

DEPARTMENT • RESEARCH & CONTENT

Governed research, synthesis, and drafting.

Three agents — research-lead, data-analyst, content-editor — doing real knowledge work on cheap model tiers, grounded in a curated knowledge base, under a propose-not-publish mandate that keeps a human on every output.

● Asphodel + Tiresias-ZT • ACME Corporation

THE STAKES

Knowledge work is where AI cost and AI risk both concentrate.

Research, analysis, and drafting run continuously and produce the language your organization acts on. Left ungoverned, that is exactly where spend leaks and where unverified text reaches a system of record.

THE COST PROBLEM

Every brief, every synthesis pass, every draft burns tokens. On a frontier tier a single governed call costs **\$0.19**. Multiply that across a department running all day and the model bill is the department budget.

THE TRUST PROBLEM

A draft with no provenance, no grounding, and no review is a liability the moment it is published. Content that shapes decisions needs a source, an author, and a reviewer — enforced, not requested.

Src • Grafana ROI dashboard (uid acme-roi-economics), observed per-call cost.

THE WORKFORCE

367 governed agents, mapped to the real Acme org.

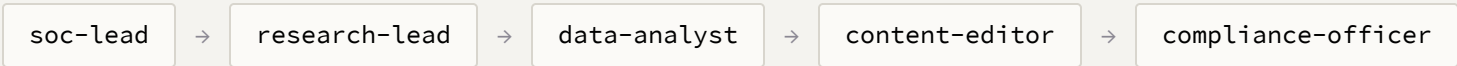
Research & Content is one department inside a live zero-trust workforce: 7 showcase agents on a topological hand-off DAG plus 360 fleet agents mapped like-for-like to the organization by department and role-family.

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

Src · CP registry (367 agent tenants, postgres) · agent-manifest.json · Samba-AD directory (120,120 humans + 205,034 NHI, 10 regions, 626 roles).

A three-agent slice of a seven-agent hand-off DAG.

The showcase workforce is a 14-edge topological hand-off DAG. Research & Content owns the middle: it receives from the security desk, refines, and passes a reviewable draft to compliance.

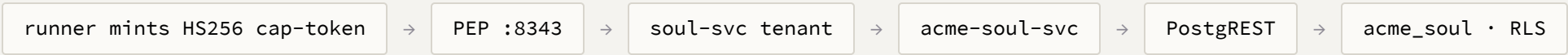


research-lead	Frames the question and gathers grounded source material Opens each cycle from a domain brief; searches the curated KB before generating.	node 4 / 7
data-analyst	Synthesizes and structures the findings Turns raw research into an analyzed, cited memory record.	node 5 / 7
content-editor	Drafts the deliverable as a proposal Writes for review — never straight to a system of record.	node 6 / 7

ONE JOB, END TO END

Every call is deny-by-default before it does anything.

A research or drafting job never touches data directly. It mints a capability token, passes the enforcement plane, and only then reaches its own isolated tenant — with the whole hop signed into an audit chain.



enforce	PEP validates api-key + capability token, evaluates policy deny-by-default On allow, forwards verbatim to that agent's backend only.	:8343
record	PEP signs the call into an Ed25519 audit chain → ships to CP Durable WAL at /data/audit-wal.	audit
meter	Usage reported to /cp/usage/report → tzt_usage → Grafana ROI board Per model: tokens in/out, cost, events.	tzt_usage

LIVE PROOF • OBSERVED

What this department actually spends.

Routing favors free and self-hosted tiers by design. Fleet agents run in rolling batches at ~1–2 governed jobs a day each, and the observed per-call cost reflects it.

\$0 GEMMA-4-31B : FREE / CALL	\$0.0015 GPT-OSS:120B / CALL	\$0.002 DEEPSEEK-V4-PRO / CALL	\$0.19 OPUS-4.8 / CALL
---	--	--	--

| The frontier tier costs \$0.19 a call. This department almost never touches it.

Src • Grafana ROI dashboard (uid acme-roi-economics) • fleet_runner.py (FLEET_BATCH=6, INTERVAL=300s, MIN_GAP=10h).

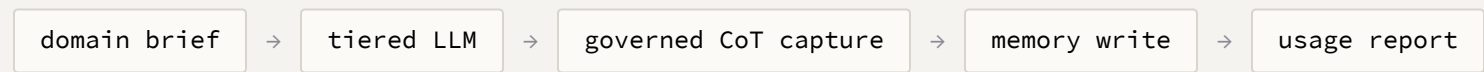
Research that starts governed, not audited after the fact.

research-lead opens each cycle from a domain brief, grounds itself in the curated KB, and runs on a cheap home tier — escalating only when the material demands it.

home	Runs on a free / self-hosted large tier by default Free-homed routing: free_large · ollama_oss · ollama_deep.	cheap tier
ground	Searches its curated KB before generating POST /v1/memory/search over its own RLS-isolated tenant.	soul-svc
escalate	Keyword triggers raise the tier only when warranted +incident / +breach / +critical → escalate toward the ceiling.	home→ceiling
degrade	Fallback-to-floor on model error Free-homed floor = ollama_oss for a reliable degrade path.	→ floor

Synthesis that leaves a durable, isolated record.

data-analyst takes the research hand-off and turns it into structured findings.
Every job is the same governed shape — brief in, reasoning and memory out,
usage metered.



REASONING IS CAPTURED

The chain of thought is written through the governed path via POST /v1/cot/capture — the analysis is inspectable, not a black box.

FINDINGS ARE ISOLATED

The result lands in acme_soul with RLS forced. Each agent is its own soul-svc tenant; cross-tenant reads require an explicit SET ROLE service_role.

Src • fleet_runner.py job shape • soul-svc endpoints (/v1/cot/capture, /v1/memory/write) • acme_soul RLS forced.

Propose, not publish.

content-editor writes the deliverable — but as a proposal, into governed memory. Nothing reaches a system of record until a human dispositions it, and the writes themselves are inspected at ingest.

<code>draft</code>	Writes the deliverable to its tenant memory as a proposal Handed downstream to compliance-officer, node 7 of 7.	<code>/v1/memory/write</code>
<code>inspect</code>	ContentInspection (enforce) rejects unprovenanced writes Requires <code>source_type</code> + <code>authored_by</code> on every write.	<code>enforce</code>
<code>review</code>	Memory and CoT are reviewed in the management console asphodel-ui — a human approves before anything is published.	<code>:3001</code>

Every draft carries a source, an author, and a reviewer — enforced at the write, not requested in a guideline.

GROUNDING & ISOLATION

Grounded in a curated KB it is not allowed to leave.

Every agent in the department reads and writes only its own slice of the data plane. The curated knowledge base is the ground truth for research — and the row-level boundary is enforced by the database, not the application.

CURATED-KB GROUNDING

Agents search their memory before drafting — GET /v1/memory and POST /v1/memory/search, embeddings via OpenAI — so synthesis stands on retrieved sources rather than model recall.

FORCED ROW-LEVEL SECURITY

_memories and _soul_api_keys carry FORCE RLS in acme_soul. Even the postgres role is subject to it; crossing a tenant boundary is impossible without SET ROLE service_role.

Src • Plane 4 (data) — acme-soul-svc → Caddy /rest/v1 stripper → PostgREST → Cloud SQL acme_soul (RLS forced).

Deny-by-default, signed, and watched for drift.

The department runs under the same enforcement as the rest of the workforce: nothing is permitted implicitly, every call is signed into a tamper-evident chain, and behavioral drift is caught automatically.

policy	Deny-by-default at the PEP; on allow, forwarded verbatim Each agent reaches only its own soul-svc backend.	:8343
audit	SoulKey Ed25519 signed audit chain shipped to the control plane Durable WAL at /data/audit-wal · SIEM via RFC5424 syslog → Promtail:1514 → Loki.	signed
watch	SoulWatch detects behavioral drift → auto-quarantine Shared postgres quarantine store.	auto
review	Analyst console for audit/verify, soulwatch, trajectory + drift analytics Role-token gated (pdp.tiresias.watch).	/admin/*

The self-host-vs-API answer, per call.

The ROI board measures observed cost per model. For a department that runs research and drafting all day, tier choice is the entire economics — and the governed floor is effectively free.

MODEL TIER	PROVIDER	OBSERVED COST / CALL	USD
opus-4.8	OpenRouter paid		\$0.19
sonnet-4.5	OpenRouter 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	OpenRouter free		\$0

Free tiers do saturate — hermes-405b:free returns chronic 429s — which is precisely the self-host case that deepseek-v4-pro and gpt-oss:120b answer at fractions of a cent.

RESEARCH & CONTENT • IN ONE LINE

Cheap tiers, curated grounding, a human on every draft.

research-lead, data-analyst, and content-editor do governed knowledge work
at fractions of a cent per call — every job deny-by-default, every write
provenanced, every deliverable proposed for review before it is ever published.

- Asphodel + Tiresias-ZT • ACME Corporation • 367 agents
live on GCP