

Portfolio Management Discipline for an LLM Risk Desk: The Bridge to AI Cost Management

Eric Allione¹

¹aigamma.com. Correspondence: eric@aigamma.com.

Abstract

This paper shows how a bridge can be built between public-equity portfolio risk management and the operational control of LLM fleet spend. Mathlimit, the system presented here, organizes metered services into positions and a book, estimates pressure from trailing evidence, and applies fixed policy through reversible restrictions alongside independent hard budgets. In the example deployment, the dashboard prompted a manual reduction in narration frequency; a dated receipt comparison shows 67.9 percent lower estimated daily cost after the intervention. Automated enforcement covered one of 19 logical positions, where the ledger exposed a discrepancy between published allowance and recorded skips. The contribution is a deployed composition of established methods and lessons about meter reconciliation, consumer contracts, and release conditions. The paper examines how spend visibility, controller state, and consumer actions can inform strategic interventions and decisions about continued fleet development.

1 Introduction

Public-equity portfolio managers govern changing exposures through positions, book-level limits, and a risk desk. Part of transferring a portfolio manager's framework is the risk-management culture that comes with it. Potentially damaging events should not be ruled out simply because they seem unlikely; contingency planning must account for their consequences. Working with clients, including universities, I have encountered a recurring fear of chatbots because of the potentially open-ended cost exposure from adversarial activity. This concern motivates the paper's scenario calculations: an adversary who repeatedly retriggers a circuit breaker, and a persistent stream of spending that remains below an instantaneous detection threshold. These calculations examine how costs accumulate under stated assumptions and why independent budget limits matter.

An LLM fleet operator faces a related problem: many services independently commit spending against a shared budget. As the fleet's operator, I needed to decide which recurring work to reduce before an account limit interrupted other work. The dashboard exposed scheduled narration as the dominant expense, prompting a cadence reduction. I transfer portfolio operating discipline to that problem; the system neither prices securities nor optimizes investment returns.

The fleet inventory includes roughly 2,300 public chatbots. The historical usage audit covers one studied property, where 53 chat turns cost a few estimated dollars. One authorized page-narration scheduler produced 99 percent of that property's estimated LLM spend. From May 11 through midday June 12, 2026, it ran at roughly three times its later cadence and 2.7 times its later estimated dollars. Twenty observed full receipt days carried about \$149 excess against the later rate; the June 12 transition date is excluded. No recorded alarm identified that schedule, and human discovery time is unknown. Authorized software exposed the gap between incurring cost and recognizing its accumulating effect.

Cost management already offers allocation, forecasting, anomaly detection and automation (Amazon Web Services 2020b,a; BerriAI 2026). The challenge is connecting graded action across a changing book to verified consumer behavior. Provider caps remain essential but can interrupt unrelated work sharing an account. Figures 1 and 2 show the control loop and workload.

The historical evidence spans 61 days, May 8 to July 8: 37,878 priced calls and an estimated \$317.93. A catalog of 63 incidents across 21 agent frameworks supplies neighboring motivation (Khan 2026); this scheduler study is not an agent benchmark.

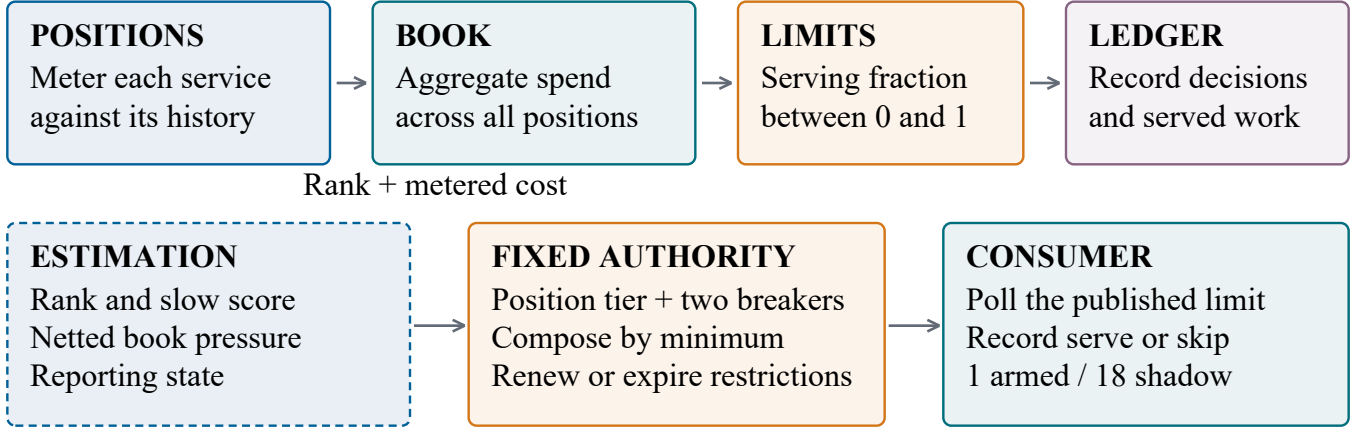
I contribute:

1. **A portfolio control architecture for LLM spend** combining position and book state, trailing estimation under fixed authority, reversible restrictions, hard budget boundaries, and a ledger that includes consumer actions.
2. **A limited-scope production account** with 19 logical positions, one enforcing consumer, a fixed decision policy, and independently re-derived measurements that separate controller state, enforced outcomes, and studied-property receipts.
3. **Deployment lessons grounded in observed failures** of telemetry completeness, consumer accounting, identifier continuity, and the relationship between sampling time and restriction exposure.

2 Deployment Setting and Design

The deployment spans four young serverless domains whose all-vendor burn approached \$400 a month. The initial book held 18 page-level LLM positions; a nineteenth and largest

Portfolio discipline becomes an operating contract



Provider pause is off. Reporting pressure does not itself set the serving fraction.

Figure 1: Portfolio concepts and actual control authority. The empirical rank and metered cost feed the limiter; slow scores and netted pressure support reporting. One consumer is armed, 18 positions are shadow, and provider pause is off.

spender joined from a sibling property on July 10. The studied property’s dominant workload was scheduled market-page narration. Retrieval and embedding traffic measured five orders of magnitude below it. Those services remain part of the architecture, but I do not separately evaluate vector-store control.

Since July 8, the server has polled real telemetry, updated the book, and written decisions. Its views informed my two interventions in Section 5.1. Actuation is narrower: one low-risk narrator is armed, 18 logical positions are shadow, and no third party operates the desk. Shadow positions record controller state without enforcing restrictions. The ledger’s 37 identifiers reflect an 18-position namespace change plus one new position, not 37 concurrent positions.

The July 7 design records established position and book jurisdiction, self-release, and human governance. Subsequent ladder specifications added the quarantine exception and minimum composition. Uncertain anomaly state fails open while an independent declared budget is intended to fail closed. Operation tested and refined these choices. No participant co-design was conducted.

A position owns metered history and a published serving fraction, its clamp. The book reports netted excess and has a separate aggregate breaker (Section 3.3). Caps, edge rules and per-entity buckets retain their own boundaries; the desk sees no IP, user or session and cannot adjudicate individual fairness.

Historical rows defaulted missing provenance to desk and recorded no exact limits-version identifier. The snapshot can align controller states and consumer verdicts by time window, but cannot reconstruct an event-complete causal join.

A separate database cost boundary motivated a provider watch. Its management token is absent and its pause actuator is disarmed; this deployment does not establish protection of

that budget.

3 Estimation and Control

3.1 What the estimator measures

The desk separates an adaptive estimate from the authority to restrict work. Let i index completed five-minute bins. The measured rate x_i is twelve times the bin’s dollar cost, expressed in dollars per hour; s_i is its seasonal reference in the same units. The normalized ratio r_i compares those rates. The recent statistic averages at most six ratios, a thirty-minute window. Its empirical rank is the fraction of older ratios strictly below that recent average. For a nonempty pool,

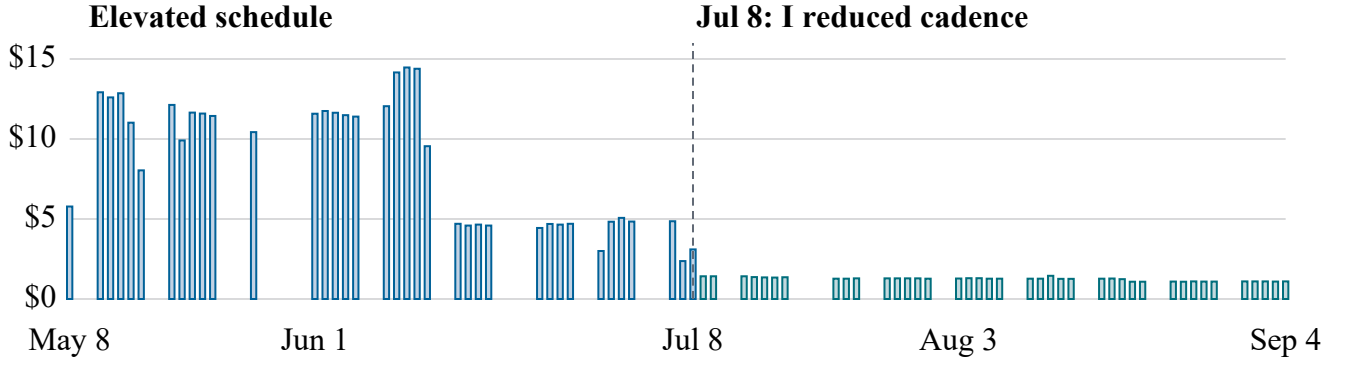
$$r_i = \frac{x_i}{\max(s_i, 0.008)}, \quad \bar{r}_i = \frac{1}{m_i} \sum_{j=i-m_i+1}^i r_j, \quad (1)$$

$$p_i = \frac{1}{|I_i|} \sum_{j \in I_i} \mathbf{1}\{r_j < \bar{r}_i\}.$$

Here m_i counts available bins, capped at six; I_i is the historical trigger pool with $|I_i|$ observations. The indicator is one if its condition holds, zero otherwise. Maximum selects the larger denominator; \$0.008 per hour prevents a vanishing reference. Ratios and p_i are dimensionless; an empty pool returns $p_i = 0$. Rank 0.98 means 98 percent of pooled ratios are lower, not a calibrated incident probability: the query averages six bins, whereas the dependent pool values are unsmoothed (Montgomery 2019; Qiu 2014; Chakraborti and Graham 2019).

The live 14-day calendar window has a causal exclusive cutoff; its two-day gap phases in after warm-up (Appendix C). This delays contamination without permanently exclud-

(a) One property: observed daily estimated spend



(b) Dated intervention comparison

Before: Jun 15 to Jul 7 (14 dates)		\$4.43/day
After: Jul 9 and 10 (2 dates)		\$1.42/day

67.9% lower estimated daily cost; service value and causal savings are unmeasured.

Missing receipt dates have no imputed cost. The post-freeze series has 40 observed dates.

Figure 2: Studied-property receipt history and my dated intervention. The lower bars compare mean estimated daily cost over fourteen baseline dates with two post-change dates. This is a descriptive comparison, not causal savings. The post-freeze series contains forty observations across 42 reference exchange sessions; missing dates are not zero-cost observations.

ing incidents. Calendar annotations support retrospective adjudication; they do not select live training data. Fast reporting follows recent dispersion; slow reporting uses a seasonal budget-scale floor. Both score bins before moment updates, preventing within-bin self-normalization while permitting later seasonal absorption. Appendix C defines the scores. Drip, leak and burst labels interpret state; they do not select the serving fraction.

3.2 Severity, duration, and renewal

Three ordinary tiers use rank thresholds P93, P97.7 and P99.5, serving fractions 0.75, 0.50 and 0.25, and renewal clocks of 18, 6 and 2 successful polls. At regular five-minute polling those clocks are 90, 30 and 10 minutes. Figure 3 puts duration beside severity. Escalation requires three consecutive qualifying polls. Once a tier is active, one qualifying poll refreshes its clock. A higher tier replaces a lower tier; at expiry, the highest currently satisfied lower tier can take over, otherwise the position opens. Repeated polls need not contain distinct data bins.

For an isolated trigger followed by quiet observations, the operational tradeoff can be expressed as withheld allowance:

$$A_k = (1 - c_k)T_k\Delta, \quad (2)$$

$$(A_1, A_2, A_3) = (22.5, 15, 7.5) \text{ min.}$$

Tier k has serving fraction c_k , withheld fraction $1 - c_k$, and

clock T_k polls, with $\Delta =$ five minutes per poll. A_k is withheld allowance in full-rate minutes, including the triggering interval. Tier 2's half allowance for thirty minutes equals fifteen full-rate minutes. At one work unit per minute, sufficient demand and successful enforcement would yield fifteen withheld units; dollars additionally require a cost model.

The dated design motivates persistence against isolated excursions, earlier reconsideration of severe restrictions, and a small probe flow during quarantine. These purposes do not establish optimal constants: the percentile landmarks are empirical ranks, not Gaussian confidence levels. A retrospective nine-policy grid crosses upper-tail multipliers 0.9, 1 and 1.1 with clock offsets of minus one, zero and plus one successful poll; serving fractions remain fixed. Across 18 positions from June 16 through July 7, each five-minute position bin receives equal weight. Commanded allowance reduction is 3.15 to 3.54 percent, with 3.22 percent for the frozen table (Appendix B). No policy reaches tier 3. This checks nearby restriction burden, not detection or benefit. Doubling clocks doubles isolated withheld allowance; renewal and fallback defeat a lifetime bound.

3.3 Which layer can restrict the consumer

The book governor nets under-spend against over-spend and reports aggregate pressure. Equal opposing deviations cancel there while a position can still tighten locally. Published authority instead comes from the position ladder, position

Stronger restrictions have shorter renewal clocks

Tier / trigger	Served	Clock	Withheld allowance
1 / P93	0.75	 90 min	 22.5
2 / P97.7	0.50	 30 min	 15
3 / P99.5	0.25	 10 min	 7.5

Withheld allowance = (1 - served fraction) x clock, in full-rate minutes.

Isolated trigger; regular five-minute polls; no renewal or lower-tier fallback.

Figure 3: The fixed ordinary ladder. Last column: isolated-trigger withheld allowance in full-rate minutes, assuming regular five-minute polls, no renewal or lower-tier fallback. Stronger tiers expire sooner but do not equalize allowance; renewal can prolong restriction indefinitely.

breaker and book breaker. For the frozen fractions, their composition is

$$c_{j,n} = \min\{\ell_{j,n}, b_{j,n}, b_{B,n}\}. \quad (3)$$

The published fraction is $c_{j,n}$. For position j at successful update n , $\ell_{j,n}$, $b_{j,n}$, and $b_{B,n}$ are ordinary, position-breaker, and book-breaker fractions. Inactive fractions are one; breakers use 0.02 while open and zero after hard isolation. Minimum selects the tightest allowance: two 0.5 limits give 0.5, not 0.25; position 0.5 with book 0.02 gives 0.02. This relates allowances, not downstream obedience. Pressure scores and provider-board limits are excluded; the pilot polls a separate position-limits endpoint.

The consumer polls its fraction, scales its local bucket or scheduled work, and fails open on error. One low-risk narrator was armed on July 13; 18 logical positions remain shadow. An environment flag controls arming and its removal is roll-back. The provider pause actuator is disarmed. The systemic tripwire requests human attention when three proportional position scores exceed its threshold; it does not add a wider throttle. Section 5.3 checks the recorded consumer actions.

3.4 Repeated floods and the limit of a guarantee

A trigger threshold answers when to act; it does not cap the rate arriving before or during that action. The flood trigger floor is

$$F(B, q_{99}) = \min\{\max(5, 25q_{99}), \max(5, 0.025B)\}. \quad (4)$$

For $B > 0$, B is the dollar budget for the analysis period; q_{99} is the historical P99 five-minute cost. The value F is a dollar threshold per five-minute bin. The coefficient 0.025 allocates 2.5 percent of the budget to a bin, with a \$5 minimum; $B = \$200$ gives $F = \$5$. The implemented absolute trigger is strictly above F ; two qualifying P99.9 control updates provide a separate position-breaker trigger. Neither condition upper-bounds an attacker's offered rate.

After an opening, continuous pressure at subsequent probes doubles the cooldown from one to two to four five-minute intervals. The third relatch enters quarantine. The probe fraction is 0.02; after 24 hours in quarantine it becomes zero, subject to continued control updates and enforcement. The conditional fluid reference, with uniform cost within each bin, is

$$L_{\text{ref}}(H) = r [1 + 0.02 (1 + 2 + 4 + 12 \min(H, 24))] . \quad (5)$$

Here r is constant offered dollars per five-minute bin and H is hours from quarantine entry to effective remediation. $L_{\text{ref}}(H)$ totals the initial full-rate bin, 1+2+4 cooldown bins, and quarantine traffic at fraction 0.02; twelve converts hours to bins. At $r = \$5/\text{bin}$, six hours gives \$12.90 and 24 hours gives \$34.50 (Figure 4). This reference scale does not imply that the strict \$5 trigger fires at equality.

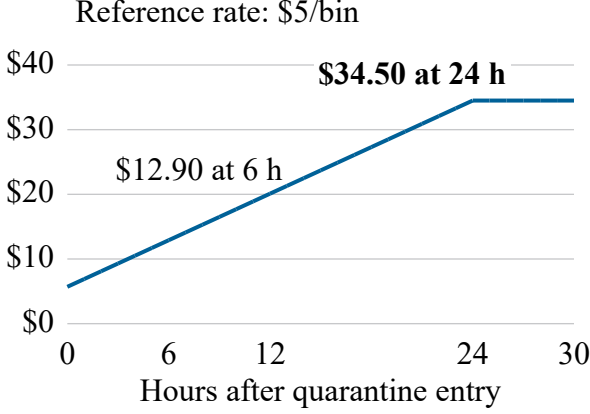
This assumes offered cost bounded by r , regular successful five-minute updates, at most one initially unmitigated bin, triggering at every probe, persistent state with no release override before effective remediation, and immediate dollar-proportional enforcement. Unequal call costs defeat an unstated conversion from call fractions to dollars. Going quiet at probes can reset relatches and permit new episodes, so the formula does not bound arbitrary repeated attacks over a calendar period.

Stealth is elementary to state even when detection is difficult. Suppose the admitted, metered cost is a constant $b > 0$ dollars per five-minute bin, below an instantaneous threshold τ . From zero initial cost,

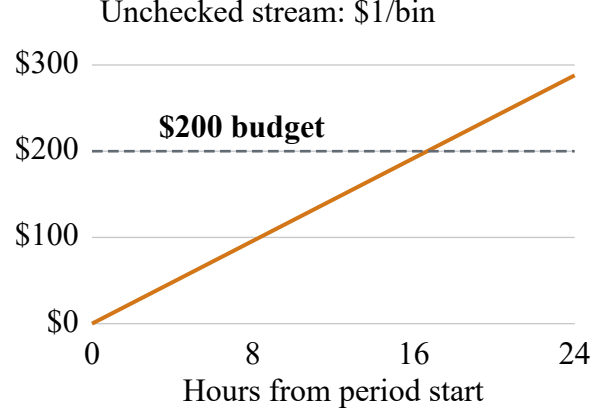
$$0 < b < \tau, \quad C_N = Nb, \quad N_B = \left\lfloor \frac{B}{b} \right\rfloor + 1. \quad (6)$$

N counts bins, C_N is cumulative dollars, and N_B is the first bin where cumulative cost exceeds budget B ; floor rounds down. A \$1/bin stream stays below \$5/bin, reaches \$200 at bin 200 and exceeds it at bin 201 (Figure 4). The slow channel can expose persistent excess relative to its seasonal

(a) Repeatedly triggered episode



(b) Subthreshold accumulation



Illustrative curves: a trigger floor does not bound cumulative cost.

Figure 4: Two distinct failure modes. Left: fixed-rate, continuously retriggered episode cost with persistent state and immediate proportional enforcement; its clock begins at quarantine, with hard zero after 24 hours. Right: a hypothetical \$1/bin billed stream accumulates despite remaining below a \$5/bin trigger. Neither panel is a measured attack or an assurance of deployed protection.

reference, but cannot identify every harmful subthreshold stream. Cost that conforms to a learned reference can still be unacceptable against a budget. Longer attacks can also teach that reference. Detection therefore needs an explicit comparison target; the period budget must be enforced independently.

The slow channel adds a second comparison. Suppose the preceding 24 measured rates equal the fixed seasonal reference s dollars per hour, then rate rises by $\delta > 0$, in the same units. With learnable history and that reference held fixed, after $n \geq 1$ new five-minute bins its score is

$$z^{\text{slow}}(n) = \frac{\min(n, 24)}{24} \frac{\delta}{0.12s + 0.008}. \quad (7)$$

The score is dimensionless. The first factor shows the two-hour buildup; the denominator is a rate tolerance of 12 percent of the reference plus \$0.008/hour. A LEAK label also requires sustained threshold occupancy, so a score crossing alone is insufficient. This reports persistent excess; it does not directly set the clamp. The fixed-reference assumption eventually fails when a long change enters the seasonal history. Appendix C gives the general score definitions.

A hard device must instead reserve a bounded cost atomically before each admission and reject work that would exceed the remaining budget. Every billed action must be covered and its charge must not exceed the reservation. Appendix C derives this contract. The pilot does not establish it. Provider caps and consumer-local ceilings are possible devices (Anthropic 2026); the archive records a \$200 provider stop but no dated export of the final configuration.

4 Evidence and Evaluation Protocol

Decision constants were frozen prospectively on July 9, 2026, at hash 9ef3001; the captured public-service metadata reports

the same hash. External timestamp witnesses bound the artifact’s existence by July 13 (Adams et al. 2001; Todd 2016). Because the repository and ledger remain under my control, they establish reproducible lineage, not continuous runtime attestation. Appendix A records the exact ledger and witness coordinates.

Policy is frozen; append-only decisions atomically record wall-clock and triggering data timestamps for desk latency. Replay estimates excess burn and suppressed work under the counterfactual policy; production verdicts separately record served or skipped units. Decision code and constants are pinned at 9ef3001527fb09be06daa6ee2f3b32e507b84b9a; Appendix A identifies the dated measurement and ledger records.

The frozen tape supports expanding-history replay of the five-week era, 30 labeled perturbations per class in clean stretches, and prospective band diagnostics on its 24-hour post-freeze tail through July 10. Production evidence separately reaches ledger sequence 11,786 at September 7 00:04:01 UTC; receipts end September 4 after 58 calendar days. Six detectors share the unmodified substrate and alert-volume ceiling. Restricted policy scenarios price fixed-reference trajectories under stated metering and enforcement assumptions. Seeded incidents are not production; state transitions are not enforced actions; 30 perturbations on one tape are not 30 deployments. Appendix B gives comparator settings, static search grids and the sensitivity protocol.

5 Results from Deployment and Replay

The prospective window begins July 9, but REST intake was incomplete until the July 10 row-cap repair, when the book and namespace also changed. Replay is unaffected; first-day live decisions and fleet aggregates spanning that boundary are not directly comparable. Figure 5 separates

lifecycle accounting from consumer behavior.

5.1 My interventions and observed workload

On July 8, the dashboard showed the narrator ran four times as often as I judged necessary. Within an hour, I reduced cadence from every 15 minutes to every 60 inside narrower market hours. A worksheet forecast 84 calls per market day and about \$17 a month. Receipts instead show 18 pages at eight hourly slots on both July 9 and July 10: 144 calls, estimated at \$1.424715 and \$1.416391. The forecast is not a measured workload. The measured comparison uses all 14 observed receipt dates from June 15 through July 7, before the July 8 intervention: roughly 457 to 144 calls per observed day, down 68.5 percent; mean daily estimated cost fell from \$4.43 to \$1.420553, down 67.9 percent. The implied monthly run rate at 21.5 observed days is \$30.54 versus \$95.18. June 22 is missing, and these two post-change observations establish neither causal savings nor net benefit. Appendix A defines every interval.

On July 10, the first fleet-wide read found a sibling worker near \$9 a day, about four times the studied property's page book. Its premium model tier was retired within the hour, but no downstream outcome was measured. The views informed both of my decisions; neither was automatic. The record does not isolate the interface's causal effect.

5.2 Controller lifecycle accounting

The final ledger contains 11,713 post-freeze rows. A start changes a tier from zero to positive and pairs with the next return to zero under the same identifier. Across 37 identifiers, the ladder records 4,852 starts and 4,834 matched returns to zero. The remaining 18 starts belong to names retired during the July 10 migration. They are censored lifecycle records, not evidence of restrictions still active in September. Six return rows under replacement names have no recorded start under those names; the audit does not invent continuity across that boundary. Appendix A gives the accounting rule.

Four other non-breaker families contribute 181 chronologically matched engage/release pairs. Three controller breaker episodes opened on ai-firehose from August 27 to 29, each relatching twice and closing before quarantine. These predominantly shadow state transitions do not establish 4,852 enforced interruptions or the absence of intervening human changes.

The separate studied-property receipts contain 5,634 calls and \$49.78 of estimated cost over 40 observed dates from July 9 to September 4. Median estimated cost per observed day is \$1.27 and the maximum is \$1.45. July 20 and 21 are absent from the 42-session reference exchange calendar; no zero cost is imputed. This series covers the property's page narrations, not the whole 19-position book. Its stability follows the pre-freeze cadence cut and cannot identify a controller-caused reduction.

5.3 Consumer accounting against diagnostic references

The one armed consumer records 177 serves and 140 skips across 317 decisions on 40 actuation days, with 242.5 work

units of published allowance. Tested but undeployed reference v2 integrates contemporaneous clamps, serving 242 and skipping 75 from zero credit without restarts. The net 65 additional recorded skips measure this diagnostic discrepancy, not proven production errors.

The reference makes the serving contract explicit. Let c_n be the fraction granted at decision opportunity n , q_{n-1} unused credit before it, and y_n the decision to serve one work unit. In exact arithmetic,

$$\begin{aligned} y_n &= \mathbf{1}\{q_{n-1} + c_n \geq 1\}, \\ q_n &= q_{n-1} + c_n - y_n, \quad S_N = \left\lfloor \sum_{n=1}^N c_n \right\rfloor. \end{aligned} \quad (8)$$

Initially $q_0 = 0$, with $0 \leq c_n \leq 1$ and at most one work unit per opportunity. S_N counts serves over N opportunities; floor rounds down. Credit stays in $[0, 1)$: 242.5 granted units permit 242 serves, leaving 0.5. This is a work-count identity, not a dollar or service-value bound. The record's quarter-unit fractions reproduce it despite the implementation's small admission tolerance. Resets discard credit and can reduce serves (Appendix C).

Historical reference v1 applies a fallen clamp to earlier ticks, turning past serves into debt repaid by skips. A post-hoc initial state matches 315 of 317 verdicts, which is consistent with this mechanism but leaves two unexplained. Restart history is also missing. Resetting reference v2 once per observed day produces 76 skips; resetting every poll produces 165. These sensitivity cases show why the 65-skip difference cannot be assigned wholly to a demonstrated production bug.

The 140 skips are work not performed, with unmeasured utility. All recorded sources are desk, historically also the default for missing provenance. Missing state and version history prevent full attribution of this operational discrepancy.

5.4 Historical schedule detection

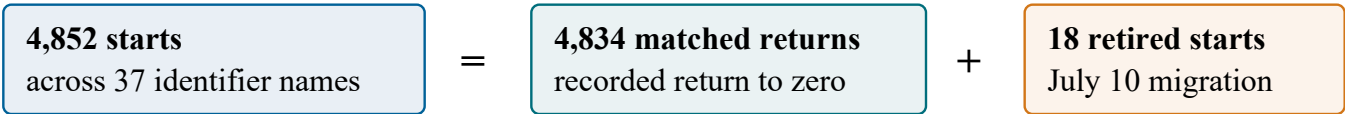
Frozen reporting replay first flags 20 minutes after the era's first armed market open (May 11, 13:30 UTC) and reaches all 18 within the half hour. It re-flags every market day of the era despite receiving three days of arming history already inside that era. The May 8 launch burst sits inside the arming window, where the desk abstains by design.

This retrospective result shows the frozen engine finds the recorded pattern without future observations. Manual discovery latency remains unknown.

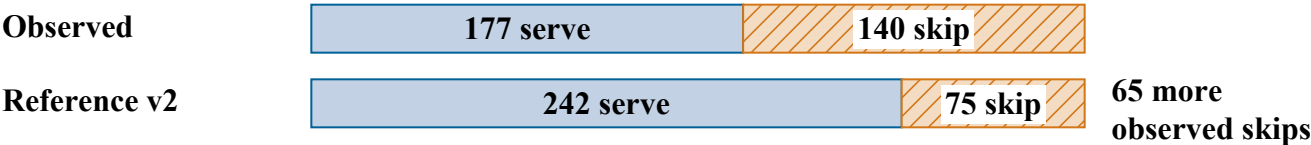
5.5 What the detector comparison establishes

The reporting detector flags LEAK or BURST, or simulated throttle intensity above 0.05. Under a shared ceiling of 17 position alerts across 396 position-days, it detects 26 of 30 twofold six-hour leaks; upper median latency is 115 minutes among detections (the larger central value for even counts). It finds 30 of 30 at threefold, fivefold, and burst conditions. The ceiling limits alert volume, not a matched false-alarm rate, and its 17 alerts contain only two event clusters. The perturbations share one real substrate.

(a) The ledger closes the lifecycle accounting



(b) One armed consumer: 317 decisions



Reference v2 is a diagnostic: zero starting credit, no restarts, 242.5 work units allowed.

Figure 5: Lifecycle records and consumer verdicts. Retired starts complete identifier-level accounting. The 317 opportunities yield 177 recorded serves versus 242 in the undeployed no-restart reference. Cause and service value remain unresolved.

On the original bounded grids, smoothed seasonal detection finds 7 of 30 and the static line 3 of 30. A broader, retrospective static search selects 0.80 times the active-bin median and finds 21 of 30 while remaining below the ceiling. Paired outcomes contain seven desk-only and two static-only detections (exact two-sided McNemar $p = 0.180$). This small sample does not establish reliable superiority or equivalence. The static median is 0 minutes among its 21 detections; comparison with the desk's 26 detections does not establish a paired speed advantage. Appendix B reports operating points and achieved alert rates. The original benchmark and frozen policy remain unchanged. The daily-budget comparator was anchored to the largest pre-evaluation day, including the elevated era; it does not represent a budget calibrated only on the later, lower cadence.

6 Deployment Lessons

6.1 Reconcile independent meters

The first telemetry pipeline reported healthy intake while undercounting by five orders of magnitude: about 1,400 hosted-store tokens versus 227 million in the application record. Ephemeral functions exported fresh cumulative counters from zero; the desk switched to calls.

The REST intake introduced a second failure: a silent 1,000-row cap hid four fifths of spend while health stayed green. Only independent receipts exposed both failures.

Cloud-billing research documents divergence and delay between provider charges and customer usage (Jellinek et al. 2014). Reconcile every intake before applying intelligence.

6.2 Include the consumer schedule in the control design

Restrictions occur at 52.1 percent of recorded polls. Comparing nights and weekends would confound work scheduling with polling phase. Within polls' observed UTC hours,

integrated restricted time is 42.0 percent, retaining repeated-hour poll weights across 316 distinct hours. Every recorded clamp agrees with reconstructed restriction at its timestamp. The difference from uniform time does not establish a jitter benefit; schedule comparisons must preserve workload opportunity and record actions separately from accounting.

6.3 Measure operating cost and service value separately

The deployment is one deterministic container with no model calls, 256 MB of memory, and a 1 GB encrypted volume. Its recorded list-price estimate is \$2.17 a month. That is infrastructure cost; my labor, maintenance, and invoice reconciliation were not measured. Comparing it with the studied property's receipt series would mix a fleet-wide controller with a narrower workload denominator.

Utility of skipped work and failure recovery were not evaluated. The reported result is a control path and fault discovery.

7 Related Work

Mathlimit combines process-control monitoring (Shewhart 1931; Page 1954; Roberts 1959; Montgomery 2019; Qiu 2014; Chakraborti and Graham 2019), financial backtesting, tail measures, and escalation (J.P. Morgan and Reuters 1996; Kupiec 1995; Christoffersen 1998; Basel Committee on Banking Supervision 1996; Artzner et al. 1999), robust estimates and intervention bands (Whalley and Wilmott 1997; Huber 1964), and reliability engineering's breakers, feedback load control, and budget discipline (Chiu and Jain 1989; Nygard 2007; Welsh, Culler, and Brewer 2001; Beyer et al. 2016, 2018). Hawkes processes motivate the amplification lens (Hawkes 1971); security, poisoning, and agent-abuse research address attacker-influenced demand (Tambe 2011; Biggio and Roli 2018; OWASP Foundation 2025; Dorsett et al. 2025; Zhou et al. 2026; Luo et al. 2026; Sivaroopan

Twofold six-hour leaks: the same 30 seeded cases

Original static grid		3 / 30	Paired comparison
Broader static grid		21 / 30	19 detected by both 7 by desk only
Frozen desk		26 / 30	2 by static only 2 detected by neither

Upper medians among detections: desk 115 min; broader static 0 min.

Exact paired McNemar $p = 0.180$: superiority and equivalence remain unestablished.

Figure 6: Reporting-detector comparison on one shared tape. Broader static search changes three detections to 21; the paired table explains the 26 versus 21 marginals. Upper median latencies use each detector’s detected subset. This does not validate the ordinary ladder.

et al. 2026; Zhang et al. 2025; Zavrak 2026; Yuan et al. 2026; Chen et al. 2026; Kelly, Glavin, and Barrett 2021).

Rate/burst accounting and staged token-bucket meters are established prior art (Bernet et al. 2002); the tier settings here are a deployment policy, not new probability theory.

Cost-management tools provide allocation, forecasting, anomaly response and budget actions; LLM gateways and routers add request limits and routing (Amazon Web Services 2020b,a; BerriAI 2026; Chen, Zaharia, and Zou 2024; Ong et al. 2025; Laptev, Amizadeh, and Flint 2015). Agent and workflow controls include runtime accounting, inference-time search and closed-loop online allocation (Khan 2026; Li et al. 2026; Wang et al. 2026). Green SARC uses a calibrated gate for predicted token and carbon cost in an agent loop (Besanson 2026). Mathlimit governs reconciled realized spend across services, without claiming per-action budget safety.

Deployed neighbors govern other objects: SLIDR controls advertising-click false positives and model updates (Muralidhar et al. 2023); AHPA bounds predictive autoscaling (Zhou et al. 2023); an enterprise assistant combines severity, evaluation and release review (Maharaj et al. 2025). Mathlimit’s contribution is the deployed spend book and inspectable ledger comparing published control with consumer behavior.

8 Limitations and Ethics

The architecture is fleet-wide; the evidence is one fleet I operate, one dominant workload, and one armed low-risk position. I own the fleet and bear its spend. Eighteen positions remain shadow and the provider pause actuator is disarmed. Usability, trust, acceptability, fairness, adoption, third-party operation, service tradeoffs, and net benefit were not evaluated. A restriction can suppress legitimate work. Because telemetry omits IP, user, and session, entity fairness remains in the serving layer.

The replay has additional limits: 30 seeded perturbations reuse one cadence-shaped tape, the alert ceiling contains two clusters, daily spend has lag-1 autocorrelation 0.556, and a

block bootstrap widens the 99th-percentile tail about 14 percent. Bands are triggers, not calibrated intervals, and dollar estimates are not invoice-reconciled. Evidence under my control is not independent replication or runtime attestation. The restricted attack scenarios are neither global worst-case proofs nor live attack benchmarks.

Reference v2 and explicit acknowledgment-source handling are tested but undeployed; historical rows lack exact limits-version correlation. The clamp surface has no complete freshness or coverage signal, so stale state can fail closed while absence can fail open. About 87 percent of the \$317.93 price map uses one assumed model tier; documented alternatives move the total by roughly 20 percent. The wider static grid is a post-audit sensitivity check. These gaps narrow dollar and actuation claims.

9 Conclusion

As a platform, Mathlimit demonstrates the broader value of connecting a portfolio manager’s mindset to a risk desk for fleets of LLMs and vector stores, moving from passive telemetry and costly surprises toward explicit limits, reversible controls, and scenario-based planning for runaway spend.

The dashboard also informed a manual cadence reduction, followed by a measured 67.9 percent decrease in estimated daily cost over the stated comparison dates. Feedback from this platform enabled me to make an informed decision to reduce the majority of my spend. It also gave me the direct control and confidence to scale my fleet without fear of runaway spend. Provider notifications separately record seven organization-wide stops in the 62 days before the desk and one in the 44 days after the freeze, with none in the last 34 days through August 21. My changes and book expansion share that interval, so the change is descriptive. One armed position supports a reconstructible integration case.

AI-assistance disclosure. An AI ensemble assisted with pressure testing, review, and validation. The author originated and verified the work and takes full responsibility.

References

- Adams, C.; Cain, P.; Pinkas, D.; and Zuccherato, R. 2001. Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP). Technical Report RFC 3161, Internet Engineering Task Force.
- Amazon Web Services. 2020a. AWS Budgets Actions. <https://docs.aws.amazon.com/cost-management/latest/userguide/budgets-controls.html>. Accessed: 2026-08-30.
- Amazon Web Services. 2020b. AWS Cost Anomaly Detection. <https://docs.aws.amazon.com/cost-management/latest/userguide/manage-ad.html>. Accessed: 2026-08-30.
- Anthropic. 2026. Rate Limits and Spend Limits. <https://platform.claude.com/docs/en/api/rate-limits>. Accessed: 2026-09-02.
- Artzner, P.; Delbaen, F.; Eber, J.-M.; and Heath, D. 1999. Coherent Measures of Risk. *Mathematical Finance*, 9(3): 203–228.
- Basel Committee on Banking Supervision. 1996. Supervisory Framework for the Use of “Backtesting” in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. Technical report, Bank for International Settlements.
- Bernet, Y.; Blake, S.; Grossman, D.; and Smith, A. 2002. An Informal Management Model for Diffserv Routers. RFC 3290, Internet Engineering Task Force.
- BerriAI. 2026. LiteLLM, Version 1.98.0: An Open-Source LLM Gateway with Per-Key Budgets and Rate Limits. <https://github.com/BerriAI/litellm/releases/tag/v1.98.0>. Accessed: 2026-08-30.
- Besanson, G. 2026. Green SARC: Predictive Cost and Carbon Governance for Agentic AI Systems. arXiv:2606.15954.
- Beyer, B.; Jones, C.; Petoff, J.; and Murphy, N. R., eds. 2016. *Site Reliability Engineering: How Google Runs Production Systems*. O’Reilly.
- Beyer, B.; Murphy, N. R.; Rensin, D. K.; Kawahara, K.; and Thorne, S., eds. 2018. *The Site Reliability Workbook*. O’Reilly.
- Biggio, B.; and Roli, F. 2018. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Pattern Recognition*, 84: 317–331.
- Chakraborti, S.; and Graham, M. A. 2019. *Nonparametric Statistical Process Control*. Wiley.
- Chen, L.; Zaharia, M.; and Zou, J. 2024. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Transactions on Machine Learning Research*.
- Chen, Y.; Chen, J.; He, C.; Li, Y.; Ji, Y.; Wu, Y.; Yang, D.; Diao, L.; Shou, L.; Zhang, H.; Li, H.; and Chen, G. 2026. Token Economics for LLM Agents: A Dual-View Study from Computing and Economics. arXiv:2605.09104.
- Chiu, D.-M.; and Jain, R. 1989. Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks. *Computer Networks and ISDN Systems*, 17(1): 1–14.
- Christoffersen, P. F. 1998. Evaluating Interval Forecasts. *International Economic Review*, 39(4): 841–862.
- Dorsett, M.; Mann, S.; Chowdhury, J.; and Mahmood, A. 2025. A Comprehensive Review of Denial of Wallet Attacks in Serverless Architectures. arXiv:2508.19284.
- Hawkes, A. G. 1971. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1): 83–90.
- Huber, P. J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35(1): 73–101.
- Jellinek, R.; Zhai, Y.; Ristenpart, T.; and Swift, M. M. 2014. A Day Late and a Dollar Short: The Case for Research on Cloud Billing Systems. In *6th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 14)*. Philadelphia, PA: USENIX Association.
- J.P. Morgan and Reuters. 1996. RiskMetrics Technical Document. Technical report, J.P. Morgan.
- Kelly, D.; Glavin, F. G.; and Barrett, E. 2021. Denial of Wallet: Defining a Looming Threat to Serverless Computing. *Journal of Information Security and Applications*, 60: 102843.
- Khan, S. 2026. Token Budgets: An Empirical Catalog of 63 LLM-Agent Budget-Overrun Incidents, with an Affine-Typed Rust Mitigation as a Case Study. arXiv:2606.04056.
- Kupiec, P. H. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives*, 3(2): 73–84.
- Laptev, N.; Amizadeh, S.; and Flint, I. 2015. Generic and Scalable Framework for Automated Time-Series Anomaly Detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1939–1947.
- Li, Y.; Deng, W.; Li, J.; and Li, X. 2026. Spend Less, Reason Better: Budget-Aware Value Tree Search for LLM Agents. arXiv:2603.12634.
- Luo, J.; Dai, J.; Liu, F.; Peng, S.; Shi, Y.; Bu, T.; Hong, G.; Pan, X.; and Zhang, Y. 2026. Autonomy Comes with Costs: Detecting Denial-of-Service Vulnerabilities Caused by Resource Abusing in LLM-Based Agents. In *35th USENIX Security Symposium (USENIX Security 26)*. Baltimore, MD: USENIX Association.
- Maharaj, A. V.; Arbour, D.; Lee, D.; Bhattacharya, U.; Rao, A.; Zane, A.; Feller, A.; Qian, K.; and Li, Y. 2025. Evaluation and Incident Prevention in an Enterprise AI Assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28): 28931–28937.
- Montgomery, D. C. 2019. *Introduction to Statistical Quality Control*. Wiley, 8th edition.
- Muralidhar, A.; Chitlangia, S.; Agarwal, R.; and Ahmed, M. 2023. Real-Time Detection of Robotic Traffic in Online Advertising. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15551–15559.
- Nygard, M. T. 2007. *Release It! Design and Deploy Production-Ready Software*. Pragmatic Bookshelf.

Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J. E.; Kadous, M. W.; and Stoica, I. 2025. RouteLLM: Learning to Route LLMs with Preference Data. In *International Conference on Learning Representations*.

OWASP Foundation. 2025. OWASP Top 10 for LLM Applications, v2025: LLM10, Unbounded Consumption. <https://genai.owasp.org/llmrisk/llm102025-unbounded-consumption/>. Accessed: 2026-08-30.

Page, E. S. 1954. Continuous Inspection Schemes. *Biometrika*, 41(1/2): 100–115.

Qiu, P. 2014. *Introduction to Statistical Process Control*. Chapman and Hall/CRC.

Roberts, S. W. 1959. Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 1(3): 239–250.

Shewhart, W. A. 1931. *Economic Control of Quality of Manufactured Product*. Van Nostrand.

Sivaroopan, N.; Thilakarathna, K.; Zomaya, A.; Manu; Guo, Y.; Plested, J.; Lynar, T.; Yang, J.; and Yang, W. 2026. SHIELD: An Auto-Healing Agentic Defense Framework for LLM Resource Exhaustion Attacks. arXiv:2601.19174.

Tambe, M. 2011. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

Todd, P. 2016. OpenTimestamps. <https://opentimestamps.org>. Accessed: 2026-08-30.

Wang, X.; Liu, Z.; Feng, S.; Yuan, P.; Li, Y.; Shi, J.; Zhang, Y.; Tan, C.; Zhang, J.; Pan, B.; Hu, Y.; and Li, K. 2026. On Time, Within Budget: Constraint-Driven Online Resource Allocation for Agentic Workflows. arXiv:2605.06110.

Welsh, M.; Culler, D.; and Brewer, E. 2001. SEDA: An Architecture for Well-Conditioned, Scalable Internet Services. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP)*, 230–243.

Whalley, A. E.; and Wilmott, P. 1997. An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs. *Mathematical Finance*, 7(3): 307–324.

Yuan, Y.; Nayak, A.; Kundu, S.; and Talati, N. 2026. Agentic AI Workload Characteristics. arXiv:2605.26297.

Zavrak, S. 2026. Content-Aware Attack Detection in LLM Agent Tool-Call Traffic: An Empirical Study of Features, Architectures, and Evaluation Protocols. arXiv:2605.11053.

Zhang, Y.; Wang, X.; Gao, H.; Zhou, Z.; Meng, F.; Zhang, Y.; and Su, S. 2025. PD3F: A Pluggable and Dynamic DoS-Defense Framework Against Resource Consumption Attacks Targeting Large Language Models. arXiv:2505.18680.

Zhou, K.; Zheng, Y.; He, Y.; Xue, M.; Gong, X.; Wang, Y.; Zhang, X.; and Lam, K.-Y. 2026. Beyond Max Tokens: Stealthy Resource Amplification via Tool Calling Chains in LLM Agents. arXiv:2601.10955.

Zhou, Z.; Zhang, C.; Ma, L.; Gu, J.; Qian, H.; Wen, Q.; Sun, L.; Li, P.; and Tang, Z. 2023. AHPA: Adaptive Horizontal Pod Autoscaling Systems on Alibaba Cloud Container Service for Kubernetes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15621–15629.