The Lyapunov Time of Delegated Work: A Closed-Form Predictability Horizon for AI Systems Under Probabilistic Error and Behavioral Drift

Bethel Kameni*

July 2026

Abstract

Chaotic systems have a predictability horizon: trajectories diverge exponentially at a rate given by the Lyapunov exponent, and forecasting beyond the reciprocal of that rate — the Lyapunov time — is futile. Weather admits roughly two weeks; the solar system, five million years. We show that AI-delegated work has a Lyapunov time, derive it in closed form, and prove that — unlike weather — part of its exponent is an engineering variable. From two axioms (per-step error is bounded away from zero because training is probabilistic; execution choices such as model switching contribute additional behavioral drift) the pipeline's deviation from its intended trajectory grows at exponential rate $\lambda = \ln(1/q)$, where $q = (1-\varepsilon)(1-\bar{\delta})$ is per-step retention. The exponent decomposes as $\lambda = \lambda_\varepsilon + \lambda_\delta$: an intrinsic component fixed by the model and an operational component set by how the work is executed. We prove that expected delivered value $W(n) = v\, n\, e^{-\lambda n}$ is uniquely maximized at the Lyapunov time

$$n^* = 1/\lambda \;\approx\; 1/(\varepsilon + \bar{\delta}),$$

so the optimal task quantum *is* the predictability horizon, and the universal $1/e$ operating point of an optimally loaded pipeline is definitional: the Lyapunov time is the e-folding time of degradation. The optimum is robust to per-step value accrual with costs and, via an absorbing Markov formulation, to abandoning independence. Part II proves λ_δ is controllable: exact $O(nk^2)$ optimization of routing policies, a structure theorem tying blocky execution to quasi-metric handoff costs, a closed-form swap threshold with a Lambert-W amortization horizon, and a perturbation-stability theorem showing handoff shocks are forgotten or frozen according to the contractivity of the successor's dynamics. We close with the Fault-Tolerance Threshold Conjecture for AI pipelines — the analogue of the quantum error-correction threshold theorem — proposing that verification below a critical cost fraction extends the Lyapunov horizon without bound, with recent million-step zero-error results as supporting evidence. All results are proven within the stated model; the exponent decomposition and the threshold structure are modeling assumptions about real pipelines, made empirically testable by the measurement protocol and falsifiable conjectures that close the paper.

*Mowa, a Blue Mountain Sky Technologies company. Houston, TX. Comments welcome.

1 Introduction

Every field that forecasts a dynamical system eventually learns its Lyapunov time. Nearby trajectories of a chaotic system separate as $e^{\lambda t}$; once separation reaches the tolerance of the application, prediction is over. The reciprocal $1/\lambda$ — the e-folding time — is the canonical predictability horizon: about two weeks for synoptic weather, millions of years for planetary orbits. The horizon is not a failure of effort; it is a property of the dynamics.

Applied AI in 2025–2026 is a forecasting field that has measured its horizon without deriving it. Long-horizon task completion is the organizing benchmark of agent evaluation: METR documents success falling from near-certainty on minutes-long tasks to under 10% on multi-hour tasks, with the 50%-reliability horizon doubling roughly every seven months [1]. The cause is undisputed — because models are trained by probabilistic methods, per-step error never reaches zero, and deviations compound: 99% per-step accuracy yields 36.6% success at 100 steps and 0.004% at 1,000 [3, 4]. Engineering programs attack the limit by decomposition and voting [2], consensus execution [3], and reliability instrumentation [5, 7]. What the field lacks is what dynamics gave meteorology: the exponent, its decomposition, and the closed-form horizon it implies.

This paper supplies them. The pipeline's per-step retention q defines a divergence rate $\lambda = \ln(1/q)$; we prove the value-maximizing task quantum is exactly the Lyapunov time $n^* = 1/\lambda$, that every optimally loaded pipeline therefore operates at precisely one e-fold ($1/e \approx 36.8\%$) of degradation, and — the consequential part — that the exponent splits into an intrinsic term λ_ε fixed by the model and an operational term λ_δ set by execution choices. Weather cannot lower its Lyapunov exponent. AI operators can, and Part II gives the exact theory of how.

Axioms. Two premises, both empirically grounded:

1. **Probabilistic error.** Training minimizes expected loss over a distribution; per-step error is bounded below: $\varepsilon > 0$ for every model trained this way [3, 4, 1].
2. **Behavioral drift from execution.** Switching models to economize tokens [12, 13], provider version changes [20], and context contamination perturb behavior beyond baseline error; switching in particular induces significant, directional drift from even a single handoff [17]. Averaged per step these contribute $\bar{\delta} \geq 0$ under typical operation. Individual perturbations are sign-indefinite — some handoffs are beneficial (Part II, Assumption 2) — so $\bar{\delta} \geq 0$ describes the net average of unmanaged execution, not every event; a policy engineered to exploit beneficial handoffs can in principle drive the operational average negative (Lemma 5.2), which is precisely the sense in which the operational exponent is controllable.

Contributions.

1. **The Lyapunov exponent of delegated work** (Definition 3.1) and its decomposition $\lambda = \lambda_\varepsilon + \lambda_\delta$ into intrinsic and operational divergence.
2. **The horizon theorem** (Theorem 3.2): delivered value is uniquely maximized at the Lyapunov time $n^* = 1/\lambda \approx 1/(\varepsilon + \bar{\delta})$; the $1/e$ operating point of optimal loading is the e-folding identity (Corollary 3.3), and value at the horizon scales as $1/\lambda$ (Corollary 3.4).
3. **Robustness**: the reciprocal-exponent law survives per-step value accrual with costs (Theorem 3.6) and correlated errors via an absorbing Markov formulation (Theorem 4.1).

4. **Controllability of λ_δ** (Part II): exact $O(nk^2)$ routing optimization (Theorem 6.1); blocky execution optimal iff log-handoff costs are quasi-metric (Theorem 6.2); a closed-form swap threshold δ^* (Theorem 7.1) and Lambert-W amortization horizon (Theorem 7.3); and a perturbation-stability theorem (Theorem 8.1): handoff shocks decay or persist according to the contractivity of the successor's dynamics, explaining the observed spectrum of switching outcomes [17].
5. **The Fault-Tolerance Threshold Conjecture** (Section 9): the analogue for AI pipelines of the quantum error-correction threshold theorem — verification below a critical cost fraction extends the horizon without bound — with million-step zero-error execution [2] as supporting evidence.
6. **A measurement program** (Section 10): falsifiable conjectures and a protocol estimating $(\lambda_\varepsilon, \lambda_\delta, H, \rho)$ from switch-matrix and fingerprint instrumentation [17, 19, 20].

We are explicit about the division of labor. The optimization calculus is elementary, and the Lyapunov vocabulary is imported, not invented; the contribution is the identification — that delegated cognitive work is a divergent dynamical process whose horizon has a closed form, whose exponent decomposes into a fixed and a controllable part, and whose controllable part admits an exact, computationally cheap theory. The dynamics framing is not decoration: Part II's contraction coefficients, perturbation bounds, and stability dichotomy are dynamical-systems objects already, and the $1/e$ universality that appears coincidental in a bare-calculus treatment is definitional in this one.

2 Related Work

Long-horizon reliability. Exponential decay of sequential success is developed in [3]; relay-style results give exponential degradation in chain depth for any positive per-agent error [4]; METR measures the empirical horizon [1]. Reliability-science frameworks propose decay curves and consistency metrics [5, 6]; TraceToChain fits agent traces to absorbing Markov chains with goodness-of-fit certificates [7]. Challenges to independence [8] and evidence that cascades attenuate some error types while amplifying others [9] motivate Section 4. MAKER completes million-step tasks by maximal decomposition with voting and red-flagging [2]; in our frame it is an error-correction scheme operating below threshold (Section 9), and the horizon n^* is the subtask size such schemes implicitly target.

Optimal sizing of compound systems. Chen et al. [10] prove non-monotone scaling in the number of *parallel* LM calls and derive the maximizing count; recent work derives closed-form optimal *token allocations* across sequential stages under budget constraints [11]. Neither optimizes the quantity of task per span, and neither carries a drift term; the horizon $n^* = 1/\lambda$ with the $(\lambda_\varepsilon, \lambda_\delta)$ decomposition appears to be new to the AI-systems literature.

Routing and switching. Cost-aware routing (FrugalGPT [12], RouteLLM [13], Hybrid LLM [14], CARROT [15]; survey [16]) optimizes per-query selection with no coupling between consecutive assignments. Hankache et al. [17] measure switch-induced drift directly: significant, directional, occasionally beneficial. Behavioral fingerprinting [19] and endpoint-stability tests [20] supply measurement instruments; [18] finds task ambiguity a stronger inconsistency driver

than model identity (Conjecture 4). Part II is, to our knowledge, the first decision theory in which the handoff perturbation is a first-class, asymmetric, sign-indefinite term.

Dynamical and physical analogues. Lyapunov exponents and e-folding predictability horizons are classical [22]; Dobrushin contraction is the standard ergodicity coefficient for Markov kernels [23]. The fault-tolerance threshold theorem of quantum computation — below a critical physical error rate, error correction sustains arbitrarily long computation at polylogarithmic overhead [24] — is the template for Section 9; we import its *structure* as a conjecture, not its Hilbert-space formalism, which does not apply.

Part I: The Horizon

3 The Lyapunov Time of Delegated Work

Definition 3.1 (Divergence rate of a pipeline)**.** A span of delegated work consists of n task units executed sequentially. Each unit preserves the integrity of the work product with probability $q = (1-\varepsilon)(1-\bar{\delta}) \in (0,1)$, with $\varepsilon, \bar{\delta}$ as in Axioms 1–2. Under independence (relaxed in Section 4), integrity after n units is

$$Q(n) \;=\; q^n \;=\; e^{-\lambda n}, \qquad \lambda \;:=\; \ln(1/q) \;=\; \lambda_\varepsilon + \lambda_\delta,$$

where $\lambda_\varepsilon = \ln \frac{1}{1-\varepsilon}$ is the *intrinsic exponent* (fixed by the model and its training method) and $\lambda_\delta = \ln \frac{1}{1-\bar{\delta}}$ is the *operational exponent* (set by execution choices). λ is the exponential rate at which the realized trajectory of the work departs from the intended one — the Lyapunov exponent of the pipeline. Under the modeling assumption that intrinsic error and operational drift act as independent multiplicative degradation mechanisms, the decomposition is exact: exponents of independent mechanisms add. Where the mechanisms interact (e.g., routers switching models precisely when context is already degraded), the decomposition holds to first order and the interaction enters empirically. For small rates, $\lambda_\varepsilon \approx \varepsilon$ and $\lambda_\delta \approx \bar{\delta}$.

Assumption 1 (All-or-nothing value, baseline case)**.** The span delivers value proportional to work attempted, conditional on integrity: $W(n) = v\, n\, e^{-\lambda n}$, $v > 0$. (Per-unit accrual: Theorem 3.6.)

Theorem 3.2 (The horizon theorem)**.** *$W(n) = v\, n\, e^{-\lambda n}$ has a unique global maximum on $(0, \infty)$ at the Lyapunov time*

$$n^* \;=\; \frac{1}{\lambda} \;\approx\; \frac{1}{\varepsilon + \bar{\delta}} \quad (\varepsilon, \bar{\delta} \text{ small}).$$

Proof. $W'(n) = v\, e^{-\lambda n}(1 - \lambda n)$, positive for $n < 1/\lambda$ and negative after; the stationary point is unique. $W(0^+) = 0$ and $W(n) \to 0$ as $n \to \infty$, so the interior maximum is global. The approximation is $\lambda = (\varepsilon + \bar{\delta})(1 + O(\varepsilon + \bar{\delta}))$ from the logarithm expansion. □

Corollary 3.3 (The e-folding identity)**.** *At the horizon, integrity is $Q(n^*) = e^{-\lambda \cdot (1/\lambda)} = e^{-1} \approx 36.8\%$ — independent of the model, the error rate, and the drift rate. This is not a coincidence of the calculus but the definition of Lyapunov time: the optimally loaded pipeline runs to exactly one e-fold of degradation. Relative to the unconstrained expected-value objective, pipelines observed above $1/e$ integrity are underloaded and below it overloaded; the identity converts telemetry into a*

load diagnostic with no free parameters. This is the optimum of Assumption 1, not an operating prescription: applications with failure externalities should impose an integrity floor τ, which truncates the horizon to $\ln(1/\tau)/\lambda < n^*$ and moves rational operation to the left of $1/e$ by design.

Proof. Immediate from $n^*\lambda = 1$. □

Corollary 3.4 (Value at the horizon)**.** *$W(n^*) = v/(e\lambda) \approx v/(e(\varepsilon + \bar{\delta}))$: achievable value per span is inversely proportional to the Lyapunov exponent. Halving the exponent doubles both the rational task quantum and the value it delivers; and since $\lambda = \lambda_\varepsilon + \lambda_\delta$, the marginal value of drift reduction equals that of error reduction, dollar for dollar of exponent.*

Proof. Substitute Theorem 3.2 and Corollary 3.3 into W. □

Remark 3.5 (Two exponents, two owners). λ_ε is, at any moment, a constant of nature for the operator: it is set by the model. λ_δ is set by routing policy, version pinning, and context hygiene. The horizon theorem therefore does more than locate an optimum; it identifies which part of the predictability limit is an engineering variable — the disanalogy with weather that makes the subject actionable. Note also that the measured 50% horizon [1] and the optimal load point are different quantities: $e^{-\lambda n} = 0.5$ at $n = \ln 2/\lambda \approx 0.69\, n^*$, so horizon-matching *underloads* relative to optimum, while intuitive "run until it usually works" operation at 80–90% reliability underloads severely: optimal loading is more aggressive than practice, tolerating 63% failure to maximize expected delivered value. Whether that tolerance is acceptable is an application question (Section 11); the theorem states the value-maximizing point, not a safety policy.

3.1 Per-step value accrual

Theorem 3.6 (Horizon under accrual and cost)**.** *If each unit costs $c > 0$ and delivers v when the chain has retained integrity, expected profit is $\Pi(n) = \sum_{t=1}^{n}(ve^{-\lambda t} - c)$. If $ve^{-\lambda} > c$, the unique maximizer is*

$$n^*_{\text{acc}} \;=\; \left\lfloor \frac{\ln(v/c)}{\lambda} \right\rfloor \;\approx\; \frac{\ln(v/c)}{\varepsilon + \bar{\delta}};$$

otherwise $n^*_{\text{acc}} = 0$.

Proof. Marginal profit $\Delta\Pi(n) = ve^{-\lambda n} - c$ is strictly decreasing with a single sign change when $ve^{-\lambda} > c$; extend while $ve^{-\lambda n} \geq c$, i.e., $n \leq \ln(v/c)/\lambda$. □

Remark 3.7. Both regimes obey the same law: *optimal load* $= \Theta(1/\lambda)$, with numerator 1 (all-or-nothing) or $\ln(v/c)$ (accrual). The reciprocal-exponent form is the invariant — as it must be, since $1/\lambda$ is the only timescale in the problem.

4 Robustness: Correlated Errors

Independence is the standard objection [8, 9]. The dynamical frame answers it naturally: model the pipeline as a stochastic dynamical system.

Let context quality be a state x_t in a measurable space $\mathcal{X}$ with absorbing failure set F; the executing configuration acts as a Markov kernel P with Dobrushin contraction coefficient $\alpha = \sup_{\nu \neq \nu'} \|\nu P - \nu' P\|_{\text{TV}} / \|\nu - \nu'\|_{\text{TV}}$ [23].

Theorem 4.1 (The exponent survives correlation)**.** *If the kernel leaks uniformly into failure —* $\inf_{x \notin F} P(x, F) \geq \kappa > 0$ *— then* $\Pr(x_n \notin F) \leq e^{-\lambda_\kappa n}$ *with* $\lambda_\kappa = \ln \frac{1}{1-\kappa}$*, and Theorems 3.2–3.6 hold with* λ_κ *as the envelope exponent:* $n^* \leq 1/\lambda_\kappa \approx 1/\kappa$*.*

Proof. Condition on survival to step t and the surviving state: one-step survival is $\leq 1 - \kappa$ uniformly in state and history (Markov property); multiply over n steps. $W(n) \leq vne^{-\lambda_\kappa n}$, so the envelope's maximizer bounds the achievable horizon. □

Exponential divergence thus requires only a uniform per-step leak into an absorbing failure mode, not independence: correlation moves the constant, not the law. Two scope notes. First, this is a worst-case *envelope*: the uniform-leak condition demands a state-independent bound, and real pipelines with strongly state-dependent failure rates will exhibit effective exponents below λ_κ, varying with task mixture and starting state. Second, the bound is one-sided; richer correlated regimes — error concentration at decision points [8], cascades that attenuate one error type while amplifying another [9] — can produce local structure the envelope does not capture. Estimation of κ from traces is the absorbing-DTMC program of [7].

Part II: The Operational Exponent Is Controllable

5 The Routing Model

We open the box on λ_δ. The dominant operational source of drift is heterogeneous execution: assigning different models to different steps to economize cost [12, 13], which induces measurable handoff perturbations [17].

Definition 5.1 (Sequential routing problem)**.** A span of n steps is executed over a pool $M = \{1, \ldots, k\}$ with per-step success $p_i \in (0, 1)$ and cost $c_i > 0$. A policy $\pi : \{1, \ldots, n\} \to M$ incurs at each consecutive pair $\pi(t) = i, \pi(t+1) = j$ a multiplicative handoff factor $h_{ij} > 0$, $h_{ii} = 1$; write $h_{ij} = 1 - \delta_{ij}$, $d_{ij} = -\ln h_{ij}$. Quality and utility:

$$Q(\pi) = \prod_t p_{\pi(t)} \prod_t h_{\pi(t),\pi(t+1)}, \qquad U(\pi) = VQ(\pi) - \sum_t c_{\pi(t)}.$$

The policy's induced operational exponent is $\lambda_\delta(\pi) = \frac{1}{n} \sum_t d_{\pi(t),\pi(t+1)}$ — the quantity entering Theorem 3.2.

Assumption 2 (Empirical form of H)**.** H is asymmetric ($h_{ij} \neq h_{ji}$) and sign-indefinite in δ (some $\delta_{ij} < 0$): handoff effects are significant, directional, occasionally beneficial [17]. Equivalently, the "distances" d_{ij} may be negative and are direction-dependent.

Lemma 5.2 (Policy-dependent horizon)**.** $Q(\pi) \leq p_{\max}^n \max(1, h_{\max})^{n-1}$ *with* $p_{\max} = \max_i p_i$*,* $h_{\max} = \max_{i \neq j} h_{ij}$*. If* $h_{\max} \leq 1$*, the classical bound follows [3, 4]. If some* $h_{ij}h_{ji} > 1$*, alternation achieves per-step log-quality* $\frac{1}{2} \ln(p_i p_j h_{ij} h_{ji})$*, exceeding* $\ln p_{\max}$ *whenever* $p_i p_j h_{ij} h_{ji} > p_{\max}^2$*: the effective exponent — hence the Lyapunov time* n^* *— is a property of the* policy*, not the model.*

Proof. The bound is termwise. Under alternation, $Q = (p_i p_j)^{\lfloor n/2 \rfloor} (h_{ij} h_{ji})^{\lfloor (n-1)/2 \rfloor} r$ with bounded parity remainder; take logs, divide by n. □

6 Optimal Policies

Theorem 6.1 (Exact optimization in $O(nk^2)$)**.** *For any $\lambda \geq 0$, the maximizer of $L_\lambda(\pi) = \sum_t [\ln p_{\pi(t)} - \lambda c_{\pi(t)}] - \sum_t d_{\pi(t),\pi(t+1)}$ over all k^n policies is computable by $D[t][j] = \ln p_j - \lambda c_j + \max_i(D[t-1][i] - d_{ij})$ in $O(nk^2)$ time; sweeping λ traces the cost–quality frontier.*

Proof. All coupling is between adjacent steps, so a suffix's value depends on the prefix only through its final model: optimal substructure holds, and induction gives $D[t][j]$ as the optimum over length-t prefixes ending in j. nk entries, each a max over k. Lagrangian duality exposes the frontier's concave envelope. □

Call a policy *blocky* if each model used occupies one contiguous segment; fix the assignment multiset to isolate ordering.

Theorem 6.2 (Blocky optimality ↔ quasi-metric structure)**.** *(a) If d is a quasi-metric ($d_{ij} \geq 0$; $d_{ik} \leq d_{ij} + d_{jk}$), some blocky ordering is optimal for every assignment multiset.*

(b) If the round-trip gain $g_{ij} = \ln(h_{ij}h_{ji}) > 0$ for some pair, blockiness fails: deliberate alternation strictly dominates for suitable multisets.

Proof. (a) Let $\sigma = WA_1XA_2Y$ place model A in two maximal blocks, X nonempty and A-free; f, l denote first/last models of segments. Merging to $\sigma' = WA_1A_2XY$ changes boundary cost by $\Delta d = d_{l(X),f(Y)} - d_{l(X),A} - d_{A,f(Y)} \leq 0$ (triangle inequality; if $Y = \varnothing$, $\Delta d = -d_{l(X),A} \leq 0$ by nonnegativity). Each merge weakly improves and strictly reduces block count; finite induction terminates blocky. *(b)* With m copies each of i, j: any blocky ordering pays one boundary, $O(1)$; alternation banks $(m-1)g_{ij} + O(1) \to \infty$. The multiset fixes $\sum \ln p$ and cost, so alternation dominates for large m. Note $g_{ij} > 0 \iff d_{ij} + d_{ji} < 0 = d_{ii}$: a triangle violation with coincident endpoints. □

Remark 6.3 (The gap is the point). Sufficiency needs the full quasi-metric; necessity concerns round trips only. Triangle violations with all $d > 0$ make *relayed* handoffs optimal — still blocky. Whether production H is quasi-metric is Conjecture 3; the shape of optimal execution hinges on it.

7 The Swap Threshold and the Amortization Horizon

At step t, model i holds the pipeline with realized prefix quality Q_{pre}; tail length $m = n - t$. Compare STAY with SWITCH to j for the tail (single-switch tails are without loss under Theorem 6.2(a)).

Theorem 7.1 (Swap threshold)**.** *$\Delta U = VQ_{\text{pre}}(h_{ij}p_j^m - p_i^m) + m(c_i - c_j)$ is affine and strictly decreasing in δ_{ij}; switching is rational iff $\delta_{ij} < \delta^*$, where*

$$\delta^* = 1 - \Big(\frac{p_i}{p_j}\Big)^m \Big[1 - \frac{m(c_i - c_j)}{VQ_{\text{pre}}p_i^m}\Big].$$

Proof. STAY yields $VQ_{\text{pre}}p_i^m - mc_i$; SWITCH yields $VQ_{\text{pre}}h_{ij}p_j^m - mc_j$; subtract; the coefficient of h_{ij} is positive; solve $\Delta U = 0$. □

Proposition 7.2 (No upper cutoff)**.** *If $s = c_i - c_j > 0$, $\Delta U \to +\infty$ as $m \to \infty$: on tails long enough that failure is near-certain either way, take the savings.*

Proof. Exponentials vanish; ms grows linearly. □

Theorem 7.3 (Single crossing; Lambert-W amortization horizon)**.** *For an equal-capability pool ($p_i = p_j = p$) with $\delta_{ij} > 0$: $f(m) = ms - VQ_{\text{pre}}\delta_{ij}p^m$ is strictly increasing with $f(0) < 0 < f(\infty)$, hence has unique root*

$$m^* = \frac{1}{\ln(1/p)}\, W\Big(\frac{VQ_{\text{pre}}\,\delta_{ij}\ln(1/p)}{s}\Big),$$

W the principal Lambert branch: switching is irrational for $m < m^*$, rational for $m > m^*$.

Proof. $f'(m) = s + VQ_{\text{pre}}\delta_{ij}p^m \ln(1/p) > 0$; boundary signs give uniqueness. Substituting $u = m\ln(1/p)$ turns $ms = VQ_{\text{pre}}\delta_{ij}p^m$ into $ue^u = VQ_{\text{pre}}\delta_{ij}\ln(1/p)/s$; the argument is positive so the principal branch applies. For $p_j < p_i$, a crossing exists by the same boundary signs; uniqueness holds under $s \geq VQ_{\text{pre}}h_{ij}p_j\ln(1/p_j)$, which forces $f' > 0$ since the negative term of $f'(m) = s - VQ_{\text{pre}}h_{ij}p_j^m\ln(1/p_j) + VQ_{\text{pre}}p_i^m\ln(1/p_i)$ is bounded by $VQ_{\text{pre}}h_{ij}p_j\ln(1/p_j)$ for $m \geq 1$. □

Remark 7.4 (Changeover structure). d_{ij} is paid once per switch; savings accrue per step: the setup-cost structure of lot-sizing [21], of which Theorem 7.3 is the LLM-routing instance.

8 Perturbation Stability: Why Some Handoffs Are Forgotten

In dynamical terms, a handoff is an impulsive perturbation to the pipeline's trajectory; its fate is decided by the stability of the downstream dynamics.

Theorem 8.1 (Perturbation stability of handoffs)**.** *Let π, π' differ only on steps $S = \{t_1 < \cdots < t_q\}$, using kernels P_i vs. P_j; let α_r be the Dobrushin coefficient at step r. Then*

$$\|\mu_n^{\pi} - \mu_n^{\pi'}\|_{\text{TV}} \leq \sum_{t\in S} \|P_i - P_j\|_{\text{TV}} \prod_{r>t,\ r\notin S} \alpha_{\pi(r)},$$

and the success-probability difference obeys the same bound.

Proof. Interpolate through hybrids $\sigma_0 = \pi, \ldots, \sigma_q = \pi'$ applying the substitution at the first ℓ elements of S. Adjacent hybrids differ at one step t_ℓ; with ν their common entering distribution and K_ℓ the common downstream composition, $\|\mu_n^{\sigma_\ell} - \mu_n^{\sigma_{\ell-1}}\|_{\text{TV}} = \|\nu(P_j - P_i)K_\ell\|_{\text{TV}} \leq \|P_i - P_j\|_{\text{TV}} \prod_{r>t_\ell, r\notin S} \alpha_{\pi(r)}$ by submultiplicativity of contraction along K_ℓ (steps in S are handled at their own stage). Sum via the triangle inequality; success is an event, so its probability difference is bounded by the terminal TV distance. □

Corollary 8.2 (Stability dichotomy; handoff robustness)**.** *A perturbation entering contractive dynamics ($\alpha \ll 1$ downstream) is* forgotten *geometrically; entering marginal dynamics ($\alpha \approx 1$), it is* frozen in*. Defining per-model handoff robustness $\rho_j = 1 - \alpha_j$, the observed spectrum of switching outcomes — benign, beneficial, damaging [17] — corresponds to variation in downstream ρ: the second measurable alongside H, estimable via absorbing-DTMC fits [7] and perturbation-response tests [20].*

9 The Fault-Tolerance Threshold Conjecture

Quantum computation faces structurally the same obstruction as delegated work: a physical process with irreducible per-step error, compounding across a long computation. Its resolution is the threshold theorem [24]: if the physical error rate is below a critical value, concatenated error correction sustains arbitrarily long computation at polylogarithmic overhead. The theorem's *structure* — a phase transition in achievable computation length as a function of per-step error and correction overhead — transfers to our setting as a precise conjecture. (The Hilbert-space formalism does not transfer, and we do not pretend it does.)

Augment the model of Section 4: after every block of b task units, a *verification step* is available at cost fraction ϕ of the block's cost, which detects an integrity violation with probability η and, upon detection, restores the last verified state (rollback), resetting the divergence clock for the surviving trajectory.

Conjecture 1 (Fault-tolerance threshold for AI pipelines)**.** *There exists a critical surface $\varepsilon_c(\eta, \phi, b) > 0$ such that:*

*(i) (*Below threshold*) if the per-unit degradation rate satisfies $\varepsilon + \bar{\delta} < \varepsilon_c$, verified execution achieves, for every target integrity τ, unbounded task length n at total cost $O\big(n \operatorname{polylog}(n/(1-\tau))\big)$: the Lyapunov horizon is extended without bound at polylogarithmic overhead;*

*(ii) (*Above threshold*) if $\varepsilon + \bar{\delta} > \varepsilon_c$, no verification schedule at bounded cost fraction sustains integrity: the horizon remains $\Theta(1/\lambda)$.*

Three considerations support the conjecture. First, the mechanism that makes the quantum theorem work — errors must be detected faster than they propagate — has its analogue here in Theorem 8.1: contractive downstream dynamics ($\rho > 0$) localize perturbations, which is precisely the condition under which block-local verification can catch them before they globalize. Second, MAKER's completion of a task exceeding one million steps with zero errors via maximal decomposition, per-subtask voting, and red-flagging [2] is an existence proof that *some* verification scheme crosses *some* threshold: its per-subtask error correction at bounded overhead sustained an effective horizon three to four orders of magnitude beyond the unverified n^* of its base model. Third, the elementary rollback calculus is already suggestive: with perfect detection ($\eta = 1$), a verified block of size b succeeds per attempt with probability $e^{-\lambda b}$, so delivered blocks follow a geometric retry process with $\mathbb{E}[\text{attempts}] = e^{\lambda b}$ and expected cost $c\, b\, (1+\phi)\, e^{\lambda b}$ per delivered block — finite for every b, with cost-per-delivered-unit $c(1+\phi)e^{\lambda b}$ minimized as $b \to 0$ but total-value-per-span maximized at block size $b^\dagger = \Theta(1/\lambda) = \Theta(n^*)$ once per-verification fixed costs are included, by the same calculus as Theorem 3.2. The optimal verified pipeline thus tiles long work into blocks of size proportional to the Lyapunov time — recovering, as a corollary of the conjectured theorem's easy direction, the empirical design wisdom of decomposition schemes. What remains genuinely open, and what the conjecture asserts, is the imperfect-detection case $\eta < 1$ with drifting verifiers, where the verifier itself is subject to Axioms 1–2 — the true analogue of fault tolerance, in which the error-correcting machinery is as faulty as the computation it protects.

Falsification path: measure $(\hat{\lambda}, \hat{\eta}, \hat{\phi})$ for a fixed pipeline family, sweep $\varepsilon + \bar{\delta}$ across the conjectured ε_c by controlled degradation (temperature, quantization, deliberate adverse routing), and test for the predicted phase transition in achievable verified horizon.

10 Falsifiable Conjectures and Measurement Protocol

Conjecture 2 (Round-trip submultiplicativity)**.** $g_{ij} = \ln(h_{ij}h_{ji}) \leq 0$ *for all production pairs. If violated, Theorem 6.2(b) and Lemma 5.2 activate: alternation strictly dominates and lowers the effective exponent below any single model's — a qualitatively new execution strategy.*

Conjecture 3 (No metric structure)**.** *d fails the quasi-metric axioms: asymmetry is established [17]; we conjecture triangle violations $d_{ik} + d_{kj} < d_{ij}$, making relayed handoffs strictly better than direct ones and defeating router designs premised on symmetric model-distance embeddings.*

Conjecture 4 (Ambiguity interaction)**.** $\delta_{ij} = \delta_{ij}^0\, \gamma(\textit{ambiguity})$ *with γ increasing [18]: structured-task measurements underestimate production drift.*

Protocol. (1) Diagnostic suite spanning structured and ambiguous tasks. (2) Estimate p_i ($\Rightarrow \hat{\lambda}_\varepsilon$) per model per task class. (3) Populate $\hat{H}$ ($\Rightarrow \hat{\lambda}_\delta(\pi)$ for any policy) via paired switch-matrix runs with permutation tests [17, 20]. (4) Test Conjectures 2–3 as inequality tests on $\hat{H}$ with multiplicity control. (5) Fit absorbing DTMCs for $\hat{\alpha}_i, \hat{\rho}_i$ [7]; validate Theorem 8.1 by regressing terminal perturbation on downstream $\prod \hat{\alpha}$. (6) Verify the e-folding identity: pipelines loaded to $\hat{n}^* = 1/\hat{\lambda}$ should exhibit terminal integrity statistically indistinguishable from $1/e$; compare delivered value against horizon-matched and unbounded baselines. (7) Evaluate the DP router (Theorem 6.1) against cost-only routers on end-to-end utility. (8) Execute the threshold sweep of Section 9.

11 Discussion and Limitations

What the frame changes. The field measures how long agents can work; this paper derives how long they should, and names the object correctly: delegated work is a divergent dynamical process, its predictability horizon is a Lyapunov time with closed form $1/\lambda$, and the exponent splits into a part nature fixes and a part operations set. The e-folding identity gives a zero-parameter load diagnostic; the swap theory gives a closed-form prohibition (never switch inside the amortization horizon m^*) that no cost-based router currently respects; and the threshold conjecture organizes the fault-tolerance program — decomposition, voting, verification [2, 3] — as error correction relative to a derivable optimal block size $b^\dagger = \Theta(n^*)$.

Limitations. (i) The optimization calculus is elementary and the Lyapunov vocabulary imported; the contribution is the identification, the decomposition, the universality and stability results, and the exact theory of the controllable exponent. (ii) The all-or-nothing value model drives the aggressive $1/e$ operating point; applications with severe failure externalities should optimize W subject to an integrity floor, which simply truncates n at $\ln(1/\tau)/\lambda < n^*$. (iii) λ_δ as a per-step average compresses heterogeneous drift sources; Part II disaggregates only routing — versioning and context contamination enter empirically. (iv) Q_{pre}, q, and hence λ are estimated, and estimation error propagates into n^*, δ^*, m^*. (v) Uniqueness in Theorem 7.3 is proven for equal capabilities, with a sufficient condition otherwise. (vi) Conjecture 1 is open; its easy direction (perfect detection) is sketched, its hard direction (faulty verifiers) is genuinely analogous to fault tolerance and correspondingly nontrivial.

References

[1] T. Kwa et al. Measuring AI Ability to Complete Long Tasks. METR, *arXiv:2503.14499*, 2025.

[2] E. Meyerson et al. Solving a Million-Step LLM Task with Zero Errors. *arXiv:2511.09030*, 2025.

[3] The Six Sigma Agent: Achieving Enterprise-Grade Reliability in LLM Systems Through Consensus-Driven Decomposed Execution. *arXiv:2601.22290*, 2026.

[4] Relay Degradation in Sequential Agent Chains. *arXiv:2603.26993*, 2026.

[5] Beyond pass@1: A Reliability Science Framework for Long-Horizon LLM Agents. *arXiv:2603.29231*, 2026.

[6] Towards a Science of AI Agent Reliability. *arXiv:2602.16666*, 2026.

[7] Measuring the Unmeasurable: Markov Chain Reliability for LLM Agents (TraceToChain). *arXiv:2604.24579*, 2026.

[8] Beyond Exponential Decay: Rethinking Error Accumulation in Large Language Models. *arXiv:2505.24187*, 2025.

[9] Hallucination Cascade: Analyzing Error Propagation in Multi-Agent LLM Systems. *arXiv:2606.07937*, 2026.

[10] L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. Zaharia, J. Zou. Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems. *NeurIPS*, 2024. arXiv:2403.02419.

[11] Toward Reliable Design of LLM-Enabled Agentic Workflows: Optimizing Latency-Reliability-Cost Tradeoffs. *arXiv:2605.23929*, 2026.

[12] L. Chen, M. Zaharia, J. Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv:2305.05176*, 2023.

[13] I. Ong et al. RouteLLM: Learning to Route LLMs with Preference Data. *arXiv:2406.18665*, 2024.

[14] D. Ding et al. Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing. *ICLR*, 2024.

[15] S. Maity et al. CARROT: A Cost Aware Rate Optimal Router. *arXiv:2502.03261*, 2025.

[16] Dynamic Model Routing and Cascading for Efficient LLM Inference: A Survey. *arXiv:2603.04445*, 2026.

[17] R. Hankache et al. Evaluating Performance Drift from Model Switching in Multi-Turn LLM Systems. *arXiv:2603.03111*, 2026.

[18] How Consistent Are LLM Agents? Measuring Behavioral Reproducibility in Multi-Step Tool-Calling Pipelines. *arXiv:2605.28840*, 2026.

[19] J. Pei et al. Behavioral Fingerprinting of Large Language Models. *arXiv:2509.04504*, 2025.

[20] Behavioral Fingerprints for LLM Endpoint Stability and Identity. *arXiv:2603.19022*, 2026.

[21] H. M. Wagner, T. M. Whitin. Dynamic Version of the Economic Lot Size Model. *Management Science* 5(1), 1958.

[22] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. 2nd ed., Westview Press, 2015.

[23] R. L. Dobrushin. Central Limit Theorem for Nonstationary Markov Chains. *Theory of Probability and Its Applications* 1(1), 1956.

[24] D. Aharonov, M. Ben-Or. Fault-Tolerant Quantum Computation with Constant Error. *Proc. 29th ACM STOC*, 1997.