GOVERIS
AI spend governance / CALYBRIS Engine

AI spend governance audit

Every AI call must justify its cost.

Audit, simulate, enforce. Control spend before tokens become budget burn.

META ONLY | PROMPTS OFF | SYNTHETIC SAMPLE

Verdict
ACTIONABLE COST LEAKAGE

Client	Synthetic Enterprise
Run	d1fc63947d4007a9
Period	500k metadata replay

Executive snapshot

Ready for shadow-mode policy replay with outcome checks. Sample uses 500,000 audited calls; paid pilot requires 500+ real calls.

Evidence maturity
89/100 Shadow-ready

pilot gate target: 500+ replayed calls

Readiness inputs

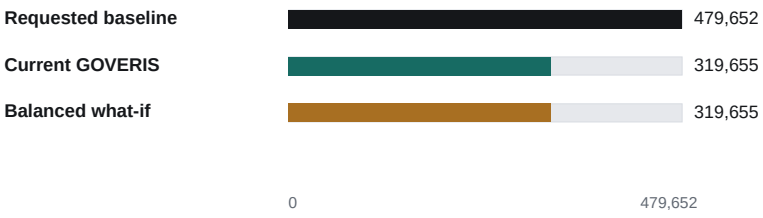
Calls	500,000
Coverage	100.00%
Verdict	Cost leakage found

TOTAL CALLS	CURRENT SPEND	AVOIDABLE SPEND
500,000	$3,196.55	$1,599.97
Audited decision rows	Policy-selected model cost	Requested minus selected

SAVINGS RATE	AT 100K CALLS/MO	AT 1M CALLS/MO
33.36%	$319.99	$3,199.94
Catalog estimate, not a claim	Linear demo projection	annualized $38,399.26

Spend shape

Requested baseline versus GOVERIS policy and a balanced replay.



Policy decision mix

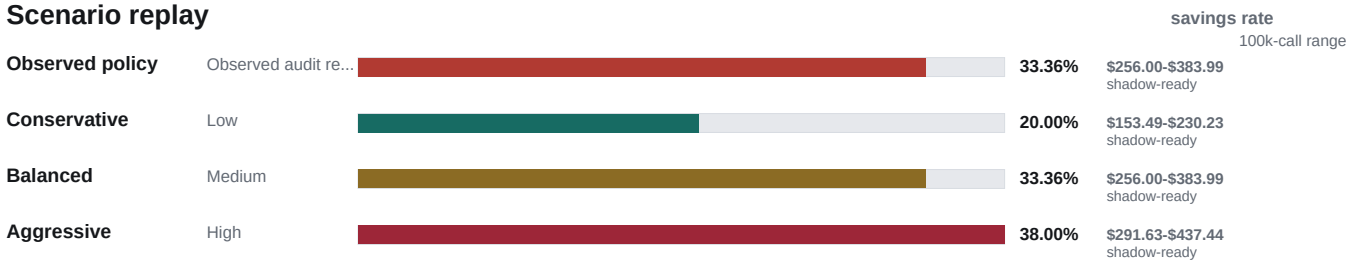
500,000 audited calls

Allow	380,241
Downgrade	116,290
Block	3,469

What-if policy simulation

Projection layer. Treat as pilot planning until replayed against real logs with outcome checks.

Scenario replay



Recommended next actions

01	Review blocked call 341580becd (enterprise-acme/support-agent): net value 11.7403 cents, policy risk or budget guardrail triggered a block decision; confirm product-owner rules before enforcement.
02	Review blocked call ca131d0dee (northstar-ai/evaluation-pipeline): net value 10.9477 cents, policy risk or budget guardrail triggered a block decision; confirm product-owner rules before enforcement.
03	Review blocked call 47e3ee0a7e (enterprise-acme/sales-assistant): net value 20.5885 cents, policy risk or budget guardrail triggered a block decision; confirm product-owner rules before enforcement.
04	Review downgraded calls first: GOVERIS found 116290 requests where a cheaper model satisfied the quality floor.
05	Review blocked calls with product owners: 3469 requests failed budget, risk, or tenant constraints.

Attribution and audit trail

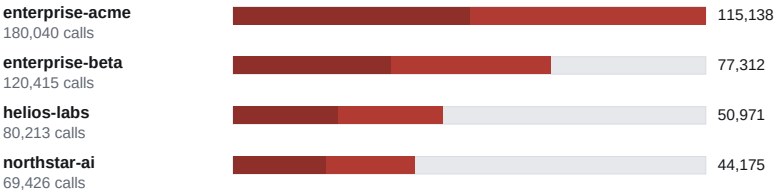
Where spend was created, where GOVERIS found avoidable cost, and what evidence was locked.

Priority calls to inspect

Call	Slice	Decision	Route	Why it matters
f56daad6fb	enterprise-acme support-agent	BLOCK	gpt-4o -> blocked	Policy risk or budget guardrail triggered a block decision conf 0.62; net 11.7403 cents; context hash phash_support_a...
cbea7af2c3	northstar-ai evaluation-pipeline	BLOCK	gpt-4o -> blocked	Policy risk or budget guardrail triggered a block decision conf 0.53; net 10.9477 cents; context hash phash_evaluatio...
853c4b5dcd	enterprise-acme sales-assistant	BLOCK	gpt-4o-mini -> blocked	Policy risk or budget guardrail triggered a block decision conf 0.64; net 20.5885 cents; context hash phash_sales_ass...
ddd1759fd5	enterprise-beta coding-agent	BLOCK	gpt-4o -> blocked	Policy risk or budget guardrail triggered a block decision conf 0.64; net $1.44; context hash phash_coding_ag...
0ed41cf4a9	enterprise-acme evaluation-pipeline	BLOCK	gpt-4o -> blocked	Policy risk or budget guardrail triggered a block decision conf 0.63; net 27.1362 cents; context hash phash_evaluatio...

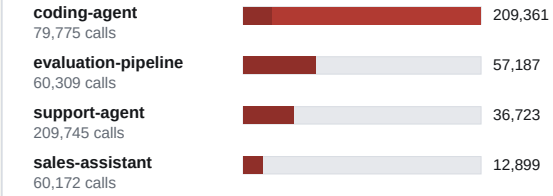
Tenant breakdown

Spend bars with avoidable-cost overlay.



Use-case breakdown

Spend bars with avoidable-cost overlay.



Governance risk strip

Counts that should drive review queues before enforcement.

PREMIUM PRESSURE	HIGH-RISK CALLS	LOW CONFIDENCE	CACHE ROUTES
227,028	49,990	20,954	44,029
requested expensive model	needs owner review	needs labels or review	separate execution path

Outcome quality gate

Cost control becomes production-ready only after quality labels or human review.

OUTCOME LABELS	HUMAN REVIEW	QUALITY DRIFT	LATENCY IMPACT
0/500,000	required	not measured	132 us policy
required before savings claims	review downgraded and blocked calls	capture task success or CSAT	track proxy and provider latency

Methodology notes

- CALYBRIS treats each LLM call as an investment decision: expected value minus model cost minus risk pressure.
- Decision score is reported as net value = expected_value_cents - estimated_cost_cents - risk_penalty_cents.
- Savings compare requested-model spend against the policy-selected model using the CALYBRIS model catalog.
- What-if rows are scenario simulations, not realized production savings.
- Projection ranges widen when the replay sample is small or audit-trail coverage is incomplete.
- Quality preservation still requires outcome labels or human review samples.

Interpretation guardrail

GOVERIS estimates cost leakage from the model catalog and decision log. What-if savings are planning projections until replayed on real logs with outcome checks.

Evidence lock

Reproducible artifact trail for exactly this audit run.

Run ID d1fc63947d4007a9

Generated 2026-06-22 17:33 UTC

Input SHA-256 1ca88f2157f7af9be609a2bc8d9868bc0c05bdc09c160afc59911423d5041355

Summary SHA-256 7625cb5c7b853967cd3ece50fba6e24266a0a8bb283e6e992839abe823e926e7

What this proves

- same input file can be replayed
- summary artifact is locked
- not a production savings claim

Evidence lock answers: which rows were audited, when, and with which summary artifact.

7-day metadata-only shadow pilot

Observe first. Keep production routing unchanged until the evidence is reviewed.

01 Deploy
Run the private GOVERIS Docker image inside the customer environment.

02 Mirror
Send metadata envelopes only; prompts, responses, and provider keys stay out.

03 Replay
Compare conservative, balanced, and aggressive what-if policies in shadow.

04 Decide
Deliver the report and promote nothing without explicit change approval.

Paid pilot intake checklist

Customer deploys	A digest-pinned private Docker image inside its VPC or controlled environment. No log export is required.	Three separate plane keys, a writable local evidence volume, and an approved retention window.
Metadata mirror	Model, input/output token counts, tenant, workflow, latency, risk, confidence, and approved outcome labels.	Prompt and response content are rejected from the shadow route and never written to the decision ledger.
GOVERIS returns	Executive PDF, local dashboard, aggregate findings, review queue, tenant/use-case attribution, and policy scenarios.	Raw events remain in the customer environment; savings remain estimates until outcomes validate them.
Pilot boundary	Seven days in observe-only mode. Production model calls continue on their existing path without GOVERIS enforcement.	Enforcement, provider proxying, and custom adapters require a separate approved production scope.

How the shadow replay starts

One private image, one local volume, one metadata mirror. No PDF, CSV, or JSON export request.

01 Pin image
Customer approves the private image digest.

02 Start shadow
Metadata envelopes mirror asynchronously.

03 Review locally
Inspect costs, risks, and proposed routes.

04 Deliver audit
Executive PDF, evidence hashes, next actions.

Pilot acceptance gates

Enforcement remains locked until evidence, quality, and rollback controls are reviewable.

01 EVIDENCE
>=99% mirror coverage
Counts reconcile with customer telemetry.

02 DATA QUALITY
No prompt capture
WAL scan and field allowlist verified.

03 QUALITY
Outcomes reviewed
Downgrades checked against task results.

04 CONTROL
Enforcement remains off
Promotion requires explicit approval.

GOVERIS

Govern. Verify. Operate.

Evidence-first AI spend governance, powered by the CALYBRIS decision engine.

01
AUDIT
Make model spend attributable.

02
SIMULATE
Replay policy before enforcement.

03
ENFORCE
Ship only reviewed controls.