

A I C O S T . A I
C L O U D I N T E L L I G E N C E . A I L L C
Irvine, California ~ Austin, Texas

Governance and Assessments for enterprise AI

The assessment. Continuous governance. Both from the same decision engines.

Twenty modules and 247 controls, connected to your real systems.
Findings priced in dollars, decisions compiled into policy your gateways load, and the result re-proved every night in a journal you can verify yourself.

W I T H A I C O S T . A I

Cost and governance are two sides of the same coin. The engines that price an AI workload are the engines that govern it.

Prepared 17 September 2026. Figures and capability facts carry the date they were captured. Scenario figures are illustrative of enterprise structure and do not represent any customer's actual spend.

1. AICost.ai in brief

What it is, before the governance part

AICost.ai is a cost and governance decision layer for enterprise AI workflows. It models the full economics of an AI workload before it exists, turns the approved decision into enforceable policy, and proves the outcome against it after go-live.

120+

deterministic decision
engines

MCP

chains: agents, IDEs, CI,
your apps

247

controls, priced and
owned

20

assessment modules,
200 questions

Nightly

continuous re-proof

MODEL

Workload economics

Token, reasoning, context, RAG and embeddings, vector DB, image, video, audio, training, fine-tuning, seats and subscriptions. Full AI TCO: build, run, evaluate, govern, maintain, observe.

DECIDE

Policy-ready outputs

Best architecture and model mix for the workload's quality, latency, compliance and scale. Budget, route, cap, gate and allowlist. Human-readable and machine-callable.

OPERATE AND MEASURE

From policy to proof

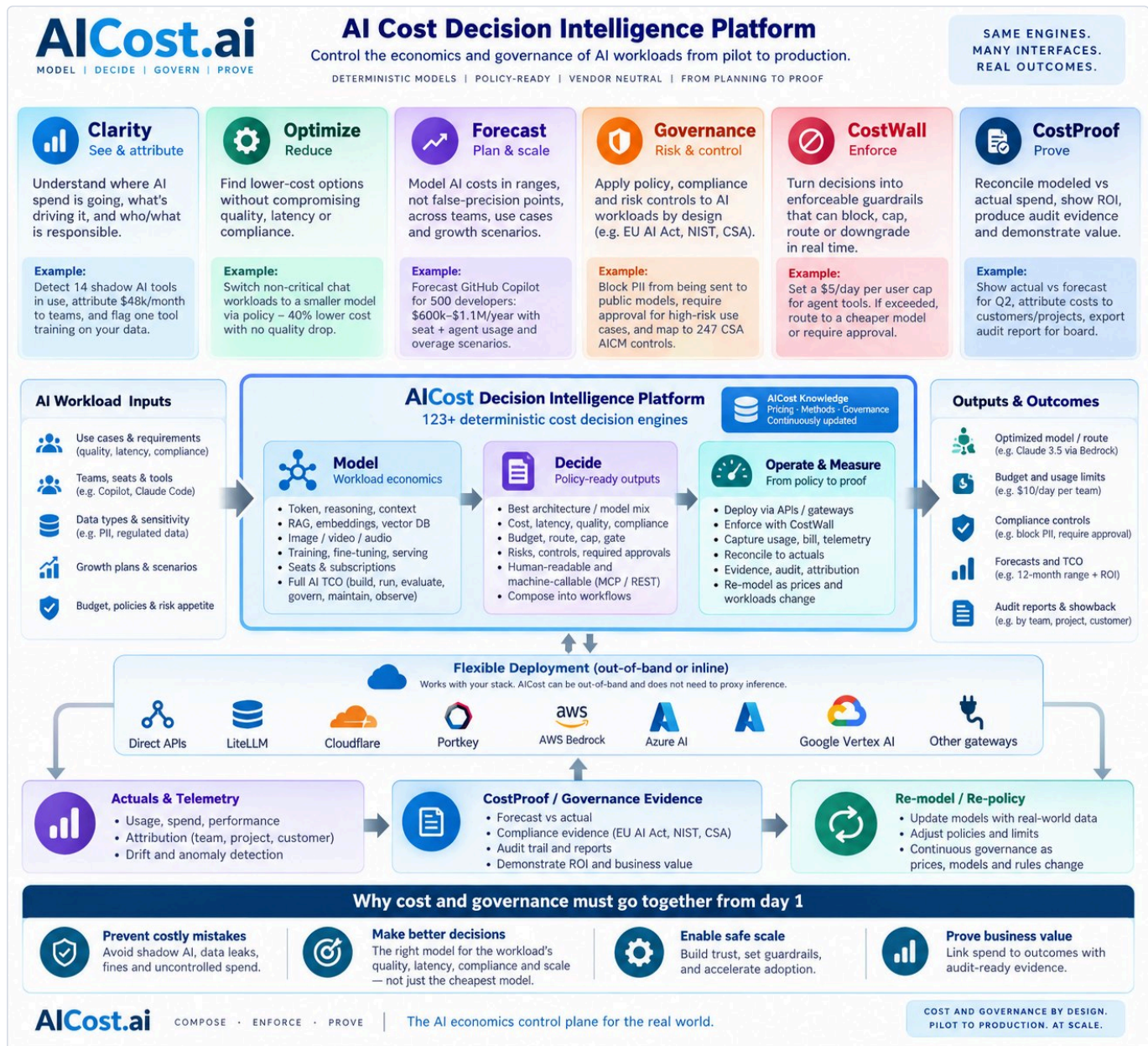
Deploy via your APIs and gateways, enforce with CostWall, capture usage and bill, reconcile to actuals, evidence and attribution, re-model as prices and workloads change.

The engines make no model calls

The decision-engine runtime is deterministic. The same inputs produce the same answer every time, with the source and date of every rate used, the assumptions made, and which engine to call next. That is what makes a figure reproducible for an auditor and safe to put in a policy. Language models are used elsewhere in the product for drafting and input suggestion, never inside the engines that produce the numbers.

Same engines, reached by a person or by a machine

A finance manager uses a calculator with three sliders. An agent, IDE assistant or CI job calls the same engine over MCP at mcp.aicost.ai and receives a typed answer with cost per unit, provenance, and the next engine to call. Because each engine names its successors, chains are assembled rather than programmed: a workflow nobody has costed before gets a chain in a working session, not a roadmap date.

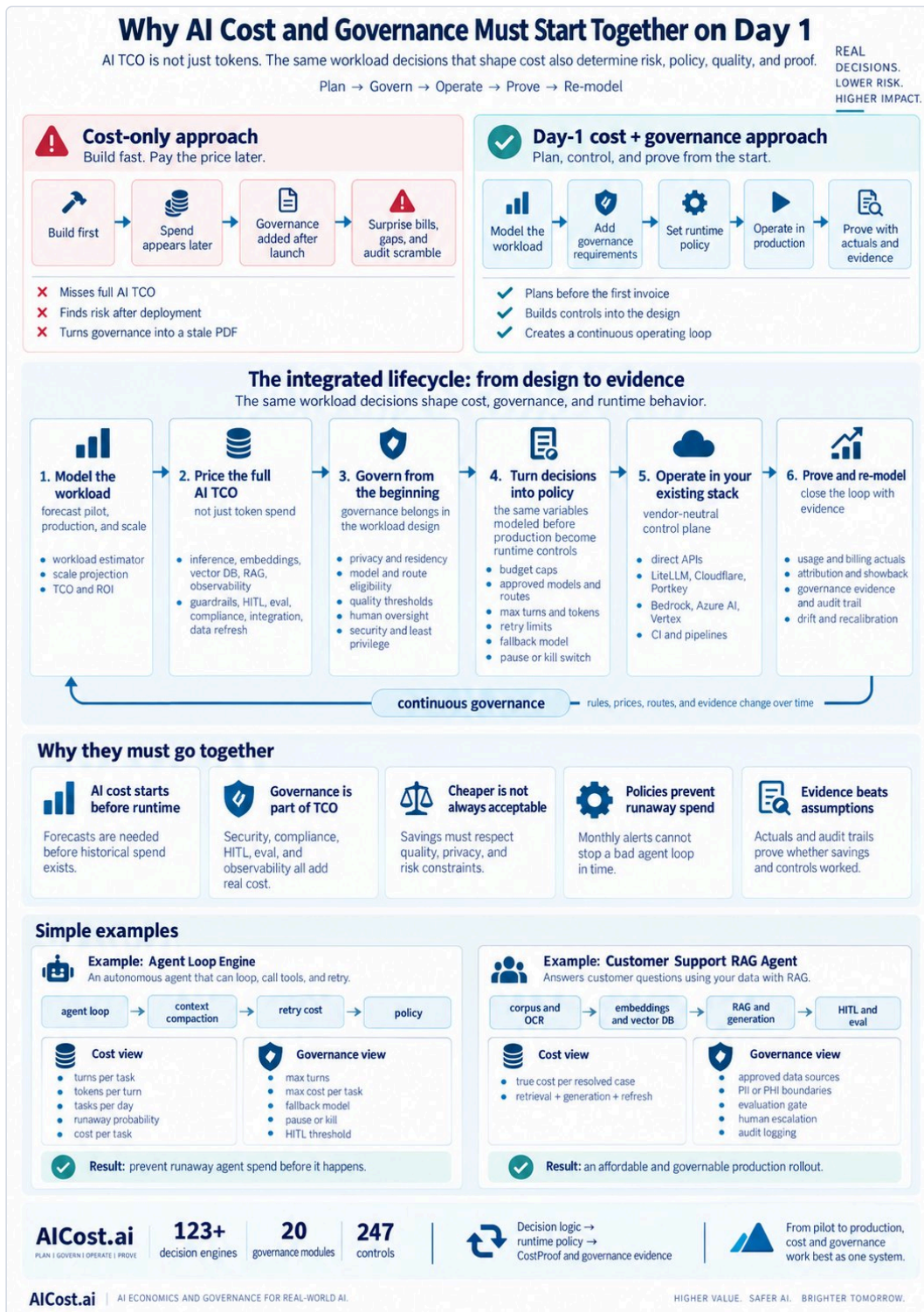


The platform: Clarity, Optimize, Forecast, Governance, CostWall and CostProof on one engine set, deployed out of band.

Why cost and governance are one decision

An agent loop has a cost view: turns per task, tokens per turn, tasks per day, runaway probability, cost per task. It has a governance view: maximum turns, maximum cost per task, fallback model, pause or kill, and the human review threshold. They are the same variables. Modelling them once and using them for both is why the policy is consistent with the economics rather than negotiated against them.

Cost-only approach	Cost and governance from day one
Build first, spend appears later, governance added after launch, then surprise bills and an audit scramble	Model the workload, add governance requirements, set runtime policy, operate in production, prove with actuals and evidence
Misses full AI TCO	Plans before the first invoice , governance premium priced as part of TCO
Finds risk after deployment	Builds controls into the design
Turns governance into a stale document	Creates a continuous operating loop: plan, control, prove, re-model

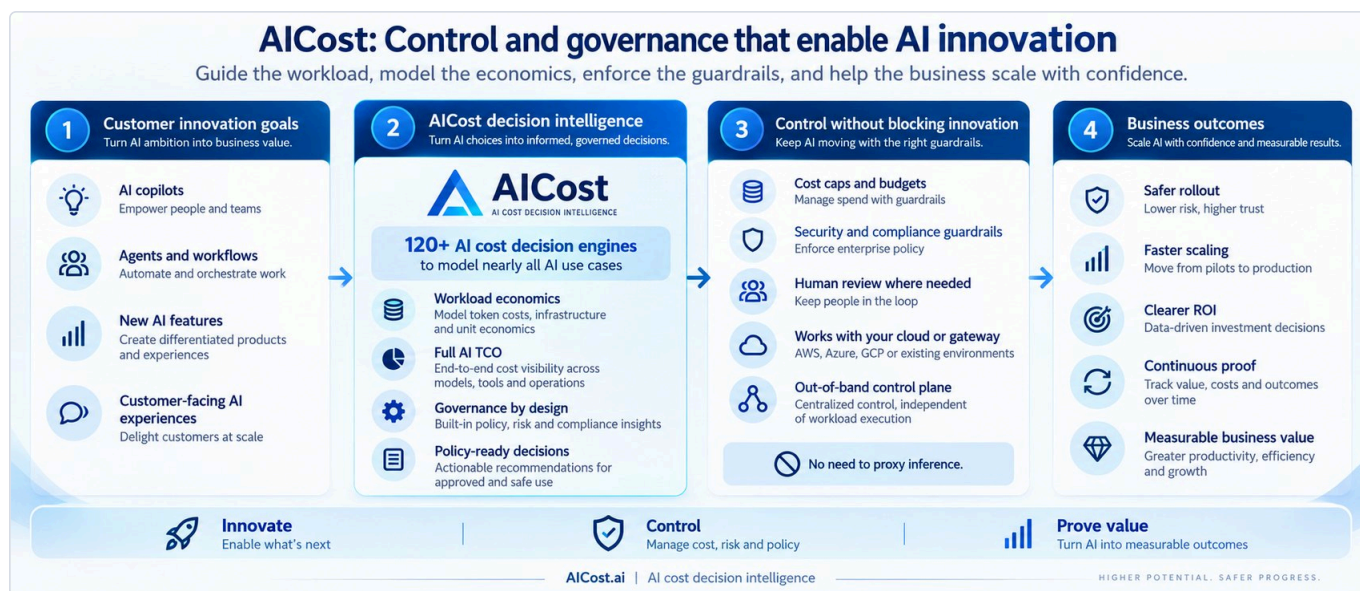


The integrated lifecycle from design to evidence, with the agent loop and support RAG examples.

2. Outcomes for enterprises

What changes when cost and governance run as one loop

Outcome	What it looks like	What produces it
You know what you have	Every AI system, owner and risk tier, including the shadow AI found on the bill and the corporate card, and the local models on laptops a proxy never sees	Inventory from metered spend, identity, endpoint and financial exports
Governance has a price	One dollar figure a CFO can budget: the Compliance Premium. Every finding priced, owned and given a remediation window	Compliance Cost, Governance Cost Pack, HITL and Guardrail engines
Decisions are enforced	Allowed models, ceilings and routes live in your gateway, not in a memo. A runaway loop at 2am is constrained, downgraded, then stopped, and it is in the ledger by morning	Policy engine, policy_patch, the graduated ladder, CostWall
Pilots reach production	Eligibility, residency and human oversight settled before build, not after the audit. About three weeks to first enforcement	Six eligibility gates, six templates, the elimination trace
Board, auditor and regulator answered	Inventory, policy estate, ceilings, exposure, cost of controls and the signed journal, all from one live source that can be compared October to January	Hash-chained journal, oversight log, recertification record, Annex IV pack
No rip and replace	FinOps, GRC, gateway and observability tools keep doing their jobs. We add the decision layer and export FOCUS, OSCAL and STAR back into them	Out-of-band control plane, FOCUS 1.x, NIST OSCAL, CSA STAR exports
Innovation keeps moving	Graduated enforcement instead of blunt caps, no-downgrade zones for regulated work, and time-boxed ceilings for experiments with automatic expiry	Warn, constrain, downgrade, gate, approve, break



Control and governance that enable innovation: from customer goals to measurable business outcomes.

What each role in the decision gets

Role	What they get
AI and architecture	A common method for deciding which workloads are worth building, which routes are eligible, and what architecture is economically sustainable at expected volumes
FinOps and finance	A bridge from provider spend to workload-level forecast, unit economics, the governance premium, showback and realised savings
Engineering and platform	Explainable thresholds promoted into gateway, CI, cloud or application policy without forcing a new inference path
Governance and security	Evidence that policy, route eligibility, human oversight and quality requirements were designed before production and remained in place afterwards
The business owner	A unit that means something in their terms: cost per resolved case, per accepted pull request, per contract reviewed. Not cost per token

P A R T A

The AI Cost and Governance Assessment

A complete, current picture of every AI system you run: what it costs, who owns it, which rules apply, where the gaps are, and what each fix costs. Available on its own, with or without continuous governance afterwards.

3. What the assessment is

Twenty modules, two hundred questions, one control register

The assessment answers one question with five parts: what may we run, what should we run, what will the governed workload cost at scale, what controls must exist, and can we prove the approved state is true. Today those parts are usually answered by different people at different times, which is where late rework and unpriced obligations come from.

The CSA AI Controls Matrix (247 controls) is the spine. An intake wizard scopes which controls apply at must, should and nice tiers for your profile, the questionnaire expands from there, and every answer carries a justification, an evidence link and an assurance level: self-attested, evidenced or verified. Answer once, and the obligations matrix reports across every framework.

Evidence first, questionnaire last. With read-only access to billing, identity, licensing and gateways, most of the 200 questions are answered from portal evidence and metered spend. Your team fills in only the short list that cannot be observed. Where a control cannot be measured, the report says so: "what we do not know" is a section, not a footnote, and an unmeasured control is never counted as a passing one.

The twenty modules

#	Module	What it answers for you	What it gives you
Discover: find, understand, get ready			
1	AI inventory and shadow AI	What AI are we actually running?	Discovery from network, financial, endpoint and identity data, not interviews. Four-state registry with a named approver
2	Data and knowledge readiness	Is our content ready for AI to read?	Source-of-truth registry per information domain; classification and licensing gaps for RAG and fine-tuning
3	AI identity and least privilege	Who and what can act as us?	Service principals, OAuth grants, app credentials, keys; who can call vs who did
4	Licensing, procurement and sprawl	What are we paying for and not using?	Seat reconciliation, unused licences, duplicate invoices, NCE and vendor subscriptions vs usage

Secure and comply: reduce risk, meet requirements

5	Security control plane	What happens if this is compromised?	Blast radius: what an agent could reach vs what it did; gateway, guardrails, DLP, prompt-injection and PII screening
6	Privacy, confidentiality and IP	Where can this data go?	Dataset by surface decisions against captured vendor facts: retention, training defaults, contractual opt-outs
7	Regulation, audit and compliance	What law applies to us?	EU AI Act classifier plus tracked US state statutes, dated obligations, control mapping across four frameworks
8	AI incident response	What do we do at 2am?	Thirteen incident classes, only one of which is security; statutory clocks attached

Prioritize and design: value and the right architecture

9	Value and workflow intelligence	What should we automate first?	Process health gate that refuses to price ROI on a broken process; the 19 percent of stages where AI genuinely helps
10	Workflow integration	How do we automate without a hundred one-offs?	Template plus adapter plus configuration plus policy; canonical capabilities, never vendor APIs
11	Model and vendor selection	Which model, by which route?	The unit is model by route; six gates applied in order, price last, with the elimination trace
12	AI cloud architecture	What should we actually build?	Bedrock vs Foundry vs Vertex vs direct, regions, provisioned capacity, self-host break-even, scored against your estate

Govern and control: turn decisions into controls

13	AI cost governance and FinOps	What does this cost, and are we making money?	Telemetry, attribution, budgets, ceilings, showback, forecast vs actual, cost levers with recovery rates
14	Agents, MCP and autonomous actions	What can our agents do on their own?	MCP server inventory, tools, auth, transport; turn limits, envelopes, runaway protection
15	Quality, evaluation and assurance	Is the output good enough?	The deterministic boundary: which exact steps AI should perform; golden sets, eval gates, drift checks before any swap
16	Vertical governance	What does our industry require?	Healthcare, legal, defense, finance, education and employment lenses; no-downgrade zones

Operate and adopt: enable people, prove value

17	Copilot readiness and value	Is it safe to switch Copilot on, and is it paying back?	Oversharing, labels, root cause; seat ROI, adoption depth, subscription vs metered per developer
18	Adoption, training and change	Will people actually use it?	Adoption against licence, training gaps, AI literacy pack, usage policy generated from what you actually have
19	Shared responsibility and operating model	Who owns what?	Append-only register of what was found, recommended and decided, by a named person, in their own words, with a review date
20	Evidence, reporting and continuous improvement	Can we prove the approved state still holds?	Ledger, oversight log, recertification, board and audit packs; the drift report and the verify command

Every finding carries a severity, a required skill level, a price, an owner and a plain-English recommended action. Slice by module, by severity, by regulation, by vendor or by workload.

4. Standards, mapped once

One control register, four frameworks, plus the 2026 US state layer

Framework	Type	Role inside the assessment	What we produce
CSA AI Controls Matrix v1.1	Technical control catalog	The spine. 247 controls across 18 domains, expanded into the questionnaire, each with a justification and assurance level. The most technical of the four, so platform and security teams can implement it	Control-by-control status, priced and owned; STAR for AI Level 1 working sheet
ISO/IEC 42001	Certifiable management-system standard	Proves governance runs as an operating system: policy, roles, impact assessment, lifecycle, supplier relationships. Procurement increasingly asks for it	Management-system gaps and audit evidence
NIST AI RMF 1.0	Voluntary US framework	Govern, Map, Measure, Manage, with the Generative AI profile. Our eval gates are Measure; the enforcement ladder is Manage	Function scorecard; findings exported as NIST OSCAL into your GRC tool
EU AI Act	Binding law	Risk classification per workload. Article 14 oversight log, Article 15 evaluation heartbeat, Article 50 transparency, Article 73 serious-incident clock, Annex IV technical documentation regenerated from the journal. Dates tracked as the Digital Omnibus moves them	Classification, obligations with dates, Annex IV pack, incident lifecycle
2026 US state layer	State statutes	California SB 53, AB 2013, SB 942 and AB 489 (healthcare), Texas RAIGA, Colorado ADMT, plus employment bias-audit rules (NYC LL144, Illinois HB 3773). Every obligation dated and watched at 90, 30 and 7 days	Dated obligation register and countdown alerts

Why one register. The same evidence answers the auditor, the regulator, the board and the security team. Where they differ in what they ask, the mapping does the translation; the underlying control, its owner and its price stay the same.

5. Connected to your real systems

Read-only. Never in your inference path. No prompts or content.

Microsoft 365 and Entra

- Entra ID roles, PIM, Conditional Access, sign-in logs
- Defender for Cloud Apps discovered AI apps; Intune detected apps (local AI a proxy cannot see)
- Licensing and NCE SKUs, Copilot seats and usage reports
- SharePoint data access governance, Purview DSPM for AI
- Thirteen Graph endpoints, one app registration, all read-only, no content scopes

AWS Bedrock

- CUR 2.0 daily export with resource IDs and caller identity
- Model invocation logging, one inference profile per agent
- Cost allocation tags at the payer account, twelve-month backfill
- Cross-account read role with a per-customer external ID, two S3 prefixes, under \$2 a month

Azure AI Foundry and GCP Vertex

- Cost Management daily export; per-deployment RBAC
- Azure Monitor token metrics, PTU showback by utilisation share
- BigQuery billing export with labels honoured
- Org-policy publisher allow-lists

Vendors, gateways and finance

- Anthropic, OpenAI, Google and other usage APIs
- LiteLLM, Portkey, Cloudflare, Kong, OpenRouter, Helicone
- Corporate card and AP exports: AI on the company card nobody expensed as AI
- Vendor trust and DPA pages, hashed and dated, with contradictory and not-stated evidence states preserved

Required	Why	Not required
Read-only cost and usage	The retrospective half: what was actually spent	Prompts, completions or any customer content
A team-scoped key convention	The team dimension must exist when spend is recorded; it cannot be added afterwards	Any change to how your teams build or deploy
A workload description	Model, token shape, volume, cache and batch share, review rate	Production access of any kind
An owner per workload	Governance needs an accountable person, not a system	A completed inventory. The inventory is an output, not an input

Coverage honesty rule: every report prints its coverage ("62 of 100 devices are enrolled; unenrolled devices are invisible to this method"). Below 80 percent, coverage itself becomes a finding.

6. How the assessment runs

Three weeks, mostly asynchronous, two live sessions

Phase	Stage	What we do	What you do	Elapsed
0	Exchange	Issue organisation and tenant identifiers, MCP keys, and the connector or export mapping for your lanes	Tell us which lanes you run, give read-only access	Day 1
1	Baseline	Run the visibility chain across the workloads in scope. Produce the first attributed number	Name an owner per workload	Week 1
2	Attribution	Map teams to workloads, keys to agents, tags to business context	Adopt the key convention	Week 1 to 2
3	Assessment	Twenty modules scored from evidence; scoped questionnaire for the remainder; findings priced and owned	Answer the short list, review the findings	Week 2
4	Policy	Run the policy engine per team. Produce the elimination traces and the first policy patch	Review the traces and challenge them	One session
5	Roadmap	Stop / fix / approve / scale roadmap with a price on every row; Compliance Premium; executive deck	Take the decisions the closing slide asks for	Week 3
6	Optional: enforce and govern	Deploy the wall, set caps, start the watchers and the journal (Part B)	Agree ceilings and the downgrade ladder	Week 3 onward

How success is measured

Five criteria, all objective, all verifiable by your own team without taking our word for any of them.

Criterion	The evidence	How you verify it independently
A per-team cost number that did not exist before	Ledger rows with team, workload and agent attached	Compare the roll-up against your vendor invoices
At least one structural saving identified	A cache, batch, routing or placement lever with the arithmetic shown	Re-run the lever against the vendor rate card
A priced, owned findings register	Every finding with severity, skill level, price and window	Pick any row and re-run the engine behind the price
A policy patch your gateway can load	The diff your security team reviewed	Your own change control
Forecast accuracy scored	Forecast against actual, reported as a percentage	The status endpoint reports it on every run

7. What you hold at the end

Eleven deliverables plus a ten-slide executive deck whose closing slide asks for decisions, not applause

Reports and registers

- 01 Executive report and risk heatmap
- 02 AI inventory and use-case catalog with owners and tiers
- 03 Obligations matrix across all frameworks
- 04 Findings register, priced and owned, with remediation windows
- 05 Compliance Premium: your governance cost in dollars
- 06 Model eligibility matrix per route with elimination trace

Policy, evidence and the plan

- 07 Policy patch for gateway or IAM, ready to load
- 08 Vendor evidence pack, hashed and dated
- 09 Target operating model and shared responsibility register
- 10 Hash-verified evidence register and journal
- 11 Stop / fix / approve / scale roadmap with a price on every row
- Plus the executive deck, and MCP access to every engine after we leave

Machine-readable exit	Goes into
NIST OSCAL findings	Your GRC platform
CSA STAR Level 1 working sheet	Your assurance program
FOCUS 1.x cost data	Your FinOps tool
model.json per workload	Re-run anything, prove everything
Annex IV technical documentation pack	EU AI Act conformity file

8. Two ways to buy: assessment only, or assessment plus continuous governance

Every step credits forward, so diagnosis is never sunk cost

01 AI Act Quick-Check Free, 2 minutes Your EU AI Act risk class and likely obligations. No signup wall.	02 Take-Stock Session One hour, fixed fee Your issues taken stock of, two workflows named and priced, a same-day written plan.	03 AI Opportunity Snapshot One week Where money and hours leak, fastest ROI, daily cost visibility, governance cleared before you build.	04 Cost, Workflow and Governance Assessment Three weeks, fixed fee Everything in Part A. Complete on its own.	05 Continuous Governance Flat monthly per tenant Everything in Part B. The assessment that re-proves itself every night.
Assessment only (steps 01 to 04)		With continuous governance (step 05)		
A complete, priced, evidence-backed picture as of delivery, with a policy patch you can load and a roadmap you can execute		The same picture, kept true: controls re-evaluated nightly against your own billing and usage, watchers, incidents, recertification		
You operate the roadmap; re-assess quarterly or annually as a project		Re-assessment becomes a re-run; the delta is the report. Monthly signed drift report with the verify command		
Right when you need a decision basis, a board answer, or an audit file now		Right when AI is in production and prices, models, vendors and laws change faster than an annual cycle		

PART B

Continuous governance

Governance that becomes operational. The cost engines compile into policy, the policy is enforced in your stack, and the evidence accumulates as a by-product of operating rather than as a project before an audit.

9. The policy engine

Governance you can load, not just read

Everything in Part A produces a decision. The policy engine turns those decisions into a rule that holds when nobody is watching. For each workload it answers which models this team may use, at what ceiling, on which route, and it answers as configuration rather than as advice.

Eleven questions, six gates, six templates

The intake tree is ordered so a low-sensitivity workload reaches an answer in three questions and a regulated one earns the long path: what data will this workload touch, does the organisation hold an attestation an undocumented flow would break, what plan or surface is in use, which cloud agreements are held, is shared inference forbidden, where must processing stay, must the vendor be barred from training, what kind of workload, does it answer while someone waits, what accelerator memory is available, what routes traffic today.

1 Residency
where processing
may stay

2 Certification
SOC 2, ISO, HIPAA
BAA, AI Act

3 Training data
vendor defaults,
contractually off

4 Tier
consumer, team,
enterprise plan

5 Capability
fits the workload
shape

6 Price
only among what
survived

Price is last on purpose. A model that fails residency is not expensive, it is ineligible, so pricing it would mislead. Eligibility is scored per route: the same model on Bedrock in Frankfurt and on a consumer plan are different answers.

```
route: legal / contract-review data_class: regulated residency: EU
gpt-5-mini FAIL gate 1 EU residency unavailable on this route
llama-3-hosted FAIL gate 3 training default on, no contractual opt-out
claude-sonnet-5 PASS bedrock-eu-central-1, BAA in place, training off
→ policy_patch: model_list, budget cap, no-downgrade zone, approval above cap
```

Template	Selected when	Posture
Open, cost optimised	Data class is public	Any vendor, any route. Optimise purely for cost and latency
Internal business	Data class is internal	Named vendors only, deny by default, training off

Regulated personal data	Data class is regulated	Agreement in place, residency constrained, training contractually off
Health or financial, agreement required	Health data plus a matching attestation	Narrow corridor, only covered surfaces reachable, kill switch among required controls
Export-controlled or defence	Export data class or attestation	Deny by default. Nothing leaves the permitted boundary
Self-host or open weight	Sufficient accelerator capacity	Open-weight branch, priced against the hosted alternative

What gets exported, and where it goes

The output is never a bare verdict: the template chosen, the models eligible per route, the compiled policy, the elimination trace with a plain-English reason for every model ruled out, and the capture date of every vendor fact relied on. The compiled part is exported to whatever your teams actually use: `policy_patch` for LiteLLM, Portkey, Cloudflare AI Gateway, Kong and OpenRouter; Bedrock IAM per-model conditions; Azure RBAC; Vertex org policy; and vendor-console allowlists, per-key budgets and ceilings for teams on direct keys with no gateway at all.

10. From policy to enforcement: the graduated ladder

The first four steps remove cost, the fifth asks a person, the sixth stops

WARN recorded and notified Cost velocity or context growth crosses a threshold. Nothing restricted.	CONSTRAIN cache, reuse, compact Above plan but legitimate. Cost removed before capability.	DOWNGRADE routed to economy Routine substeps move to a cheaper route that passed eligibility.	GATE quality gate The cheaper route must still pass the accepted evaluation.	APPROVE requires a person Spend beyond the envelope is a recorded human decision.	BREAK pause or kill Loop outside the hard envelope. Smallest useful scope, journaled.
---	--	---	--	---	---

No-downgrade zones. A contract review, a production incident response or a regulated decision can be marked so the ladder skips the economy route entirely and goes from constrain to approve. A blunt cap treats a legitimate spike and a pathological loop identically; a graduated ladder with protected zones does not.

The downgrade step. When spend crosses the soft cap the wall does not block; it changes the model and work carries on at a fraction of the cost. On the proving run the hard cap took considerably longer to arrive because the cheaper model stretched the same budget further, and the ledger recorded both models, so the downgrade is visible in the data rather than asserted.

The ledger records	Why it matters
Model, team, workload, key, agent and dollars, per call	Showback and chargeback become a by-product of the control, not something reconstructed from an invoice
Enforcement events, refused and downgraded calls	What the ceiling prevented becomes a reportable line, not an anecdote
Hash-chained, signed entries	Tamper-evident. An auditor can test it rather than take it on trust
Forecast against actual, scored as a percentage	The forward model is graded, including when the grade is unflattering
FOCUS 1.x export	Lands in the FinOps tooling already in place, in the standard format

11. The cadence: the assessment that re-proves itself

Controls evaluated against live data on a schedule; evidence accumulates as a by-product

Cadence	What runs	Evidence produced
Daily	Cost actuals ingest, bill vs gates, customer-eye engine audit	Per tenant, workload and agent showback; untagged spend as a named residual
Weekly	Seven watchers: Compliance Premium drift, model eligibility, vendor evidence freshness, regulatory countdown at 90/30/7 days, inventory hygiene, findings burn-down, actual bill vs gates	Journalized sweep, drift rows, alerts to Slack or webhook
Monthly	Branded drift report	Chain-verified, HMAC-signed, with the verify command you run yourself
Quarterly	Access recertification	Who can call vs who did; unused grants and unknown actors; attestation table
Continuous	EU AI Act Article 73 incident clock, Article 14 oversight log, Article 15 evaluation heartbeat, live inventory	Annex IV pack regenerated on demand from the journal

Drift caught in days, not at renewal

A model repriced, a vendor certificate aging out, an owner leaving, an agent tripling its retries: named and dated. Serious incidents open with the 15, 10 or 2-day deadline attached; overdue renders red before a regulator asks.

Cost and governance in one loop

Watcher seven reconciles the actual bill against the enforced policy. FinOps tools have half of this; GRC tools have neither. Audits, insurers and procurement get current evidence instead of last year's document.

12. Enterprise workflows under continuous governance

Engine chains over MCP, the cost view and the governance view side by side

Each workflow below is a chain composed from the engines, gates and enforcement points already described. The engines come first and the limit comes last: a ceiling traced to a specific cost driver, with the quality condition that permits the optimisation stated alongside it, is a policy an engineering organisation will approve and deploy.

A. Distributed dev teams with agentic AI

per-key attribution → agent-loop-cost → context-compaction → prompt-cache-roi → multi-model-router → eval-gate → agentic-variance-reserve → **cost-firewall**

Cost view

Turns per task, tokens per turn, tasks per day, runaway probability, cost per task. Base loops and the runaway tail priced separately, because agent workloads exceed budgets through variance rather than the average. Seat vs metered break-even per developer, not one policy for everyone.

Governance view

Maximum turns, maximum cost per task, fallback model, pause or kill, human review threshold. Platform engineering keeps the frontier model with a per-run envelope; feature teams get mid-tier with a per-key daily budget; data and ML get batch-only; client-facing teams are residency-gated before price.

Result: four teams, four different correct policies, one rule set, one gateway config diffed before it loads. A loop that misbehaves at 2am is constrained, downgraded, then stopped, with every step in the ledger by morning.

B. Customer support with RAG, fine-tuning and human review

corpus-onboarding-cost → data-freshness-cost → chunking-optimizer → prompt-cache-roi → fine-tuning-cost → hitl-review-cost → eval-gate + anti-hallucination → **support-deflection-roi**

Cost view

Eight priced layers: corpus onboarding, freshness, retrieval shape, inference at your cache hit rate, retries billed in full, guardrails, evaluation, human review. The unit is cost per resolved case, which support leaders already report.

Governance view

Approved data sources, PII and PHI boundaries, evaluation gate, human escalation threshold, audit logging. A cheaper model that escalates more often moves cost from the token bill to the reviewer; the eval gate refuses to call that a saving.

Result: an affordable and governable production rollout, with the 30 percent delivered by chunking and cache hit rate rather than by degrading answers.

C. Marketing and multimodal AI

image-generation-cost → video-generation-cost → card-and-AP-import → model-eligibility (open template) → **campaign showback**

Cost view

One cost per approved asset across vendors on one rate catalogue, with the shadow vendors on the corporate card pulled into the same total.

Governance view

Public data class, so the widest eligible vendor set optimised purely for cost and speed, brand and content policy attached, seat and overage alerts in the vendor console.

Result: the function with no governance burden still appears in the same enterprise total as the one carrying the heaviest.

D. Legal contract review, regulated

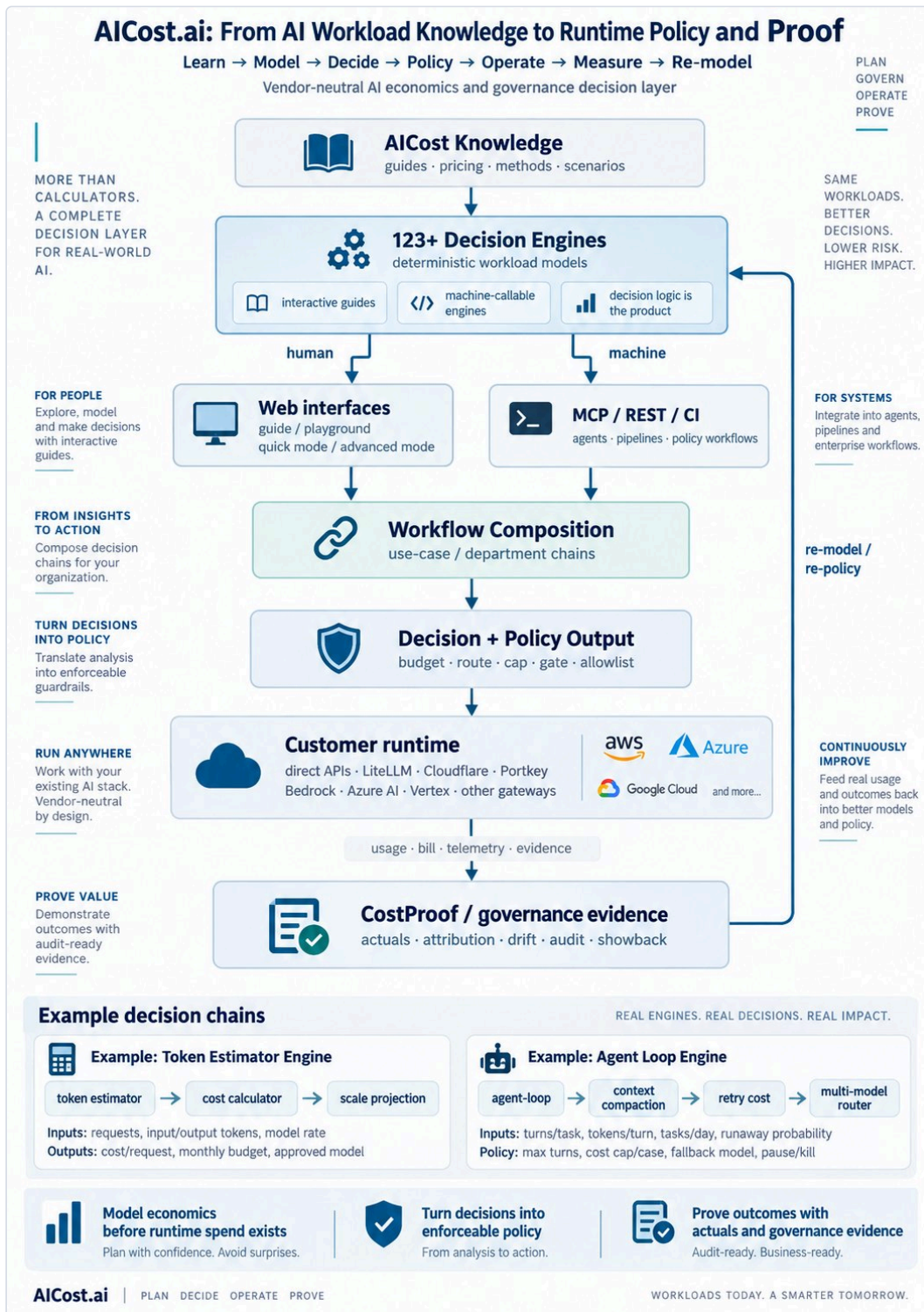
model-eligibility (6 gates) → jurisdiction-risk-scorer → compliance-cost → hitl-review-cost → **no-downgrade zone + approval above cap**

Cost view

Cost per contract reviewed, with the residency premium and the human review rate priced as controls rather than argued as principles.

Governance view

Certified surfaces only, agreement in place, retention contractually off, no-downgrade zone, approval above cap, oversight logged. Defensible to the same auditor that approves marketing's open template.



From workload knowledge to runtime policy and proof: human and machine reach the same engines; the loop closes on evidence.

One enterprise, one rule set, eleven policies

Org unit and workload	Data class	Eligible routes	Ceiling	Enforced at
Platform engineering, agentic refactor	Source code	Frontier eligible, geo routing	Per-run envelope, turn limit	Gateway
Feature teams, code assist	Source code	Mid-tier only	Per-key daily budget	Gateway
Data and ML, offline evaluation	Internal	Batch tier only	Monthly budget	Gateway
Solution engineers, client documents	Client confidential	Residency-gated, zero retention	Per-workload cap, no downgrade	Gateway plus vendor console
Customer support, RAG assistant	Internal business	Economy primary, mid-tier on escalation	Daily soft and hard cap	Gateway
Support quality, evaluation pipeline	Internal business	Batch tier, judge model fixed	Monthly budget	Direct key
Marketing, content generation	Public	Widest eligible set	Seat plan, overage alert	Vendor console
Legal, contract review	Regulated	Certified surfaces only	No-downgrade, approval above cap	Gateway plus approval
Finance, analysis assistant	Confidential	Mid-tier, retention off	Monthly budget	Gateway
HR, policy assistant	Personal data	Narrow allowlist, oversight logged	Per-workload cap	Gateway plus oversight log
Innovation, agent experiments	Internal	Any eligible route, time-boxed	Hard monthly ceiling, auto-expiry	Gateway

Nobody negotiated with anybody. Each policy follows from the data class and the attestation, which is why they can differ without being arbitrary. The whole table is one file that compiles to gateway configuration, versioned and diffed before it loads.

AICost.ai Decision Intelligence

Governed AI cost control across integrated enterprise workflows

123+ DETERMINISTIC DECISION ENGINES
Human-readable and machine-callable

1 MODEL THE DECISIONS THAT DEFINE A WORKLOAD

Representative engines

CLARITY

Where is AI money going and what is driving it?

- AI Cost Calculator
- Token Estimator
- Business Bill Diagnose
- Context Window Cost
- Reasoning Token Cost
- Retry Cost
- AI Seat ROI
- Pricing History
- Vendor Concentration

OPTIMIZE

What can we change while preserving required quality?

- Prompt Cache ROI
- Multi-Model Router
- Batch vs Realtime
- Context Compaction
- Token Reduction
- Stack Savings
- GPU Spot Savings
- Local vs Cloud
- Self-Host Break-even

PLAN & FORECAST

What happens at production and at 10x scale?

- Scale Projection
- Annual Forecast
- Quarterly Forecast
- Overage Forecast
- Budget Planner
- Agentic Variance Reserve
- Voice Agent Stack
- RAG Pipeline
- TCO / ROI Builder

GOVERN

What conditions must remain true for this workload to be acceptable?

- Model / Route Eligibility
- Jurisdiction Risk
- Compliance Cost
- Guardrail / Security Cost
- HITL Review Cost
- Eval / Hallucination Cost
- Data Freshness
- Model Maintenance
- Integration Cost

CONTROL & PROVE

How do we keep production inside the approved envelope and prove the result?

- Cost Firewall
- Policy Manager
- Gateway Policy
- Observability
- CostProof
- Forecast vs Actual
- Workload Showback
- Attribution / Evidence
- Drift / Recalibration

2 COMPOSE THE RIGHT ENGINES FOR EACH ENTERPRISE WORKLOAD

Illustrative workload chains

CODING AGENT

Seat economics → Agent loop → Cache → Eval → Cost policy

SUPPORT RAG

RAG pipeline → Agent loop → HITL → Eval → CostProof

VOICE AGENT

Speech → LLM / tools → Agent turns → HITL → Unit economics

MULTIMODAL MARKETING

Image / video → Variations → Human review → Forecast → ROI

LEGAL / REGULATED

Eligibility → Jurisdiction → HITL → Compliance → Evidence

ENTERPRISE COPILOT

Seat ROI → Usage → Routing → Governance → Showback

HUMANGuides · Playgrounds · Quick · Advanced

MACHINEMCP · REST · CQ/CD · Enterprise agents

ONE WORKLOAD. MULTIPLE DECISIONS. ONE GOVERNED ECONOMIC ENVELOPE.
Model before scale. Apply policy in production. Reconcile actuals, quality and business outcomes.

AICost.ai | Governed AI economics from pilot to production

Illustrative engine groupings and workflow chains

Representative engines by job, and the chains composed for six common workloads.

13. How this differs from the alternatives

Assessment, not paperwork

The usual governance assessment	AICost
Questionnaire and interviews; findings rest on what people remember	Connected to real systems. Entra ID, Defender, Intune, licensing, Bedrock CUR and logs, Azure and GCP exports, gateways, card and AP exports. Evidence first; a scoped questionnaire only for what cannot be read
Maturity heatmap in five colours	The Compliance Premium in dollars. What governing your AI costs, per control and per system, computed by engines you can re-run
A policy memo nobody can enforce	A policy engine. Six gates per route with the elimination reasoning shown, exported as gateway config or IAM conditions
Numbers estimated by an analyst	120+ deterministic engines in MCP chains. No model calls inside the math; same inputs, same answer, auditable
A PDF that is true on delivery day	Continuous. Watchers, heartbeats, incidents and recertification keep it true; every event HMAC-signed
Governance as a separate program from cost	One loop. The ceiling a team lead agreed to in a meeting is the ceiling enforced in production, produced by the same code

You could use	What it gives	What it cannot do	What AICost adds
A spreadsheet	Fine for one assessment	Does not persist between cycles; cannot prove what was decided a year ago	The second, tenth and fortieth cycle; the shared responsibility register
A GRC platform	Forms and attestation storage	Knows nothing about AI spend, models or agents	Cost telemetry as the sensor; violations named with dollars and rows; OSCAL export back
A FinOps tool	Cloud cost visibility	No controls, no obligations, no evidence chain, nothing before spend exists	Cost and governance reconciled in one loop; forward modelling; FOCUS export back
A consultancy	Expert judgment, a report	Decays from day one; evidence assembled by hand at audit time	The assessment, and then a machine that keeps checking it
An in-path AI gateway	Runtime enforcement	Sits in the inference path: latency, availability, key custody	Out-of-band decision and audit; enforcement stays in your infrastructure, exported as its config

12 Core AICost Philosophies

How AICost differs from traditional FinOps for AI workloads

Control the economics and governance of AI workloads from pilot to production.

 **The economic decision is not just which model is cheapest. It is which architecture and model mix satisfies the workload's **quality, latency, compliance, and scale** requirements at the best economics.**

AI TCO includes build, run, evaluate, govern, maintain, observe, and remediate. Beyond token costs.

1 Workload first Start with the AI use case, not only the cloud bill. Model the workload, user behavior, quality needs, latency targets, and compliance boundaries. Example: Customer support RAG, coding agents, or marketing content all have different economics. ✓ Built for AI workloads 	2 Full AI TCO Model more than inference token costs. Include reasoning, context, embeddings, vector DB, OCR, evals, observability, HITL, guardrails, integration, retraining, and maintenance. Example: A support bot cost per resolved case includes retrieval, generation, review, and refresh. ✓ Beyond token costs 	3 Plan before spend exists Forecast pilot, production, and scale before the first invoice arrives. Use workload assumptions to estimate TCO, ROI, and variance ranges. Example: Estimate development cost for 500 developers using AI coding tools before rollout. ✓ Plan before rollout 
4 Cost and governance together Governance starts on day one. Security, privacy, model eligibility, human oversight, and audit needs are part of the workload design. Example: A PHI or PII workload may require approved models, residency rules, and human review. ✓ Governance by design 	5 Deterministic decision engines Use reproducible cost decision models that are auditable and machine callable. The same inputs should yield the same output, assumptions, and evidence. Example: A token estimator or agent loop engine can be re-run with pinned pricing and identical logic. ✓ Auditable decision logic 	6 Policy ready outputs Turn decisions into operational guardrails such as budget caps, model allowlists, route rules, turn limits, token limits, and approval gates. Example: Set an agent cost cap per task and fall back to a cheaper approved model when needed. ✓ Operationalize decisions 
7 Out of band control plane AICost can guide and govern without sitting in the inference path. It can stay vendor neutral and export decisions into existing runtime systems. Example: Apply policies through LiteLLM, Portkey, Bedrock, Azure AI, Vertex AI, or direct APIs. ✓ Works with your stack 	8 Compose, enforce, prove Chain engines into workflows, enforce selected controls in runtime, and then prove outcomes with actuals and evidence. Example: Token estimator plus router plus CostWall plus CostProof creates a full operating loop. ✓ Full operating loop 	9 Forecast to actual Close the loop between modeled economics and real usage. Reconcile bills, telemetry, attribution, and showback against the original plan. Example: Compare forecast vs actual by team, project, customer, model, or provider. ✓ From forecast to proof 
10 Vertical and function aware Different departments and industries need different economics and controls. AICost adapts to marketing, engineering, healthcare, finance, legal, and support workflows. Example: Healthcare prior authorization agents, marketing video generation, and engineering copilots should not be modeled the same way. ✓ Economics by context 	11 Partner with the customer AICost is not only generic SaaS. Workflows, costs, and governance are intertwined with each customer's business processes, data, and operating model. Example: Co-design a workflow for a regulated claims process or a secure enterprise coding environment. ✓ Partnership-led delivery 	12 Customize where it matters Use the platform as a base, then tailor engines, parameters, policy packs, and MCP runtime rules to the customer's unique requirements. Example: Create a custom engine with customer specific variables and policies for their approval flow, budget rules, and governance needs. ✓ Configured to fit you 

What this means for customers



Better architecture choices
Choose the right model mix for quality, latency, and cost.



Safer pilot to production rollout
Reduce risk with built-in governance and controls.



Clearer ROI and unit economics
Make data-driven investment decisions.



Continuous governance and proof
Reconcile, show value, and keep improving.



AI cost decision intelligence | [Compose. Enforce. Prove.](#)

Real AI outcomes start with better decisions.

The twelve operating principles behind every engine and every policy.

14. Questions we get

Straight answers

Question	Answer
Do you see our prompts or customer content?	No. Read-only cost and usage, a team-scoped key convention, a workload description and an owner per workload. No prompts, completions or customer data, and no production access.
Do you sit in our inference path?	No. Traffic keeps flowing application to gateway to vendor as it does today. Policy goes out as configuration; actuals come back to the ledger.
Are the numbers generated by an AI model?	No. The decision engines make no model calls. Same inputs produce the same answer with the source and date of every rate.
Can we buy the assessment without continuous governance?	Yes. Part A is complete on its own: findings, policy patch, roadmap, deck and evidence pack. Part B is the option to keep it true afterwards, and it credits from the assessment.
What if you cannot read our systems?	Then we ask, precisely. A scoped questionnaire is generated from the 247 controls for your profile, must-tier first. Every answer carries a justification, an evidence link and an assurance level.
How do we know you never edited the record?	Every monthly report ships with the command you run yourself against the hash-chained, signed journal. If a single entry had changed, verification fails.
Does this replace our FinOps or GRC platform?	No. Each keeps doing its job. None of them owns the complete workload economic decision, which is the gap AICost fills, and it exports back into each of them.
What does a pilot look like?	Connector exchange on day 1, first attributed number in week 1, policy patch in one session, caps live by week 2 to 3. Five objective success criteria your own team verifies.

Start here

One workload, one exhaust path, first graded report in 48 hours

Free AI Act Quick-Check

Two minutes, no signup wall. Your risk class and likely obligations.

aicost.ai/governance/quick-check

Book a Take-Stock Session

One hour, fixed fee. Two workflows named and priced, a same-day written plan. Credits into the assessment.

aicost.ai/info

Talk to us

Enterprise, MSP resale and investors:

partner@cloudintelligence.ai.

Product and MCP beta keys:

support@cloudintelligence.ai.

949-378-7860 (Irvine) · 469-951-8383 (Austin)

Everyone else sells you a photograph of your compliance. We sell you the camera, running nightly. Governance here is a dollar figure, not a colour chart, and every dollar traces to an engine you can re-run. And here is the command you run yourself to prove we never edited the record.

CloudIntelligence.ai LLC, Irvine, CA and Austin, TX. NVIDIA Inception member. aicost.ai · mcp.aicost.ai · mspcost.com. Prepared 17 September 2026.

Appendix. The one-page summaries

The two flyer pages, for forwarding



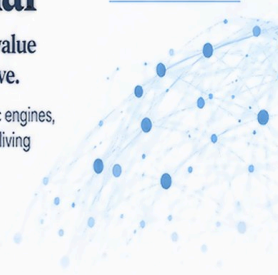
AICost.ai
MODEL | DECIDE | GOVERN | PROVE
Irvine, California | Austin, Texas

Governance that becomes operational

AI Cost & Governance helping enterprises turn AI into governed production value with connected evidence, enforceable policy, and continuous proof after go-live.

AI Cost does more than assess. It connects to real systems, computes priced findings with deterministic engines, compiles decisions into policy, and keeps governance current through recurring re-proof. The result is a living operating capability for AI cost, risk, and accountability.

CONTINUOUS AI GOVERNANCE



Connected to your real systems

Evidence first. Read-only access where available. Detailed questionnaires only where system data cannot be read automatically.



Microsoft 365 and Entra

- ✓ Entra ID roles and access
- ✓ Defender discovered AI apps
- ✓ Licensing and Copilot usage
- ✓ Purview and SharePoint signals



AWS Bedrock

- ✓ CUR and cost allocation tags
- ✓ Invocation logs and agent usage
- ✓ Cross-account read role
- ✓ Per-model visibility



Azure AI Foundry and GCP Vertex

- ✓ Cost exports and deployment RBAC
- ✓ Token and usage metrics
- ✓ BigQuery billing export
- ✓ Org-policy and allow-lists



Vendors, gateways and finance

- ✓ OpenAI, Anthropic, Google usage APIs
- ✓ LiteLLM, Portkey, Cloudflare, Kong, OpenRouter
- ✓ Corporate card and AP exports
- ✓ Vendor trust and DPA evidence



Where we cannot read it, we ask precisely

AI Cost generates a scoped customer questionnaire mapped to the relevant controls and evidence requirements. Responses are tied to justification, evidence links, and assurance level, so the output is more than a checklist.

The policy engine

Governance you can load, not just read.

1

2

3

4

5

6

Residency

Certification

Training data

Tier

Capability

Price

```
route: legal / contract-review
data_class: regulated
residency: EU
claude-sonnet-5 PASS Bedrock EU, training off, approved
policy_patch: model list, budget cap, no-downgrade zone, approval above cap
```

- ✓ Eleven intake questions, longer only when the workload requires it
- ✓ Plain-English elimination trace for every model ruled out
- ✓ Exports as configuration for gateways, IAM, or vendor-console controls

Continuous governance cadence



Daily

Cost actuals ingest, bill versus policy, engine audit



Weekly

Watcher sweep for drift, eligibility, evidence, and inventory hygiene



Monthly

Branded drift report with signed evidence



Quarterly

Access recertification and ownership review



Continuous

Incident clock, oversight log, and evaluation heartbeat

Seven built-in watchers

Premium drift

Model eligibility

Vendor evidence

Regulatory deadlines

Inventory hygiene

Findings burn-down

Bill vs policy

Why continuous beats annual



Drift is caught in days, not at renewal.



Audits stop becoming fire drills.



Legal and oversight clocks stay visible.



Cost and governance stay in one operating loop.

What customers hold at the end

Executive report and risk heatmap

AI inventory and use-case catalog

Obligations matrix across frameworks

Findings register with owners and remediation windows

Compliance Premium in dollars

Model eligibility matrix

Policy patch ready to load

Vendor evidence pack

Target operating model

Evidence register and signed journal

Stop / fix / approve / scale roadmap

Executive deck and MCP access to the engines

By the numbers

20

governance modules

247

controls

120+

deterministic engines

11

deliverables plus executive deck

Nightly

continuous re-proof



AICost.ai
MODEL | DECIDE | GOVERN | PROVE
Irvine, California | Austin, Texas

Continuous governance for enterprise AI.

partner@aicost.ai | 949-378-7860



Scan for AICost.ai/info

**Connected evidence.
Operational policy.
Real Governance.**

AICost.aiMODEL | DECIDE | GOVERN | PROVE
Irvine, California | Austin, TexasAI COST &
GOVERNANCE CONTROLS

Continuous Governance of Enterprise AI

AI Cost & Governance helping enterprises move from pilot to production with enforceable policy, continuous proof, and controlled AI spend.

AICost connects to real systems, turns findings into operational policy, and keeps governance current after go-live. This is not a one-time assessment. It is a living governance capability for AI cost, risk, and operating readiness.

-  **PLAN**
with clarity
-  **GOVERN**
with confidence
-  **OPERATE**
with control
-  **PROVE**
real outcomes

Why enterprises need this now

What enterprises struggle to sustain in 2026 is not intent. It is continuous governance that survives real-world change.



Production AI creates cost drift

Usage grows across teams and tools, making spend harder to predict and control.



Policies often remain documentation, not enforcement

Good intent does not prevent risky or noncompliant use in live environments.



AI inventories go stale quickly

New models, users and tools appear constantly, and manual inventories fall behind reality.



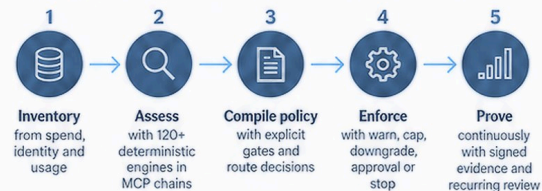
Auditors and boards need current proof, not last year's deck

Point-in-time artifacts are no longer sufficient for regulators, auditors and executive oversight.

What makes AICost different

Typical assessment	AICost continuous governance
Interviews and questionnaires only	Connected to real systems first: Entra, Bedrock, Azure, GCP, gateways, licensing and billing exports
Maturity heatmap	Deterministic governance and cost engines with repeatable outputs
Policy memo	Policy engine that turns decisions into enforceable controls
Snapshot in time	Continuous governance with watchers, heartbeats and nightly re-proof
Generic findings	Priced findings, owned actions and customer-specific roadmap

How governance becomes operational



Where data cannot be read automatically, AICost uses a detailed customer questionnaire scoped to the relevant controls and evidence requirements.

Key differentiators



Connected to real systems for evidence-first assessment



120+ live deterministic engines in MCP chains



Policy engine that makes governance operational



Continuous governance, not a one-time deliverable



Detailed questionnaire only where system data is unavailable

Customer outcomes



Faster path from pilot to production



More controlled AI spend and fewer surprises



Governance that stays current after go-live



Audit, board and regulator readiness with current evidence



Clear accountability across AI systems, owners and policies

By the numbers

20

governance modules

247

controls

120+

deterministic engines

11

deliverables plus executive deck

∞

Continuous re-proof after go-live

AICost.ai
MODEL | DECIDE | GOVERN | PROVE**Governance that operates in the real world.**

partner@aicost.ai | 949-378-7860



Scan for AICost.ai/info

**Connected evidence.
Operational policy.
Real Governance.**