The Architecture of Sustainable Scale

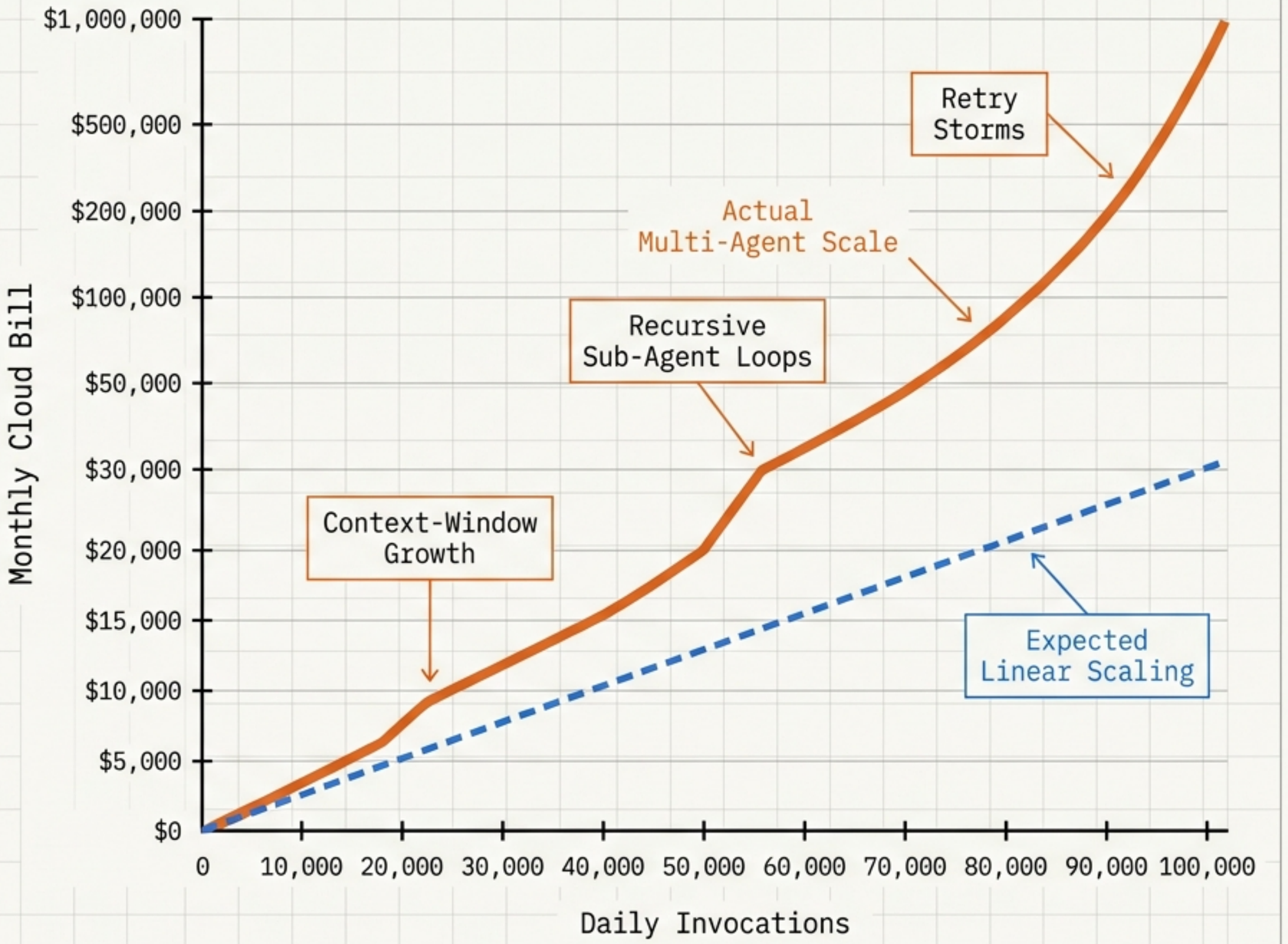
Operational Levers and Engineering Controls for Multi-Agent Systems

GEMINI ENTERPRISE AGENT PLATFORM | FINOPS & RELIABILITY RUNBOOK

NotebookLM

The Linear Scale Fallacy

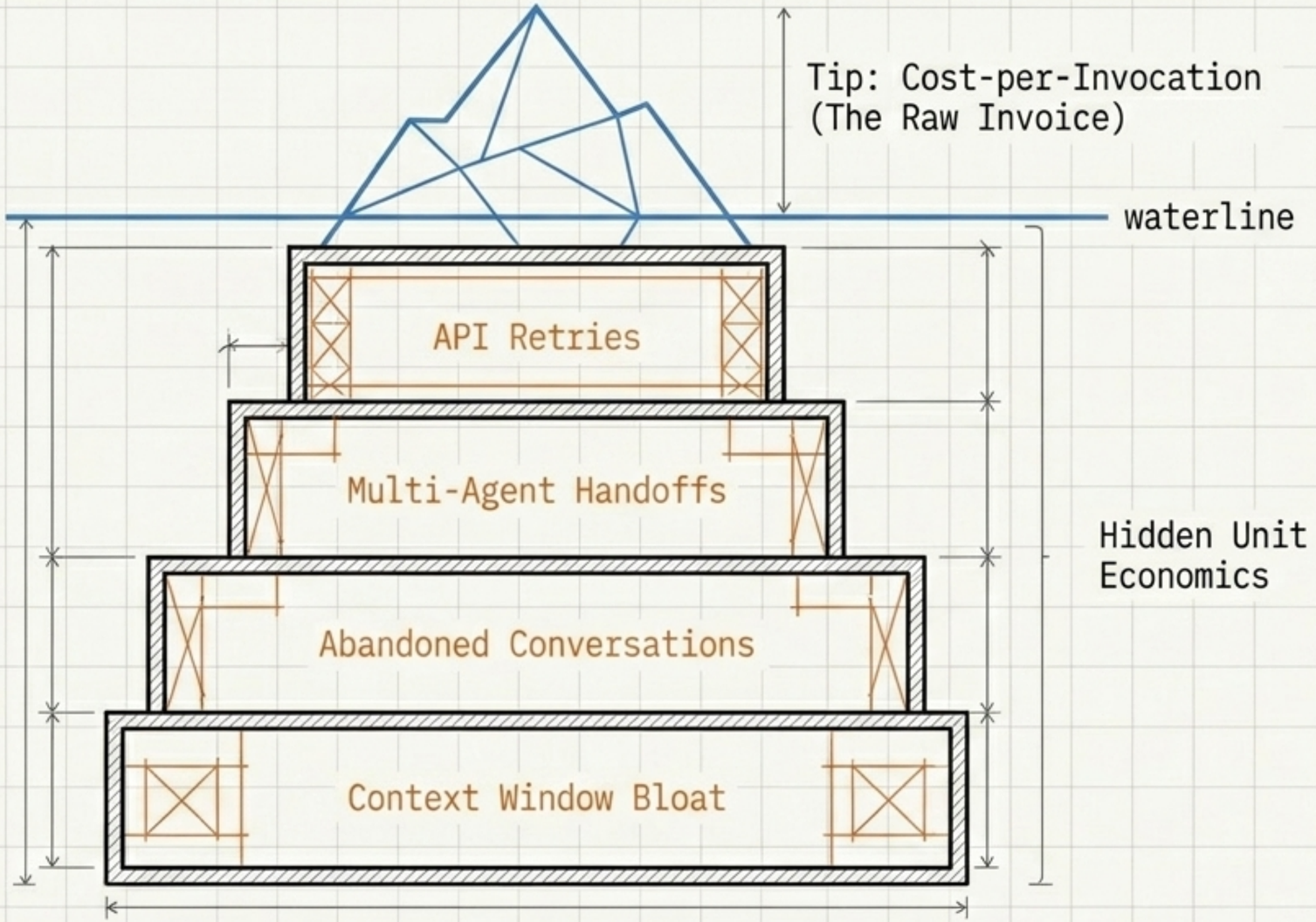
A multi-agent prototype costing $200/month at 1,000 invocations a day will not cost $20,000 at 100,000 invocations. At scale, super-linear traps surface.



Cost management requires structural circuit breakers, not just negotiated token rates.

NotebookLM

The Unit Economics That Actually Matter



Metrics Glossary

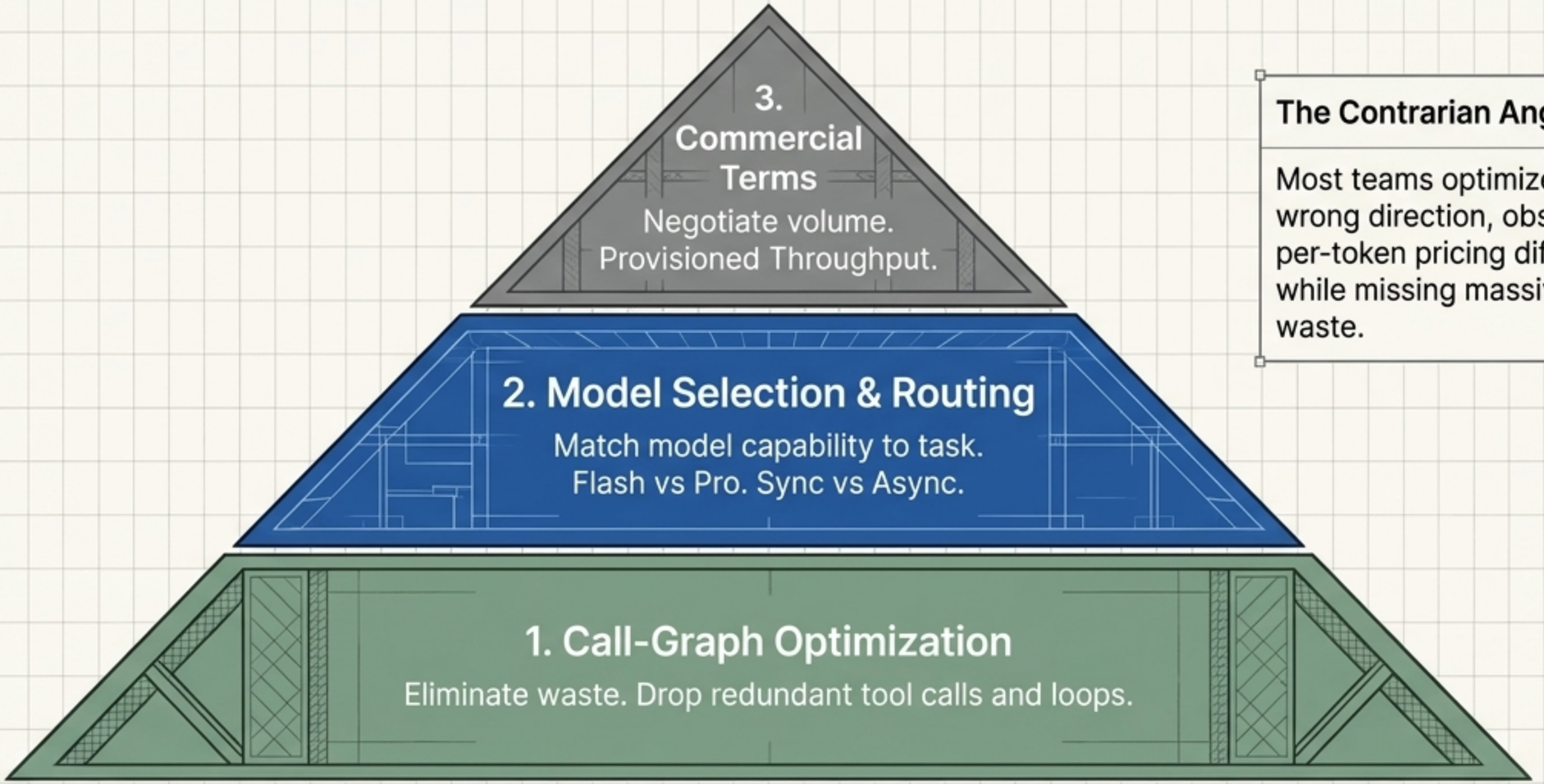
cost-per-invocation
Engineering hygiene.
Engineering hygiene. Easy to calculate, but hides operational truth.

cost-per-resolved-task
The operational truth. Corrects for retries, handoffs, and abandoned work. The ratio of bill to useful outcomes.

cost-per-active-user-month
The procurement-conversation metric.

NotebookLM

Cost is an Architectural Decision, Not a Billing Outcome



The Contrarian Angle

Most teams optimize in the wrong direction, obsessing over per-token pricing differences while missing massive structural waste.

EXAMPLE MATH: If a request costs $0.10 and redundant tool calls account for $0.062, eliminating the redundant calls drops cost-per-invocation by 62%—before changing a single model.

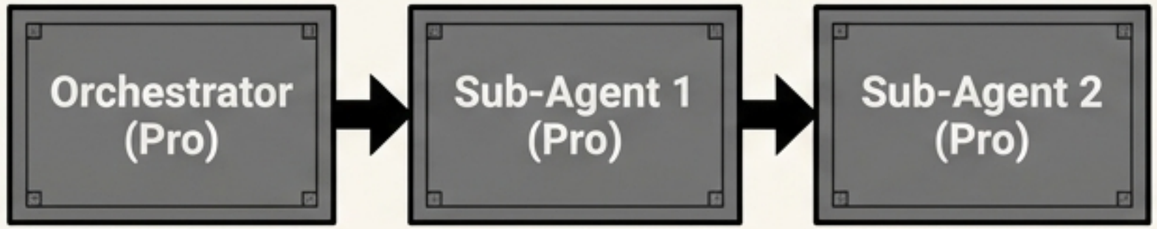
NotebookLM

Model Selection is the Largest Single Lever

Workload: 3-agent invoice pipeline | 60,000 invoices/month | **Pricing:** Flash is 13x cheaper than Pro

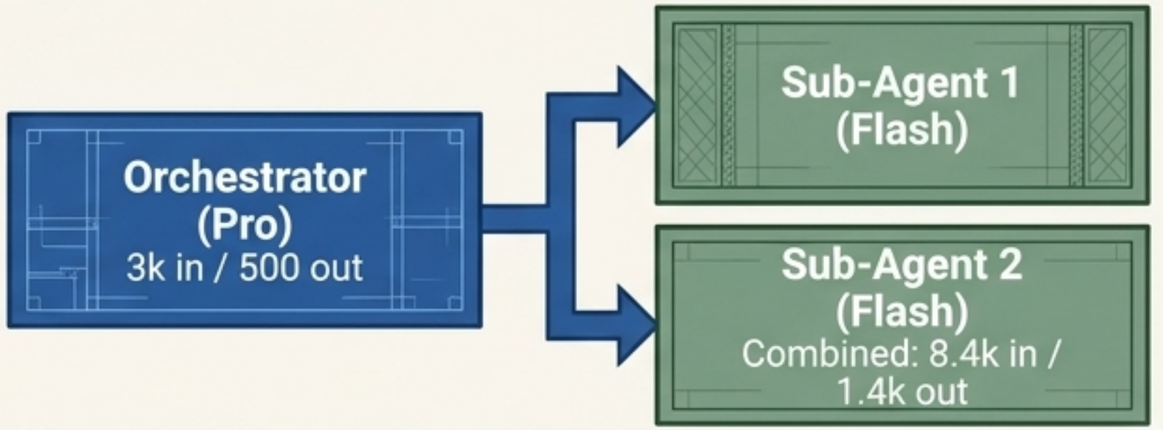
Token Shape Breakdown

Path A (Uniform Pro Strategy)



Total Output: $0.038 per invoice / $2,280 monthly.

Path B (Heterogeneous Routing)



Total Output: $0.012 per invoice / $726 monthly.

RESULT: Rebuilding the token shape with Pro everywhere is 3.1x more expensive for the exact same throughput.

NotebookLM

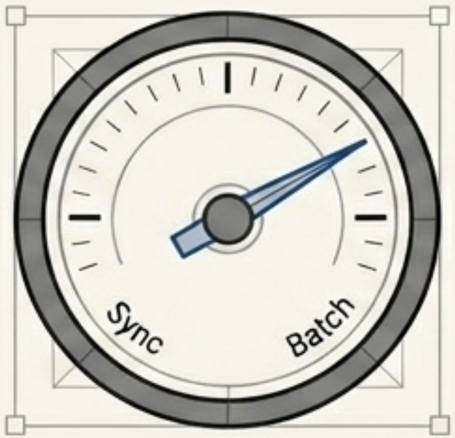
The Five Operational Levers

Precise engineering controls to keep production economics defensible.

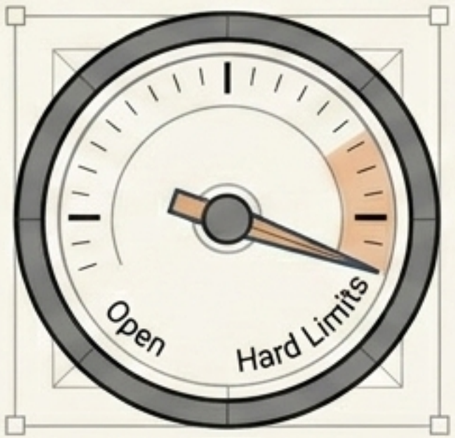
Lever 1: Throughput

Lever 2: Routing

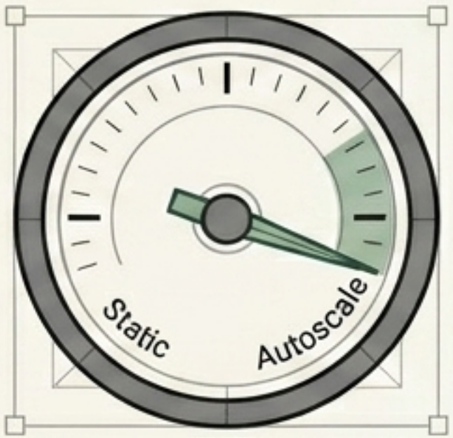
Lever 3: Execution



Lever 4: Guardrails

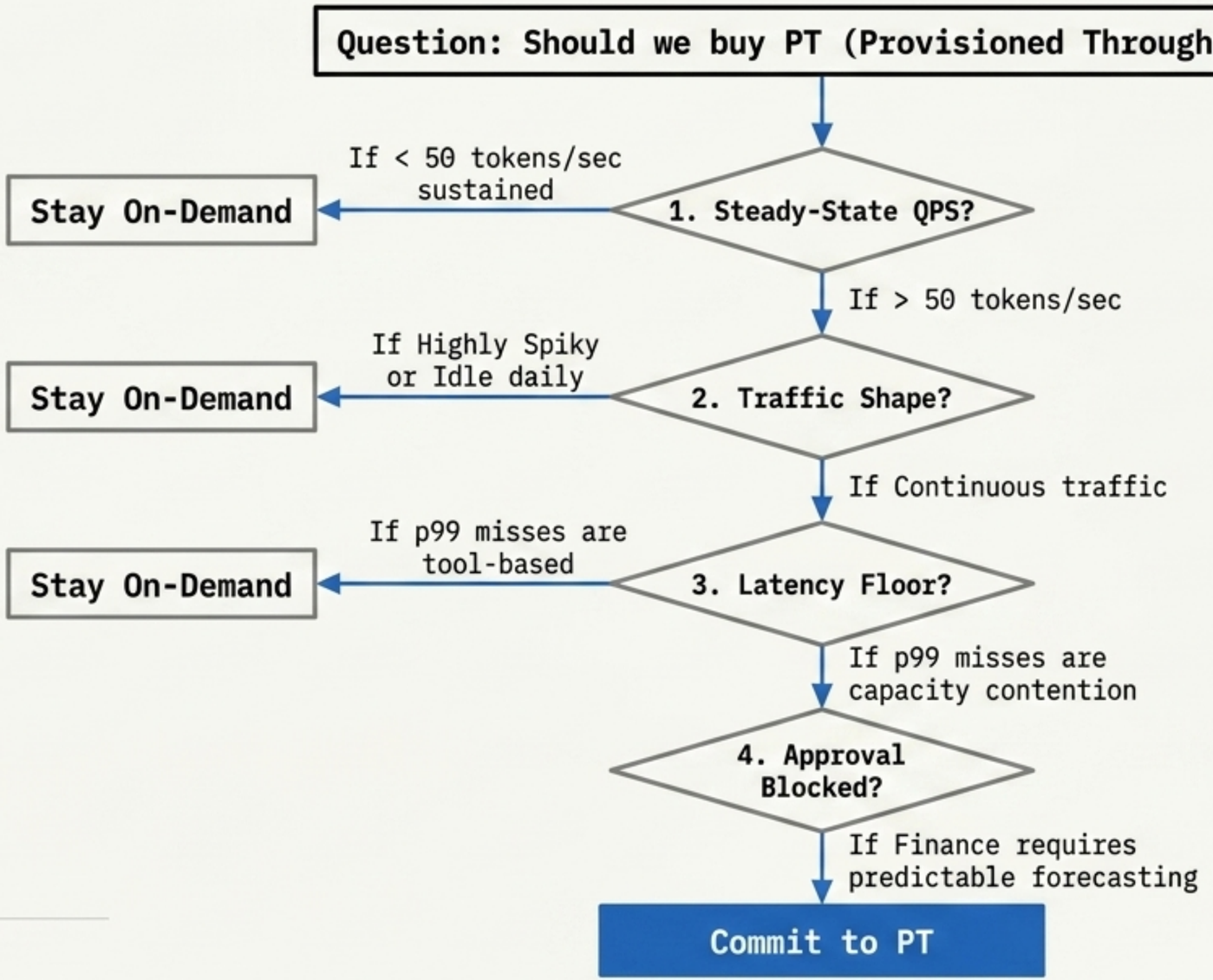


Lever 5: Compute



NotebookLM

Lever 1: The Provisioned Throughput Algorithm



PT is sold in 100-token/second generation units (~$35k/mo).

Most workloads stay on-demand until reliability or procurement constraints outweigh raw math.

NotebookLM

Lever 2: Routing & Endpoints

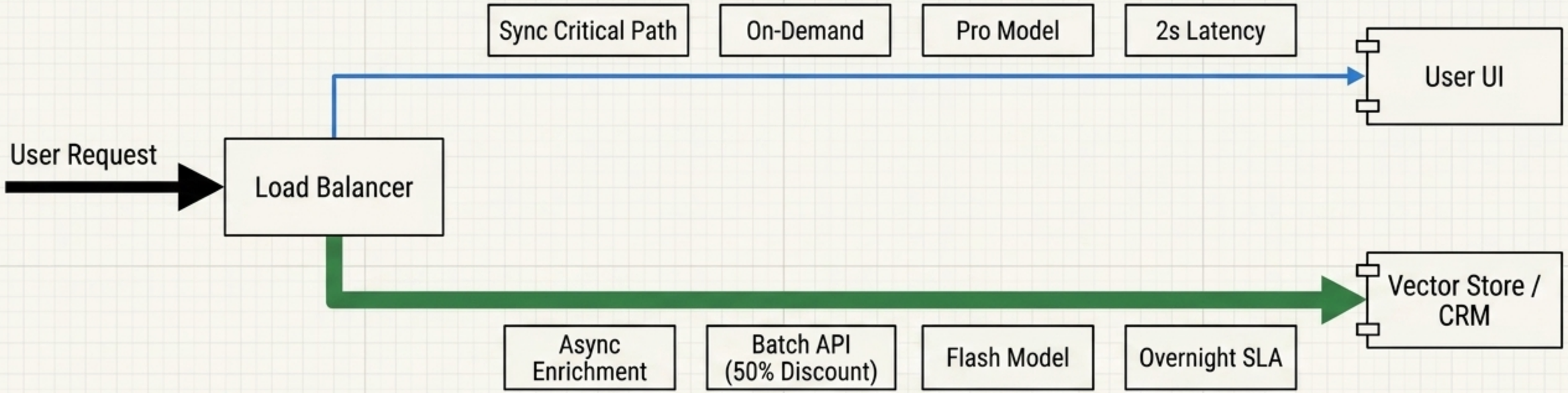
Constraint	Regional Endpoint	Global Endpoint
Compliance demands provable residency	[x] Required	[ ]
ML processing must stay in controlled geography	[x] Required	[ ]
Multi-region failover desired	[ ]	[x] Optimal
Highest possible pooled throughput	[ ]	[x] Optimal
Lowest possible latency for one geography	[x] Co-located	[ ]

THE SUBTLE TRAP: Deploying an agent regionally for residency, but calling the global model endpoint. The runtime is regional, but inference processing is uncontrolled.

RULE: Reject deployments that mix a residency-sensitive agent with a global endpoint.

NotebookLM

Lever 3: Batch Execution Asynchrony

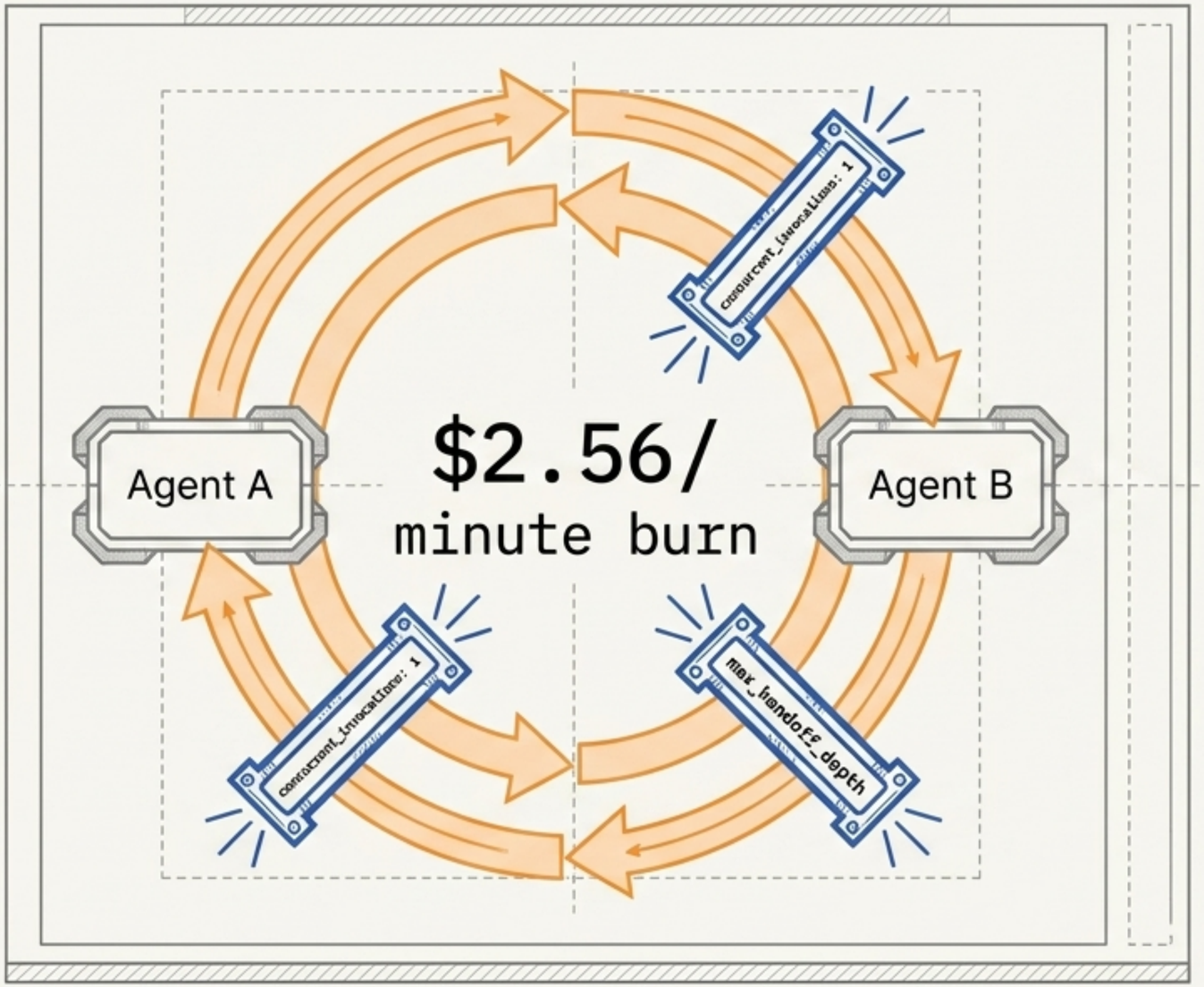


THE MATH:
Batch prediction cuts the token rate by 50%. Compounded with Flash pricing, batch runs at $0.075 per 1M input tokens.
Competitive with self-hosted open-source.

KEY INSIGHT:
The bill is the silent cost of simplicity. Split ‘need-it-now’ reasoning from ‘nice-to-have’ enrichments.

NotebookLM

Lever 4: Guardrails & Runaway Loops



THE INCIDENT ARITHMETIC:

40 invocations/min.

22k input tokens, 2.5k output tokens per invocation.

Model: $2/M input | $8/M output.

Result: The loop burns $2.56 every minute.

The Three Quota Tiers

1. **Project-level:** Caps total GCP throughput.
2. **Agent-level:** Rate limits to stop one rogue agent.
3. **Per-tenant:** Gateway limits preventing one user from starving others.

NotebookLM

Lever 5: Autoscaling Compute

Configuration A:
min_replicas: 0

- **Cost Profile:** Zero idle cost (extreme weekend savings).
- **Latency Impact:** First-request pays cold-start startup penalty.
- **Best For:** Internal tools, asynchronous batch processors.

Configuration B:
min_replicas: 3, idle_timeout: 3600

- **Cost Profile:** Fixed baseline runtime cost (paying for idle).
- **Latency Impact:** Consistently tight p99; zero startup latency.
- **Best For:** Customer-facing synchronous chat.

KEY INSIGHT:
The cost-vs-latency tradeoff lives entirely in the min_rlicas and idle_timeout knobs. You are purchasing a tighter p99 reliability with idle infrastructure spend.

NotebookLM

Cross-Vendor Benchmarking

True Cost Benchmark Matrix

Vendor A	
List Price:	$8/M out
Output Length:	2,100 tokens
First-pass Success:	90%
Raw Cost:	$0.0168
Retry-Adjusted True Cost:	$0.0187

Vendor B	
List Price:	$15/M out
Output Length:	1,400 tokens
First-pass Success:	97%
Raw Cost:	$0.0210
Retry-Adjusted True Cost:	$0.0216

RULE: Vendor pricing pages are **insufficient ground truth.** Benchmark on your real workload.

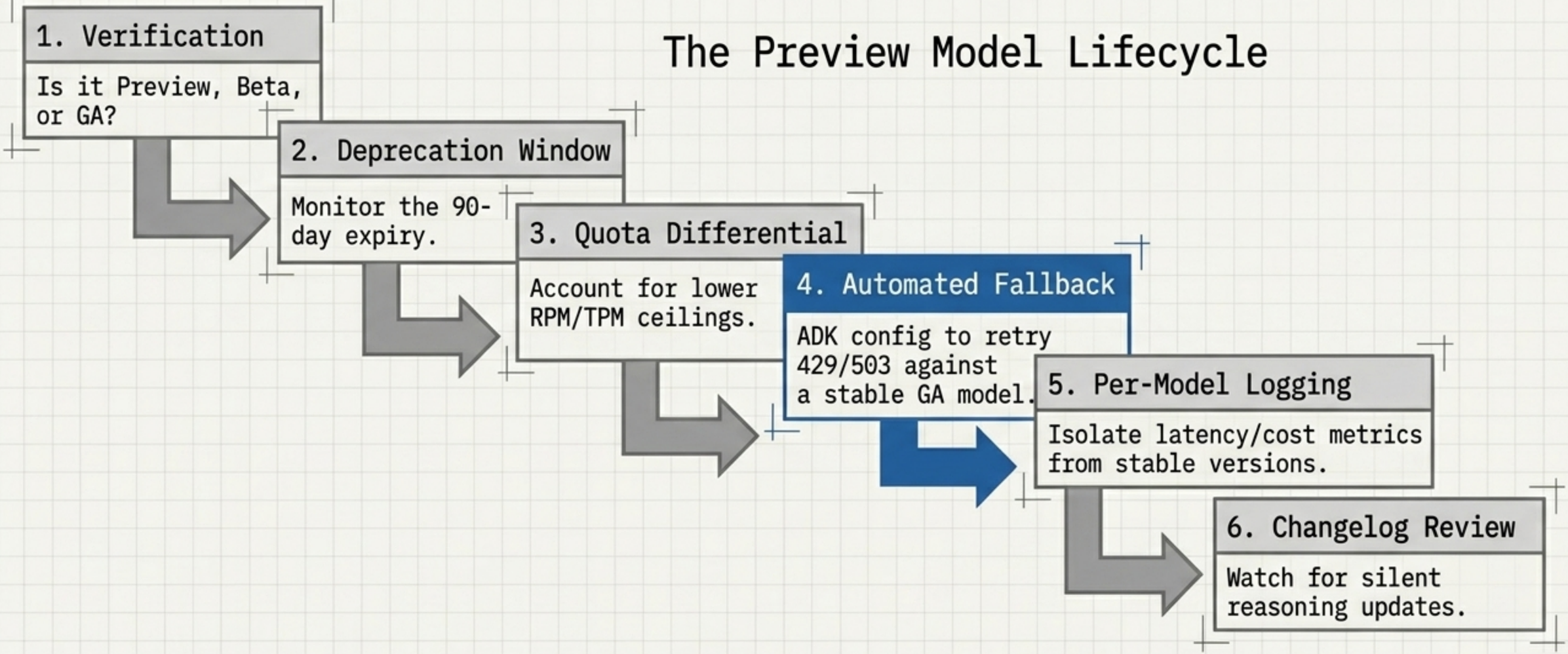
KEY INSIGHT: List price is only one of four cost dimensions. Total unit economics depend on token **efficiency**, **tool-call efficiency**, **retry rate**, and reasoning-quality-per-dollar.

NotebookLM

Safely Deploying Preview Models

Preview models offer top-tier reasoning but introduce lifecycle risk. Production requires a rigorous checklist.

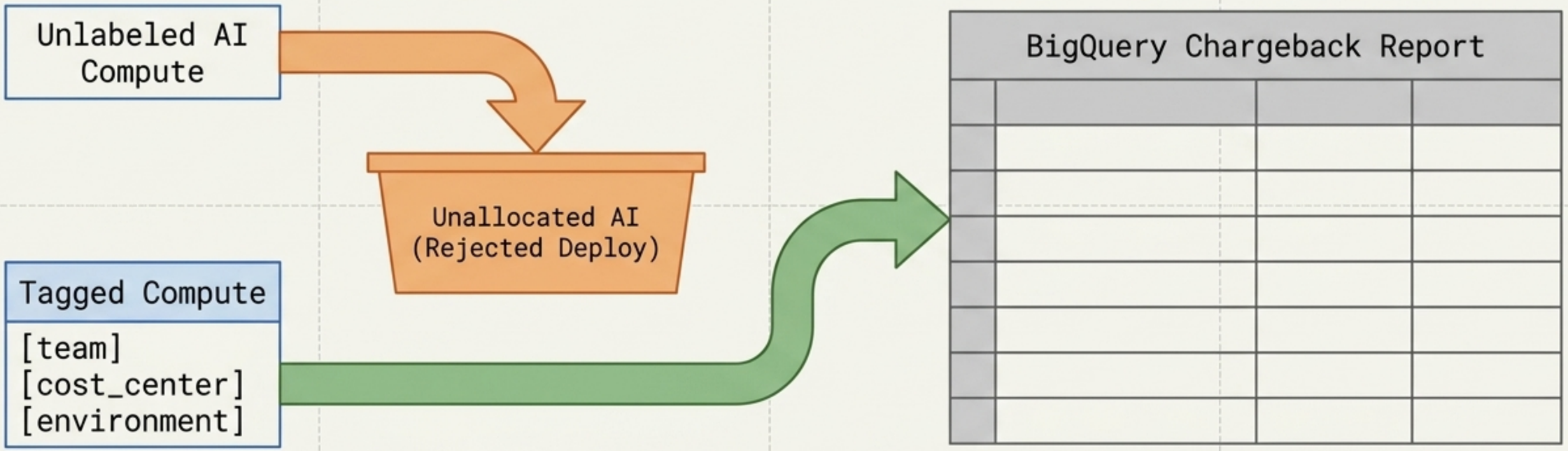
The Preview Model Lifecycle



NotebookLM

Cost Attribution & Governance

The Tagging Ledger



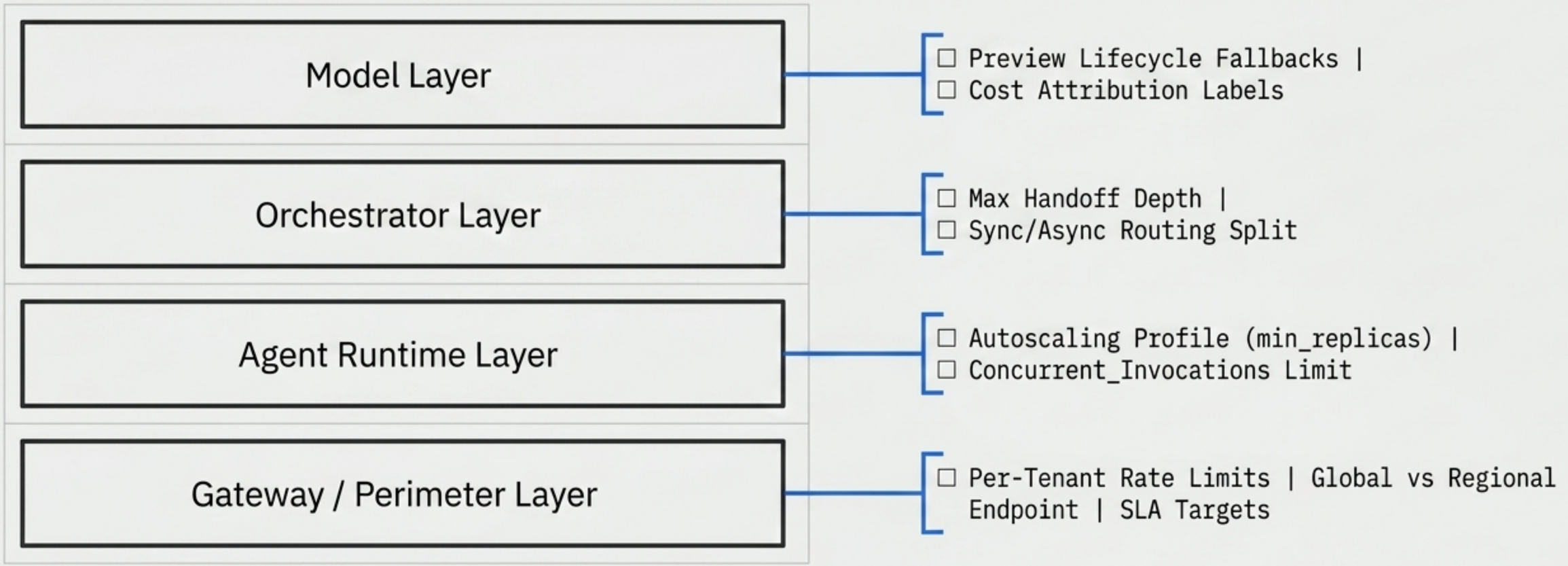
THE FINANCE QUESTION: "Which team pays for what?" This is only answerable if labels are wired at deploy time.

ACTIONABLE RULE: Make team, cost_center, and environment mandatory fields in the deploy template. Route untagged spend to a visible unallocated report until metadata is fixed.

NotebookLM

The Complete Production Runbook

Cost, reliability, and security are managed by the exact same engineering controls.



NEXT STEPS: Combine this architectural runbook with your observability and security controls for CISO and Finance sign-off.

NotebookLM