \tag{33}$$

For m independent parents the divergence is m times this quantity. Le Cam's two-point method then bounds the sum of testing errors away from zero whenever $m\epsilon^2$ is bounded by a sufficiently small constant. Hence $m = \Omega(\epsilon^{-2})$ is necessary, completing the matching rate. □

Scope: uncontrolled offspring and cross-cascade pooling. The Poisson likelihood applies to *uncontrolled* offspring counts (the intended use is distinguishing an intrinsic $R = 1 \pm \epsilon$). Under active parent-level containment the count Z_u (zero w.p. u, else $\text{Poisson}(R_S)$) is zero-inflated Poisson with $\mathbb{E}[Z_u] = (1-u)R_S = R_{\text{eff}}$ and $\text{Var}(Z_u) = (1-u)R_S + u(1-u)R_S^2$, so it is overdispersed by a factor $\text{Var}/\text{mean} = 1 + u\,R_S = 1 + u\,R_{\text{eff}}/(1-u)$ (e.g. $R_S{=}2, u{=}0.5$ gives mean 1, variance 2). Certifying an *already-controlled* system from post-intervention counts therefore requires a zero-inflated likelihood analysis. When testing controlled means near $R_{\text{eff}} = 1$, the factor $1 + u\,R_{\text{eff}}/(1-u)$ diverges as $u \to 1$ (intrinsic R_S must then grow), so a controlled-count rate statement needs explicit assumptions on u and the testing regime—theorem 11 does not establish constants uniform in u; at fixed intrinsic R_S the factor instead approaches the finite $1 + R_S$. The m labeled parents are collected across independent immigrant-seeded cascades: a subcritical cascade has *expected* total progeny $(1 - R_{\text{eff}})^{-1}$, but this is a mean, not an upper bound—realized trees can be arbitrarily large—so a prespecified exploration-and-restart protocol across independent cascades collects the required parents without selecting only unusually large or surviving trees.

B.7 Proof of Proposition 13

Proof. In the Poisson cluster construction, an event at time t generates children according to an inhomogeneous Poisson process on (t, ∞) with rate $s \mapsto \phi(s-t)$. The total number of children is therefore Poisson with mean

$$\int_t^\infty \phi(s-t)\,ds = \int_0^\infty \phi(v)\,dv = R_S. \tag{34}$$

Independent containment changes the mean offspring number to $(1-u)R_S$ exactly as in theorem 5. The branching tree, total progeny, and extinction boundary depend only on that mean and offspring generating function. The temporal locations of the children retain the normalized density $\phi/\|\phi\|_1$, so duration and finite-window observability can depend on the kernel shape even though the progeny threshold does not. □

Zero-latency assumption. This kernel-shape invariance assumes containment acts at parent instantiation (synchronous pre-execution gating). With a non-zero inspection latency $\delta > 0$, offspring emitted during $(t, t+\delta]$ execute before containment takes effect, and the escaping mass $\int_0^\delta \phi(v)\,dv$ over that window depends on the kernel shape and concentration. The invariance then holds only up to this shape-dependent leakage, so under asynchronous auditing the effective offspring mean is $(1-u)R_S + u\int_0^\delta \phi$ rather than $(1-u)R_S$.

B.8 EMPIRICAL ANALYSIS DETAILS

The primary proxy labels a file-read attempt when its path contains one of

$$\{\texttt{secret}, \texttt{credential}, \texttt{config/}, \texttt{password}, \texttt{.env}, \texttt{token}, \texttt{api_key}, \texttt{database}\}, \tag{35}$$

and labels a text-search attempt when its query contains one of

$$\{\texttt{password}, \texttt{secret}, \texttt{token}, \texttt{key}, \texttt{credential}, \texttt{api_key}, \texttt{database_url}, \texttt{connection_string}, \texttt{auth}\}. \tag{36}$$

The narrow proxy retains only direct secret or credential paths and explicit password, secret, or credential searches. The broad proxy additionally marks configuration-like file extensions, network terms, cryptographic terms, and identity-related paths. These interface-agnostic rules operate on attempted operation type and arguments; they do not inspect the model's prose answer.

For trace r with labels $Y_{r,1}, \ldots, Y_{r,n_r}$, the adjacency statistic is

$$T(Y) = \sum_{r} \sum_{t=2}^{n_r} Y_{r,t-1} Y_{r,t}. \tag{37}$$

The primary randomization partitions each trace into bins $b_{r,t} = \lfloor 4(t-1)/n_r \rfloor$ and independently permutes labels within every (r, b) stratum. Thus each permutation preserves the exact event count of each coarse phase of each task. With M permutations, the one-sided Monte Carlo value is

$$p = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{T(Y^{(m)}) \geq T(Y)\}}{M+1}. \tag{38}$$

Provider-level values are adjusted by Holm's step-down procedure. Repeating the same construction with two and eight phase bins tests sensitivity to the coarse discretization.

The complementary history model uses call index only. If $x_{r,t} = (t-1)/(n_r - 1)$ and

$$H_{r,t}(\beta) = \sum_{j<t} e^{-\beta(t-j)} Y_{r,j}, \tag{39}$$

the alternative is

$$\operatorname{logit} \Pr(Y_{r,t} = 1) = a_r + \gamma_1 x_{r,t} + \gamma_2 x_{r,t}^2 + \gamma_3 x_{r,t}^3 + \eta H_{r,t}(\beta), \qquad \eta \geq 0. \tag{40}$$

We select β from $\{0.15, 0.30, 0.60, 1.00, 1.80\}$ and charge both η and the selected decay as additional parameters in the reported AIC difference. This model is an ordinal-time sensitivity analysis, not a Hawkes branching-ratio estimator.

C OPERATIONAL UNITS, JOINT FEASIBILITY, AND DEPENDENT MARKS

This appendix makes the model's operational assumptions and units explicit.[1]

Notation across theoretical and empirical settings. The conditional containment probability q appears in $u = pq$. Empirical $\widehat{q}_{\text{fwd}}$ and q_{repl} instead denote forwarding and post-replacement error frequencies; their conditioning sets and denominators differ. A generic $\widehat{q}$ in a handoff panel is its panel-specific transmission frequency, not containment effectiveness. Multiplying an empirical frequency by planned fan-out b is a finite-workflow reference and does not establish a causally identified R_S.

[1]These are standard consequences of the stated branching and mean-rate model, included as clarifications rather than new branching-process theory. They do not establish empirical capacity calibration or deployment safety.

C.1 Discrete genealogies and continuous-time rates

The branching primitive is the number of direct children caused by one parent. Thus R_S, p, q, and R_{eff} are dimensionless and apply to an independently reproducing discrete-generation genealogy. Embedding that genealogy in continuous time additionally assigns offspring delays: in the Hawkes construction, $\phi(s)$ has units of inverse time and $\int_0^\infty \phi(s)\,ds = R_S$. Under the instantaneous, independent containment assumption, changing the offspring delays leaves the genealogy and its mean threshold unchanged. Inspection delay or state-dependent selection can change the offspring law and requires a different model.

The mean rate $\bar{\lambda} = \mu/(1 - R_{\text{eff}})$ additionally requires an immigrant rate μ and the stationary, subcritical regime; μ, $\bar{\lambda}$, and B must use the same clock unit. A discrete model with ν immigrant events per generation instead has mean $\nu/(1 - R_{\text{eff}})$ events per generation, not events per second. If a workflow fixes an interval Δ and independently establishes sustainable capacity of κ inspections per interval, then $\mu = \nu/\Delta$ and $B = \kappa/\Delta$ provide a units conversion. Our per-generation cap κ alone establishes neither Δ nor sustainable throughput. Variable generation durations, checker latency, retries, and queueing must be measured separately. A token or call budget likewise is not a time rate. Selecting at most κ parents can also make the realized observation probability depend on frontier size, unlike the constant independent p used in the theorem. If a deployed checker also inspects outputs outside this risk-event model, their inspection work must be included before using the load calculation as an estimate of total demand.

C.2 Joint subcriticality and mean inspection capacity

For this calculation allow the boundary values $R_S \geq 0$ and $\mu \geq 0$, with $B > 0$ and $p, q \in [0, 1]$. Write $a = \mu/B$ and $D = 1 - R_S + pqR_S$. The paper's two strict conditions are equivalent to the single inequality

$$R_{\text{eff}} < 1, \quad L_S < 1 \quad \Longleftrightarrow \quad D > pa \quad \Longleftrightarrow \quad 1 - R_S + p(qR_S - a) > 0. \tag{41}$$

Indeed, when $D > 0$, $L_S = pa/D$; conversely $D > pa \geq 0$ implies both strict conditions. At fixed p and $R_S > 0$, the equivalent intervention requirement is

$$u > \frac{R_S - 1 + p\mu/B}{R_S}, \qquad u = pq \in [0, p]. \tag{42}$$

The capacity term tightens the stability-only inequality when $p\mu > 0$. The right side is a strict boundary, not an attained minimum; if it is negative, $u = 0$ may already be feasible. In particular, p and u cannot be optimized independently when q is fixed.

For clarity, set $c = qR_S - a$ and hold q fixed. The complete feasible observation set within $[0, 1]$ is:

- If $R_S < 1$, all p are feasible when $c \geq 0$; when $c < 0$, feasibility requires $p < (1 - R_S)/(-c)$.
- If $R_S = 1$, feasibility requires $q > a$ and $p > 0$.
- If $R_S > 1$, a feasible p exists exactly when $c > R_S - 1$; the feasible values satisfy $p > (R_S - 1)/c$.

These sets retain strict endpoints. For $R_S = 0$, no division by R_S is needed: feasibility is simply $p\mu/B < 1$, independent of q. For $q = 0$, feasibility requires $R_S < 1$ and $p\mu/B < 1 - R_S$; observation without containment cannot stabilize a critical or supercritical genealogy. If $\mu = 0$, mean load is zero in the subcritical regime and the remaining condition is $D > 0$. The empty process with no immigrants is also possible outside that regime, but its zero activity does not establish that a seeded cascade is subcritical. Finite seeded bursts can require substantial inspection work even when their long-run time-average arrival rate is zero.

On the subcritical domain $D > 0$, with q, R_S, μ, B fixed,

$$\frac{\partial L_S}{\partial p} = \frac{\mu}{B} \frac{1 - R_S}{(1 - R_S + pqR_S)^2}. \tag{43}$$

For $\mu > 0$ this is positive for $R_S < 1$, zero for $R_S = 1$, and negative for $R_S > 1$. In the last case, greater observation can shrink the stationary cascade enough to reduce total inspection demand.

This comparison is defined only where the stationary mean exists. If $\mu = 0$, the derivative is zero throughout that domain. As $D \downarrow 0$, load diverges provided $p\mu$ remains bounded away from zero; this qualification matters. For example, when $R_S = 1$ and $q > 0$, $L_S = \mu/(Bq)$ for every $p > 0$, although $D = pq \downarrow 0$ as $p \downarrow 0$.

Here $L_S < 1$ is a mean offered-work condition under the stated arrival and containment model. It does not, by itself, prove a queueing theorem, bound waiting times, or guarantee that inspection finishes before a parent propagates. Such claims require a service and scheduling model, latency assumptions, and appropriate stationarity/integrability conditions. The synthetic inspection-count caps do not empirically calibrate these time-rate quantities.

C.3 When marks affect reproduction

The scalar harm-rate formula assumes reproduction-independent marks. If near-miss and harm events reproduce differently, treat them as event types N and H. Let $\phi_{ij}(s)$ be the kernel for type-i children of a type-j parent, and define $A_{ij} = \int_0^\infty \phi_{ij}(s)\, ds$. Columns therefore index parents and rows index children. With type-specific independent observation p_j and zero-offspring containment probability q_j conditional on observation, the mean controlled offspring matrix is

$$C = A \operatorname{diag}(1 - p_N q_N, 1 - p_H q_H), \qquad R_{\text{eff}}^{\text{type}} = \rho(C). \tag{44}$$

This is column scaling because containment suppresses all children of the selected parent, not only children of a particular type. The standard multitype subcriticality criterion is $\rho(C) < 1$. With immigrant-rate vector $\boldsymbol{\mu} = (\mu_N, \mu_H)^\top$, its stationary mean is

$$\bar{\boldsymbol{\lambda}} = (I - C)^{-1}\boldsymbol{\mu}, \qquad \bar{\lambda}_{\text{harm}} = \bar{\lambda}_H, \qquad L_S^{\text{type}} = \frac{p_N \bar{\lambda}_N + p_H \bar{\lambda}_H}{B}. \tag{45}$$

The load expression assumes a shared monitor with equal inspection work per event; unequal service requirements require corresponding work weights. In general, multiplying total rate by a fixed immigrant harm fraction does not give the stationary harm rate. As a consistency check, if $v = (1 - h, h)^\top$, $A = R_S v \mathbf{1}^\top$, $\boldsymbol{\mu} = \mu v$, and $p_j = p$, $q_j = q$, then these expressions reduce to $\bar{\boldsymbol{\lambda}} = v\mu/[1 - (1 - pq)R_S]$ and the scalar burden formula. Mark-dependent data and interventions require estimating the matrix entries and type-specific control effects; the present experiments do not establish those deployment quantities.

D Model sensitivity to residual repair and observation noise

Scope and prospective design. These new numerical experiments address the consequences of two departures from the idealized control and measurement assumptions. They are *model-only*; their simulated counts are not additional agent responses, and agreement with the stipulated formulas does not validate those formulas for deployed agents. Before collection, we fixed the complete grids, seed 2026092301, censoring, analysis and displays, and hashed all seven source/protocol/test files. Thirteen implementation tests passed. Collection and analysis each ran once, with no exclusions, tuning or scientific retry. All 60 residual-repair cells and 81 observation-noise cells are retained, including null, critical, adverse and censored cases.

D.1 Residual repair changes both mean and offspring variance

Each parent is independently repaired with probability u. Conditional offspring are $\text{Pois}(R_{\text{res}})$ after repair and $\text{Pois}(R)$ otherwise, with fixed parameters. Consequently,

$$m = (1 - u)R + uR_{\text{res}}, \tag{46}$$

$$\text{Var}(Z) = m + u(1 - u)(R - R_{\text{res}})^2, \tag{47}$$

$$f(s) = (1 - u)e^{R(s-1)} + ue^{R_{\text{res}}(s-1)}. \tag{48}$$

The variance follows total variance, and f is the mixture PGF; replacing the entire mixture by $\text{Pois}(m)$ would erase its overdispersion. Given a generation of z parents, our simulator first draws $k \sim \text{Binom}(z, u)$ and then $Z' \sim \text{Pois}((z - k)R + kR_{\text{res}})$, exactly aggregating conditionally independent offspring. For $R > 1$ and $R_{\text{res}} < 1$, strict subcriticality requires $u > (R-1)/(R-R_{\text{res}})$. If

both endpoint means are at least one, no mixture is strictly subcritical. Conversely, if $R < 1 < R_{\text{res}}$, increasing repair past $(1 - R)/(R_{\text{res}} - R)$ destabilizes the process. Thus a blanket impossibility claim based on $R_{\text{res}} \geq 1$ alone would omit the uncontrolled regime.

We cross $R \in \{0.8, 1.2, 1.8\}$, $R_{\text{res}} \in \{0, 0.4, 1, 1.2\}$ and $u \in \{0, 0.25, 0.5, 0.75, 1\}$. Each cell has 100,000 independently sampled parent counts and 2,000 cascades from one root: 6 million parent counts and 120,000 cascades in total. Trajectories stop at extinction, generation 50, or a cumulative count of at least 10,000. A still-live trajectory censored before generation 50 has unknown later survival. Across all cells, 20,266 trajectories are censored; none is silently counted as extinct or surviving. The all-cell CSV includes adverse survival bounds, their nominal Wilson 95% enclosure, and the exact finite-generation probability $1 - f^{\circ 50}(0)$. Asymptotic nonextinction, computed separately from the smallest PGF fixed point, is a different quantity.

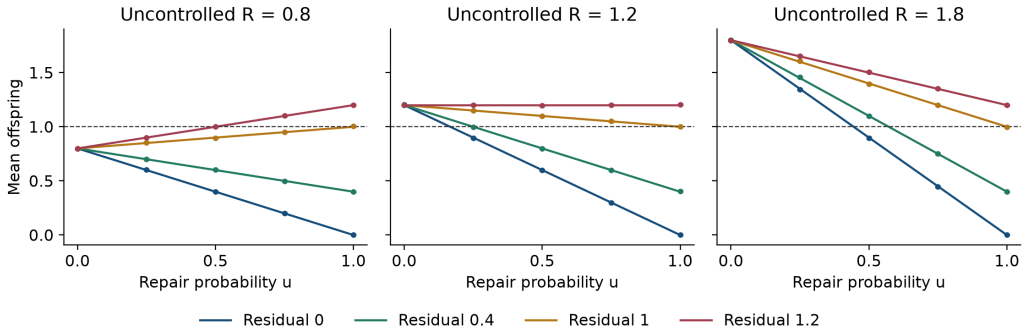

Figure 2: All 60 fixed residual-repair cells. Lines are analytic means; points are simulated means with nominal 95% Monte Carlo intervals (100,000 parents/cell). These are pointwise implementation diagnostics, not simultaneous evidence for an agent model. The dashed line marks mean one.

Table 15: Prespecified $R = 1.8, u = 0.5$ anchors. Mean intervals are ± 1.96 Monte Carlo SEM. Generation-50 survival bounds retain censored trajectories as unknown; no infinite-horizon empirical survival claim is made. Offspring means and SEM use 100,000 independent parent counts per cell; survival and censoring use 2,000 cascades per cell.

R_{res}	Analytic m	Sample mean (half-width)	Survival bounds	Censored
0	0.90	0.8939 (0.0081)	0.05–0.05%	0
0.4	1.10	1.0999 (0.0078)	8.65–13.10%	89
1	1.40	1.3955 (0.0077)	0–47.40%	948
1.2	1.50	1.5058 (0.0079)	0–58.05%	1,161

At the same 50% dose, hard containment produces $m = 0.9$ whereas residual mean 0.4 produces $m = 1.1$. Their respective stability thresholds are 0.4444 and 0.5714; the historical dose near 0.5 cannot be treated as sufficient under arbitrary repair semantics. The prespecified reverse-regime anchor, $R = 0.8, R_{\text{res}} = 1.2, u = 1$, gives a sample mean 1.2016 ± 0.0068 and 631/2,000 censored cascades. These findings illustrate model limits; the observed, dose-dependent agent replacement frequencies do not identify a fixed R_{res}.

D.2 Observation noise can obscure criticality

Suppose all n sampled parent opportunities are known, including parents with no children. For each independent parent, the true count is $N \sim \text{Pois}(R)$, each child link is detected independently with probability d, and the number of independently misattributed children is $\text{Pois}(\eta)$. Here d denotes *link detection*, not the main paper's monitoring probability p. Poisson thinning and addition give

$$X \sim \text{Pois}(dR + \eta), \qquad \widehat{R}_{\text{naive}} = \overline{X}, \qquad \widehat{R}_{\text{corr}} = \frac{\overline{X} - \eta}{d}, \quad \text{Var}(\widehat{R}_{\text{corr}}) = \frac{dR + \eta}{nd^2}. \tag{49}$$

The correction requires externally known $0 < d \leq 1$ and $\eta \geq 0$. Its variance inflation relative to ideal labeled sampling is $(dR + \eta)/(d^2 R)$; missing links alone yield $1/d$. Negative corrected estimates

are retained, without clipping. Dependence, missing parents, state-dependent ascertainment and estimated correction parameters are not covered.

The following analytic extension was added during reporting; it did not select the simulation grid or change any outcome.

Proposition 15 (Testing with a known observation channel)**.** *Under eq. (49), testing $R = 1 - \epsilon$ against $R = 1 + \epsilon$, $0 < \epsilon < 1$, with worst-case error at most $1/4$ requires and suffices, up to absolute constants, to use*

$$n = \Theta\left(\frac{d + \eta}{d^2\epsilon^2}\right).$$

This statement assumes independent known parent opportunities and known d, η; it is not a passive-trace identifiability guarantee.

Proof. Let $\lambda_\pm = d(1 \pm \epsilon) + \eta$. Threshold the corrected mean at one. Chebyshev gives worst-case error at most $\lambda_+/(nd^2\epsilon^2)$, so $n \geq 4\lambda_+/(d^2\epsilon^2)$ suffices. Conversely, $D(\mathrm{Pois}(\lambda_-)\|\mathrm{Pois}(\lambda_+)) \leq (\lambda_+ - \lambda_-)^2/\lambda_+ = 4d^2\epsilon^2/\lambda_+$, using $\log x \leq x - 1$. Pinsker's inequality then gives total variation strictly below $1/2$ when $n < \lambda_+/(8d^2\epsilon^2)$, so equal-prior average error is strictly above $1/4$, and the worst-case error is no smaller. Since $d + \eta \leq \lambda_+ \leq 2(d + \eta)$, these bounds give the stated rate. □

Unknown-channel nonidentifiability. The subcritical world $(R, d, \eta) = (0.8, 0.5, 0.3)$ and the supercritical world $(1.2, 0.5, 0.1)$ both induce exactly $\mathrm{Pois}(0.7)$ observations. Their joint observed distributions coincide for *every* n. Consequently no procedure using these counts alone can distinguish their criticality with worst-case error below $1/2$. Additional samples cannot repair an unidentified observation channel. This explicit construction demonstrates possibility, not prevalence in agent logs.

Fixed-grid numerical study. We cross $R \in \{0.8, 1, 1.2\}$, $d \in \{0.25, 0.5, 1\}$, $\eta \in \{0, 0.2, 0.6\}$ and $n \in \{100, 500, 2000\}$, using 2,000 replicate datasets per cell (162,000 total). Each retained total is sampled from its exact sufficient-statistic distribution $\mathrm{Pois}(n(dR + \eta))$. We report bias, RMSE, variance and point-decision errors; $R = 1$ cells instead report above-one fractions, without labeling either side erroneous. Classification uses exact decimal thresholds to avoid floating-point tie errors. Wilson intervals describe Monte Carlo uncertainty only, not simultaneous coverage, model uncertainty or confidence-qualified testing. These numerical point decisions need not meet the proposition's conservative finite-sample bound.

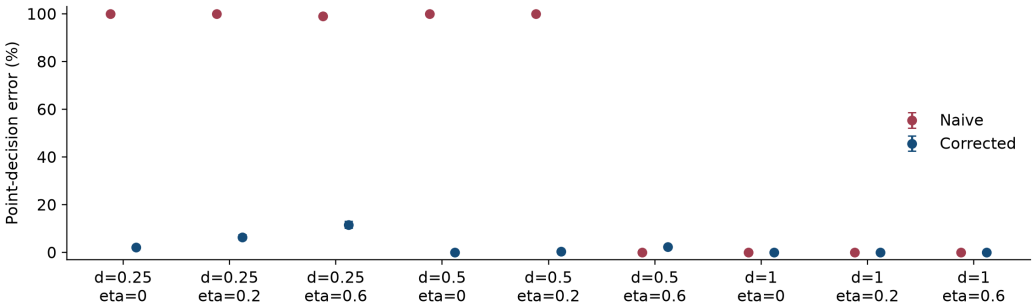

Figure 3: All nine observation-channel combinations at the prespecified $R = 1.2, n = 500$ display slice. Each point uses 2,000 datasets and a nominal Wilson 95% interval. Spurious links can offset missing links in a point decision without making the naive estimator unbiased. Correction assumes known channel parameters. All other 72 cells remain in the local all-cell CSV.

With $R = 1.2, d = 0.5, \eta = 0, n = 500$, the naive rule misclassifies all 2,000 replicates (95% interval $[99.81, 100]\%$), while correction misclassifies 2/2,000 (0.10%, $[0.03, 0.36]\%$) and doubles analytic estimator variance. With $R = 0.8, d = 1, \eta = 0.6, n = 500$, naive error is likewise 100% ($[99.81, 100]\%$), versus 0/2,000 corrected errors ($[0, 0.19]\%$) and variance inflation 1.75. Both anchors were specified before collection. They quantify the observation model's consequences, not the performance of a practical parentage estimator.

Provenance and unresolved empirical requirements. The local model-limits package (version 1, 23 September 2026) contains the frozen design/code, every raw replicate, both all-cell CSVs, figures, and single-run analysis receipts. It uses no provider calls, tools or external actions and adds USD 0 to the combined accounting. These studies address quantitative sensitivity, not measured soft-repair calibration, specialized-baseline superiority, a faithful public-task emulator, independent judge validity, real monitoring throughput, or unavailable historical response archives. Those empirical requirements remain open.