JETSTREAM

The AI Budget Black Box

A Four-stage FinOps Framework for Enterprise AI

AI adoption is accelerating faster than the budgets behind it. Here is how to see every dollar, control it, and bring it down, without putting the brakes on the teams using it.

5201 Great America Parkway
Suite 346
Santa Clara, CA 95054
go.jetstream.security | sales@jetstream.security

V1-061026

Table of Contents

Executive Summary 03

Stage 1: Transparency 04

Stage 2: Cost Control 05

Stage 3: Cost Savings 06

Stage 4: Cost Optimization 07

One Platform, 4 Stages 08

Executive summary

Enterprise AI spend is growing faster than the controls around it. The same trust gap that stalls AI adoption shows up in the budget. Leaders cannot say who is spending what, on which models, for which outcomes, and the invoice is the first place they find out. **This paper sets out a four-stage FinOps framework** that treats AI spend as a governed control surface rather than a number discovered after the fact. The stages run in dependency order. Transparency comes first, because you cannot govern spend you cannot see. Cost control comes next, because ungoverned access produces runaway bills. Savings and optimization follow, because they are only safe and durable once the first two stages hold. The through-line is the one we apply to every part of AI governance. Governance is the accelerator, not the brake.

The bill nobody can read

Most enterprises moved faster on AI adoption than on AI accounting. Budgets were approved, pilots became production, and agents started doing real work. The financial picture did not keep up. Finance sees a set of provider invoices. Engineering sees a set of dashboards. **Nobody sees the whole bill broken down the way a CFO needs it, by the people and projects that drive the spend.**

A FinOps-governed enterprise can answer these Qs: What did it cost? Who owns it? Was the spend necessary? Did it stay on budget? The four stages below show how an organization earns the right to answer all four, in the order the answers depend on each other. Transparency is the foundation, control is the guardrail, savings is the return, and optimization is the refinement.

Stage 1. Transparency

The problem. No enterprise runs its AI on a single provider. Even a committed Microsoft shop ends up using AI capabilities from other services, because no one vendor is best at everything and teams pick the tool that fits the task. Mergers and acquisitions deepen the sprawl. A company that standardized on one stack inherits a different one the day a deal closes. The mixing happens inside a single workflow too. It is now common to write code with one model and review it with another, Claude on one side and Codex on the other. Each choice is reasonable. Together they produce an AI estate that no single provider invoice can describe. Spend that cannot be traced to a specific, authorized use leaks quietly, and a shared or compromised credential turns that leak into shrinkage no one can see.

What it looks like. When Uber rolled Claude Code out to its 5,000 engineers, the company also signaled it would add a second provider for the same work. Multi-vendor is not an edge case. It is the default end state of a serious AI program, and it guarantees that the real bill lives in pieces.

No enterprise is a one-stop shop

A single enterprise, on a single Tuesday. Marketing drafts copy on one provider. Engineering writes code with a second model and reviews it with a third. A newly acquired business unit runs a fourth stack the parent company has never inventoried. A shared API key from last year's pilot is still live on someone's laptop. **Five sources of spend, one of them forgotten, and zero unified bills.**

The solution. Transparency starts with a global roll-up of all AI cost, with breakdowns by user, team, department, and project. **The JetStream SAIG Platform™** (Security-first AI Governance) produces that roll-up by first discovering the full AI estate, including the shadow tools no one registered, and then binding every unit of spend to an accountable identity and to the approved design recorded in **JetStream AI Blueprints™**. The breakdown becomes a property of the system rather than a quarterly reconstruction. You cannot control, save, or optimize spend you cannot first see and attribute.

Stage 2. Cost Control

The problem. Visibility tells you where the money goes. It does not stop the money from going. The second stage is control, because ungoverned AI access produces bills that arrive faster than anyone can react to them. A demo runs hot for two hours. A team burns its annual token budget a third of the way into the year. By the time the invoice lands, the spend is already gone, because access was handed out with no budget attached.

What it looks like. Uber burned through its entire 2026 AI budget in four months, driven by Claude Code adoption that climbed from 32% to 84% of itsengineers. The tool was not broken. It worked exactly as designed, and that was the problem. Token-based billing meant individual engineers ran $500 to $2,000 a month, and Uber's own CTO spent $1,200 in a single two-hour demo session. In the most extreme case reported to date, an AI consultant told Axios that a client spent around half a billion dollars in a single month on Claude licenses after failing to set usage limits on employee accounts. None of these were the result of a bad decision. Each was the result of no decision.

Giving AI to developers and citizen developers without budget controls is like giving your fourteen-year-old a credit card with the same limit as the parents. Irresponsible at best, reckless in reality.

The solution. Control means budgets and rate limits enforced at the moment of use, not reconciled at the end of the quarter. The **JetStream SAIG Platform™** applies that enforcement at the platform layer, scoped to identity and to the approved Blueprint, so a runaway workflow is stopped before the invoice rather than explained after it. Anomalous burn is detected and blocked as it happens. With the **JetStream Key Broker™**, every credential is a scoped, revocable virtual key, so a single leaked or shared key is revoked on its own while every other workflow keeps running.

Cautionary tales: when access has no limit

Uber burned its entire 2026 AI budget in four months as Claude Code spread to 84% of 5,000 engineers. Individual engineers ran $500 to $2,000 a month, and the CTO alone spent $1,200 in a single two-hour demo. (Uber CTO Praveen Neppalli Naga, via StartupFortune and The Information)

Agents get stuck in recursive loops and run up a bill all weekend while no one is watching. (reported industry pattern)

A client spends around half a billion dollars in a single month on Claude licenses after failing to cap employee accounts. (Axios)

Stage 3. Cost Savings

The problem. With spend visible and controlled, an enterprise can finally pursue savings without flying blind. The instinct is to negotiate the rate card. The bigger lever is using the right model for the right task, and proving it with data rather than guessing. Rate cards mislead because they are flat. They price per token, but token counts vary by model, and a newer model can generate noticeably more tokens for the same input. A newer flagship can ship with a tokenizer that produces up to 35% more tokens for the same text than its predecessor. At identical per-token prices, two models then produce very different bills for the exact same prompt, and the gap compounds across a large corpus.

What it looks like. A flat rate card cannot tell you that 80% of your employees' prompts are short classification tasks where a small model is fine, while 20% are long reasoning tasks where a larger model earns its cost. Nor can it see that a stateless agent resends its full history on every tool call, so by step 20 of a loop the input on a single call can exceed 50,000 tokens. At a $3 per million input rate, that one late step costs about $0.15, and a 50-step task can pass $5. The waste is invisible until someone measures it.

The solution. The **JetStream SAIG Platform™** brings a set of techniques to bear on the same goal, routing work to the right model and cutting the cost of the work that remains.

Five ways to cut the AI bill

Right model for the task, not the rate card. Reuse answers to repeated intent. Inventory and retire shadow spend. Batch the work that can wait. Prune the context that does not earn its tokens.

Semantic caching recognizes when a new prompt matches the intent of one already answered and avoids sending it to the model again.

Shadow AI elimination maintains a complete inventory of every AI tool, subscription, and API in use, with monthly cost attached, so savings opportunities surface instead of hiding.

Prompt sampling runs a representative sample of real prompts against each candidate model to expose the true cost distribution rather than the headline rate.

Batch processing moves work that does not need a real-time answer, such as data enrichment, bulk classification to asynchronous batch APIs that run around half the cost.

Context window pruning keeps a sliding window of the relevant recent steps instead of the full history on every call, so late-loop steps stop ballooning.

Stage 4. Cost Optimization

The problem. Savings reduce the cost of the work you run. Optimization reduces the work itself. Every extra round trip to a model is tokens spent and money gone, and most of those round trips exist because the first answer missed context the model never had.

What it looks like. The cost of a task is rarely one call. It is the loop. A request that should take one pass takes four and that difference is pure waste no rate card will ever show you.

The solution. The SAIG Platform's approach to this stage is StreamContext, which runs at the JetStream AI Hub. It is the evolution of the dynamic context the platform already applies, which shapes a prompt to the invoker's identity and entitlements. StreamContext appends the context a prompt needs before it reaches the model, so the first answer is closer to the intended one and the loop is shorter. That is the difference between spending less on each call and needing fewer calls.

StreamContext: fewer iterations, conformant by default

A citizen developer asks for a function. Without help, the first draft ignores half the organization's standards, and three more rounds of prompting follow. With StreamContext, the AI Hub attaches the SDLC and coding guidelines to the prompt before it reaches the model. The first answer already conforms. A loop that used to take four iterations takes one.

One platform, four stages

The four stages are a sequence, not a menu. Transparency makes control possible. Control makes savings safe. Savings make optimization worth pursuing. The JetStream SAIG Platform delivers them as one control plane rather than four tools. The platform discovers and inventories the AI estate, binds every action to an accountable identity and an approved Blueprint, enforces budgets and policy inline at the AI Hub, issues revocable virtual keys through the Key Broker, and shapes prompts with StreamContext so the work itself gets leaner. Run in order, the four stages turn AI spend from a black box into a managed line of the business.

Transparency
makes control possible

Control
makes savings safe

Savings
make optimization worth pursuing

Optimization
makes scale affordable

When every dollar has an owner, AI adoption becomes sustainable. When no dollar has an owner, AI adoption becomes a liability. You can't scale what you can't afford.

Talk to us

The teams that get ahead of this will treat AI economics as part of AI governance, not as a finance clean-up exercise after the fact. Talk to us about how the JetStream SAIG Platform governs AI spend, from transparency through optimization, so your organization can adopt AI at speed without the budget becoming the thing that stops you.

JetStream AI Blueprints™ give enterprises the operational truth about how AI behaves. JetStream AI Manifest™ ensures that truth rests on complete and current inventory. Together **with AI Drift Detection™**, **AI Hub™**, **Identity Broker™** and **Key Broker™**, JetStream provides the visibility, control, and accountability required to scale AI responsibly.

Contact us at sales@jetstream.security

For more information about JetStream Security

5201 Great America Parkway
Suite 346
Santa Clara, CA 95054
go.jetstream.security | sales@jetstream.security

JETSTREAM