From cost per token to value per token in banking

Cost-down strategies miss the real economic lever. Outcome density is the metric that separates AI activity from AI value.

Most banks know what their AI workloads cost. Few know what their workloads produce. That gap is the difference between AI activity and AI value, and it is a huge blind spot in financial services today.

What is outcome density and why banks need it now

- A token spent on boilerplate is not the same as a token spent triaging a fraud alert that prevents a six-figure loss.
- Outcome density expresses business value produced per million tokens, in the currency of the workload's intended outcome.
- Three disciplines unlock the metric: outcome instrumentation, outcome attribution, and outcome valuation.
- Workload classification ensures governance and model choice match consequence and volume, rather than blanket policy.

Why cost per token is the wrong primary question

In our 2nd article of the 'Tokenomics' series, What bank CFOs must know about AI unit economics, we established that unit economics tells the bank what a workload costs. In this article, we pick up the harder half of the equation: what does a workload actually produce? The pivot matters because banks default to the wrong reflex. When finance sees an AI run rate, procurement is asked to negotiate it down, not asked what it produces. Once a bank accepts that cost per token is the wrong primary question, the real question surfaces immediately, and it needs its own metric.

The reflex is familiar. A generation ago, banks measured offshore captives by per-full-time equivalent (FTE) cost arbitrage and missed the strategic question of what those captives could become. In the early cloud era, banks demanded vendor invoices be cut by 20 percent and missed the question of what cloud-native architectures could unlock. The cost-obsession trap fails for the same reason every time. It asks the wrong primary question.

What does a token produce? The right economic question for banks

The right primary question is not "what does a token cost". The right primary question is "what does a token produce". Outcome density is defined as the business value produced per unit of token consumption, measured in the currency of the workload's intended outcome. The currency is workload-specific. The same wealth advisor copilot we costed in the previous article produces value in one of two currencies: qualified lead conversion, or client meeting preparation time saved. Outcome density measures which. That measurement is what allows a bank to decide whether the workload is worth the tokens it consumes.

For a fraud triage agent, the currency is loss avoidance per thousand tokens. For a complaint resolution agent, it is regulatory escalation rate and customer satisfaction per thousand tokens. For a developer copilot, it is shipped engineering throughput per thousand tokens, net of defect rate. Outcome density is harder to measure than input cost. That is precisely why most banks default to measuring input cost only. The result is a distorted economic picture.

Three disciplines that make AI outcome density credible

Discipline one: Instrumenting AI workloads for outcomes
Every workload must be instrumented to capture not only its token consumption but its business outcomes. At the per-call level, with attribution back to a specific customer, transaction, or process instance. Instrumentation is engineering work. It must be designed in from the start of the workload, not bolted on later. Platforms that make instrumentation a default capability rather than a per-workload custom build collapse months of retrofit work into days.

Discipline two: Attributing outcomes causally, not correlatively
The business outcome must be causally attributable to the AI workload, not merely correlated with it. That requires a control design. A/B tests for low-stakes workloads. Before-and-after analyses with defined counterfactuals for higher-stakes workloads. External validation by the bank's analytics or model risk function. Banks that skip this step report outcome density numbers that do not survive contact with the audit committee.

Discipline three: Valuing AI outcomes in P&L terms
The outcome must be expressed in financial terms that are coherent with the bank's existing P&L statement. Finance must lead here. A loss avoided of one dollar is not the same as a revenue dollar earned. A minute of relationship manager time saved is not the same as a minute of operations staff time saved. A regulatory escalation avoided has a probabilistic value that depends on prior loss history and current regulatory posture.

Tokenomics requires Finance to publish a workload valuation grid that converts business outcomes to economic value with auditable assumptions. The output of these three disciplines is an outcome density score for each workload, typically expressed as economic value per million tokens consumed. That score is what allows leadership to compare workloads, allocate the next dollar, sunset workloads that are not earning their token spend, and challenge vendor proposals with rigor.

How to classify AI workloads by consequence and volume

Treating all AI workloads as a single category is the most common mistake bank finance functions make. Workloads fall into structurally different economic categories, and the right governance and economics differ across them. A four-quadrant classification, organized by outcome consequence and consumption volume, gives the bank a usable lens.

Category	Workloads	Cost driver	Governance approach
High consequence, high volume	Credit decisions, anti-money laundering (AML) triage, complaint handling, trade surveillance	Cost of being wrong; per-error cost is everything	Governance overhead often reaches 30 to 50 percent of total cost in high-consequence workloads
Low consequence, high volume	Document summarization, meeting transcripts, email drafting, generic developer assist	Raw token consumption	Model right-sizing, prompt optimization, caching, batching
High consequence, low volume	Credit memo drafting, investment committee prep, senior client communication	Quality over throughput	Higher-end reasoning models acceptable; per-call governance intensive
Low consequence, low volume	Long-tail experiments, innovation pilots, individual productivity experiments	Portfolio budget, typically 5 to 15 percent of discretionary AI spend	Graduate or sunset based on outcome density

The discipline of classification is what allows a bank to apply proportionate governance to each workload. Without it, banks either over-govern everything, which kills experimentation and slows delivery, or under-govern everything, which produces the regulatory exposure profile seen in many institutions today.

How outcome density changes AI investment decisions

The shift is conceptual before it is operational. Cost obsession asks how to make AI cheaper. Outcome density asks how to make AI more valuable per unit of consumption. The two questions produce different decisions. The cheap question retires workloads with high token spend. The valuable question retires workloads with low outcome density, even when their token spend is low. The cheap question pushes every workload to the smallest model. The valuable question pushes high-consequence workloads to the model that produces the highest outcome density, even when that model is more expensive per token. The valuable question is the right question for a bank. The cheap question is the right question for a workload that should not exist in the first place.

Isn't outcome density subjective?

The metric invites gaming. Self-reported outcomes will bend toward the reporter, and audit committees know it. That is precisely why outcome density needs the same independent validation banks already apply to credit and market risk models.

How to activate outcome density in your bank now

- Publish an outcome density score for every production workload within the next reporting cycle, even if the first version is rough.
- Sunset any workload whose outcome density cannot be measured after one full quarter of attempt.
- Classify every workload into the four-quadrant model before applying governance, not after.
- Mandate Finance ownership of the workload valuation grid, with auditable assumptions reviewed at least annually.

A series for the agentic banking era

This is the third article in a seven-part series on tokenomics for banks. The next article explores the structural lever that makes outcome density scale, the platform cost curve and the agent as cost center.

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