WITNESS AI

AI FinOps: Risks, Costs, Outcomes

Understand AI spend by intent, control it, and prove the return.

The board wants proof your AI spend pays off. You can't show it.

AI bills are climbing fast, and almost no one can say which tokens are adding value. Enterprise AI spend is forecast to reach $2.52 trillion in 2026, up 44% in a year, yet most organizations cannot see where it goes: shadow subscriptions multiply, vendor invoices show an island of tokens, and nothing shows the purpose of outcomes for all the spend. Agents make it worse, where loops burn through monthly budgets overnight.

Cost is only one of three questions the board is asking. The CISO cannot confirm what data reaches which models or produce audit evidence on demand, or show the ROI for their security investments. The CAIO cannot tie AI usage to outcomes: 88% say they use AI, but under 40% show measurable impact and 62% stay stuck in pilots. Risk, cost, and adoption are critical and interrelated metrics. Without governance in the traffic path, these metrics rest on data you cannot see.

The AI spend problem, in numbers:

73%
Organizations exceeded their AI budget this year

44%
Organizations have a functional AI FinOps practice

214x
Variation in token cost across providers

50-70%
Share of enterprise AI traffic that is personal or non-work use

Reduced risk, quantified

The value of exposure events prevented.

Risk avoided: Blocked exfiltrations, leaks, and injections convert to dollars of risk avoided, using multipliers you set.

Proprietary IP protected: Source code and IP that never escaped to a third-party model, counted and priced as loss prevented.

Compliance cost avoided: Audit evidence on demand cuts audit-prep hours and shrinks exposure to non-compliance penalties.

Reduced cost, controlled

AI tokens saved by preventing wasteful spend, both internal and B2C.

Route to the right model: Intent-based routing sends each prompt to the cheapest model that clears the quality bar, up to 80% less cost.

Find shadow subscriptions: Discovery catalogs every AI tool by team, where real spend often runs 3x the enterprise licensed tools.

Cut non-work prompt waste: Per-prompt classification filters personal and off-topic use before consumes tokens.

Adoption & outcomes, proven

Productive AI you can prove, separate from the noise.

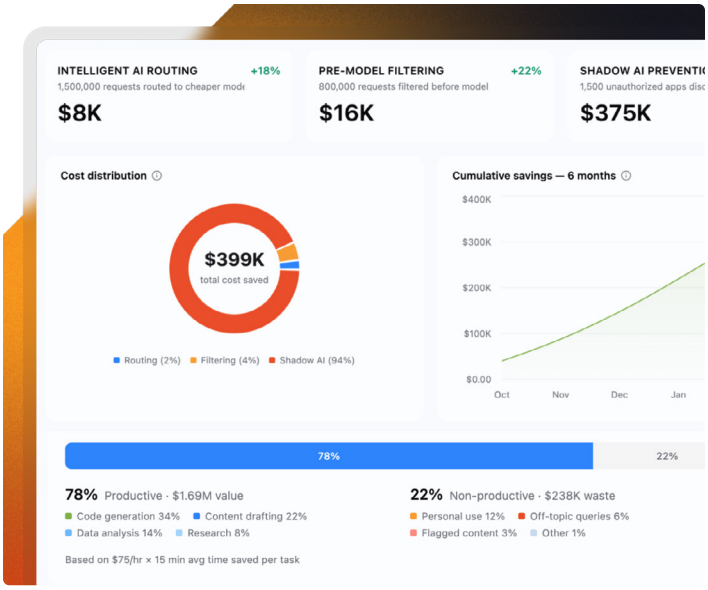
Adoption by team and intent: See who uses AI, for what, by department: active users, usage intensity, and growth over time.

Productive vs. wasted use: Classification separates real work from personal and off-topic, with evidence.

Prove governance coverage: Show what share of AI usage is governed, connecting outcomes to controls.

Capabilities across all three lenses

- **Unified AI ROI dashboard:** One view that rolls risk avoided, cost avoided, and adoption into a powerful dashboard across employees, models, applications, and agents, so the board, CFO, and CISO read one AI FinOps number from the same screen.
- **Configurable ROI multipliers:** Set your own dollar values for cost per security incident, per data leak, per proprietary-code leak, per prompt, and employee hourly rate, so every figure reflects your business, not a vendor default.
- **Intelligent model routing:** Prompts are scored for risk, complexity, and purpose, then routed to the cheapest model that meets the quality bar. Independent research shows routing cuts inference cost 40-80% with no measurable quality loss on routine work.
- **Shadow AI discovery & spend:** A continuously updated catalog surfaces every unsanctioned app, agent, and MCP server on the network, with no endpoint client required, plus an estimate of the annual spend those tools represent.
- **Proprietary-leak prevention:** Runtime guardrails track and block source-code and IP exfiltration, prompt injection, and jailbreak attempts in routed traffic, then convert each prevented event into the cost of the risk avoided.
- **Productive-use classification:** Intent classification separates employee business and personal use, and customer use from relevant or abuse, so tokens are spent only on what matters.
- **Multi-provider metering:** Sitting in the traffic path across every provider, the platform meters human and agent token spend and attributes it to teams and purposes. It gives you centralized spend visibility, and control, across all your AI vendors.



Intelligent AI routing: the built-in cost engine

The right model for every conversation.

Frontier models cost far more than lightweight ones for the same task. A routine ‘document this code’ prompt headed to a premium model gets inspected and rerouted to a cheaper one that clears the bar, for employees and agents.

How routing controls spend:

- **Score every AI request:** Risk, complexity, and purpose are scored for each prompt, human or agent, in real time.
- **Pick the right model:** Route to a cheaper model when it meets the quality bar; keep premium models for work that needs them.
- **Filter before inference:** Block irrelevant, non-work, and B2C chatbot-abuse prompts before they burn a token of paid inference.

About WitnessAI

WitnessAI is the confidence layer for enterprise AI. Our AI-native platform provides network-level visibility and intent-based controls to govern your entire workforce, human employees and AI agents alike. We deliver runtime security for models, applications, and agents, deployable via network integration or lightweight APIs. Our single-tenant architecture ensures data sovereignty and compliance.

Learn more at witness.ai.