AgentOps 2.0: Evals, Observabillity & Cost

Engineering the Production
AI Control Plane

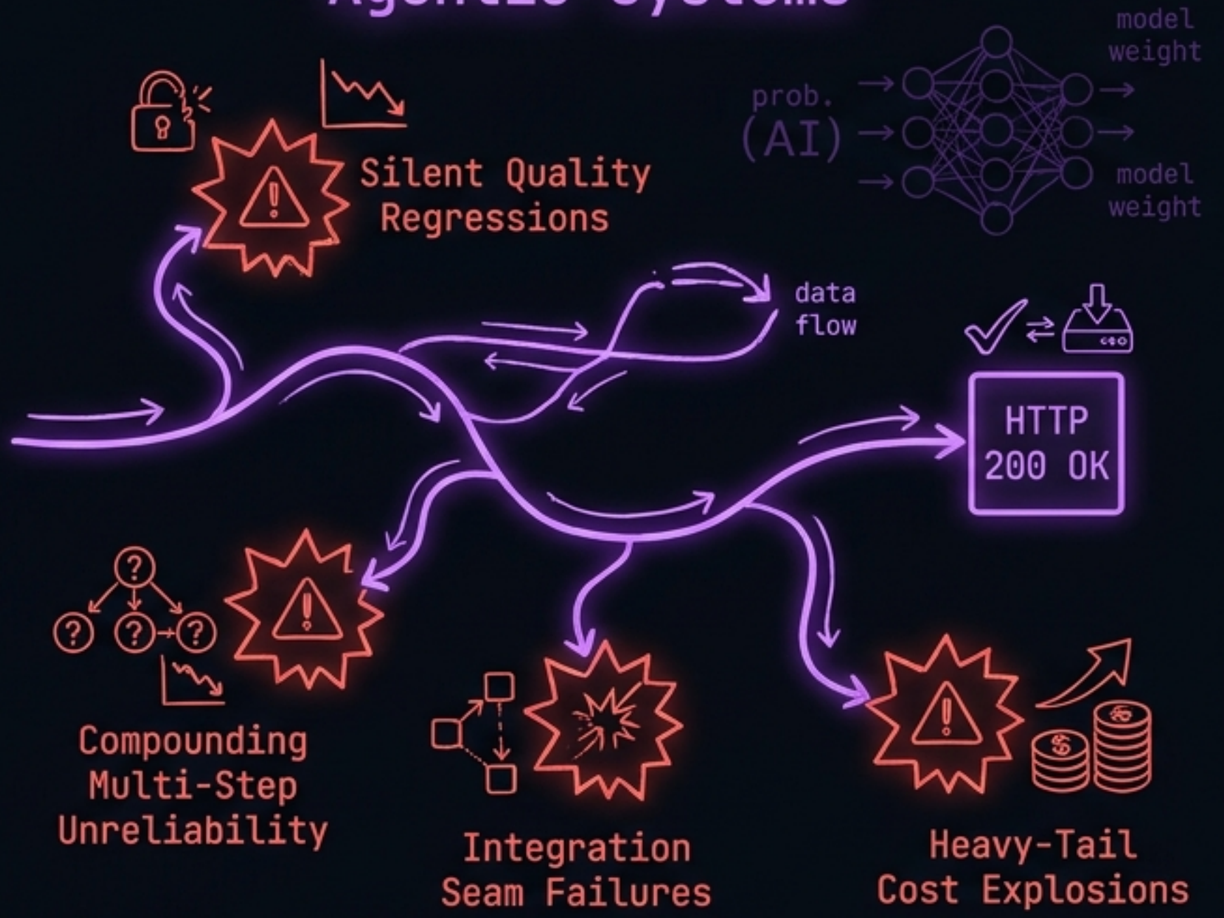


The Illusion of HTTP 200 OK

Traditional DevOps



Agentic Systems



Classical verification cannot detect “right answer, wrong trajectory” policy violations.

The Mental Model: Deterministic Containment



Prompts are probabilistic components. Your observability, simulation, and resilience stack form the deterministic containment vessel around them. Measure and price variance at every boundary.

The 3-Tier Grading Hierarchy

Tier 3: Human Domain Experts

Use to build golden sets
and calibrate Tier 2.

— Hours/Days · $50-$200+/hr

Tier 2: Calibrated Autoraters (LLM-as-Judge)

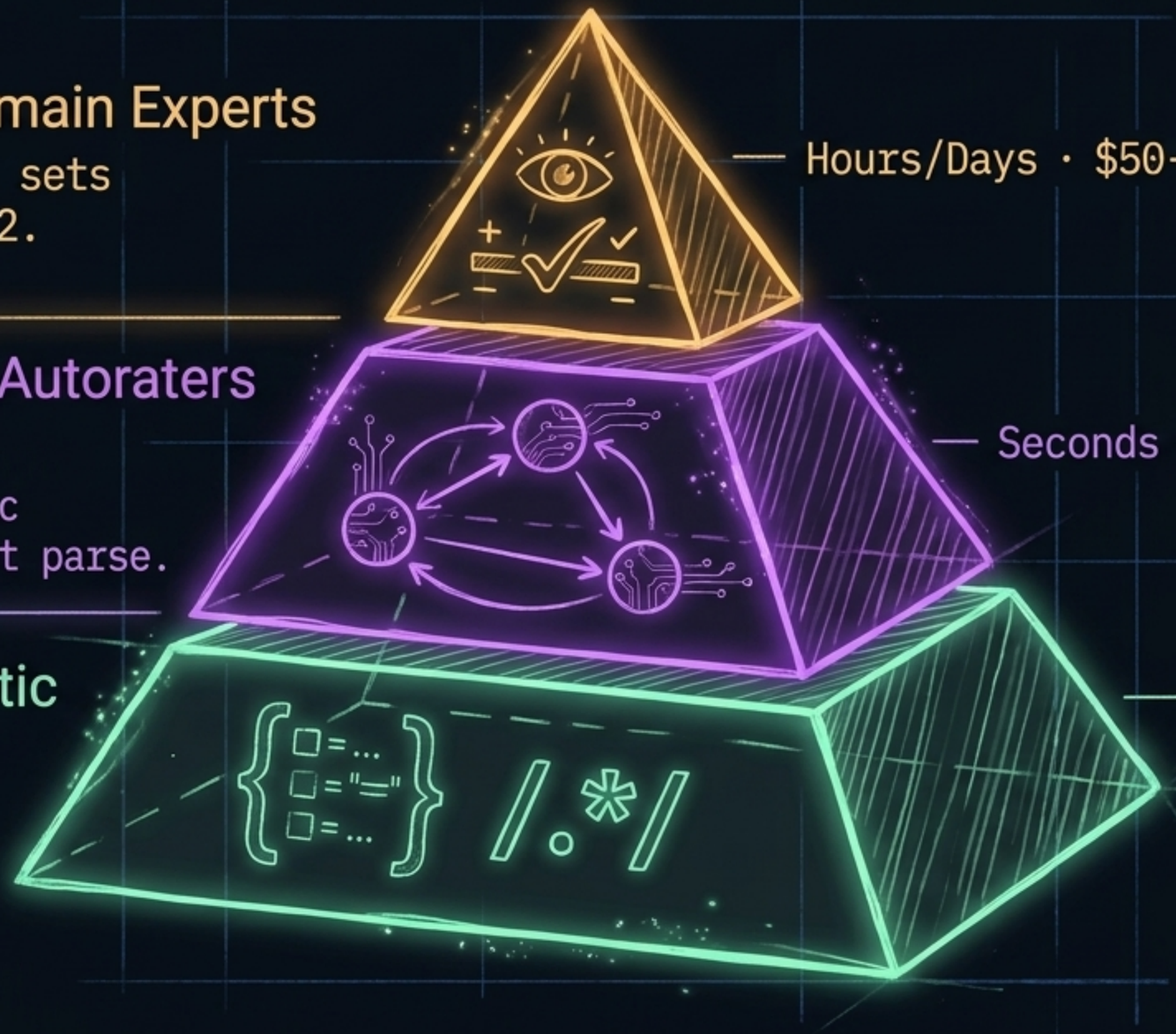
Use only for semantic
qualities code cannot parse.

— Seconds · Token Cost

Tier 1: Deterministic Code Graders

Always run first.
Gate every commit.

— Milliseconds · $0 ·
100% Reproducible



Mitigating LLM-as-Judge Measurement Biases

Verbosity Bias

Rewarding long, polite fluff.

Explicitly penalize length; grade atomic factual precision.

Positional Bias

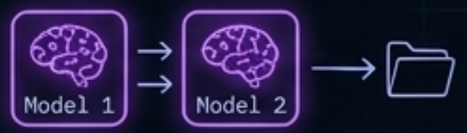
Favoring Slot A (primacy) or Slot B (recency).

Run every A/B comparison twice with swapped order. Split decisions tie.

Self-Preference Bias

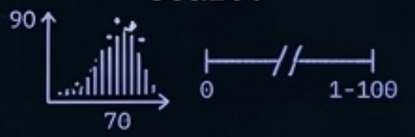
Overlooking its own blind spots.

Grade with a different/higher-capability model tier in a fresh, isolated context window.

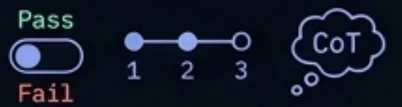


Score Inflation

Arbitrary clustering around 7–9 on a 1-100 scale.



Enforce binary Pass/Fail checks or 1-3 anchored scales. CoT rationale must precede the score.



Validation Target: Inter-annotator agreement (Cohen's Kappa) must be $\kappa \geq 0.75$

The 4-Way Evaluation Tooling Matrix

Local Repo	CI/CD Gates	Cloud Batch	Live Monitors
Unit/trajectory regression.	PR blocking gates, release diffs.	Managed scale, Model migration, 2-Step Simulation.	Live prod trace scoring.
Dataset Contract: Single session *.test.json	Dataset Contract: Upgraded EvaluationDataset JSON.	Dataset Contract: Tabular data / BigQuery.	Dataset Contract: OpenTelemetry EVENT_ONLY (no ground truth).
Cost Driver: Local token overhead.	Cost Driver: Agent + judge tokens (sampling count 1-32).	Cost Driver: Batch inference + Judge calls.	Cost Driver: Zero agent overhead, but 5-31 judge calls per trace!

WARNING: You cannot run `trajectory_exact_match` on live production traffic–it requires a golden reference trace. Live traffic requires reference-free rubrics.

Gen AI Metric Catalogue & The Compliance Trap

Deterministic

`trajectory_in_order_match`
`trajectory_precision`

($0 Cost)

Adaptive Rubrics

`TOOL_USE_QUALITY`
`MULTI_TURN_TASK_SUCCESS`

(Synthesizes pass/fail dynamically. Multi-call LLM.)

Static Rubrics

`HALLUCINATION`
`SAFETY`

(Fixed criteria. Note: **safety_v3** runs 31x calls per trace!)

Custom Code

`CodeExecutionMetric`

(Regex, JSON validation.)

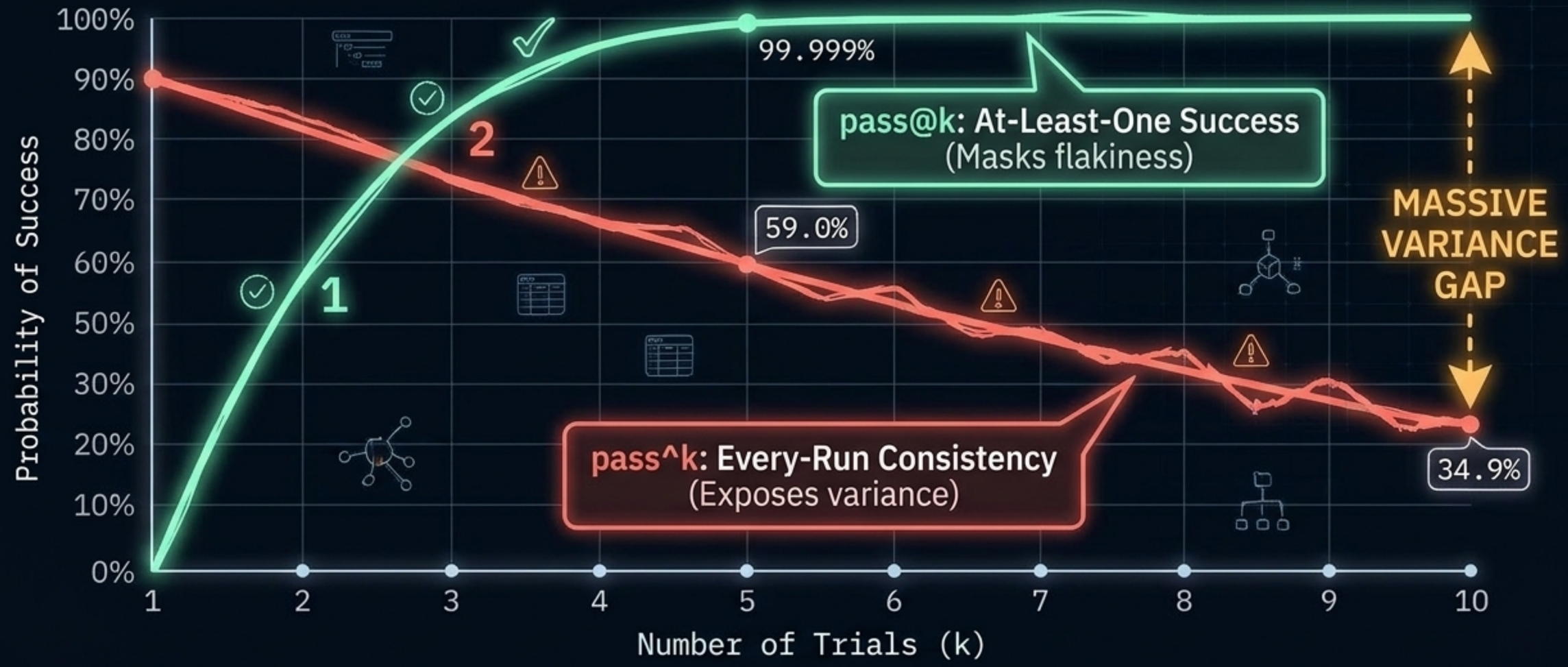
The 'Right Answer, Wrong Path' Trap



Grading solely on **`FINAL_RESPONSE_QUALITY`** passes non-compliant agents.

Always pair outcome autoraters with **`trajectory_in_order_match`**.

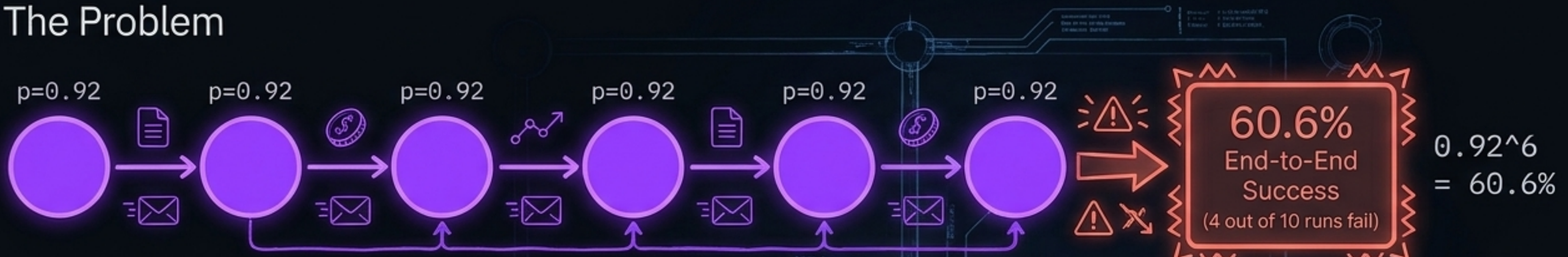
The Exponential Decay of Consistency



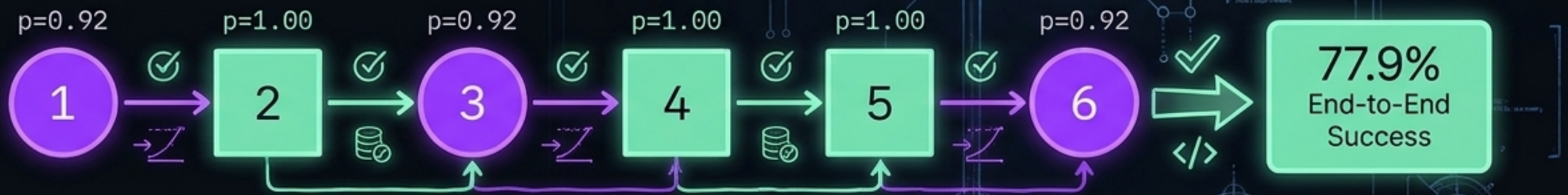
For customer-facing agents and mutating tool pipelines where every run must succeed, measuring `pass@k` is actively dong dangerous. You must gate releases on worst-case consistency: `pass^k`.

Compounding Multi-Step Accuracy

The Problem



The Architectural Remedy

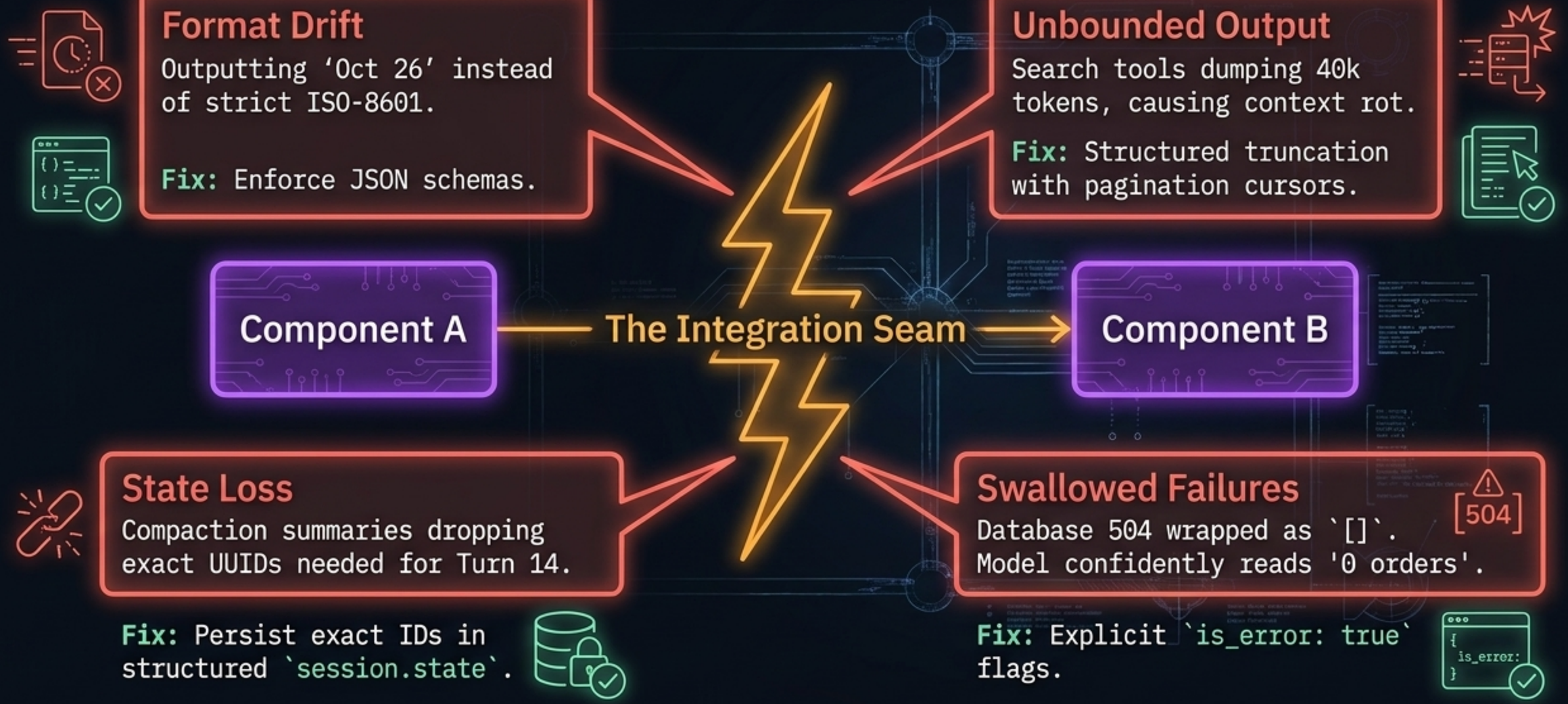


Replace probabilistic steps with deterministic workflow nodes.
Bound step counts and catch schema errors between steps.

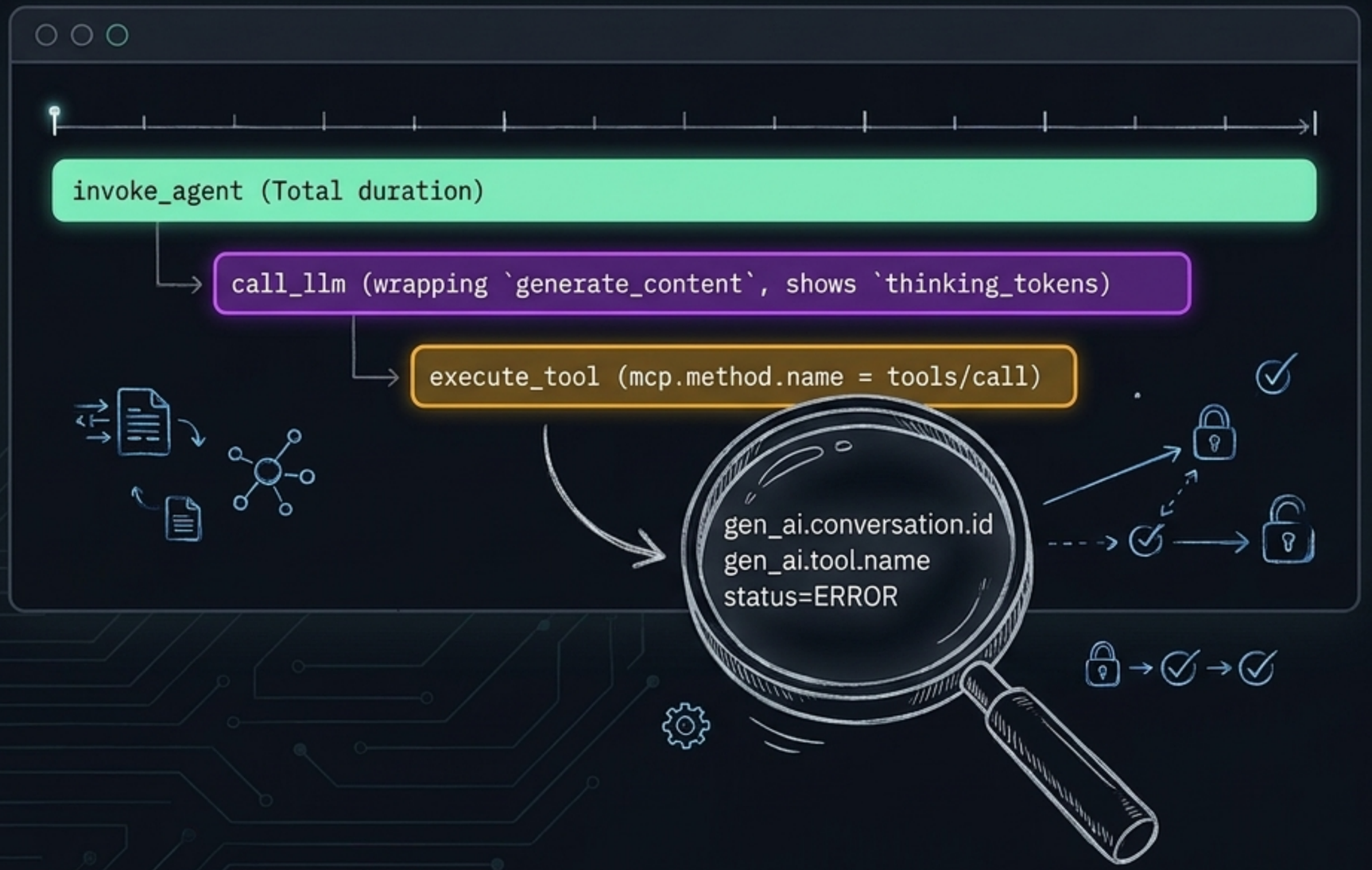
The 4-Level Agent Test Pyramid



Integration Seams: Where Agents Die



Observability & OpenTelemetry (OTel)



Critical Guardrails

1. **Dual-Attribution Audit Logging:** Logs tagged with `is_agentic=true` capture both the **SPIFFE Agent Identity** and the human principal.

2. **Media Offload:** Set `OTEL_INSTRUMENTATION_GENAI_COMPLETION_HOOK=upload` to route large large payloads to **buckets**, not **trace spans**.

Triage: Runaway Loops & Cascading Failures



`max_iterations` cap
Hard limit on the `LoopAgent`

`ReflectAndRetryToolPlugin`
Limits retries on a specific tool to 3, providing structured hints instead of vague errors.

Terminal Route mapping
Routing `DEFAULT_ROUTE` to a fallback node, never back into the cycle.

Triage Insight: **A loop must never end on model judgment alone. An iteration cap, a terminal route, or a budget must mathematically end it.**

Quality Decay: Input Drift vs. Model Drift

Input Drift



Symptoms:

New product SKUs, user phrasing changes.
Infrastructure is green, but retrieval similarity scores drop and Out-of-Domain queries rise.



Fix:

Refresh RAG corpus, expand golden set via failure clustering.

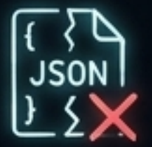


Model Drift



Symptoms:

A sudden step-change in JSON errors, refusal rates, or tool errors rates, or tool selection across all topics simultaneously.
Unpinned aliases (e.g., `-latest`).





Fix:

Pin immutable exact model snapshot IDs. Run automated Level 2/3 evals before any cutover.





The LM Feedback Flywheel

1. Capture: Online Evaluation Monitors flag traces (1-5% sampled live traffic).



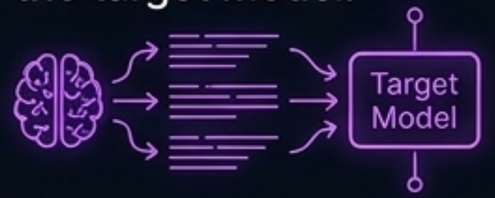
2. Scrub: Sensitive Data Protection (SDP) strips PII/credentials.



3. Analyze: Failure Cluster Analysis groups regressions by Frequency × Severity.



4. Optimize: Data-Driven Prompt Optimizer mutates instructions/demos against the target model.



5. Validate: Test against a Held-Out Split to ensure the prompt hasn't just memorized the training set.



Core Principle: Fix failure classes programmatically; stop playing manual "prompt whack-a-mole".

The Agentic Runtime Matrix



Agent Runtime (Turnkey)

Fit:

- Stateful ADK orchestrators.

Features:

- Built-in Sessions
- Memory Bank
- Code Execution sandbox

Ops:

- Lowest overhead, `min_instances` to absorb cold starts.



Cloud Run (Stateless)

Fit:

- Custom Streamable HTTP MCP servers, polyglot webhooks.

Features:

- Serverless scale-to-zero
- Ephemeral RAM (requires external Firestore/Redis).



GKE (Isolation & Control)

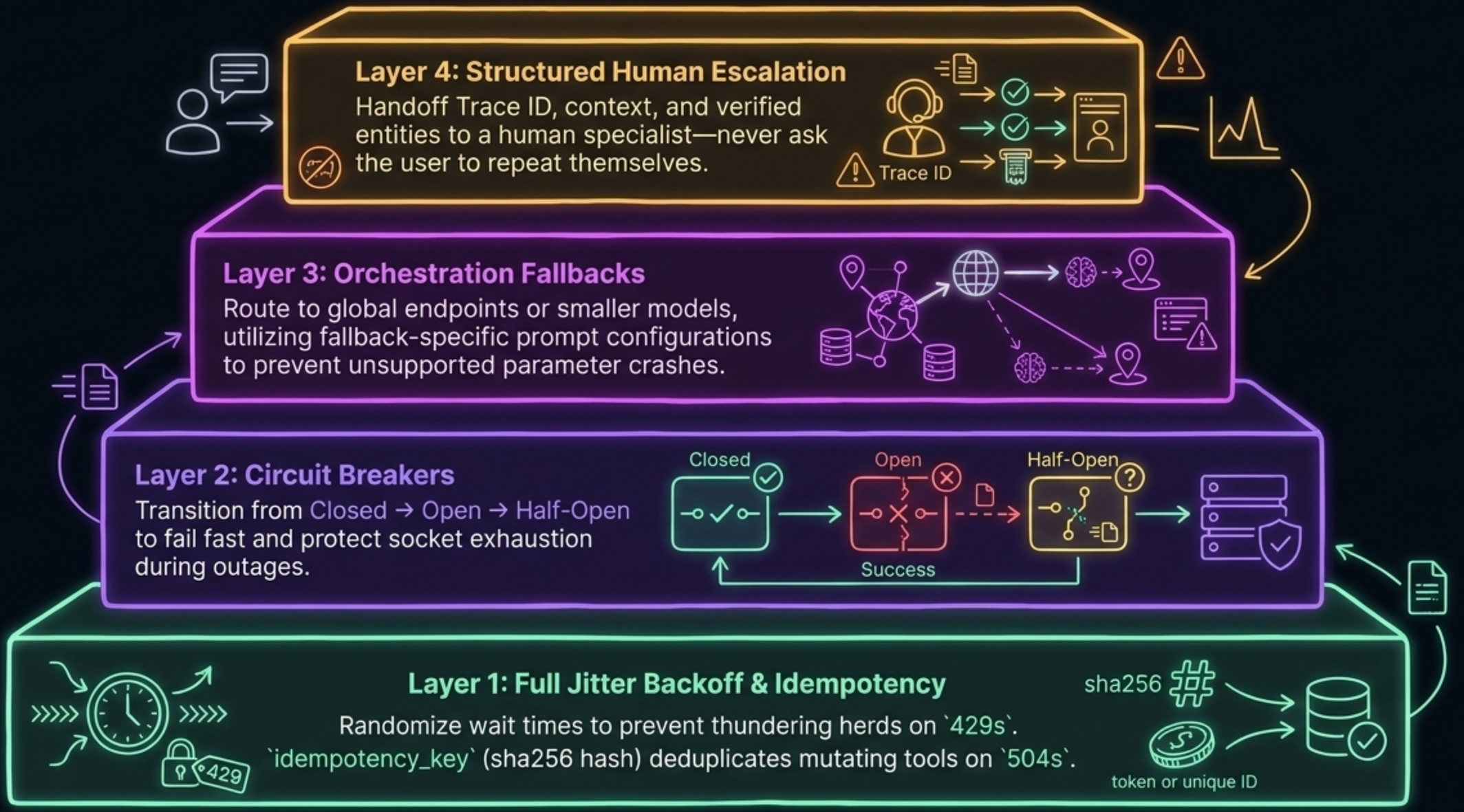
Fit:

- Dense multi-agent meshes, self-hosted open weights (Gemma 4) on GPUs.

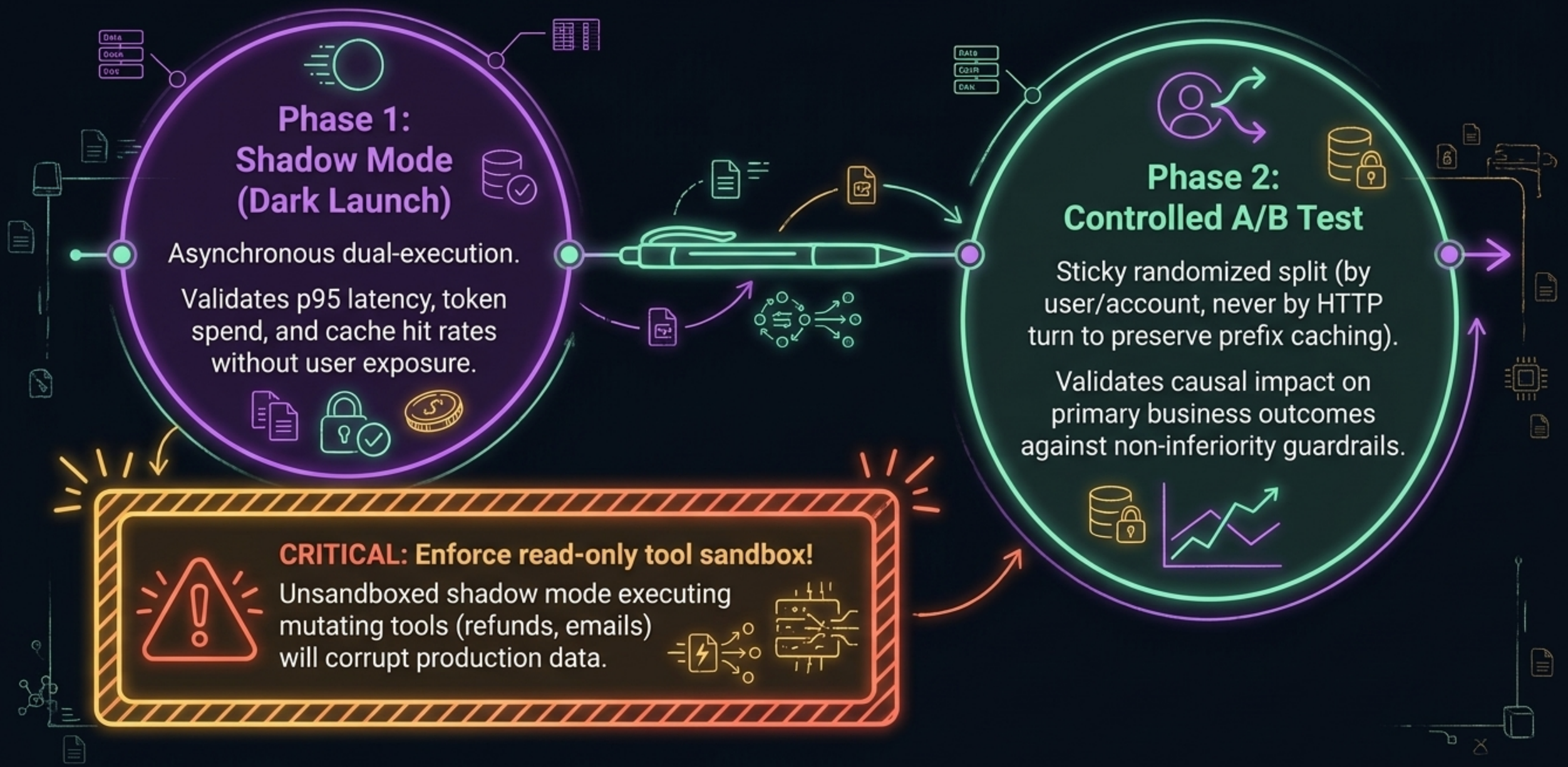
Features:

- Agent Substrate (actor-based <1s restore) or Agent Sandbox (gVisor kernel isolation for untrusted code).

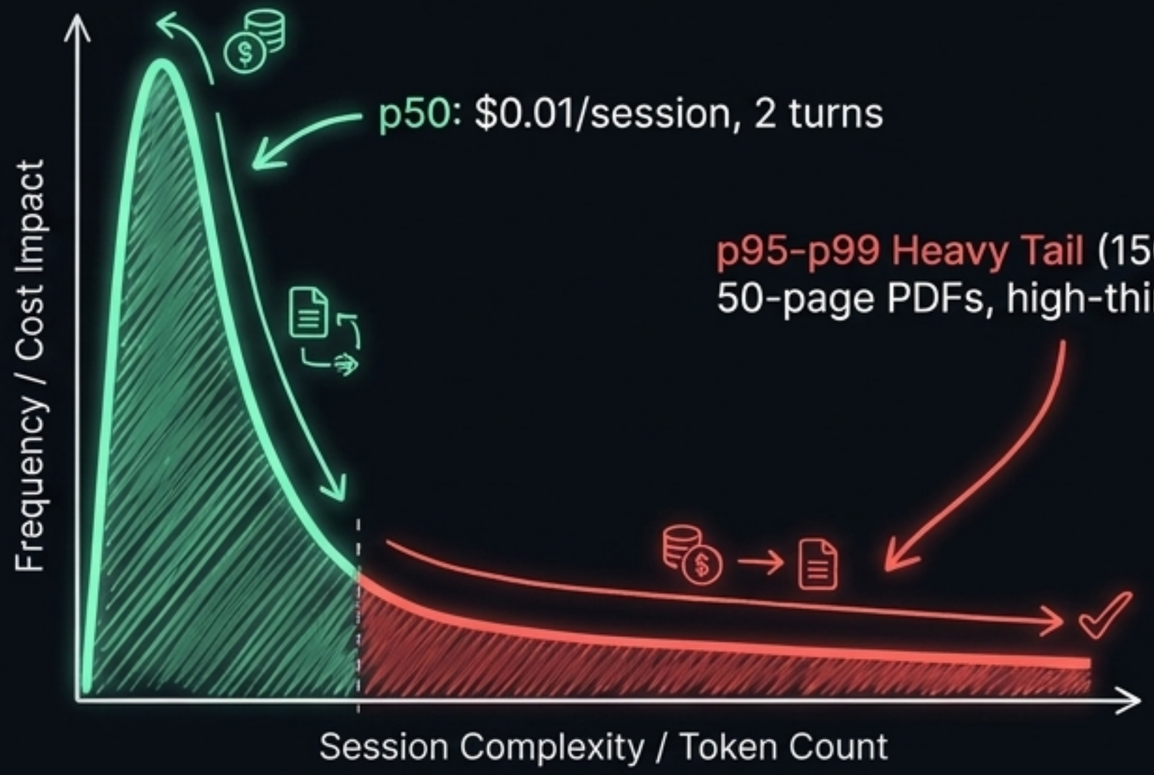
The 4-Layer Resilience Stack

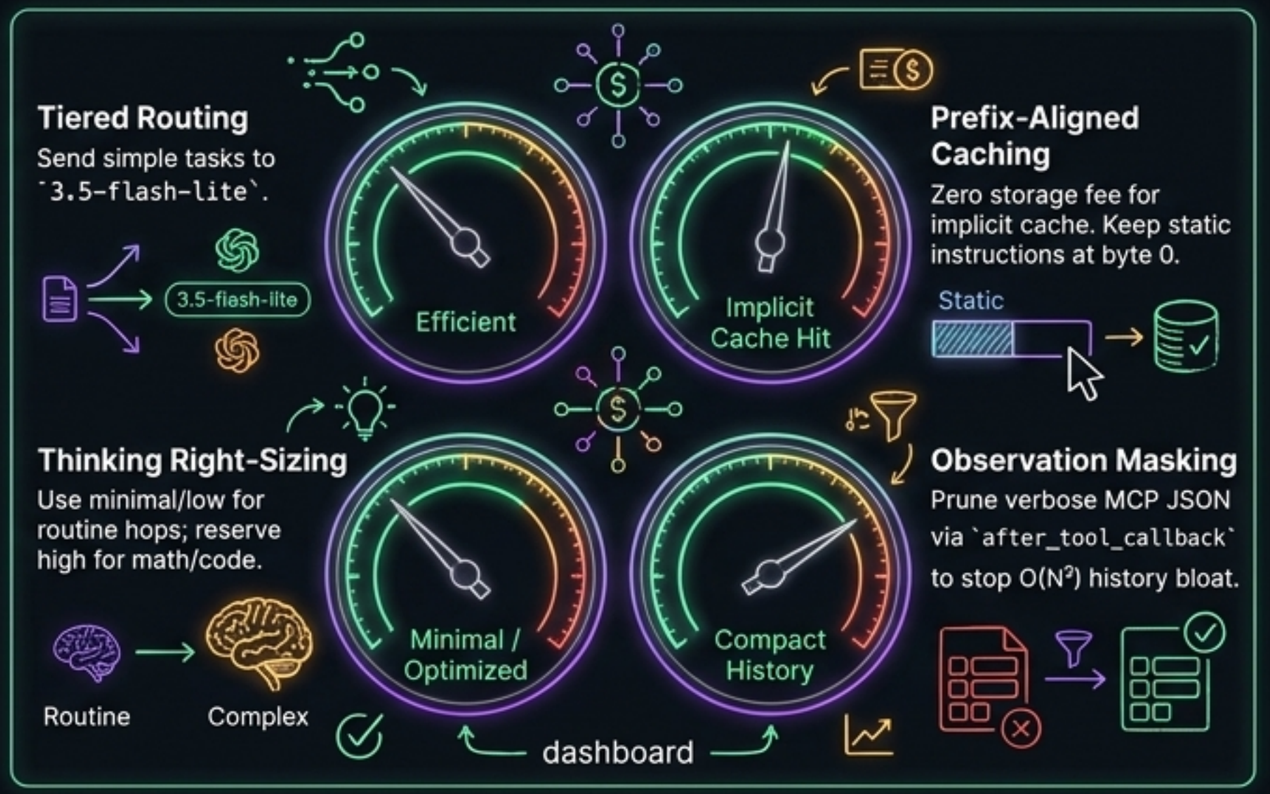


Safe Rollouts: Shadow Mode vs. Controlled A/B



Heavy-Tail Token Economics & FinOps





Step 4 ROI & Launch Readiness Triage

The Financial Ledger (TCO)

1	Gross Savings ("Mint", "token …")	→ +$480k
2	AI OpEx ("Amber", "comp …")	→ -$30k
3	**The Human-Review Tax** **-$315k** (Manual audit time + specialist rework)	→

If mandatory audit time approaches baseline manual time, ROI is negative. Fix via higher pass^k precision.

Launch Readiness

Go	Bounds met. E.g., Internal IT agent with fallback.
Conditional Go	High latency/risk. **Fix:** Async background execution + Human-in-the-Loop.
No-Go	Demands 100% determinism <5ms. **Fix:** Keep LLMs off the critical path.



Production Failure Diagnostic Matrix

Symptom	Root Cause	Remediation
97% task success, but refunds sent without ID checks.	“Right answer, wrong path.”	Add `trajectory_in_order_match`.
Cloud Monitoring billing spike on tokens.	High-cardinality `session_id` label.	Strip IDs; use BigQuery/Log Analytics for per-session p99.
Alert fatigue on 30s 503 blips.	Static single-window alerts.	Multi-window SLO burn-rate alerts (1-hr 10x fast burn, 24-hr 2x slow burn).
Live online eval causes 31x judge billing spike.	100% trace sampling on `SAFETY` v3.	Cap at 5% random sampling + max samples/run.

Trust the containment vessel, not the probabilistic core.