GOVERNED AI WORKFORCE · CIO

367 agents. One number per call. A bill you can defend.

At fleet scale, the risk is not a runaway model — it is cost sprawl no one can account for. This walkthrough makes every governed call legible: routed, metered, and priced.

Asphodel + Tiresias-ZT · ACME Corporation

THE STAKES

Cost sprawl is the real exposure

A general-purpose fleet drifts to the most expensive model that answers first. Multiply an unmanaged per-call cost across hundreds of agents running every day, and the spend is real long before anyone can attribute it.

THE FAILURE MODE

No per-agent metering. No per-model attribution. Premium models used by default. The invoice arrives as one opaque line item — impossible to challenge, impossible to forecast.

WHAT A CIO NEEDS

Every call attributed to an agent, a department, and a model — with the dollar cost recorded at the point of use, not reconstructed after the fact.

> The question is never "how much AI did we buy?" It is "which agent, which model, what did it cost — and why?"

WHAT IS DEPLOYED

A governed workforce mapped to the real org

367 governed agents run under zero-trust policy on a live GCP instance — not a sandbox. 360 fleet agents are mapped like-for-like to Acme's actual organization, plus a 7-agent showcase hand-off DAG.

367	360	28	325,154
GOVERNED AGENTS	FLEET, ORG-MAPPED	DEPARTMENTS COVERED	DIRECTORY PRINCIPALS

Source: registry (367 agent tenants, postgres) · agent-manifest.json (360 fleet) · Samba-AD directory — 120,120 humans + 205,034 NHI, 28 departments, 10 regions, 626 roles

THE ARCHITECTURE SLICE THAT MATTERS TO YOU

Cost is recorded at the point of enforcement

Every governed call passes the Policy Enforcement Point. On allow, the runner reports usage — tokens and dollars — which lands in a durable postgres table and surfaces on the ROI board. Nothing is estimated later.

`runner job` → `PEP :8343` → `soul-svc tenant` → `/cp/usage/report` → `tzt_usage` → `Grafana ROI`

Each row carries tenant, agent, model, tokens_in / tokens_out, and cost_usd, at hour and day granularity. That is the audit trail behind the invoice.

Source: tzt_usage schema · Postgres datasource acme-tzt-usage → Grafana :3000

LIVE PROOF · WORKLOAD

1–2 governed jobs per agent, per day

The fleet runs as independent rolling batches — not a synchronized burst. The cadence is deliberate and measurable, which is exactly what makes the spend forecastable.

6	300s	10h	~1–2
AGENTS PER BATCH	BATCH INTERVAL	MIN GAP PER AGENT	JOBS / DAY / AGENT

Each job is a full governed unit of work: domain brief → tiered LLM call → governed chain-of-thought + memory write → usage report.

Source: fleet_runner.py — FLEET_BATCH=6, INTERVAL=300s, MIN_GAP=10h (acme-fleet-runner, live)

WALKTHROUGH · ROUTING

Free and self-hosted first. Paid is the exception.

Routing policy favors free and Ollama tiers by default. Premium paid models are reserved for genuine escalation — an incident, a breach, a critical keyword — not for routine work.

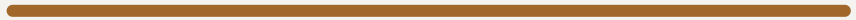
HOME	Default tier is free / self-hosted free_large, ollama_oss, ollama_deep — where the routine fleet work lands	favored
FLOOR	Free-homed agents degrade to ollama_oss a reliable self-hosted floor rather than a paid fallback	ollama_oss
CEILING	Paid tiers reserved for escalation keyword triggers +incident / +breach / +critical raise the tier	on escalation

Source: config.json (9 tiers, 3 providers) · fleet_runner.py routing · personas.json escalation keywords

LIVE PROOF · ECONOMICS

The same job, priced across the roster

This is the ROI board's per-call economics. The spread between a premium model and a self-hosted one is not a rounding difference — it is more than two orders of magnitude.

MODEL	TIER	OBSERVED COST / CALL	USD
opus-4.8	paid		$0.19
sonnet-4.5	paid		$0.03
deepseek-v4-pro	ollama cloud		$0.002
gpt-oss:120b	ollama cloud		$0.0015
Llama-3.3-70B	huggingface		$0.0012
gemma-4-31b:free	free		$0

Source: Grafana ROI dashboard (uid acme-roi-economics) · per-model $/1M-token economics from tzt_usage.cost_usd

WALKTHROUGH · THE ARITHMETIC

Why the default tier is the whole story

Routing the fleet to the escalation ceiling instead of its free home would multiply the per-call cost by more than 100x — for identical work. Governance is what keeps the default cheap.

ROUTINE CALL, GOVERNED HOME

$0 – $0.0015

free / self-hosted tiers — where 1–2 jobs per agent per day actually run.

SAME CALL AT THE CEILING

$0.19

opus-4.8 — justified only on a flagged incident, never as a default.

> Deny-by-default policy plus free-first routing means the escalation price is one you choose to pay, per call, on the record.

Source: config.json tiers · ROI board (uid acme-roi-economics)

LIVE FINDING · THE SELF-HOST CASE

Free is not free at scale

The board surfaced a real saturation signal: hermes-405b:free returns chronic 429 rate-limit errors under fleet load. The free tier cannot carry production volume — which is precisely what makes the self-host break-even calculable.

SIGNAL	hermes-405b:free — chronic 429 rate-limited under sustained fleet demand, observed on the ROI board	saturated
DEGRADE	Free-homed floor is ollama_oss self-hosted floor absorbs the load a free API cannot	reliable
BREAK-EVEN	Self-host vs. API becomes a number gpt-oss:120b at $0.0015/call sets the reference point	$0.0015

> The free tier's ceiling is the strongest argument for owning your inference — and the board proves it with data, not projection.

Source: ROI dashboard free-tier saturation panel (uid acme-roi-economics)

GOVERNANCE & AUDIT

Every dollar traces to a signed call

Cost legibility rests on enforcement. No governed call reaches a model without a validated capability token, a deny-by-default policy check, and a signed entry in the audit chain. The usage row and the audit row describe the same event.

IDENTITY	Per-agent HS256 capability token minted by the caller, validated at the PEP alongside its tenant API key	verified
POLICY	Deny-by-default, forwarded verbatim on allow site acme-site · 367-agent tenant slice, ETag-cached	enforced
AUDIT	Ed25519-signed audit chain → control plane durable WAL at /data/audit-wal; SoulWatch drift → auto-quarantine	immutable
METER	Usage reported to /cp/usage/report tenant, agent, model, tokens, cost_usd → tzt_usage	durable

Source: PEP enforcement plane (acme-tiresias-zt, :8343) · signed audit chain → CP · tzt_usage

THE ECONOMICS, SUMMARIZED

Legible by construction, not by cleanup

Attribution is not a monthly reconciliation exercise. It is recorded at the point of enforcement, for every one of the 367 agents, in the same durable store that powers the ROI board.

100x+	$0	$0.19	per call
CEILING VS. HOME SPREAD	ROUTINE FREE-TIER CALL	ESCALATION CEILING / CALL	COST ATTRIBUTION

Source: Grafana ROI dashboard (uid acme-roi-economics) · tzt_usage.cost_usd · config.json tiers

CLOSE

A cost you can defend.

367 agents, free-first routing, deny-by-default enforcement, and a per-call dollar figure recorded at the point of use. When finance asks what the AI workforce cost, the answer is a query — not an argument.

- Asphodel + Tiresias-ZT · ACME Corporation · live on GCP