Tokenomics: The new economic discipline for banking AI

Banks are managing AI with the wrong vocabulary. Tokenomics closes the gap between record investment and realized value.

Most banking enterprises now have an AI strategy. The gap between spend and realized value isn't the tech. It's the absence of an economic discipline for intelligence. That discipline has a name: tokenomics.

What every bank must know about AI economics right now

- AI spending in banking exceeded $73 billion in 2025, yet only four of the top 50 banks reported realized ROI from use cases.
- Banks still manage AI through vocabulary built for FTEs and licenses, hiding what tokens realistically cost or produce.
- Tokenomics is the discipline of designing, optimizing, and governing the economics of intelligence inside a financial institution.
- It introduces five primitives that translate AI consumption into a governable, finance-grade discipline for boards and operators.
- Institutions that build this discipline early will define the cost, speed, and safety frontier of banking for the next decade.

Is there an ROI paradox at the heart of banking AI?

Investment is at record highs and adoption is near universal. On every meaningful measure of value realization, however, the industry remains in early innings. McKinsey's Global Banking Annual Review 2025 estimates AI adoption could trim banking industry costs by up to 20 percent. Generative AI alone, by McKinsey Global Institute estimates, could create $200 billion to $340 billion in annual value in banking, equivalent to 9 to 15 percent of operating profits. The execution picture is more sobering. Industry analysis indicates that 95 percent of generative AI implementations in financial services remain in pilot rather than scaled production. Only four of the 50 largest banks reported realized ROI from AI use cases in 2025. 70 percent of financial services organizations are deploying or exploring agentic AI, but only 14 percent have achieved full-scale implementation. The conventional answer points to data quality, talent gaps, regulatory uncertainty, and integration complexity. All of those are real. None is the deepest cause.

> **The deepest cause is structural. Banks are attempting to manage the economics of intelligence using a financial and operational vocabulary that was built for a different era.**

That vocabulary, built around FTEs, projects, software licenses, and transaction volumes, has no clean place to put what AI actually is, what it costs, or what it produces.

Three forces making tokenomics a board-level consideration

A token, in the language of large language models and agentic systems, is the atomic unit of work performed by an AI system. Every prompt, every retrieval, every model call, every agent action, every drafted credit memo, every triaged complaint is metered, priced, and consumed in tokens.

Tokens are to the agentic bank what kilowatt-hours are to a manufacturing economy and what API calls were to the early cloud. They have pricing, demand elasticity, utilization patterns, capacity constraints, and value attribution that must be managed deliberately, or they will manage the institution. Three forces make this a board-level concern, not a future one.

Force one: The scale of spend has crossed a threshold

The AI in banking market is forecast to reach $45.6 billion in 2026 and $143.6 billion by 2030, at a compound annual growth rate exceeding 30 percent. JPMorgan Chase reportedly operates hundreds of AI models enterprise-wide, with approximately 150,000 employees using large language models every week. Bank of America has invested several hundred million dollars across 20 AI projects. Goldman Sachs has identified six core operating processes targeted for AI-driven reshaping. This is no longer an IT line item but one of the largest discretionary spend categories on the bank's P&L, and the trajectory is steepening.

Force two: Consumption volatility breaks traditional forecasting

AI spending does not behave like any spend category most CFOs have managed before. IDC's FutureScape 2026 warns that by 2027, large enterprises will face up to a 30 percent rise in underestimated AI infrastructure costs. The cause is not excess spending but systemic under-forecasting and the opacity of consumption models. A single agentic workload deployed at scale can generate thousands of dollars of token cost in minutes. Multiplied across the dozens of agents a bank may deploy in origination, servicing, surveillance, and middle-office processes, the financial control problem becomes structural.

Force three: The EU AI Act is closing the compliance window

The EU AI Act's high-risk provisions apply to financial services from August 2026. Credit scoring, automated lending, and AML risk profiling systems must meet strict requirements around transparency, human oversight, auditability, and bias detection. The Act has extraterritorial reach. Any institution serving the EU market is in scope regardless of headquarters location. Non-compliance penalties reach up to 7 percent of global annual turnover. Over half of organizations still lack a systematic inventory of the AI systems they operate, which is the fundamental prerequisite for compliance. Getting tokenomics wrong is a regulatory liability.

Why banking is structurally different from every other sector

Tokenomics matters for every enterprise. It matters more, and matters differently, for banks. Five reasons explain the gap.

i. **Banks carry asymmetric downside exposure:** A consumer goods company whose AI mis-categorizes a marketing audience loses some advertising spend. A bank whose AI misprices a credit decision, mishandles a complaint, fails an AML escalation, or generates an unsupervised customer communication faces regulatory action, customer detriment, headline risk, and direct legal liability.

ii. **Banks operate under model risk regimes that predate LLMs:** SR 11-7 in the United States and similar frameworks in the UK, EU, Singapore, and India were designed for credit scorecards, market risk models, and CCAR-style stress models. Generative and agentic systems do not fit cleanly into this regime, and model risk teams know it.

iii. **Banks have a quarter-century of unit-economics muscle in the wrong shape:** Banks already think in cost per acquisition, cost to serve, and cost per transaction. The reflex is to slot AI spend into these existing units, which produces a comforting but misleading picture.

iv. **Banks are platform-economy laggards in their internal operations:** Where digital-native firms have built genuine internal platforms, most banks have built portfolios of point solutions wrapped in governance. The cost-curve advantage of a real agentic platform is not yet visible inside most banks because the platform itself does not yet exist.

v. **Banks sit on regulatory-grade data assets that most enterprises do not:** Customer records, transaction histories, KYC and AML data, market data, internal policies, complaint logs, and audit trails are a remarkable source of agentic value. The economics of using these assets through token-consuming systems is a first-class part of tokenomics that pure FinOps frameworks do not fully address.

The cost-obsession trap that banks must phase out

A recognizable pattern is playing out in banks that have started to pay attention to AI economics. A capability is piloted. A wealth advisor copilot, a developer coding assistant, an underwriter document summarization tool. The pilot succeeds in narrow terms. Someone in finance asks for the unit economics. The answer comes back: each user is consuming 'x' tokens per day at 'y' cost. A spreadsheet is built. The spreadsheet shows a meaningful run rate. Procurement is asked to negotiate it down. The transformation team is asked to cap it. The conversation with the model vendor becomes a price negotiation. The economics dialogue is reduced to a single question: how does the bank make this cheaper?

> **The cost-obsession trap fails because it asks the wrong primary question. The right question is not "what does a token cost?" The right question is "what does a token produce?"**

A token spent generating throwaway boilerplate is not the same as a token spent triaging a fraud alert that prevents a six-figure loss. A token spent on a vague summarization is not the same as a token spent inside an agent that closes a complaint and avoids a regulatory escalation. Tokens are fungible in their procurement. They are not fungible in their business consequence. This is the conceptual pivot at the heart of tokenomics, and it is one that bank CFOs and CIOs need to make together, deliberately, with shared frameworks and shared metrics.

The five primitives of tokenomics for banks

Tokenomics rests on five economic primitives. Each has a definition, a measurement approach, and a set of design choices that flow from it.

i. **Unit economics:** The token is the atomic unit of AI work. Banks must be able to answer four questions for any workload at any time: how many tokens are consumed per unit of business work, at what blended cost, at what utilization, and attributable to which business unit and owner.

ii. **Outcome density:** The business value produced per unit of token consumption, measured in the currency of the workload's intended outcome. Loss avoidance per thousand tokens for fraud triage. Customer satisfaction per thousand tokens for complaint handling. Outcome density is the most important number in bank tokenomics.

iii. **Workload classification:** Not all tokens are created equal. A four-quadrant model organized by outcome consequence and consumption volume allows proportionate governance and economics for each workload type.

iv. **The platform cost curve:** The marginal cost of the next AI workload depends on whether the bank operates in project mode or platform mode. Banks that fund AI as a portfolio of discrete projects will pay for the same plumbing dozens of times. Banks that invest in a coherent platform will see the cost of the tenth workload drop dramatically below the first.

v. **The agent as cost center:** Agentic AI breaks the attribution model of human-operated tools. Each meaningful agent must be treated as a named cost center, with a business owner, a technical owner, a risk owner, an assigned budget, published outcome targets, and a documented sunset trigger.

Together these primitives form the language a bank needs to budget for, contract for, and report on AI consumption with the same rigor it brings to capital, liquidity, and operational risk.

> **Is tokenomics just repackaged FinOps with a new name?**
> **The principles overlap, but banking adds asymmetric downside, model risk, and regulatory liability that generic FinOps frameworks were never engineered to govern.**

How banking leaders should activate tokenomics now

- Reframe the AI conversation at the board level. Move it from “how much are we spending on AI?” to “what is our outcome density per million tokens?” within the next reporting cycle.
- Name a single accountable executive for AI economics. If no single name exists, tokenomics is aspiration, not discipline.
- Stop negotiating per-token prices as the lead strategy. Build the inventory, instrumentation, and outcome attribution that make price negotiations evidence-based rather than reflexive.
- Fund the platform, not the project portfolio. The cost-curve advantage compounds, and it cannot be recovered later.

A series for the agentic banking era

This article opens a seven-part series on tokenomics for banks. The articles that follow explore unit economics for the bank CFO, outcome density as the metric that matters, the platform versus project decision, the 90-day implementation sprint, the governance and ownership question, and the workforce and culture shift the discipline demands.

SEO Metadata

Keywords: tokenomics for banks, AI economics in banking, generative AI ROI banking, FinOps for AI

Meta Title: Why banks need a new economic discipline for AI

Meta Description: Tokenomics gives banks the language to govern AI spend, measure outcomes, and bend the cost curve before competitors do.

About Brillio

Brillio is The Enterprise AI Accelerator helping Fortune 1000 companies move from AI ambition to scaled impact, faster. Powered by our AI accelerator platform – Agentic Data and Application Management (ADAM), Brillio is one of the fastest-growing digital technology service providers, delivering transformation across five core workstreams: business-led transformation, customer experience transformation, AI and data engineering, digital engineering, and infrastructure engineering.

With 14 delivery locations across North America, Europe, and Asia and a team of over 6,000 customer-obsessed professionals, Brillio combines deep industry expertise, modern engineering, and accelerators to deliver measurable outcomes. Headquartered in Dallas, Texas, Brillio serves clients globally with a commitment to speed, scale, and measurable impact.

https://www.brillio.com/
Contact Us: info@brillio.com

brillio