aciinfotech



The Invoice Arrives Too Late: Governing Enterprise AI Economics at Runtime

Why falling token prices raise AI bills, and how an economic control plane turns consumption into cost per accepted outcome

Executive summary

Enterprise AI spend is now set by software at the moment it runs: how much context a workflow assembles, which model answers, how long it reasons and how often an agent loops. None of those choices is made by the people who own the budget, and the invoice that reports them arrives weeks later, organized by vendor rather than by workflow or outcome.

The gap is widening.

24-fold

Goldman Sachs Research expects global token consumption to grow 24-fold between 2026 and 2030 (Goldman Sachs, May 2026).

40%+

Gartner expects more than 40 percent of agentic AI projects to be canceled by the end of 2027, citing escalating costs and unclear business value (Gartner, Jun 2025).

73%

And 73 percent of enterprise respondents are not confident, or only slightly confident, that they can tie AI spend to an outcome their CFO would accept (Tokenomics Foundation, Sep 2026).

Our argument is that AI cost cannot be governed from the invoice. It must be governed in the request path, by an **economic control plane** that meters, attributes, bounds and routes every call and joins cost to evaluated quality, making **cost per accepted outcome** the unit of management.

THREE DECISIONS COME FIRST:

01

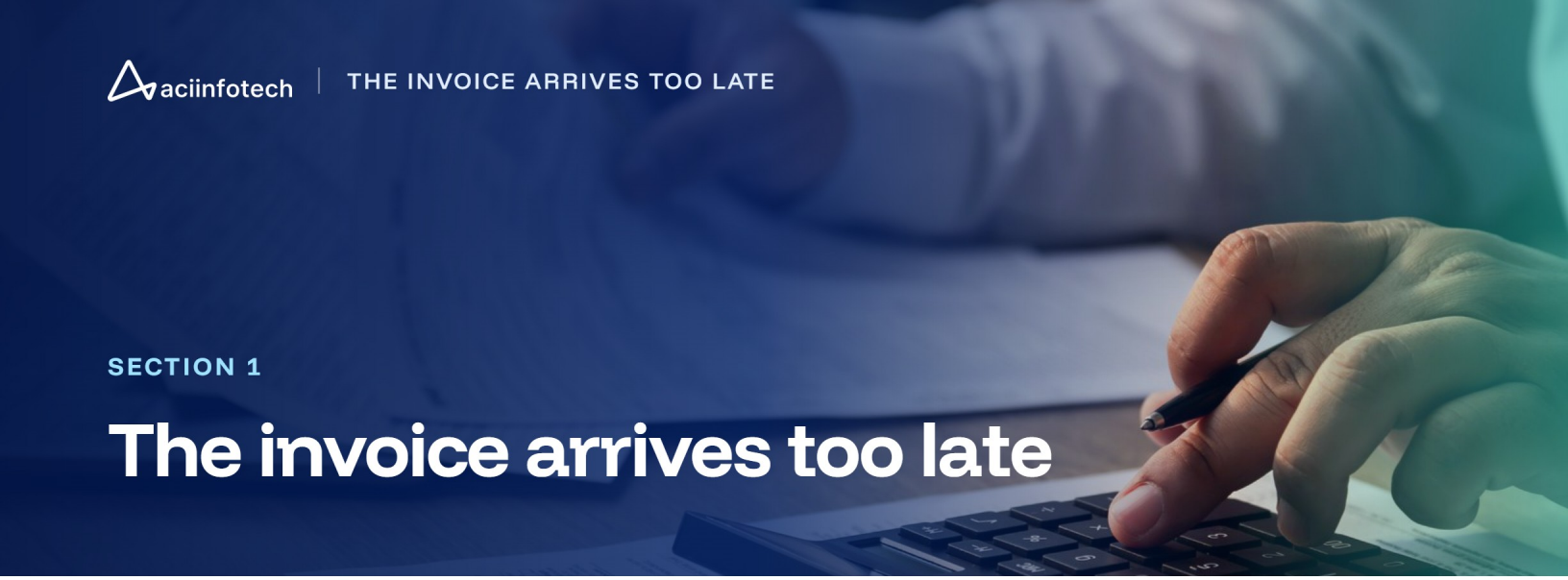
Name one accountable owner for AI unit economics

02

Require telemetry and attribution before any workload reaches production

03

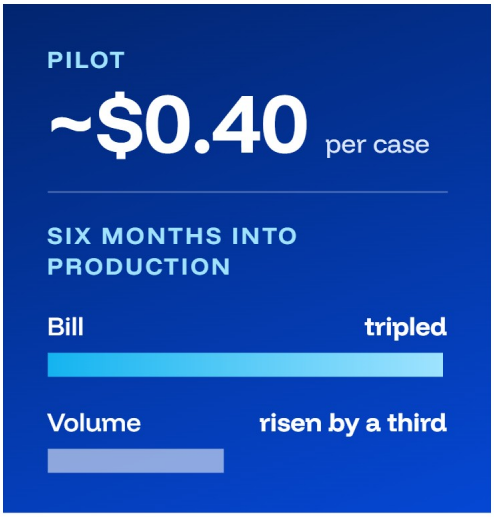
Put hard runtime budgets on every agent



SECTION 1

The invoice arrives too late

Consider a composite case. A claims-intake assistant clears its pilot at about forty cents per case. Six months into production, the bill has tripled while volume has risen by a third. The retrieval corpus doubled, a timestamp added to the system prompt broke prompt caching, a flaky API triggered agent retries, and the default moved to a newer reasoning model. **Each change was reasonable; none passed through a budget process.**



Cloud FinOps was built for decisions made at provisioning time, few in number and stable for months. AI cost is decided per request, by code or by the model itself. The industry has noticed: 98 percent of State of FinOps 2026 respondents now manage AI spend, up from 31 percent in 2024 (FinOps Foundation, 2026), and in June 2026 the Linux Foundation announced plans for a Tokenomics Foundation to set open standards for AI cost (Linux Foundation, Jun 2026). **What remains open is the architecture.**

98%

of State of FinOps 2026 respondents now manage AI spend



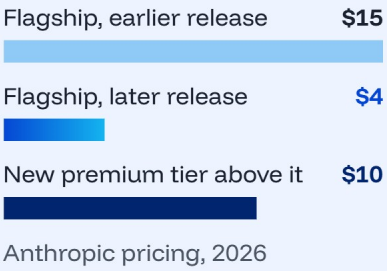
Source: FinOps Foundation, 2026

SECTION 2

Cheaper tokens, larger bills, thin proof

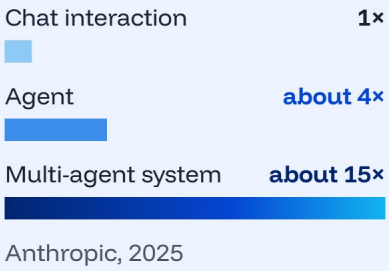
The price of a fixed capability keeps collapsing: the cost of GPT-3.5-level performance fell **more than 280-fold** in two years (Stanford HAI, 2025). But enterprises buy the frontier for their hardest work, and the frontier moves. Anthropic's flagship input price fell from $15 to $4 per million tokens across releases while a new $10 premium tier appeared above it, and its catalog now spans a **100-fold** price range (Anthropic pricing, 2026). **Consumption is rising faster than prices fall.** Agents use about four times the tokens of a chat interaction, and multi-agent systems about fifteen times (Anthropic, 2025).

Flagship input price
US dollars per million tokens



Anthropic pricing, 2026

Tokens per interaction
Relative to a chat interaction



Anthropic, 2025



Proof lags spend.

Gartner forecasts $2.67 trillion of worldwide AI spending in 2026, up 49.5 percent (Gartner, Sep 2026), yet only 37 percent of McKinsey's respondents report any EBIT impact from AI (McKinsey, Aug 2026). In the Tokenomics Foundation survey, 53 percent are at least moderately confident in their visibility into AI spend, but only 26 percent are equally confident they can tie it to an outcome a CFO would accept. Just 4 percent asked providers for lower prices.

SPEND

$2.67 trillion

worldwide AI spending in 2026, up 49.5 percent

PROOF

37%

of McKinsey's respondents report any EBIT impact from AI

Tokenomics Foundation survey



The binding constraint is **measurement**, not price.

SECTION 3

Why the FinOps playbook breaks down

01

Cost is set at execution.

Output length, reasoning tokens, tool calls and retries are unknown until a request completes, so monthly showback reports the damage after it compounds.

02

Decision rights are split.

Engineers choose models and context; finance sees totals. Not one surveyed organization without an owner for AI economics could tie spend to a CFO-grade outcome.

03

Attribution breaks.

One agentic request fans out across tools and shared gateways. Even the FOCUS 1.5 billing specification, due December 3, 2026, names the caller but not the beneficiary (FOCUS, Oct 2026).

04

Units fragment.

AI arrives as API charges, marketplace draws, SaaS credits and GPU leases, and 64 percent of respondents use AI embedded in other products.

05

Consumption is an attack surface.

OWASP ranks unbounded consumption, including "denial of wallet," among the top ten risks for LLM applications (OWASP, 2025). Spending limits are security controls, and the CISO should co-own them.



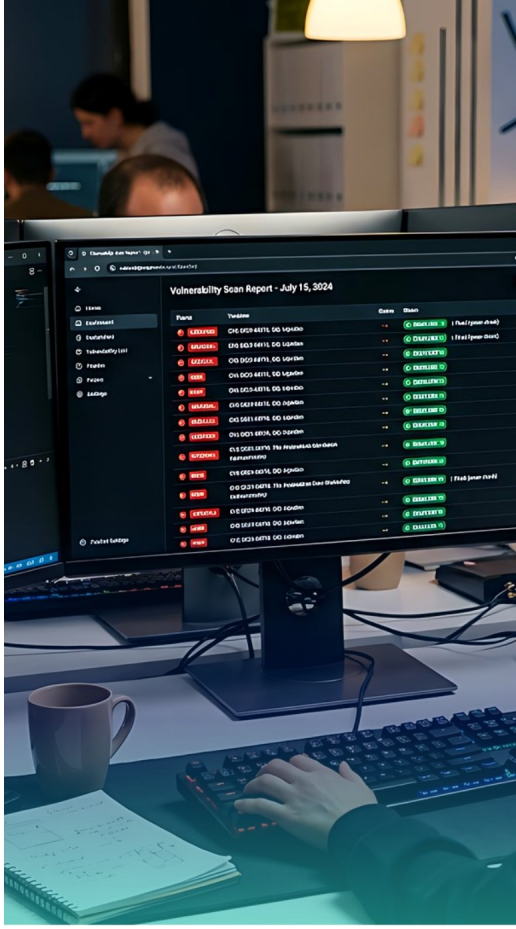
SECTION 4

Where cost is actually decided

The unit that matters is not the token but the accepted outcome:

Cost per accepted outcome =

(Model calls + Tools + Retrieval + Human review) / Accepted outcomes



Six architectural levers move the numerator:

1. Context re-sent at every agent step
2. Prompt layout, which decides whether input bills at cache rates as low as a tenth of the base price (Anthropic docs, 2026)
3. Model choice
4. Hidden reasoning
5. Agent loops
6. Whether work runs on demand, in batch or on fixed capacity

Each has an owner who rarely sees the bill.

Routing is the largest lever and the least proven: RouteLLM cut costs by 35 percent to more than 85 percent on public benchmarks (LMSYS, Jul 2024), but production savings hold only where every route is evaluated.

RouteLLM cost reduction, public benchmarks

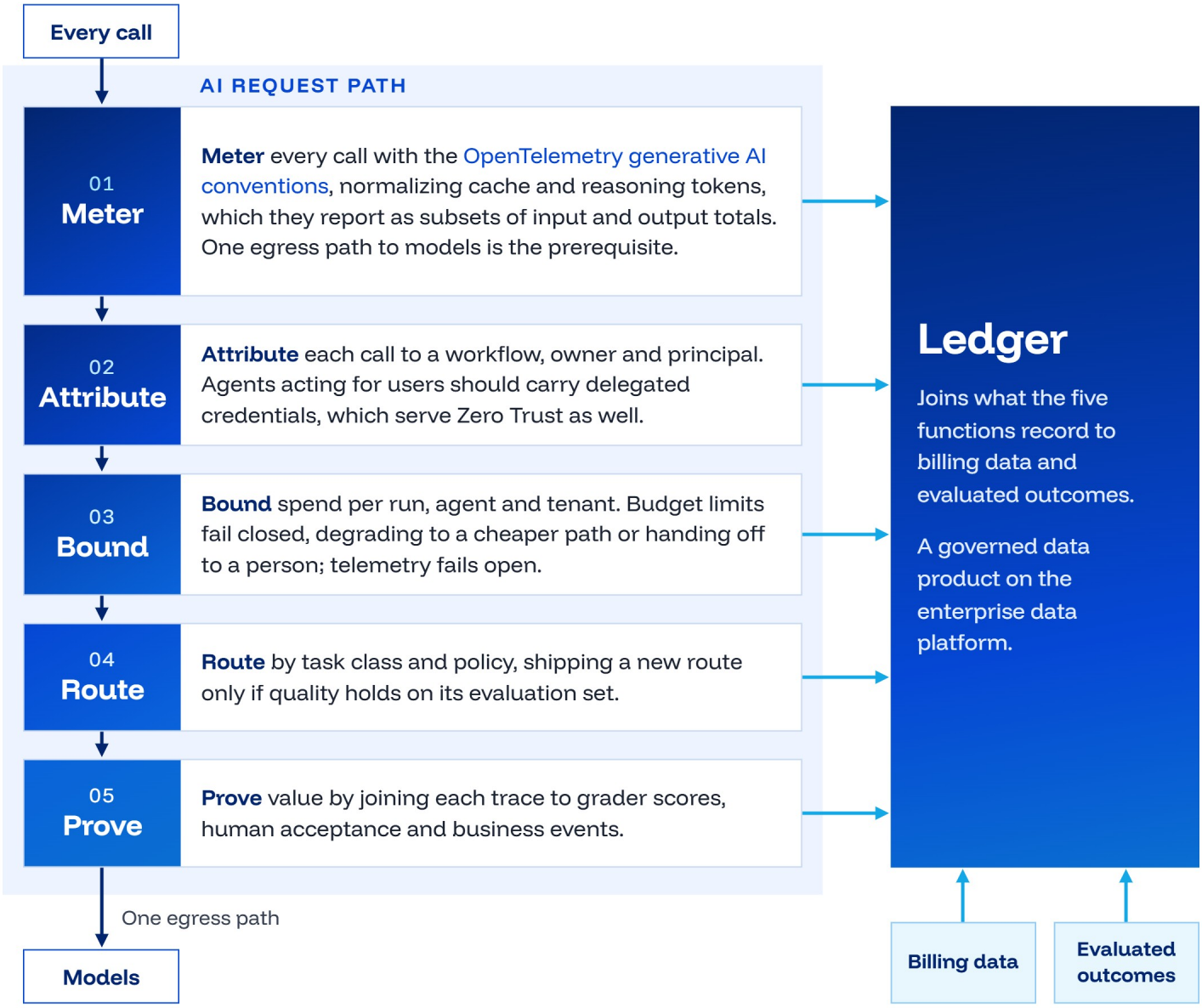


LMSYS, Jul 2024

SECTION 5

The economic control plane

We propose an economic control plane: five functions in the AI request path and a ledger that joins what they record to billing data and evaluated outcomes. It is a reference architecture proposed in this paper, built largely from components enterprises already run.



The ledger belongs on the enterprise data platform as a governed data product. Keep OpenTelemetry and FOCUS at every boundary to avoid gateway lock-in. **AI embedded in SaaS products bypasses the plane;** there, control is contractual: capped credit pools, administrator budgets and usage-data rights.

SECTION 6

Proving value: cost per accepted outcome



A hypothetical document-review workflow shows why token price misleads.

$60 / hour
Reviewer time costs $60 an hour

1 minute
Each AI output gets a one-minute check

8 minutes
A rejected output takes eight minutes of rework

Option	Token cost per case	Acceptance rate	Review and rework	Cost per completed case
A. Frontier model for every case	$0.30	90%	$1.80	**$2.10**
B. Small model for every case	$0.03	70%	$3.40	**$3.43**
C. Routed: 70% small (90% acceptance), 30% frontier (88%)	$0.11	89.4%	$1.85	**$1.96**

Token cost per case Review and rework

SMALL MODEL
Token bill **−90%**
Cost of a completed case **+63%**

ROUTING
Tokens **−63%**
Total cost **only −7%**

EACH POINT OF ACCEPTANCE
$0.08
a case

The small model cuts the token bill by 90 percent and raises the cost of a completed case by 63 percent. Routing cuts tokens by 63 percent but total cost by only 7 percent, because tokens are about one-seventh of fully loaded cost and a slight drop in acceptance gives back a quarter of the saving. Each point of acceptance is worth $0.08 a case. **Quality is the larger lever**, so measure against a pre-AI baseline, use holdout groups where possible, and count review and rework, not only tokens.



SECTION 7

Roadmap and executive agenda

The program runs about twelve months in four phases, each ending at a decision gate. **Do not optimize what you cannot attribute, and do not claim savings you have not evaluated.**

Program timeline

Months from start



PHASE 0
Baseline
Weeks 0 to 6

Inventory workloads, API keys and SaaS meters; name an owner and set a cost and outcome baseline for each workflow

Exit gate
A: owned

PHASE 1
Instrument
Months 2 to 4

One egress path, OpenTelemetry on every AI call, FOCUS billing ingest, budgets on all agents

Exit gate
B: attributable

PHASE 2
Control
Months 4 to 8

Routing and effort policies gated on evaluation, cache-aware prompts, batch for asynchronous work

Exit gate
C: quality-neutral

PHASE 3
Prove and govern
Months 8 to 12

Ledger in production; a cost-per-outcome target in every stage gate

Exit gate
D: governed

B Gate B, when most spend traces to a workflow and an owner, carries the most weight; savings claimed before it rarely survive a budget review.

90 In the next 90 days

- [ ] Name one accountable owner with authority over platform standards
- [ ] Mandate a single metered egress path, starting with the highest-spend workloads
- [ ] Put hard budgets on every agent
- [ ] Inventory every AI meter in SaaS and cloud contracts before renewals

EXECUTIVE AGENDA

Four decisions need executive sponsorship:

01 **Who owns AI economics**

02 **How much autonomy agents receive**

03 **Which workloads may override fail-closed limits**

04 **Whether to build, buy or assemble the control plane**

SECTION 8

Risks and limits

Teams can game acceptance rates by loosening criteria, so keep evaluators independent of the teams they judge. Budgets that ignore value will throttle the most productive users; McKinsey finds AI high performers hit cost constraints on coding agents about three times as often as their peers. Prompt telemetry can carry sensitive data, so capture metadata by default. Standards are still moving, routing evidence is mostly benchmark-based, and the control plane carries its own cost. **Some workloads should simply stop, and the ledger's most valuable output may be that decision.**

ABOUT

3×

McKinsey finds AI high performers hit cost constraints on coding agents about three times as often as their peers

Conclusion

The first era of enterprise AI economics was about price: which model, which discount, which provider. The next belongs to enterprises that can state, for every workflow, what an accepted outcome costs, why it costs that, and what would make it cheaper without making it worse. That capability is built into the request path, not bolted onto the invoice.

The invoice will always arrive.
The goal is that nothing on it is news.

ABOUT ACI INFOTECH

Enterprise Data and AI, Engineered and Run in Production.

ACI Infotech starts with data. We modernize enterprise data platforms across Databricks, Snowflake and Microsoft Fabric, embedding data quality, observability and governance to create trusted foundations for AI.

Building on that foundation, we help enterprises move beyond AI pilots and into production with ArqAI, our composable enterprise AI platform.

AI is also embedded in how we work. More than 40% of our software development is AI-assisted, while our growing team of certified data and AI specialists brings experience from thousands of engagements. This combination of engineering expertise, AI-assisted delivery and platform-led innovation helps accelerate delivery and drive greater value from every engagement.

Headquartered in Somerset, New Jersey, with delivery hubs worldwide, ACI Infotech is an enterprise data and AI engineering firm serving organizations across financial services, healthcare, retail, manufacturing and energy.

We build the data foundation, put AI on top of it, and run both in production.