The Assurance Freshness Gap: Risk-Bounded Revalidation and Queue-Aware Execution Control for Tool-Using AI Agents

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: https://doi.org/10.68050/JAMS.2026.769

Received: 02 October 2026; Accepted: 12 October 2026; Published: 20 October 2026

ABSTRACT

Tool-using artificial intelligence agents increasingly act on evidence that can change between verification and execution. Existing output-focused assurance and permission gates do not by themselves establish that previously checked preconditions remain valid at action time. This paper develops the assurance freshness gap, a temporal assurance framework combining signed evidence receipts, private and shared-source change hazards, probabilistic revalidation, queue-sensitive risk and conditional atomic commit. Under independent baseline-error and generating-shock assumptions, the paper derives an exact invalid-execution probability and evidence lifetime; a conservative union-bound certificate applies to enumerated dependent events with defensible marginal bounds. A stationary M/M/1 authorization queue yields a 5%-budget utilization ceiling of 85.93% under explicit illustrative parameters. The revision extends the analysis to nonstationary hazards and general M/G/1 service-time distributions. At identical 85% utilization and mean service capacity, theoretical invalid-execution risk is 2.99% with deterministic service, 4.73% with exponential service and 9.48% with high-variance hyperexponential service. Independently replicated synthetic queue simulations corroborate these transform predictions; age-dependent hazard and uncertain-hazard calculations further show how nominally acceptable mean risk can breach the target. The paper provides executable controller pseudocode, formal proofs, sensitivity results and reproducible scripts, while distinguishing conditional probabilistic guarantees from atomic state-transition protection. The evaluation is entirely synthetic and does not establish real-world safety effectiveness.

Keywords: agentic AI; evidence freshness; time-of-check-to-time-of-use; runtime assurance; queueing theory; transactional verification

1. INTRODUCTION

Tool-using language-model agents can retrieve documents, call application programming interfaces, inspect database records, and initiate consequential transactions. ReAct and Toolformer illustrate the increasing coupling of reasoning with external actions [1,2]. This coupling creates a temporal vulnerability: a statement may be accurate when retrieved yet false when an agent acts upon it. A payment authorization may be revoked, a record edited, an inventory unit sold, or a deployment policy changed during the interval separating validation from execution. The relevant question is therefore not only whether evidence was correct, but whether it remains fit for action at commit time.

Classical computer security distinguishes time of check from time of use, while distributed systems theory addresses ordering and concurrent operations [4-6]. Recent security benchmarks further demonstrate that tool-connected agents operate over potentially untrusted data [3]. Nevertheless, provenance, content correctness, policy authorization, and temporal freshness are different control properties. Cryptographic integrity can identify a document version without proving that its contents are true or current. A human approval can confirm an action without guaranteeing that the reviewed preconditions have not changed.

Related contributions by Tan (2026) develop policy-gated runtime assurance [10], organizational verification debt [11], cascading error control in multi-agent networks [12], and the unreliability of treating correlated votes as independent evidence [13]. Those studies principally address autonomy control, accumulated

verification weakness, network propagation, and information dependence. The present paper isolates an additional failure mode: evidence becomes stale during the delay between its verification and an agent's external action. Its novelty is the composition of explicit time-dependent invalidation models, dependency-aware evidence receipts, queue-induced decay, and conditional atomic commit.

The term assurance freshness gap denotes temporal invalidation of a verified action precondition between a trusted check and the linearization point of the protected state transition. This definition does not claim priority over established time-of-check-to-time-of-use (TOCTOU) security: the contribution lies in deriving risk and queue-capacity consequences for autonomous execution under an explicit threat model.

The research addresses four questions. RQ1: How can the probability that an action rests on invalid preconditions be expressed as a function of evidence age and shared change hazards? RQ2: What conservative certificate remains possible when error events are dependent? RQ3: When does a nominally stable authorization queue violate a prescribed mean-risk budget? RQ4: Which software controls are needed to connect the mathematical certificate to actual transactions?

Three contributions follow: (i) an exact, assumption-conditional freshness equation and dependence-robust bound; (ii) a closed-form authorization capacity criterion obtained by integrating queueing and temporal reliability; and (iii) a reference execution protocol separating information checks, risk assessment, revalidation, and atomic or compensating actions. Numerical results are deliberately labeled synthetic and serve to verify calculations and sensitivity rather than to imply operational effectiveness.

2. RELATED WORK AND THEORETICAL POSITIONING

2.1 Tool-using agents and their assurance surface

The performance advantages of tool use arise partly from delegating factual retrieval or computation to external systems [1,2]. The same interface expands attack and failure surfaces. AgentDojo explicitly studies prompt injection against agents processing untrusted tool outputs [3]. Such attacks may cause instruction hijacking, but a separate, non-adversarial failure can occur when the agent follows a legitimate instruction with previously valid, now obsolete evidence. The latter persists even under perfect prompt-injection defense.

2.2 Security and distributed-systems foundations

Saltzer and Schroeder's protection principles motivate narrow privilege boundaries and complete mediation [4]. Lamport's account of event ordering shows why a timestamp alone does not necessarily establish a cross-system ordering relation [5]. Herlihy and Wing's linearizability criterion provides the relevant abstraction for reasoning about the instant at which a covered state transition takes effect [6]. The proposed gate treats trusted evidence versioning and a compare-and-commit operation as software-system properties, not properties that a language model can infer from natural language.

2.3 Governance, verification capacity, and the gap in the literature

NIST's AI Risk Management Framework and its generative AI profile establish useful risk identification and control principles [8,9], without prescribing the particular freshness formula developed here. Tan's policy-gated runtime assurance [10] offers a relevant control context; verification debt [11] explains persistent underinvestment; spectral cascade control [12] addresses error transmission; and the independence-illusion analysis [13] motivates explicit representation of shared source dependencies. The present contribution is narrower and complementary: it formalizes the aging of checks between evidence acquisition and transaction commitment and quantifies the cost of waiting for approval.

3. METHODOLOGY AND SYSTEM MODEL

3.1 Design-science and analytical approach

The research follows a formal design-science sequence: identify the temporal execution failure, define an evi-

dence receipt and threat model, derive mathematical risk constraints, design a transaction gate, and evaluate analytically and by synthetic Monte Carlo verification. It includes no human participants, confidential operational logs, production deployment, or observed system outcomes. Its equations are new modelling proposals assembled from established probability and queueing primitives; empirical calibration is left for future research.

3.2 Evidence bundles, clocks, and invalidation events

Let an action A require m necessary evidence predicates. For each item i, an accountable verifier records source identity, content hash or object version, verification time t_i, claim scope, and estimated false-accept probability ε_i. The action attempt occurs at time t_{act}, and evidence age is given by Equation (1). A receipt is considered invalid once a material invalidating change occurs, even if subsequent state transitions could incidentally recreate an equivalent value; rechecking or version comparison is needed to restore assurance. For cross-domain clocks, ages require a trusted monotonic-clock or conservative uncertainty interval; a wall-clock string alone is not a trustworthy elapsed-time measurement.

$$a_i = t_{\text{act}} - t_i, \qquad a_i \geq 0 \tag{1}$$

Let a baseline check pass despite a false precondition with probability ε_i. Define Q_0, the probability that all baseline checks are valid, under the temporary assumption that their false-accept events are independent. This assumption is relaxed in Section 4.2.

$$Q_0 = \prod_{i=1}^{m} (1 - \varepsilon_i) \tag{2}$$

Material changes are represented by private source-change point processes and, where appropriate, shared group shocks. In the exact model, independent Poisson processes generate private and common shocks, but receipt failures resulting from one common shock are correlated. A group g is defined so that a common shock invalidates every affected member whose verification precedes the shock, and hence at least one necessary predicate if a shock occurs after the earliest group verification. Thus group exposure is the age of the oldest affected receipt. Independence of the underlying shock processes must not be confused with independence of the resulting predicate failures. Equation (3) sums shock exposures once per source rather than once per correlated receipt.

$$H(\mathbf{a}) = \sum_{i=1}^{m} h_i \, a_i + \sum_{g \in \mathcal{G}} h_g \max_{i \in g} a_i \tag{3}$$

This group formulation is intentionally conservative about a common event affecting any necessary member. It is not an estimate of pairwise correlation and cannot be recovered from agent votes alone. Nonstationary hazards may be accommodated by replacing each constant hazard-times-age term with its integrated intensity, subject to sound upper estimates.

Table 1. Principal notation and operational interpretation.

Symbol	Definition	Interpretation / units
A; m	Proposed action; number of necessary claims	Execution decision; count
t_i; a_i	Verification timestamp; elapsed age	Clock time; minutes
ε_i; Q0	Baseline false-accept rate; joint validity	Probabilities
h_i; h_g	Item and group invalidation hazards	Events per minute
H(a)	Cumulative invalidation exposure	Dimensionless
δ	Maximum admissible probability of invalid execution	Probability

λ; μ; W	Queue arrival rate; service rate; sojourn time	Per minute; per minute; minutes
C_invalid	Consequence of executing with invalid preconditions	Application-defined cost
S; L_S(s); SCV	Service time; Laplace transform; squared coefficient of variation	Minutes; dimensionless; dimensionless
κ; $\bar{h}$; α	Age-hazard ramp; conservative upper hazard; μ − λ	Per minute squared; per minute; per minute

3.3 Threat boundaries and certificate assumptions

The stochastic results require a complete inventory of necessary action predicates and shared dependencies, receipt binding to a precisely scoped action, trustworthy verification and commit times, and defensible source-change and verifier-error estimates. Exact Equation (4) assumes independence between baseline false-acceptance events and the underlying independent private and group shock processes. Receipt outcomes within a dependency group are not independent, because a single common shock can invalidate several receipts. For Equation (5), only separately justified marginal event bounds are required; arbitrarily dependent failures are allowed among the enumerated events. An adaptive attacker can deliberately invalidate state after a check, defeating exogenous Poisson modelling. Such cases require transaction-level deterministic controls, and if neither those nor credible event bounds are available, the gate must abstain. Clock uncertainty must be added to age conservatively.

3.4 Proposed freshness-aware architecture

Figure 1 shows a five-stage path. An agent proposes a scoped action; a trusted verifier records source-bound evidence receipts; a freshness gate checks the action-specific probability or conservative bound; a commit guard rechecks versions or obtains a transaction lock; and a logging component records the decision and observed conflicts. A model-generated confidence score may inform prioritization but cannot substitute for receipts or machine-enforced policy.

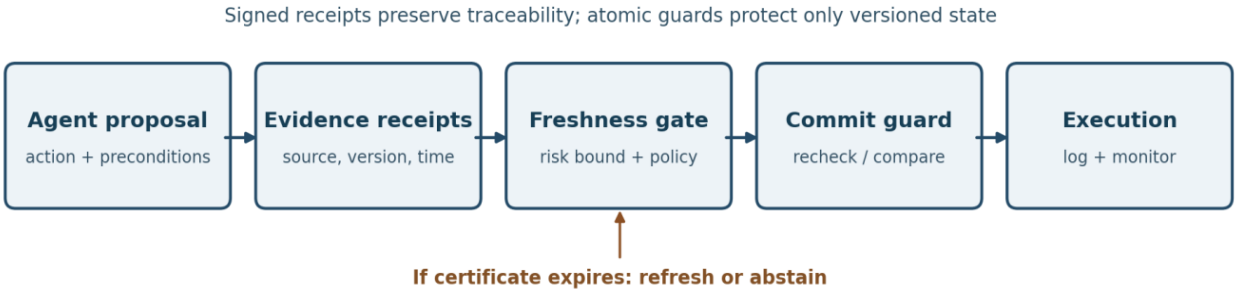


Figure 1. Reference architecture for freshness-aware tool execution. The abort or refresh branch is mandatory when the admissible certificate cannot be established.

4. MATHEMATICAL RESULTS

4.1 Proposition 1: Exact independent-shock invalidity

Proposition 1. Suppose all m baseline checks and all item and group invalidation processes are mutually independent, with the latter following homogeneous Poisson laws at the specified rates, and all necessary predicates must hold at action time. Then the probability that at least one precondition is invalid when the action is attempted is Equation (4).

$$P_{\text{bad}}(\mathbf{a}) = 1 - Q_0 \exp[-H(\mathbf{a})] \tag{4}$$

Proof. The baseline probability that all predicates are valid is Q_0. A Poisson process with cumulative exposure x experiences no invalidating event with probability exp(−x). Independence multiplies the no-event probabilities over private and group processes, yielding Q_0 exp[−H(a)]. The complement is the probability that at least one necessary precondition fails. The result is an assumption-conditional reliability identity, not a distribution-free safety guarantee.

4.2 Proposition 2: Dependence-robust conservative certificate

Proposition 2. If each baseline false-accept probability is upper-bounded by ε_i and each private or group invalidating-event probability is no larger than its integrated-hazard bound, the union inequality provides Equation (5) without assuming independence among the events. Dependence and common-cause failures therefore cannot invalidate this upper bound, although unknown hazards, omitted preconditions, or compromised verifiers can.

$$P_{\text{bad}}(\mathbf{a}) \leq \min\left\{1, \sum_{i=1}^{m} \varepsilon_i + H(\mathbf{a})\right\} \tag{5}$$

Proof. Apply the probability union inequality to the collection of baseline false-accept events and all private and group invalidation events. For any nonnegative x, 1−exp(−x) is no larger than x. Substitute the corresponding private and group exposures and cap the result at one. Define the operational estimated invalidity certificate as a calibrated bound no smaller than the right-hand side of Equation (5), with conservative hazards and check errors. A controller may authorize probabilistic execution only if this certificate satisfies Equation (6).

$$\hat{P}_{\text{bad}}(\mathbf{a}) \leq \delta \tag{6}$$

The adjective “robust” applies only to arbitrary dependence among enumerated events with defensible marginal upper bounds. It does not imply immunity to arbitrary adversarial behavior or unobserved event classes.

4.3 Proposition 3: Evidence lifetime for synchronized receipts

Proposition 3. If all items and groups are verified simultaneously, every age equals a common delay τ, all exact-model assumptions hold, and δ is at least the baseline invalidity $1-Q_0$, the largest admissible delay is Equation (7), provided the aggregate hazard is positive. A negative numerator means that the requested risk level is infeasible even at zero delay; a zero aggregate hazard makes temporal expiration irrelevant within this model.

$$\tau_{\max} = \frac{\ln\left(\frac{Q_0}{1-\delta}\right)}{\sum_i h_i + \sum_g h_g} \tag{7}$$

Proof. Substitute equal ages into Equation (4), require invalidity no larger than δ, and solve the resulting exponential inequality. The formula exhibits the central assurance freshness gap: even an initially reliable evidence bundle has a finite useful lifetime when potentially invalidating changes continue.

4.4 Proposition 4: Queue-induced assurance decay

Consider a stationary first-come, first-served M/M/1 authorization server. An evidence certificate is sealed immediately before a request enters the queue, then an action is released only after authorization service. Assume Poisson arrivals λ, exponential authorization service times with rate μ, λ < μ, and independent, exponentially distributed invalidation time with hazard h. In this special one-bundle model the certificate has

baseline error ε and initial age zero. Standard queueing theory gives the distribution of total queue sojourn time W in Equation (8) [7].

$$\Pr(W > w) = \exp[-(\mu - \lambda)w] \tag{8}$$

Integrating certificate survival over that random sojourn yields Equation (9). This is a mean risk across arrivals, not an upper bound for every individual action, and it excludes additional post-service propagation or network delays.

$$\mathbb{E}[P_{\text{bad}}] = 1 - (1 - \varepsilon)\frac{\mu - \lambda}{\mu - \lambda + h} \tag{9}$$

Proposition 4. Suppose the target is an average invalid-execution risk no greater than δ, with $\varepsilon < \delta < 1$ and $h > 0$. Then the request arrival rate must satisfy Equation (10). Feasibility additionally requires that its right-hand side is nonnegative. Thus the conventional queue stability condition $\lambda < \mu$ is necessary but not sufficient for this mean-risk target.

$$\lambda \leq \mu - \frac{(1 - \delta)h}{\delta - \varepsilon} \tag{10}$$

The proof appears in Appendix A. Since the constraint holds only in stationary M/M/1 operation, real implementations should instead enforce the per-action certificate in Equation (6), using measured ages, while using Equation (10) for preliminary capacity planning. High-priority actions, bursty arrivals, multi-server pools, and correlated service delays require separate modelling.

4.5 Proposition 5: Per-action risk exceedance under a mean-safe queue

A further distinction is between the average invalid-execution rate and the share of individual requests whose conditional invalid-execution probability exceeds δ. For $\delta > \varepsilon$ and positive h, Equation (11) defines the greatest admissible receipt age. Since the queue sojourn has the exponential distribution in Equation (8), the tail probability of exceeding this per-action risk threshold is exactly Equation (12). This is the fraction of action attempts whose calculated risk is too high, not the fraction of actions that actually fail.

$$\tau_\delta = \frac{1}{h}\ln\left(\frac{1 - \varepsilon}{1 - \delta}\right) \tag{11}$$

$$\Pr(P_{\text{bad}} > \delta) = e^{-(\mu - \lambda)\tau_\delta} \tag{12}$$

The consequence is operationally important: a system may satisfy an aggregate mean-risk constraint while exposing many particular actions to conditional risks above that same threshold. A controller must therefore evaluate actual receipt ages at its commit gate. Rejection, revalidation, or protected snapshot execution changes the admission process and cannot be evaluated by the uncontrolled M/M/1 formula without a new queue model.

4.6 Revalidation under bounded computation and changing clocks

For each group, let M denote the oldest receipt age and let B denote the oldest age among its members excluding the item to be refreshed; these quantities are indexed by group and, for B, by item in Equation (13). If one evidence item i could be revalidated instantaneously, and all other certificate ages were fixed, the reduction in the exposure function H is Equation (13). The group term is positive only if item i is a maximally old member in the relevant shared-dependency group; refreshing another copy of the same shared source may yield little benefit. Ties require careful group evaluation. The expression is an exposure reduction, not directly a probability reduction.

$$\Delta_i = h_i a_i + \sum_{g:\, i \in g} h_g \left(M_g - B_{g,i}\right) \tag{13}$$

Under zero-duration rechecks and known costs c_i, let S denote the chosen set of items to revalidate. Choosing a minimum-cost refresh set subject to a target risk has the generic discrete formulation in Equation (14). When group overlap, nonzero recheck duration, concurrency, and baseline revalidation error are present, optimal scheduling becomes more complicated. A simple ratio of exposure reduction to verification time should therefore be treated as a heuristic, not an optimality theorem.

$$\min_S \sum_{i \in S} c_i \quad \text{s.t.} \quad \hat{P}_{\text{bad}}\left(\mathbf{a}^{(S)}\right) \le \delta \tag{14}$$

A safe controller evaluates the resulting action-time ages, including time consumed by all scheduled checks, before authorization. If a resource exposes reliable versions, the preferred control is conditional compare-and-commit or an equivalent atomic transaction. For state outside the transaction's scope, probabilistic or pessimistic locks and explicit compensation policies remain necessary.

4.7 Nonstationary hazards and distinct dependence mechanisms

Proposition 6 (time-varying private and common shocks). Let the private and group point processes be mutually independent nonhomogeneous Poisson processes with deterministic nonnegative intensity functions h_i(u) and h_g(u), measured in inverse minutes. Under the baseline independence and group-effect assumptions of Proposition 1, cumulative exposure takes the integrated form in Equation (15), where t_a is the action time. A common event is counted only once in the group integral, beginning when the oldest affected receipt was verified.

$$H_{\text{ns}}(\mathbf{a}) = \sum_{i=1}^{m} \int_{t_i}^{t_a} h_i(u)\, du + \sum_{g \in \mathcal{G}} \int_{\min_{i \in g} t_i}^{t_a} h_g(u)\, du \tag{15}$$

With independent baseline checks, the exact probability of at least one invalid necessary predicate is Equation (16). Its proof follows by multiplying non-event probabilities exp[−∫h(u)du] for the independent generating processes and complementing joint survival. Dependence between resulting receipt invalidities is already represented by the shared process. Equation (16) becomes an upper certificate if its baseline-validity factor and each integrated intensity are replaced by valid conservative bounds. The inequality of Equation (5) remains available even when the generating events themselves are dependent, provided their marginal bounds hold.

$$P_{\text{bad}}^{\text{ns}}(\mathbf{a}) = 1 - Q_0 \exp[-H_{\text{ns}}(\mathbf{a})] \tag{16}$$

This distinction is central: source dependence concerns how many receipts rely on the same state, whereas event dependence concerns whether different underlying invalidations co-occur. A private-source hazard and a shared-group hazard cannot be estimated simply from pairwise agreement among agents. If group membership is uncertain, unmodelled common causes remain an uncovered risk, not a reason to treat Equation (16) as distribution free.

4.8 General service-time distributions: an M/G/1 extension

Proposition 7 (queue-sensitive mean freshness risk). Consider a stationary first-come, first-served M/G/1 queue with Poisson arrival rate λ, independent and identically distributed nonnegative service time S, finite mean, and arrival intensity ρ below one. The time W is total time in system, including both waiting and service; evidence is sealed at arrival. Define the service-time Laplace transform L_S(s) and traffic intensity by Equation (17).

$$\mathcal{L}_S(s) = \mathbb{E}[e^{-sS}], \qquad \rho = \lambda\mathbb{E}[S] < 1 \tag{17}$$

The Pollaczek-Khinchine transform for stationary waiting time, multiplied by the independent service-time transform, yields the total-sojourn transform in Equation (18), for h > 0. The expression holds under stationary FCFS M/G/1 assumptions and does not rely on exponential service [7].

$$\mathcal{L}_W(h) = \frac{(1-\rho)h\mathcal{L}_S(h)}{h - \lambda + \lambda\mathcal{L}_S(h)} \tag{18}$$

If baseline validity is independent of queue time and exogenous invalidation has constant hazard h, average invalid-execution risk equals Equation (19). Taking S exponentially distributed with rate μ reduces Equation (18) to (μ−λ)/(μ−λ+h), recovering Equation (9). With other service distributions, the 85.9% illustrative capacity boundary does not generally hold. The service-time squared coefficient of variation (SCV) is zero for deterministic service, one for exponential service, and four in the hyperexponential stress test. The formula is exact within each stipulated queue model, but is not a per-action authorization guarantee.

$$\overline{P}_{\text{bad}}^{\text{M/G/1}} = 1 - (1-\varepsilon)\mathcal{L}_W(h) \tag{19}$$

4.9 Age-dependent hazards and parameter uncertainty

A second, complementary departure keeps exponential queue sojourns but allows the conditional invalidation hazard to increase with evidence age w. Equation (20) uses an illustrative affine hazard, with h0 in min^{-1} and κ in min^{-2}. This specifies risk as a function of evidence age rather than assuming the hazard changes with the wall-clock time in the same way across all requests.

$$h(w) = h_0 + \kappa w, \qquad H(w) = h_0 w + \frac{\kappa w^2}{2} \tag{20}$$

For the stationary M/M/1 sojourn rate α = μ − λ, integrating the corresponding age-dependent survival curve produces Equation (21). It reduces to Equation (9) when κ equals zero. We evaluate the one-dimensional integral by deterministic numerical quadrature rather than pretending an exponential-hazard closed form remains valid.

$$\overline{P}_{\text{bad}}^{\text{ramp}} = 1 - (1-\varepsilon)\alpha \int_0^\infty e^{-(\alpha+h_0)w - \kappa w^2/2}\, dw \tag{21}$$

Finally, when only a defensible upper hazard $\bar{h}$ is available, monotonicity of the transform yields Equation (22). Under an independent bounded constant hazard, the worst-case M/M/1 mean-risk certificate is obtained by substituting $\bar{h}$ into Equation (9). This is a bound conditional on the stated interval being valid, not a statistical confidence guarantee about an uncalibrated source. For a conservative per-action gate, upper bounds must apply to the actual outstanding receipts, including clock and retry delays.

$$\overline{P}_{\text{bad}}(h) \le \overline{P}_{\text{bad}}(\bar{h}), \qquad 0 \le h \le \bar{h} \tag{22}$$

5. EVALUATION DESIGN

5.1 Transparent synthetic calibration

The principal experiment tests the closed-form queue result with a fully specified distribution. It is synthetic by construction and neither emulates a complete live queue nor measures deployment performance. Baseline false-accept probability is ε = 0.005; invalidation hazard h = 0.08 per minute; authorization service rate μ = 12 requests per minute; and the mean-risk budget δ = 0.05. Six offered utilizations are used: 0%, 50%, 75%, 85%, 90%, and 95%. For each utilization, independent equilibrium sojourn times are drawn from an exponential distribution with rate μ−λ. Each episode is labelled valid with Bernoulli probability (1−ε)exp(−hW).

There are 200,000 independently generated episodes per utilization level, for 1,200,000 model episodes in total, with pseudorandom seed 20261012 (NumPy default_rng). The zero-utilization row represents the limiting case of a tagged request without background arrivals, not a steady stream of arrivals at zero rate. The simulation compares observed invalid-execution fractions with Equation (9). Wilson 95% intervals quantify Monte Carlo sampling variation, not structural-model uncertainty. The companion script and CSV disclose the complete reproduction procedure.

5.2 Analytical stress tests and illustrative bundle

For sensitivity analysis, the invalidation hazard is varied across 0.04, 0.08, and 0.16 per minute, holding service capacity and baseline error fixed. A separate three-evidence worked example uses ages (2, 1, 0.5) minutes; baseline errors (0.003, 0.004, 0.005); private hazards (0.01, 0.02, 0.04) per minute; and a shared group {1, 2} with hazard 0.015 per minute. These values are illustrative assumptions, not inferred from an organization.

5.3 Robustness protocol and statistical dependence

The revised study varies service-time variability while holding Poisson arrival rate, mean service capacity $\mu = 12\ \text{min}^{-1}$, invalidation hazard $h = 0.08\ \text{min}^{-1}$ and baseline error $\varepsilon = 0.005$ fixed. Three FCFS M/G/1 systems are compared: deterministic service ($C^2S = 0$), exponential service ($C^2S = 1$), and a balanced two-phase hyperexponential distribution ($C^2S = 4$). In the final model, the two mixture probabilities are $p = (1 + \sqrt{0.6})/2$ and $1 - p$, with component rates $2p\mu$ and $2(1 - p)\mu$ respectively. All three distributions have identical mean service time $1/\mu$. Theoretical risk is computed from Equation (19), independently of the simulations.

To corroborate the transform calculations at 85% utilization, a discrete-event queue simulation processes 24 independent replications per service distribution, each with 25,000 warm-up jobs followed by 75,000 retained jobs. This gives 1.8 million evaluated actions per service type, or 5.4 million total. Arrivals are exponential interarrival samples; completion times are recursively constructed by a single-server first-come, first-served rule; action validity is then sampled conditionally on the realized sojourn. Importantly, successive actions in a queue are dependent. Consequently, reported 95% intervals are Student-t intervals calculated from the 24 independent per-replication rates, not binomial Wilson intervals calculated over correlated jobs. The original independent-episode experiment in Section 5.1 is retained as a separate unit check.

The nonstationary stress test holds utilization at 85% and evaluates Equation (21) with $\kappa = 0$, 0.02 and 0.06 min^{-2}. Hazard uncertainty is examined over $h \in [0.06, 0.10]\ \text{min}^{-1}$. These intervals and distributions are stipulated counterfactuals, not estimated operational uncertainty, and results should not be interpreted as effects measured in deployed AI agents.

6. ANALYSIS AND RESULTS

6.1 Queue risk rises before the stability limit

Table 2 displays the exact mean invalid-execution probability, Monte Carlo estimate, and Wilson interval. The simulation tracks the mathematical predictions closely; the maximum absolute discrepancy across six settings is below 0.00038. This agreement validates the implementation of the stipulated distribution but cannot establish that M/M/1 arrivals, exponential service, or constant hazards hold in real systems.

Table 2. Synthetic queue-risk experiment: exact theory versus 200,000 trials at each load.

Utilization	Analytical risk	Monte Carlo risk	Wilson 95% interval
0%	1.16%	1.16%	[1.117, 1.211]%
50%	1.81%	1.81%	[1.749, 1.865]%
75%	3.08%	3.06%	[2.984, 3.135]%

85%	4.73%	4.70%	[4.605, 4.790]%
90%	6.72%	6.72%	[6.615, 6.835]%
95%	12.21%	12.18%	[12.033, 12.320]%

At 85% utilization, the predicted mean invalid-execution risk is 4.734%, below the stipulated 5% limit; at 90% it is 6.719%, above the limit. Equation (10) yields an illustrative arrival ceiling of 10.311 requests per minute, corresponding to approximately 85.93% utilization. The queue itself remains mathematically stable up to, but not including, 100% utilization. Accordingly, a controller sized only for queue stability can have an unacceptable freshness-risk profile.

The per-request threshold analysis exposes an additional weakness of averaging. Equation (11) gives a permissible age of only 0.579 minutes, approximately 34.7 seconds. Applying Equation (12), 35.3% of arrivals at 85% utilization and 49.9% at 90% utilization have calculated per-action risks exceeding 5%. These are modelled threshold exceedances, not realized invalid-action frequencies. A policy that promises each action a 5% upper risk bound must therefore refresh or block a substantial share of requests, even at a nominally mean-safe utilization.

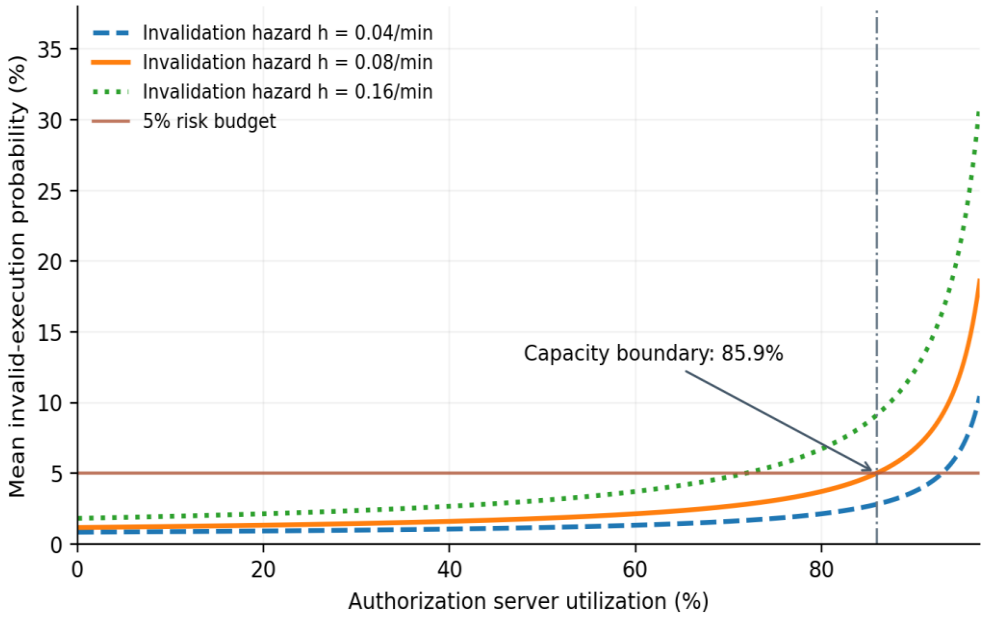


Figure 2. Analytical mean invalid-execution probability as authorization utilization increases, for three assumed hazards. The vertical marker shows the capacity boundary under the middle hazard and a 5% budget.

6.2 A shared-shock example

For the three-item bundle, Equation (3) produces an exposure of 0.09 and Equation (4) gives an invalid-execution probability of approximately 9.699%. The dependence-robust upper bound of Equation (5) is 10.2%. Hence neither certificate meets a 5% action-risk budget at the current ages. The gap between the exact value and conservative upper bound is expected and should not be construed as a measured safety margin. The group exposure correctly uses the oldest of its affected receipts, rather than double-counting a single shared invalidating event.

A targeted dependence sensitivity keeps all three private hazards, ages and baseline-error assumptions fixed while varying the shared group {1, 2} hazard from 0 to 0.015 to 0.030 min^{-1}. Exact invalidity from Equation (4) increases from 6.949% to 9.699% to 12.368%, whereas the dependence-robust union certificates from Equation (5) are 7.2%, 10.2% and 13.2%. These values illustrate how common-source exposure should be

counted once per generating shock; changing the shared hazard is not equivalent to imposing a specified Pearson correlation on verifier errors.

6.3 Comparative operational implications

The control comparison is analytical rather than an empirical contest between production policies. A fixed time-to-live can be adequate under uniform, well-bounded hazards but cannot differentiate high-risk and low-risk claims. A per-item age threshold can neglect common-source invalidation. A model-only confidence check does not constrain external state change. A dependency-aware freshness gate can use Equation (5) for per-action bounding and Equation (10) for capacity planning, but it still needs a trusted commit mechanism to close the final check-to-execution window.

Table 3. Threat-model coverage and the limits of different controls.

Failure mechanism	Primary proposed control	Residual exposure
Stale but intact document	Signed receipt + source-specific expiry	Unknown or nonstationary hazard
Shared source update	Dependency-group exposure bound	Undocumented dependency graph
Queue backlog	Risk-constrained admission + late check	Burst arrivals and service-model error
Concurrent write before action	Versioned compare-and-commit	Cross-system operations outside transaction
Malicious tool output	Least privilege + input boundary controls	Prompt injection not solved by freshness
False initial verification	Calibration + independent challenge	Correlated and unknown baseline failures

6.4 Beyond M/M/1: service variability changes the safety conclusion

Table 4 compares three service-time distributions at 85% utilization. Although mean authorization throughput and invalidation hazards are identical, service-time variability changes the distribution of evidence age. The deterministic system meets the illustrative 5% mean-risk budget by a comfortable margin, the exponential baseline barely meets it, and the high-variability hyperexponential system breaches it substantially. Replicated discrete-event results agree with the transform predictions within the reported replication-based 95% intervals. These are simulated model checks, not field-validation trials.

Table 4. Stationary M/G/1 sensitivity at 85% utilization. Estimates use 24 independent simulation replications per law; each law is given the same mean service time.

Service time (SCV)	M/G/1 exact risk	Queue simulation risk	95% replication interval
Deterministic (0)	2.99%	2.97%	[2.930%, 3.012%]
Exponential (1)	4.73%	4.70%	[4.613%, 4.791%]
Hyperexponential (4)	9.48%	9.47%	[9.176%, 9.764%]

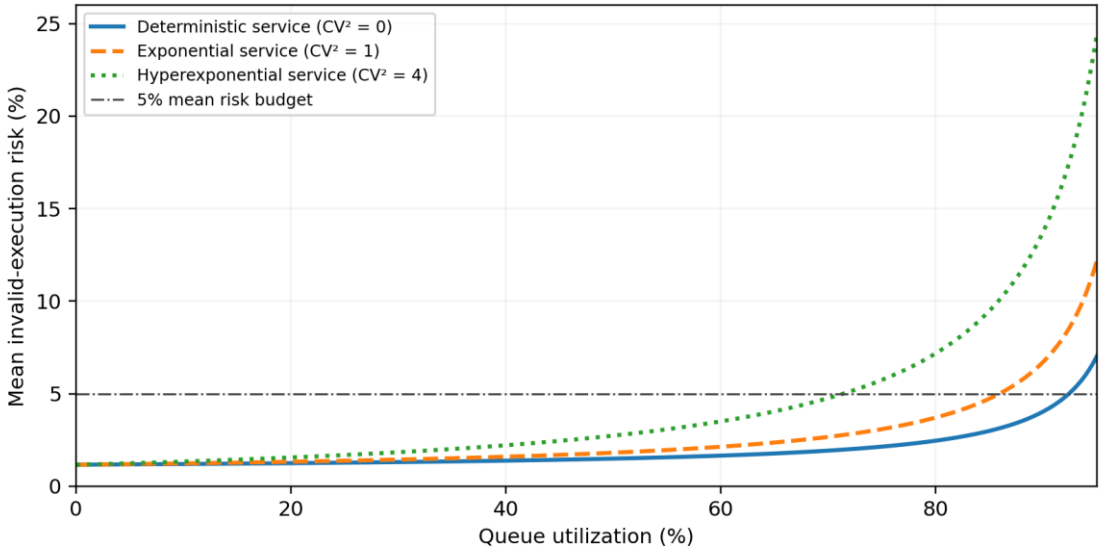


Figure 3. Exact mean risk from the M/G/1 transform across utilization and three service-time distributions. Unlike Figure 2, which changes only hazard, this figure varies service-time dispersion while fixing its mean.

The result should not be read as proving that high-variability tasks invariably produce greater risk for every possible service distribution. It is a counterexample to transferring an exponential-service utilization threshold unchanged across heterogeneous verification workloads. An additional workload-specific capacity boundary must instead be obtained by solving Equation (19), where a stationary model is empirically justified.

6.5 Nonstationary changes and uncertainty-aware safety margins

Table 5 demonstrates that mean-risk feasibility can reverse even with the queue unchanged. Under the age-ramping hazard, increasing κ from 0 to 0.02 min^{-2} raises theoretical risk beyond 5%; further increase raises it more. If the constant hazard is only known to lie in [0.06, 0.10] min^{-1}, the mean-risk estimate spans 3.710% to 5.737%. Therefore a nominal 4.734% result at h = 0.08 does not certify the budget under the entire hazard interval.

Stress condition at 85% load	Mean risk	Interpretation
Constant h = 0.06 min^{-1}	3.71%	Lower end of hypothetical hazard interval
Constant h = 0.08 min^{-1}	4.73%	Original calibration
Constant h = 0.10 min^{-1}	5.74%	Conservative upper-hazard case
Age ramp κ = 0.02 min^{-2}	5.26%	Initial hazard h0 = 0.08 min^{-1}
Age ramp κ = 0.06 min^{-2}	6.28%	Faster increase with certificate age

Table 5. Exact stationary M/M/1 risk under an assumed constant-hazard interval and deterministic quadrature for increasing age-dependent hazards. Values are counterfactual model calculations.

A worked capacity computation further answers the reviewer's request. With $\delta = 0.05$, $\varepsilon = 0.005$, h = 0.08 min^{-1} and μ = 12 min^{-1}, Equation (10) requires $\mu - \lambda \geq [(1 - 0.05)\times 0.08]/(0.05 - 0.005) = 1.688889$ min^{-1}. Thus $\lambda \leq 10.311111$ min^{-1} and utilization $\rho \leq 10.311111/12 = 0.859259$, or 85.93%. The requirement is an average-risk operating ceiling in the chosen M/M/1 model, not a bound on each realized request. At 85% utilization, mean sojourn is $1/(12 - 10.2) = 0.555556$ minutes, and average invalidity is $1 - 0.995 \times 1.8/(1.8 + 0.08) = 4.734\%$.

7. DISCUSSION

7.1 A time-sensitive definition of warranted autonomy

Autonomy should not be treated as an unchanging property of an agent or task. A delegated action may be warranted immediately following validation and unwarranted minutes later, even when the model's reasoning and tool permissions have not changed. The assurance freshness gap therefore makes temporal evidence validity an explicit part of autonomy policy. This augments the governance and runtime models of Tan [10,11] without reusing their empirical claims.

7.2 Security engineering and transactional scope

Version-bound compare-and-commit is a strong control when the resources being tested and changed are within one atomic transaction, as suggested by classical linearizability reasoning [6]. It does not automatically make a distributed workflow atomic: a bank transfer, customer email, and external API call may have distinct commit points. Sagas, compensation logic, idempotency tokens, and delayed execution are plausible responses, but this paper does not prove their sufficiency. For irreversible operations, the controller may require a newly verified condition, bounded delay, independent human approval, or abstention. These are complementary controls rather than interchangeable substitutes.

7.3 Why probabilistic assurance is conditional

An attacker can change a targeted record immediately after seeing a check. Such adaptive timing invalidates a stationary exogenous Poisson hazard. Likewise, permissions may be revoked instantaneously, service may become heavily bursty, and silent third-party updates may be missing from the dependency inventory. In those settings, the only honest certificate may be "not established." The approach is consequently a risk-accounting framework for bounded change processes, not a mathematical guarantee of safety against unconstrained attackers.

7.4 Engineering implementation and governance

An implementation should log evidence identifiers and versions, timestamps and uncertainty, action identifiers, hazard-model versions, risk bounds, refresh decisions, compare-and-commit outcomes, and any aborted transactions. Sensitive content should be minimized in logs, and signatures should be authenticated without exposing secret keys to the language model. Continuous calibration should compare forecast validity with observed changes, while fairness, accountability, and privacy review should assess whether stricter gates disproportionately delay important services. This is consistent with the operational risk-management emphasis of NIST [8,9].

For management decisions, total expected loss can be expressed by Equation (23), balancing invalid action consequences against delay and verification cost. This objective is descriptive until costs and probability distributions are measured; the paper does not claim that its proposed gate globally optimizes this expression. In Equation (23), c_b is the cost of an invalid action, c_d is delay cost per minute, and c_v is verification cost per request.

$$\mathbb{E}[L] = c_b \mathbb{E}[P_{\text{bad}}] + c_d \mathbb{E}[W] + c_v \tag{23}$$

7.5 An executable control protocol and its safety boundary

Algorithm 1 is a reference controller rather than a claim of a verified implementation. The agent cannot alter the trusted verifier, receipt issuer, hazard-policy register, monotonic timer or transaction coordinator. Risk certification is attempted only for inventoried predicates and bounded source-change processes. Time-of-check-to-time-of-use exposure cannot be eliminated by probability estimates alone: the final action must use

transactional version checks for all covered preconditions. Uncovered irreversible side effects are refused unless a separately specified containment and authorization strategy is available.

Step	Trusted controller operation
1	Receive proposed action A with immutable scope, idempotency key and required predicate list.
2	Fetch authenticated receipt for every necessary predicate; verify signature, object version, scope and dependency closure.
3	Calculate conservative elapsed ages including monotonic-clock uncertainty and anticipated commit latency.
4	If any required source lacks a calibrated upper hazard or verified commit mechanism, select protected snapshot execution or abstain.
5	Compute dependence-robust risk bound using upper baseline errors, private and shared-group exposure.
6	If bound exceeds δ, revalidate selected predicates, issue new signed receipts and recompute all ages; abort at retry limit.
7	Begin atomic transaction over all covered state; compare current source versions with receipt versions.
8	If any comparison fails, roll back, refresh and re-evaluate risk; never commit against a failed compare.
9	Recalculate residual risk for anticipated commit age; commit covered action atomically only when both version and risk gates pass.
10	Write audit log and final status; external non-transactional side effects must be separately gated or left unexecuted.

Algorithm 1. Freshness-aware compare-and-commit gate, with bounded retries and fail-closed handling of uncovered state.

A trusted source-version comparison can prevent the specific covered state from changing between verification and the associated atomic commit, provided that the compare and mutation share a linearization point. It does not prove that source information was initially correct, that unrelated predicates remain true, or that external APIs participate in the transaction [4-6]. Where global atomicity is unavailable, a formally specified saga or compensation procedure may reduce consequence but is not automatically equivalent to rollback. This distinction prevents a probabilistic freshness estimate from being mistaken for a deterministic distributed-systems guarantee.

8. LIMITATIONS AND RESEARCH AGENDA

First, the exact results assume independent stationary Poisson invalidations; bursty, self-exciting, seasonal, and adversarial changes warrant alternative survival processes or bounded worst-case analysis. Second, baseline false-accept parameters and source hazards are treated as externally known or conservatively bounded, where-

as learning them online under distribution shift requires calibration studies. Third, source-specific clocks and event ordering may be incomplete in asynchronous systems [5]. Fourth, the M/M/1 model is a one-server equilibrium idealization and does not capture priority scheduling, heterogeneity, cross-service retries, or non-Markovian verification tasks. Fifth, refreshing a receipt consumes time and may itself create new cross-item staleness; the discrete revalidation model makes no unsupported global-optimality claim. Finally, the synthetic evaluation checks internal mathematics rather than adversarial robustness, user outcomes, or deployment-level effectiveness.

A strong follow-up program should evaluate (i) change-hazard calibration with time-stamped enterprise audit logs; (ii) attacks that deliberately exploit check-to-commit latency; (iii) replayable distributed testbeds with deterministic version-conflict injection; (iv) priority and multiserver queue models; and (v) outcome-based comparisons of late revalidation, atomic snapshot execution, and compensating workflows. The primary

falsifiable question is whether per-action certificate calibration predicts observed invalid preconditions under realistic dependency drift and contested writes.

Sixth, the newly added variability scenarios remain mathematical counterfactuals. They demonstrate parameter sensitivity and queue-transform agreement under stipulated M/G/1 assumptions rather than yielding externally generalizable service-level agreements. Seventh, the 24-replication confidence intervals quantify synthetic queue-simulation variability only and do not capture hazard-estimation uncertainty or unmodelled system failures. Finally, the controller pseudocode has not undergone model checking, concurrency fuzzing, penetration testing or operational deployment.

9. CONCLUSION

Correct evidence at verification time is not sufficient evidence for correct action at execution time. This paper formalizes the assurance freshness gap through receipt ages, baseline verification errors, independent and shared invalidation hazards, conservative dependence-robust bounds, and risk-constrained action gates. The queueing result shows why mathematically stable authorization services may still operate outside a specified mean-risk budget, with an illustrative 5% threshold reached well before the usual stability limit. The architecture connects risk estimation to source receipts, revalidation, conditional commit, and accountable abstention. Its immediate value is a testable engineering framework, not a claim that existing agents have been made safe. Future empirical work must establish when the underlying hazards can be trusted and whether the proposed controls improve real-world execution reliability.

The robustness extensions sharpen the principal result: ordinary queue stability and even a mean-risk criterion can be insufficient for action-level safety, and a utilization ceiling calibrated to exponential service may not transfer to high-variability verification work. The resulting framework is a falsifiable assurance design with transparent assumptions, rather than a universal performance guarantee.

DECLARATIONS

Research design and data availability. This is an original theoretical and computational study using wholly synthetic, explicitly stated input parameters. The reproduction script and numerical output accompanying the manuscript provide all synthetic calculations. No proprietary, human-subject, or confidential data were used.

Ethics. No human participants or personal data were collected for this study; human-subject ethical approval was therefore not applicable to the reported computational work.

Funding and competing interests. The author is responsible for verifying and supplying any applicable funding acknowledgments or competing-interest disclosures before submission.

REFERENCES

1. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations. https://mlanthology.org/iclr/2023/yao2023iclr-react/
2. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36. https://doi.org/10.52202/075280-2997
3. Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramer, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems, 37, 82895-82920. https://doi.org/10.52202/079017-2636
4. Saltzer, J. H., & Schroeder, M. D. (1975). The protection of information in computer systems. Proceedings of the IEEE, 63(9), 1278-1308. https://doi.org/10.1109/PROC.1975.9939
5. Lamport, L. (1978). Time, clocks, and the ordering of events in a distributed system. Communications of the ACM, 21(7), 558-565. https://doi.org/10.1145/359545.359563

6. Herlihy, M. P., & Wing, J. M. (1990). Linearizability: A correctness condition for concurrent objects. ACM Transactions on Programming Languages and Systems, 12(3), 463-492. https://doi.org/10.1145/78969.78972
7. Kleinrock, L. (1975). Queueing systems, Volume I: Theory. John Wiley & Sons. ISBN 978-0-471-49110-1.
8. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
9. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
10. Tan, K. H. (2026). Runtime assurance for enterprise agentic AI systems: A policy-gated control model with quantitative autonomy-risk scoring. World Journal of Advanced Research and Reviews, 31(1), 512-522. https://doi.org/10.30574/wjarr.2026.31.1.1872
11. Tan, K. H. (2026). The coordination compression trap: A dynamic model of AI-enabled productivity, verification debt, and organisational resilience in global knowledge firms. Journal of Global Economics, Management and Business Research, 18(3), 256-278. https://doi.org/10.56557/jgembr/2026/v18i310956
12. Tan, K. H. (2026). Containing cascading AI errors: A spectral branching-process framework for verification allocation in multi-agent systems. International Journal of Progressive Research in Engineering Management and Science, 6(8), 374-383. https://doi.org/10.58257/IJPREMS54572
13. Tan, K. H. (2026). Five AI agents, one hidden source: The independence illusion in multi-agent decisions. Journal of Advanced Multidisciplinary Studies, JAMS Autumn Conference 2026, 164-178. https://doi.org/10.68050/JAMS.2026.489

Appendix A. Proof Details and Reproduction Logic

A.1 Dependence-free union inequality. Let the elementary adverse events include every baseline false acceptance and every invalidating private or shared source change since the relevant receipt time. The event that some necessary precondition fails is a subset of their union. No independence assumption is needed for the standard union inequality (Equation (24)). The bounds on individual event probabilities then give Equation (5).

$$\Pr\left(\bigcup_{j=1}^{k} B_j\right) \leq \sum_{j=1}^{k} \Pr\left(B_j\right) \tag{24}$$

A.2 Stationary authorization sojourn. In a stable M/M/1 first-come, first-served queue, the unconditional system time for an arriving job has exponential rate μ−λ [7]. If the evidence remains valid after time w with probability (1−ε)exp(−hw), the expected survival factor is the Laplace transform of that system-time distribution (Equation (25)).

$$\mathbb{E}[e^{-hW}] = \int_0^{\infty} e^{-hw}\,(\mu-\lambda)e^{-(\mu-\lambda)w}\,dw \tag{25}$$

Evaluating the integral establishes Equation (26), which substituted into the survival complement gives Equation (9). Solving the inequality $1-(1-\varepsilon)(\mu-\lambda)/(\mu-\lambda+h) \leq \delta$ for μ−λ gives Equation (10), requiring $\delta > \varepsilon$ and nonnegative feasible capacity.

$$\mathbb{E}[e^{-hW}] = \frac{\mu-\lambda}{\mu-\lambda+h} \tag{26}$$

A.3 Reproduction pseudocode. For each selected utilization ρ, set λ = ρμ and α = μ−λ. Repeat N times: draw W from the exponential distribution of rate α; draw U uniformly from [0, 1]; classify the episode as valid if U

< (1−ε)exp(−hW). Report the empirical invalid fraction, its Wilson binomial interval, and the analytical value from Equation (9). The exact implementation accompanies the manuscript.

A.4 Full derivation of synchronized evidence lifetime

Writing $r = \sum_i h_i + \sum g h_g$, simultaneous verification makes $H(a) = r\tau$. The exact risk constraint $1 - Q0 \exp(-r\tau) \le \delta$ is equivalent to $\exp(-r\tau) \ge (1 - \delta)/Q0$. If $Q0 < 1 - \delta$ the inequality fails already at $\tau = 0$. For $Q0 \ge 1 - \delta$ and $r > 0$, taking the natural logarithm and dividing by $-r$ (reversing the inequality) yields $\tau \le \ln[Q0/(1 - \delta)]/r$, which is Equation (7). If $r = 0$, no time-based restriction follows from this model; baseline validity still requires $Q0 \ge 1 - \delta$.

A.5 Explicit mean-risk capacity calculation

Set $\alpha = \mu - \lambda > 0$. From Equation (9), $1 - (1 - \varepsilon)\alpha/(\alpha + h) \le \delta$ implies $(1 - \delta)(\alpha + h) \le (1 - \varepsilon)\alpha$, hence $(\delta - \varepsilon)\alpha \ge (1 - \delta)h$. For $\delta > \varepsilon$, $\alpha \ge (1 - \delta)h/(\delta - \varepsilon)$. Replacing α by $\mu - \lambda$ gives Equation (10). If $\delta \le \varepsilon$ and $h > 0$, the target is infeasible. A nonnegative admissible arrival rate requires μ to be at least the derived minimum margin; otherwise the queue cannot meet the budget even for an isolated tagged request.

A.6 Independent generators versus correlated effects

Let J_g be the event that a common shock for group g occurs during the interval beginning with its oldest receipt verification. Under the group-effect definition, J_g invalidates at least one necessary group predicate and may invalidate several. The events of those receipt invalidities are therefore dependent even though the generating Poisson processes may be independent across groups. For each group, $\Pr(J_g) = 1 - \exp(-Hg)$. The exact no-shock product gives Equation (4) only when the underlying private and common generator processes, and baseline correctness, satisfy the independence assumptions. Under arbitrary dependence, the elementary union inequality bounds joint failure by the sum of valid marginal bounds; treating one J_g as several independently sampled event types is unjustified.

A.7 Derivation of the M/G/1 transform and hazard-ramp integral

For a stationary FCFS M/G/1 queue, let V denote stationary pre-service waiting time and S the arriving job's own service. PASTA and the standard Pollaczek-Khinchine transform give $E[\exp(-sV)] = (1 - \rho)s/[s - \lambda + \lambda L_S(s)]$. As the arriving service time is independent of the workload, $W = V + S$ and $L_W(s)$ is the product of their Laplace transforms, giving Equation (18). Conditional probability of unchanged evidence after W is $(1 - \varepsilon)\exp(-hW)$; taking expectations yields Equation (19). For $h(w) = h0 + \kappa w$, the cumulative age-exposure is the integral $\int_0^w (h0 + \kappa u)du = h0w + \kappa w^2/2$. With W exponential of rate α, integrating $(1 - \varepsilon)\exp[-H(W)]$ against its density $\alpha\exp(-\alpha w)$ yields Equation (21). All transforms characterize mean risk over arrivals rather than per-action upper risk bounds.