What bank CFOs must know about AI unit economics

Annual budgets and point estimates break under token economics. Finance needs a new posture before the next reforecast cycle.

Bank CFOs have managed cloud, captives, and offshoring with the same forecasting muscle. Tokens defy that muscle. Institutions that recognize the shift early will avoid the structural budget shocks others have been quietly absorbing.

Why unit economics is the place to start for CFOs

- Most banks struggle to answer four basic questions about token consumption for most of their AI workloads in production today.
- Token pricing varies by model, provider, context window, latency tier, reservation, hosting mode, and region across every workload.
- Pricing is moving downward at the per-token level but offset by larger context windows and reasoning modes that consume more.
- Annual budgets with point estimates fail within a quarter. Tokenomics requires ranges, elasticity assumptions, and quarterly review cycles.

Unit economics is the foundation of every AI investment decision

- As stated in the first article of the 'Tokenomics' series, Tokenomics: The new economic discipline for banking AI, tokenomics rests on five economic primitives.
- Unit economics is where the discipline begins because the input side is the foundation that every other primitive sits on.
- Outcome density tells the bank what the output is worth. Workload classification tells the bank how much governance to apply. The platform cost curve tells the bank how to bend the unit cost down over time. The agent-as-cost-center model gives the accountability structure that holds the rest together.
- None of those primitives function without unit economics underneath. A bank that cannot say what a workload costs cannot say what it returns, what governance is proportionate, or whether the platform investment is bending the curve.
- Banking's asymmetric downside, its established model risk regimes, and its regulatory-grade data assets, all described in the first article of this series, only sharpen why unit economics matters more here than in any other sector.

What unit economics needs: Four questions to ask

Every meaningful AI workload in a bank reduces, at the level of consumption, to tokens. There are input tokens, output tokens, tokens consumed by foundation models, tokens consumed by smaller specialist models, tokens consumed by embedding and retrieval systems, and tokens consumed by orchestration layers. The discipline of unit economics requires the bank to answer four questions for any workload, at any time.

1. **How many tokens does the workload consume per unit of business work?** Per customer interaction, per loan application processed, per complaint resolved, per code commit, per credit memo drafted.
2. **At what blended cost per thousand tokens?** Across the mix of models, providers, and hosting modes the workload uses.
3. **At what expected utilization?** Including idle time, retry overhead, evaluation overhead, and the consumption of safety and observability tooling.
4. **Attributable to which business unit, product, and outcome owner?** With a chain of accountability that is auditable.

Most banks today cannot answer these four questions for most of their AI workloads. The State of FinOps 2026 report identified granular monitoring of AI spend, including tokens, LLM requests, and GPU utilization, as the single most-requested tooling capability in the entire survey, with 98 percent of FinOps teams now responsible for managing AI spend.

The four questions in practice, applied to one workload

Consider a wealth advisor copilot deployed across 5,000 relationship managers.

i. The first question is straightforward on the surface and revealing underneath. Each advisor interaction consumes roughly 8,000 to 12,000 tokens depending on document context. That range, not a point estimate, is the honest answer.

ii. The second question exposes the mix. The workload uses a top-tier reasoning model for portfolio analysis, a smaller specialist model for meeting summarization, and an embedding model for retrieval. The blended cost per thousand tokens is not one number. It is a weighted average that shifts as the model mix evolves.

iii. The third question surfaces the overhead. Retry logic, evaluation traffic, and safety guardrail calls add 15 to 25 percent to the raw consumption number. A CFO who budgets against raw consumption is under-forecasting by that margin.

iv. The fourth question is where governance either holds or breaks. Is the workload's cost attributed to Wealth Management, or spread across Retail, Wealth, and Private Banking? Is the outcome owner the Chief Wealth Officer or the digital product lead? Without a clean answer, the workload is a shared cost with no accountable owner.

> **None of these questions has a hard answer. All of them have defensible ranges. That is the point.**

Why is AI budgeting harder than cloud budgeting?

Bank CFOs already have muscle memory for cloud FinOps. The reflex is to slot AI spend into the same discipline. That reflex undershoots the problem in three specific ways.

i. Cloud consumption scales with predictable inputs: users, transactions, storage volumes. AI consumption scales with model behavior, which is opaque. A single reasoning-mode call can consume 20 times the tokens of a standard call for the same user query. Forecasting against user counts alone misses this by an order of magnitude.

ii. Cloud pricing is stable enough to reserve capacity confidently against a 12-month horizon. Token pricing is moving quarterly, with new models, new context window sizes, and new reservation tiers reshaping the cost curve inside every planning cycle. Cloud-style annual reservations against a moving price will strand capacity or overpay against alternatives.

iii. Cloud costs are largely attributable through account tags. AI costs are attributable only if the workload was instrumented to carry business context end to end. A workload without that instrumentation is invisible in the chargeback layer, regardless of how mature the FinOps tooling is. AI FinOps is not cloud FinOps with tokens added. It is a new discipline that requires cloud-style rigor applied to a fundamentally more volatile category.

The token pricing realities that break annual budgets

Token prices vary by model, by provider, by context window length, by inference latency tier, by reservation commitment, by hosting mode, and by region. Pricing is changing rapidly. The per-token level is almost always moving downward, but the trend is often offset by larger context windows and more expensive reasoning modes that consume far more tokens per call.

The blended cost curve is moving. Finance teams that try to fix a per-token assumption in an annual budget will find the assumption wrong within a quarter. The implication for bank finance functions is direct. AI budgeting cannot be done annually with point estimates. It must be done quarterly with ranges, with explicit elasticity assumptions, and with formal review of the model and provider mix. That is a meaningful change in posture for most bank CFO offices.

Four habits of a credible AI function in practice

A finance function operating with credible AI unit economics demonstrates four habits.

1. **It publishes a workload cost dashboard,** not a spend report: Per-workload tokens consumed, blended cost, utilization, and trend lines. The dashboard updates daily, not monthly.

2. **It runs quarterly reforecasts with ranges:** A high case, a base case, and a low case for token consumption per major workload, with elasticity assumptions documented and challengeable.

3. **It reviews the model and provider mix on a defined cadence:** Pricing changes, new model releases, and reservation opportunities are evaluated against the workload portfolio, not in isolation.

4. **It treats the platform as the unit of investment, not the workload:** Per-workload cost is reported, but funding decisions and forecasting are anchored at the platform level, because that is where the cost curve actually moves.

How the primitives reinforce each other

Unit economics, outcome density, workload classification, the platform cost curve, and the agent as cost center are not independent. Unit economics gives the input side. Outcome density gives the output side. Classification tells the bank how much governance and what model mix is appropriate. The platform cost curve tells the bank how to bend the unit cost down over time.

The agent-as-cost-center model gives the accountability structure. Together they form the answer to a question bank executives are increasingly asked by their boards. The answer to “how do you manage AI spend” is no longer “we have a budget” or “we use FinOps tools” or “we are negotiating with vendors.” The answer is “this is managed as a coherent economic discipline, with the same rigor as capital, liquidity, and operational risk.”

> ### Isn't the smart play to wait for prices to fall further?
> **With per-token prices falling steadily, will building a unit economics discipline now over-invest in a problem the market will solve in time? The argument holds until you look at total spend. Consumption is rising faster than prices are falling, and bank AI budgets are up despite cheaper tokens.**

What unit economics means for the finance function

- Move AI budgeting from annual point estimates to quarterly ranges with documented elasticity assumptions inside one cycle.
- Publish a per-workload cost dashboard that updates daily, not a monthly spend report that reflects last quarter's reality.
- Build a model and provider review cadence into FP&A, treating the mix as a live portfolio decision rather than a procurement event.
- Anchor funding decisions at the platform level, because that is where the unit cost curve actually moves.

A series for the agentic banking era

This is the second article in a seven-part series on tokenomics for banks. The next article explores outcome density, the metric that makes unit economics worth measuring at all.

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