Financial Services

The Executive Guide to AI Cost Visibility and Tokenomics in Financial Services



splunk> by CISCO

The economics of AI are changing fast in **financial services**

The shift toward agentic AI is changing the underlying economics of enterprises across the financial services industry. The upside is substantial: McKinsey's 2026 research shows that agentic AI can accelerate source-of-wealth assessments in banking by 30% to 50%. But realizing those efficiency gains at scale requires visibility into the cost of delivering the intelligence itself.

In the AI era, the token is the operating currency of intelligence. Much like the transition to cloud computing, AI has introduced variable spend that can rapidly outpace traditional budgets if left unchecked. The organizations that will maintain a competitive edge are those that treat AI tokenomics as the new frontier of financial operations (FinOps).

The stakes are higher than they appear. Gartner indicates that by 2027, 40% of agentic AI projects will be canceled due to escalating costs, unclear business value, or inadequate risk controls. The remaining 60% of these projects will be those that successfully match their architectural choices to actual business needs — and these projects will define industry leaders. Financial services leaders who can connect AI consumption to infrastructure costs, performance, and business outcomes will be better positioned to scale AI efficiently and demonstrate its value.

To achieve this, you'll have to overcome the black box of AI spending. This means understanding and improving the hidden drivers of AI value — adoption, productivity, and business impact — rather than optimizing only for visible spending. Relying solely on volume metrics hides the true cost of AI initiatives, creating a dangerous gap between innovation and profitability.

As an executive, your goal is not to slow down AI innovation, but to provide it with the guardrails of transparency. Transparency comes from evaluating and tracing what agents actually do; control comes from the runtime limits that stop spend before it compounds. Mastering the intersection of token costs, output quality, and infrastructure utilization ensures that AI strategy remains a high-ROI asset rather than a budget risk. The question for leadership teams is no longer "can we build this," but rather "is what we're building worth what it costs to run?"

Read on to learn how to achieve greater cost visibility, validate your AI investments, and guide your financial services organizations through the AI token economy.



The **new economics** of agentic AI

The era of subsidized “developer-preview” AI is over. As financial services organizations transition from static chatbots to agentic workflows, AI is no longer just a productivity tool — it is a massive, unpredictable accelerant of variable operating expenses. This shift necessitates tokenomics, defined as the emerging discipline of measuring, attributing, and governing the variable costs associated with AI. In *Machine Data Outlook: Operationalizing AI in Financial Services*, 61% of respondents reported fully deploying or rolling out custom autonomous AI agents, while 64% have done the same with vendor-provided AI tools.

While per-token costs have dropped 90% since 2023, enterprise AI spend has surged by 320% (Market Analysis, Q2 2026). This divergence also signals rapid growth requiring heavy back-end investment as the shift from experimental to high-volume agentic usage accelerates. With token consumption projected to grow 24x by 2030, the “cheap” era is ending. Some companies have terminated flat-rate plans in favor of market-rate, metered API usage — shifting high-volume fleet workloads to costs exceeding $1,500 per engineer per month. For financial institutions, that spending volatility compounds as AI reaches into transaction processing, fraud detection, compliance, and customer operations — areas where cost, latency, reliability, and risk are tightly linked.

To mitigate these API costs, many industries are moving to open-weight models running on internal hardware. However, this means invoice surprises give way to infrastructure complexity. With high-bandwidth memory sold out through 2026, firms are hitting physical supply constraints. Whether you buy tokens or host the model yourself, you remain responsible for measuring the cost without losing sight of the data, context, and controls required to support regulated workloads.

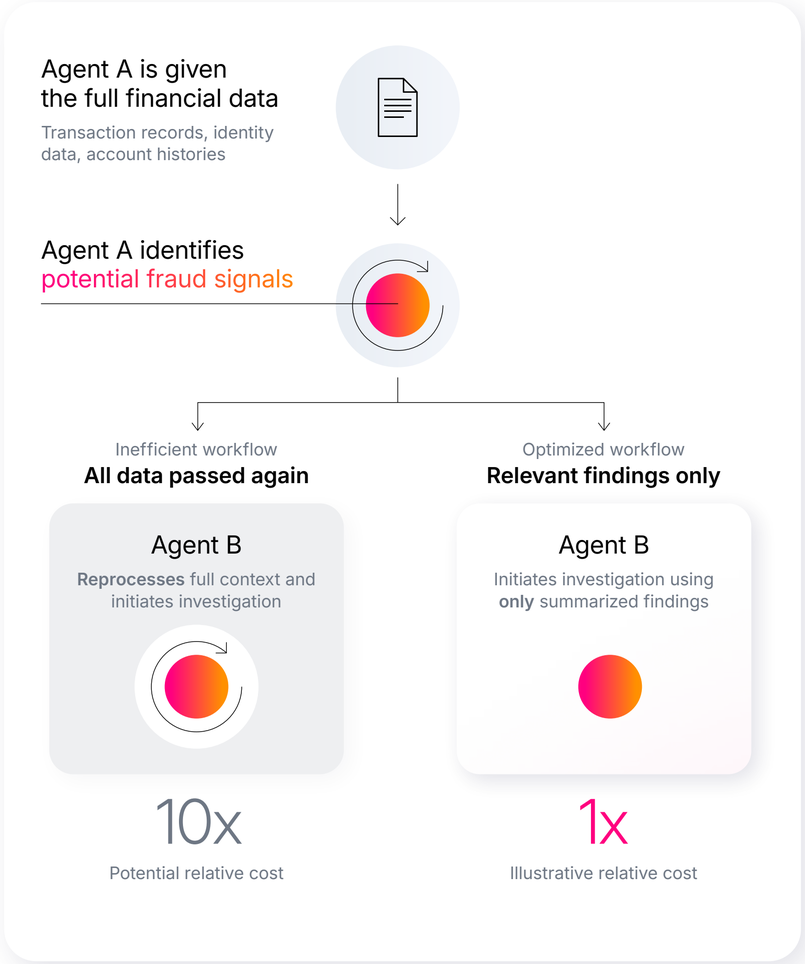
This new reality has forced a shift in the corporate discipline of AI. It is why the Linux Foundation launched the Tokenomics Foundation in June 2026 — to provide the frameworks necessary to treat AI intelligence as a measurable, accountable line item. If you cannot trace a token to a specific business outcome, your AI strategy is currently unmanaged. Moving forward, the goal is not to stop the intelligence but to quantify the cost per outcome before the bill arrives.

Why AI spend is so **hard to see**

For the modern executive, AI visibility functions more like a significant data blind spot. You receive a monthly invoice for raw token consumption, but this metric is a misleading yardstick that hides the actual cost of your intelligence. To understand where your budget is going, look below the application layer at the compute, memory, and orchestration spend it takes to run financial workflows.

Single-model implementations often operate in silos. When performance wavers, teams can struggle to pinpoint why, leaving your ROI targets vulnerable to unexplained slippage. Additionally, multi-agent systems introduce far greater complexity and numerous new failure points. Because agents can take different routes to reach the same goal — choosing different tools or sequences — traditional testing fails, making it difficult to assess whether an agent is truly efficient. Visibility into every decision point is needed, yet most organizations lack the telemetry to see how actions like specific tool calls and context choices influence the final result.

As such, the “share-everything” context is one of the most common budget leaks. Consider the agentic workflow in the following chart.



To master AI spend, teams should shift from tracking raw tokens to calculating **effective cost or cost per outcome.** This is the true economic cost of intelligence, derived by looking below the application layer, where the hidden drivers of spend reside. A holistic effective cost model accounts for the entire stack:

- **Orchestration overhead:** The compute spent on handoffs, verification, and context management between agents.
- **Infrastructure utilization:** The real-time cost of graphics processing units (GPUs), memory, and vector databases, which remain invisible in traditional API billing — especially for open-weight models that have no external invoice to reconcile.
- **Caching efficiency:** The counter-intuitive realization that larger, well-cached prompts often result in lower total costs than shorter, uncached ones.
- **Model selection and quality trade-offs:** The strategic choice of model size versus task performance, ensuring you aren’t paying for over-indexed intelligence on simple tasks or low-quality outputs that require expensive, iterative retries.

Until you reconcile these variables, the wrong metric is driving the budget. Your goal is to unearth the hidden cost underneath the invoice and to confirm that spend is funding high-performance intelligence rather than paying for redundant compute cycles.

The pillars of **AI cost visibility**

Remember: Effective management requires measurement, and optimization depends on visibility. Let's explore the three pillars of AI cost visibility — measure, observe, and optimize — to move from passive budget uncertainty to active financial control across increasingly AI-enabled financial services operations.

Measure: Know what your AI actually costs

To regain control, move from reactive invoice reconciliation to a FinOps-style attribution model. Your first step is to meter spend at the team, application, and workflow level rather than viewing it as a monolithic expense. For a financial institution, that means understanding what AI costs not only by platform or business unit, but by outcomes such as fraud investigation, compliance review, payment processing, or customer servicing.

This step requires redefining your primary metric from raw token counts — which ignore the underlying compute — to "effective cost." An effective cost model aggregates the full stack: GPU utilization, memory footprint, and vector database overhead alongside your inference tokens. By integrating these infrastructure lines into a single view, you can implement internal chargebacks that hold specific teams accountable for the intelligence they deploy. You can also identify waste before it appears on your monthly bill.

Measuring this effectively also requires accounting for the "coordination tax" inherent in multi-agent systems. In complex financial workflows, multiple agents may be coordinating across transaction data, identity signals, compliance rules, and operational systems at once. And each new agent adds communication channels and cascading memory requirements.

When coordination overhead exceeds the value of specialization — as seen when systems default to a "share-everything" context approach — you face a 10x cost premium. Instrument your systems to track these overhead drivers, choosing between selective context sharing, summarization, or full-context passing. If the cost of specific handoffs, verifications, and cross-agent communication cannot be traced, the resulting data does not measure AI performance. It merely measures a tendency toward technical debt.



Observe: See spend and quality in real time

To move beyond reactive budgeting, you should integrate full-stack correlation into a single executive view that maps requests, latency, tokens, and quality signals side-by-side. The goal is to see the true cost of intelligence alongside quality indicators, like hallucination and bias, all powered by a monitoring suite that tracks action completion, tool selection quality, and API failures. Because call_model (LLM real-time processing) is the primary driver of both cost and latency, the dashboard must identify system-level patterns — such as redundant API calls that can be batched or agents stuck in repetitive loops due to coordination failures — before the bill arrives.

Monitoring costs is a disciplined operational routine, not a one-time setup. Your team should review daily performance spikes to catch immediate malfunctions and weekly token and redundancy patterns to optimize batching. Monthly reviews can track ROI and align agent output with KPIs, such as payment reliability, compliance throughput, and customer impact. Reinforce this process with a tiered alerting system. When an alert triggers, teams verify the root cause using trace logs, apply a fix, and "close the loop" using observability metrics to confirm the system has returned to optimal performance.

Beyond mere detection, this pillar introduces the proactive defense of preventive guardrails and anomaly management. Unlike systems that only offer after-the-fact alerts, these guardrails act as real-time circuit breakers, stopping runaway agents before they inflate your infrastructure costs. By correlating your data cost with these guardrails, you gain the ability to enforce financial discipline at the request level, ensuring your AI initiatives scale with efficiency and control rather than just volume.

Optimize: Spend less for the same result

True optimization requires an "evidence, not guesswork" framework that maps technical strategy to specific business constraints. Before committing to the complexity of a multi-agent system, your leadership teams should apply a rigorous validation sequence: can better prompt engineering solve the problem? Are subtasks truly independent? And is your latency tolerance measured in seconds rather than milliseconds?

If the answer is yes, you can achieve substantial cost reduction by routing workloads to the cheapest model capable of the job. By leveraging empirical performance data, you can replace expensive frontier models with smaller, high-accuracy models wherever the quality delta is negligible.

Managing AI costs is a matching problem: route each workload to the most cost-effective environment that still meets your standards. Light, routine requests can reuse prior answers instead of paying to reprocess them. Heavy, complex work should be broken into contained pieces so costs stay predictable. And where privacy rules or infrastructure limits apply, that same routing logic lets you shift demanding workloads from metered public services onto infrastructure you control — without losing performance. For financial services organizations handling sensitive customer, transaction, and risk data across hybrid environments, that flexibility is as much a governance decision as a cost one.

Optimization is a continuous loop, not a one-time build. Instead of all-or-nothing automation, let the system handle what it is confident about and escalate the rest to people. In higher-risk financial workflows, that threshold should reflect policy, transaction limits, and regulatory expectations — not confidence scores alone. Reusing prior work, running retrieval steps in parallel, and matching each task to the right-cost infrastructure turns optimization from manual troubleshooting into a predictable driver of ROI. And when latency stays high, that is the signal to change models. The goal is simple: every AI dollar buys reliable, controlled intelligence — not overhead.

The governance layer:
Run AI like every other line of spend

Governance is the bridge between technical optimization and predictable business ROI. To run AI like any other line of variable spend, you should transition from ad-hoc experimentation to a formal FinOps discipline. This starts with clear accountability: a joint coalition of finance, infrastructure, engineering, security, risk, and compliance leadership should govern AI spend. Establish clear attribution policies that map costs directly to specific teams, applications, and workflows. Then align them to business risk and regulatory obligations. You move AI from an invisible utility into a transparent, high-ROI business asset.

Control comes from pairing spending limits with behavioral limits. Budget caps stop runaway costs, but they don't stop an agent doing something it shouldn't. Architectural guardrails address that risk by defining what data an agent can touch, what actions it can take, how far it can run unsupervised, and when a decision must go to a person. In financial services, those boundaries are your governance policy made enforceable: if an agent behaves unexpectedly, it is stopped programmatically before it breaches a cost, privacy, compliance, or risk-control threshold.

By integrating these guardrails into your continuous-improvement rhythm, you create an environment where open banking and payments or cybersecurity and financial teams can innovate within defined constraints. Your monitoring routine generates the data, your governance policy defines the limits, and your guardrails ensure that those limits are never breached. Just as importantly, the resulting audit trail helps teams understand and validate how AI reached a decision or took an action — a critical requirement in regulated financial workflows. This is how you transform AI from a budget-breaking uncertainty into a scalable, predictable engine for your financial institution.

95%

of financial services CISOs are responsible for AI governance and risk management

Source: Splunk, *The CISO Action Guide for Financial Services*, 2026

How Splunk **helps**

Splunk by Cisco helps organizations evaluate, observe, and control agent behavior and token costs through Splunk Agent Observability, an AI command center for agentic workflows. The solutions help solve the “cost of monitoring” paradox by powering our evaluations with Luna evaluation models, proprietary small language models (SLMs) that operate at a 98% lower cost than frontier models. This efficiency allows teams to evaluate 100% of agentic traffic — eliminating the blind spots created by the 10% sampling rates typical of general-purpose tools. With this granular fidelity, the question of whether models are truly “good enough” can be answered with hard numbers, not estimates.

Beyond visibility, Splunk Agent Observability offers built-in, eval-powered guardrails that actively block hallucinations, personally identifiable information leakage, and tool misuse in real time. With deep-stack correlation, affordable 100% traffic coverage, and automated runtime control, workloads can be routed to the most cost-effective models with confidence. The transition goes beyond just observing AI to actively governing the cost, quality, and security of every intelligent workflow across the enterprise.



The financial services executive **action plan**

You now have the framework to move beyond the experimental phase of AI and into the era of operational maturity. The goal is to provide innovation with the fiscal guardrails necessary to transform AI from a variable budget risk into a high-ROI business asset. By establishing FinOps-style attribution, prioritizing effective cost over raw volume, and instrumenting the entire stack for real-time observability, teams can take the wheel of enterprise AI strategy.

Next steps:

Audit your AI blind spots.

Identify your most critical AI workflows across areas such as fraud, compliance, payments, and customer operations, and determine if you can currently attribute their spend to a specific team, app, or business outcome.

Establish the baseline.

Audit current inference spending and identify where "share-everything" context or model-routing inefficiencies are leaking budget.

Implement preventive guardrails.

Move from after-the-fact alerting to proactive governance by setting hard budget caps and automated runtime controls on your most active agent fleets. Set clear boundaries for data access, tool use, and human escalation in higher-risk workflows.

Standardize your context strategy.

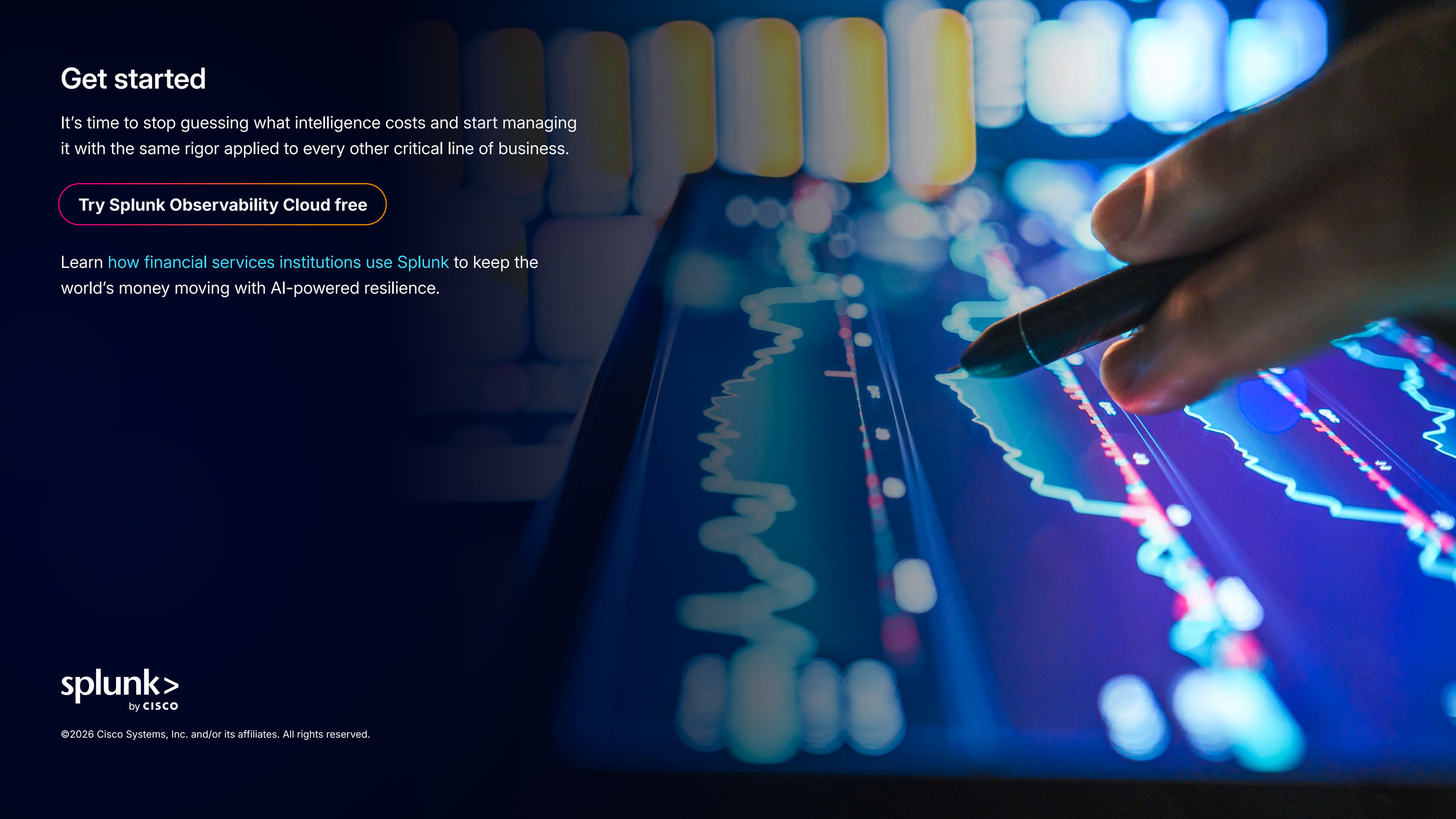
Review your agentic workflows against The 2026 Caching Playbook for Agents to see where simple architectural shifts — like prompt caching or model routing — can yield immediate cost reductions without sacrificing the context, privacy, or control required for financial services use cases.

Get started

It’s time to stop guessing what intelligence costs and start managing it with the same rigor applied to every other critical line of business.

Try Splunk Observability Cloud free

Learn how financial services institutions use Splunk to keep the world’s money moving with AI-powered resilience.



splunk> by cisco

©2026 Cisco Systems, Inc. and/or its affiliates. All rights reserved.