

Tokenomics

and the New Economics
of Agentic AI



Contents

03

Introduction

05

How Agentic AI Changes
the Inference Equation

07

Why Value Per Token Matters
More Than Cost Per Token

10

Workload Placement and the
Hybrid AI Economy

13

The Hidden Levers: Utilization,
Optimization, and Cost Intelligence

15

Governance, Guardrails,
and a Token-Smart Operating Model

18

Scaling Agentic AI With
Predictable Economics

As enterprises move from AI pilots to agentic workflows, tokens are becoming the measure of AI value. Leaders need to look beyond just token cost and consider how each workflow creates value, where it should run, how it should be governed, and how hybrid AI infrastructure can help scale it efficiently.



AI's next cost surprise may not come from building models; it may come from using them.

Many organizations have tested generative AI through pilots, productivity tools, and one-off assistants — as evidenced in Enterprise Strategy Group's [research](#), which found that among organizations that have adopted, are testing, or plan to adopt AI, 70% already have AI in production.

"AI has moved from interesting pilots to being usable in real operating environments," said Fuzz Hussain, portfolio marketer at Dell Technologies. That shift is changing the enterprise conversation from whether generative AI can produce useful outputs to how AI can be governed, scaled, and applied to real workflows with predictable economics.

The technical foundation has also matured. Models are more capable, and the surrounding AI stack is more enterprise-ready. Advances in retrieval, orchestration, and local or on-premises inference have made it more practical to run AI systems with the performance, governance, and cost profile enterprises need.

With that foundation in place, AI is moving from experimentation into production, and leaders are asking a harder question: How much business value can AI create, and what does it cost to create that value? That question increasingly comes down to tokens. Tokens are the small units of text, data, and context that AI models process each time they interpret information, reason through a task, or generate an output. As AI systems become more capable, tokens are becoming a practical measure of the intelligence those systems produce.

The more useful intelligence an AI system creates, the more value it can deliver. But that value only scales if tokens are produced, routed, governed, and consumed efficiently. That is the role of token economics, or tokenomics: understanding how tokens are generated, what value they help create, and how efficiently organizations can turn them into business outcomes.



AI has moved from interesting pilots to being usable in real operating environments.

FUZZ HUSSAIN

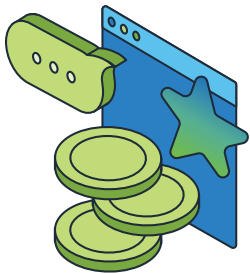
Portfolio marketer at Dell Technologies

A simple chatbot interaction may use a manageable number of tokens. A user asks a question, the model returns an answer, and the workflow ends. Agentic AI works across a longer chain of steps. An agent may need to gather context, make decisions, interact with tools, and evaluate its own progress before it can complete the task. That added autonomy is what makes agentic AI valuable, but it also means token usage can accumulate across steps the user may never see. Each additional step can add compute demand, latency, and cost.

Once agents are embedded into high-volume business workflows, token usage becomes more recurring, distributed, and difficult to forecast.

Tokens are often treated as a technical detail. But at operational scale, they become a measurable driver of AI cost, performance, and value. That is why tokenomics is becoming a business discipline. It helps leaders understand, forecast, and optimize how tokens are created and consumed across the AI estate.

IN PRODUCTION, TOKENOMICS HELPS ANSWER QUESTIONS THAT GO BEYOND USAGE ALONE



What value does this workflow create?



What does it cost to complete?



Where should it run?



Which model does it use?



How often does it retry?



How much context does it need?



Does the completed outcome justify the spend?

Cloud cost management became essential when cloud moved from experimentation to production. Tokenomics follows a similar path as AI moves from isolated use cases into always-on business workflows. A token-smart strategy requires more than tracking usage after the fact. It requires an architecture that can place workloads where they make the most sense across desktide systems, edge environments, data centers, and cloud-connected services.

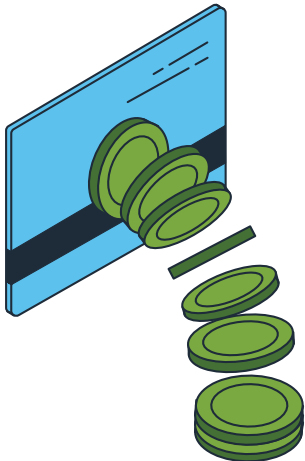
The **Dell AI Factory with NVIDIA** can help organizations make that shift. Many enterprises already have the ambition and use cases for agentic AI;



Teams often underestimate how much invisible work agents do. Tokens are spent across the full workflow, not just the final answer

TYLER COX

Distinguished Engineer,
Dell Technologies



the harder problem is scaling across disconnected systems, scattered data, and devices that were not built for persistent autonomous workflows. Dell provides a full-stack foundation for that work through its modular server architecture, Dell AI Data Platform with NVIDIA, and Deskside Agentic AI, helping organizations move from fragmented pilots to governed, measurable production environments.

For organizations scaling agentic AI, the goal is to make AI spending more predictable, measurable, and connected to business value.

How Agentic AI Changes the Inference Equation

In production, agentic workflows can make token consumption harder to predict. As Tyler Cox, distinguished engineer at Dell Technologies, explained, teams often underestimate “how much invisible work agents do.” Tokens are spent across the full workflow, not just the input and final answer — including retrieval, tool calls, retries, validation, routing, and follow-up actions.

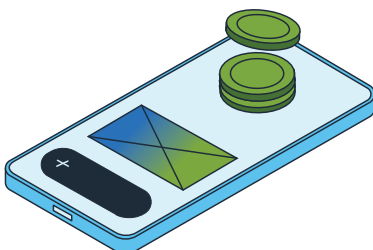
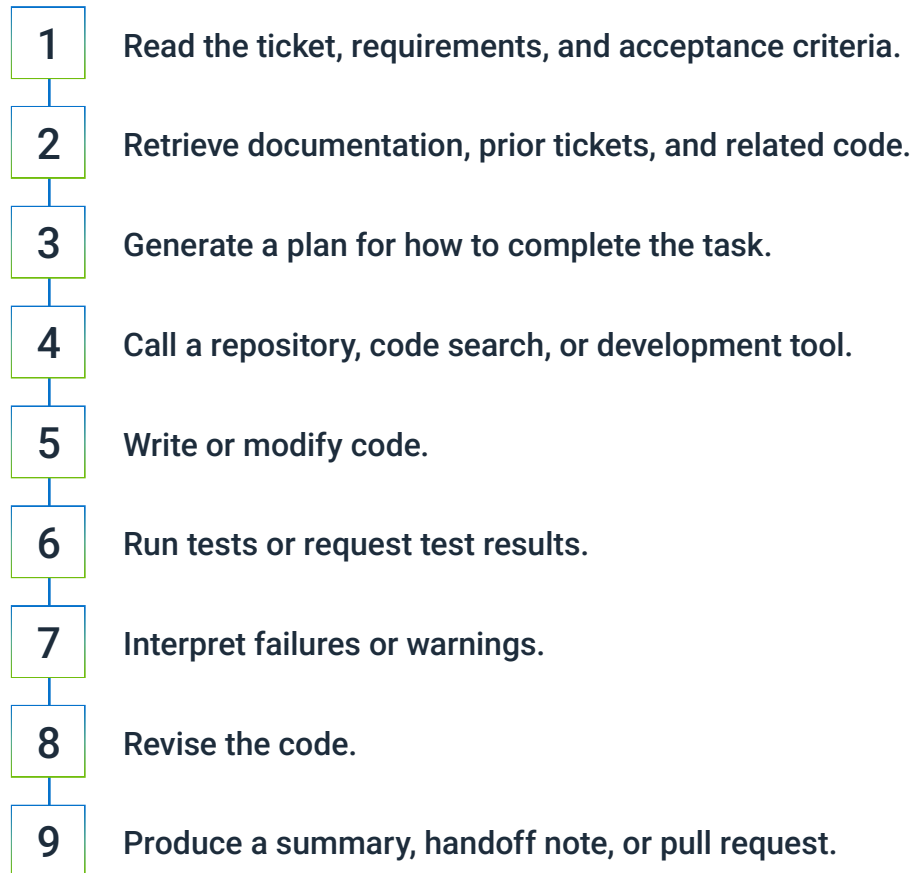
That shift matters because inference is becoming the everyday workload of AI. Training creates the model, but inference is where business value is realized every time AI is used inside a real workflow. As organizations operationalize copilots, reasoning systems, and autonomous agents, inference becomes the recurring activity that produces value, consumes tokens, and drives infrastructure demand.

Agentic AI adds another layer to that equation. The model is only one part of the system. Agents also rely on an orchestration layer — sometimes described as a harness — that helps coordinate how the agent gathers context, selects tools, accesses data, works with memory, and moves through the steps needed to complete a task. That harness is what allows agents to do more than answer a single question, but it also makes token consumption less visible than a simple prompt-and-response interaction.

As a result, token consumption becomes a workflow-level issue. Much of the cost can come from the orchestration around the full task: the reads, retrieval, routing, validation, retries, and follow-up actions that happen before the user sees the completed outcome.

A conventional chatbot may involve one prompt and one response. In contrast, a software development agent, for example, may move through a sequence like this:

HOW A SOFTWARE DEVELOPMENT AGENT MAY MOVE THROUGH A SEQUENCE



Each stage may involve more than the visible prompt and answer. Input and output tokens are only part of the load; system prompts, retrieved context, tool responses, and validation loops can also add to the total. A single completed task may therefore contain many model interactions that are not visible if one only looks at the final answer.

Inference costs are falling quickly, driven by improvements in model design, hardware, infrastructure efficiency, chip utilization, and specialized silicon. Gartner predicts that inference on a 1 trillion-parameter LLM will cost GenAI providers more than 90% less by 2030 than in 2025. But lower provider costs do not automatically mean lower enterprise AI spend when agentic workflows require more context, retrieval, tool use, validation, routing, and follow-up actions before they produce a completed outcome.

Lower-cost commodity tokens may make routine AI capabilities easier to embed into everyday software. Advanced reasoning is different. Frontier models still depend on scarce compute, while agentic workflows can consume far more tokens before they complete the work.

WHAT CHANGES WHEN AI BECOMES AGENTIC

ACROSS AGENTIC WORKFLOWS, THE COST DRIVERS USUALLY APPEAR IN FOUR PLACES:

**More context**

Agents often need larger context windows and multi-turn history to complete work.

**More tool use**

Each tool call can add a model decision, system request, response handling, and follow-up prompt.

**More retrieval**

Retrieval-augmented generation (RAG) can improve accuracy by grounding outputs in enterprise data, but it also increases the amount of context the model must process.

**More retry risk**

Failed calls, weak retrieval, incomplete data, or looping behavior can raise token consumption quickly.

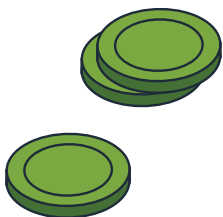
At production scale, these patterns can affect both cost and performance. Agentic AI can do more valuable work than a one-shot chatbot, but its economics need to be managed at the workflow level.

Why Value Per Token Matters More Than Cost Per Token

Price per token is a useful metric, but it only tells part of the story. A low-cost token can still lead to an expensive workflow if the agent uses too much context, takes too many steps, retries too often, or relies on a model that is more powerful than the task requires.

“The most important shift is moving from cost per token to cost per outcome,” Hussain said. Leaders should not ask only what each token costs; they should ask whether the AI system is completing meaningful work in a way that is cheaper, faster, more scalable, or more predictable than the current alternative.

That shift changes how organizations evaluate token use. A support workflow that resolves an issue on the first try, a contract review that reduces legal rework, or a coding agent that produces usable output faster may justify more token consumption than a lower-value task. In that sense, tokenomics is not about minimizing token use at all costs. It is about maximizing the value created by each workflow while controlling the cost of producing that value.





For executive teams, the priority should be cost per outcome, fit-for-purpose models and context windows, and workload placement.

TYLER COX

Distinguished Engineer,
Dell Technologies

Cost per outcome gives leaders a better way to evaluate AI economics. It asks what it costs to complete a business task, from resolving a customer issue to reviewing a contract, generating code, producing financial analysis, or handling a supply chain exception. That metric connects AI spending to the work the business cares about.

The outcomes that matter most are the ones tied to real business performance: faster resolution, higher employee productivity, improved cycle times, better service quality, and stronger governance.

Failed or low-quality outputs should also be part of the calculation because they create extra review, rework, and delay. A contract review workflow offers a simple example. A cheaper model may cost less per call, but require repeated prompts, longer human review, or more corrections before legal teams can use the output. A more capable model may cost more per call, but if it produces a reliable first pass and reduces review cycles, its cost per completed review may be lower.

That is why cost per outcome has to be measured against business performance, not token price alone. The outcomes that matter most are the ones that improve how the business operates, such as faster resolution, higher employee productivity, shorter cycle times, better service quality, and stronger governance.

WHAT BELONGS IN COST PER OUTCOME

Cost per outcome is only useful if teams measure the full path from request to completed work. A complete view should account for:

- Input and output tokens
- Model selection
- Context size
- Tool calls and retries
- Concurrency and latency requirements
- Infrastructure utilization
- Data movement
- Governance requirements and human review
- Outcome quality and rework
- Business value of the completed task

Cox said leaders often miss these workflow-level variables. For executive teams, he said the priority should be cost per outcome, fit-for-purpose models and context windows, and workload placement.

Quality data also affects value per token. Poor retrieval can force agents to process irrelevant context or repeat steps that better data would have avoided. The Dell AI Data Platform with NVIDIA helps address that problem by giving teams a governed way to prepare, index, and orchestrate enterprise data for AI workloads.

For production AI workloads, the platform can deliver up to 200% faster data streaming and 12x faster vector indexing with NVIDIA acceleration. It also helps reduce manual data preparation through automated end-to-end data orchestration, turning work that can take weeks into minutes. For teams building RAG and agentic AI workflows, that can improve retrieval quality, reduce wasted context, and make token use easier to control.

That data foundation matters because token efficiency is shaped before the model ever generates an answer. If the retrieval layer is noisy, the model may receive too much context, miss the right context, or require additional attempts to complete the task.

Model selection can change the economics of the entire workflow. Not every task needs the largest model or the longest context window. The better choice is the model, context size, retrieval pattern, and infrastructure placement that completes the task reliably at an acceptable cost.

This is why tokenomics should not be treated as a mandate to use fewer tokens in every situation. Some high-value workflows may justify more context, more capable models, or additional validation steps. The goal is to use the right amount of intelligence for the task, then produce that intelligence as efficiently as possible.

Cost per outcome also gives leaders a way to compare AI investments with business KPIs. Technical telemetry becomes easier to evaluate when it is linked to operational measures: tickets resolved, contracts reviewed, code generated, claims processed. This makes it easier to decide which workflows deserve more investment and which need redesign before they scale.

The same logic applies at the AI factory level. Enterprise Strategy Group's Dell-commissioned economic validation found that the Dell AI Factory with NVIDIA can deliver 1,225% ROI over four years and 269% ROI in Year 1. Those findings give the cost-per-outcome argument a concrete business benchmark, showing how coordinated infrastructure, data, governance, and workload placement can translate into measurable AI value.

QUESTIONS LEADERS SHOULD ASK BEFORE SCALING AN AGENTIC WORKFLOW

Before moving an agentic workflow into production, leaders should ask:

- What business outcome does this workflow complete?
- What value does that outcome create?
- Which models, tools, and data sources does it use?
- What causes retries, failed attempts, or human rework?
- How much context does the agent really need?
- What level of latency is required?
- What governance controls are needed before the agent acts?
- Where should the workload run for the right balance of cost, performance, security and governance, and control?

Workload Placement and the Hybrid AI Economy

Where tokens are generated and consumed can matter as much as how many are used. As Hussain put it, key placement criteria include “data sensitivity, latency, model and context requirements, workload predictability, and cost control.”

Those questions are becoming more flexible as AI infrastructure matures. NVIDIA NemoClaw and OpenClaw can support secure agent runtime capabilities, while confidential computing helps protect how data, models, prompts, reasoning, and actions are handled while AI is running. OpenShell can add sandboxing and policy guardrails, which matters for autonomous agents because the runtime itself becomes part of the security boundary.

As more open, proprietary, and frontier models become practical across cloud, data center, edge, and desktop environments, hybrid AI becomes less about where AI can run and more about where each workload should run. The decision comes down to the right balance of data sensitivity, latency, model access, governance, utilization, and cost.

HOW COMMON AI ENVIRONMENTS CAN MAP TO DIFFERENT WORKLOAD NEEDS

Environment	Best Fit	Why it works
 Deskside systems	Local agentic workflows, developer assistants, personal productivity agents, and sensitive-data experimentation on Dell Pro Precision with GB10 or GB300 and Dell Pro Precision Fixed Workstations	Keeps work close to users, devices, and local data while supporting local agent runtimes and secure experimentation
 Edge	Low-latency, bandwidth-sensitive, or privacy-sensitive use cases in locations such as factories, stores, field sites, and healthcare facilities	Reduces data movement and supports local processing where work happens
 Data center	Governed, repeatable, high-volume enterprise workflows such as internal knowledge assistants, code assistants, and document processing	Supports policy enforcement, performance control, access to enterprise data, Dell AI Data Platform with NVIDIA, and modular scaling with Dell PowerRack and AI
 Cloud or API services	Bursty demand, experimentation, highly variable usage, or specialized model access	Provides flexibility and access to rapidly changing model options
 Hybrid	Mixed AI portfolios with different traffic, data, latency, and governance needs	Allows each workload to run where economics and controls fit best, supported by the Dell AI Factory with NVIDIA across deskside, edge, data center, and cloud-connected environments

Agentic workloads are not always a single process running in one place. Different stages of an agent's workflow can place different demands on infrastructure. Model inference and reasoning may depend heavily on GPUs. Tool execution may depend on CPUs. Security and data movement may depend on DPUs. Memory and retrieval depend on storage. Orchestration depends on the network fabric connecting it all. Workload placement is therefore not just about choosing a location; it is about matching each workflow, and sometimes each stage of the workflow, to the right mix of resources.

AI ENVIRONMENTS



Deskside systems

Deskside systems can support agentic AI close to users and local data. In Dell's updated portfolio, that includes Dell Pro Precision with GB10 and GB300 and Dell Pro Precision Fixed Workstations. These systems give developers, technical teams, and knowledge workers a way to test local agents, keep sensitive data closer to the user, and understand what a workload needs before moving it into a broader production environment.

The deskside agentic AI stack also includes NVIDIA NemoClaw for always-on agent runtime support and OpenClaw for persistent, multi-step workflows. OpenShell adds sandboxing and policy guardrails, while CrowdStrike Falcon provides continuous protection at the OS and BIOS levels. Nemotron models can support reasoning and complex task execution for local agentic workflows.

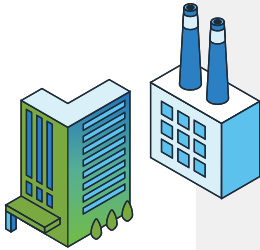


Edge

Edge environments are useful when data is generated outside the central data center. That can include sites such as factories, stores, field locations, and healthcare facilities, where local processing may be valuable. In those cases, sending data back to a central environment can slow the workflow, increase cost, or create added security and compliance review.

Edge placement can help agents act closer to where work happens. A visual inspection agent in a factory, for example, may need low-latency access to camera feeds. A field operations agent may need to process local data even when connectivity is limited. These workloads may not fit neatly into a centralized model, and in some cases, deskside agentic AI may also be a practical option when space, power, or edge-server constraints limit what can be deployed on site.

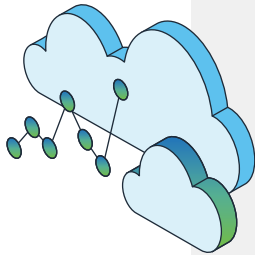
AI ENVIRONMENTS



Data center

Data centers are often a strong fit for governed, repeatable, high-volume enterprise workflows. Internal knowledge assistants, code assistants, and document processing systems are common examples, especially when predictable performance and access to enterprise data are required.

Control is a main advantage. In the data center, teams can more directly manage security, agent access to data, performance monitoring, and infrastructure use. High-volume does not always mean perfect utilization. Even so, predictable, governed workloads are easier to optimize than sporadic demand because teams can plan capacity, tune infrastructure, and improve cost per completed task over time.



Cloud and API services

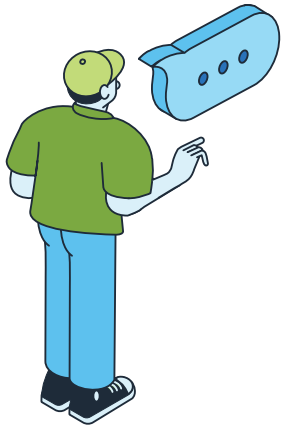
Cloud and API services remain important. They can support burst capacity, rapid experimentation, and access to specialized model services. At the same time, specialized and frontier capabilities are no longer exclusively a cloud conversation, as more model options become available for controlled or on-premises environments.

Cloud can also give organizations room to evaluate new models before they commit to a longer-term placement decision. The point is to place each workload according to its traffic pattern, data profile, governance needs, and economic model.

PLACEMENT DECISIONS SHAPE LONG-TERM CONTROL

Data movement should factor into the placement decision. Repeatedly moving large enterprise datasets into external services can add cost, slow the workflow, and increase compliance review. In some cases, keeping the model closer to the data can reduce both cost and risk.

Workload placement is also a question of long-term control. As agents mature, their prompts, retrieval patterns, workflow logic, model choices, and tool integrations become part of how the organization works.



If that operating logic lives only inside one environment or provider, the organization may have less flexibility when prices, governance requirements, or architecture plans change.

Most enterprises will split agentic workloads across environments. Some work will stay close to users or data; some will belong in the data center; and some will remain better suited to cloud or API services. Timing can also shape the hybrid approach. Batch workloads may run during lower-demand windows, while real-time workflows may need faster environments and more predictable capacity.

Through the [Dell AI Factory with NVIDIA](#), Dell Technologies and NVIDIA support a flexible AI architecture for placing agentic workloads across desktide systems, edge environments, data centers, and cloud-connected environments.

The Hidden Levers: Utilization, Optimization, and Cost Intelligence

After placement, economics depend on how efficiently each environment produces useful tokens. Two organizations can consume similar token volumes and see very different costs if one has better workload discipline, utilization, and optimization practices than the other.

This is where tokenomics moves from planning into operations. Once AI is running in production, the question is not only where the workload lives, but also how efficiently that environment can turn infrastructure, energy, data, and model capacity into useful output. Energy is an important cost driver, especially as inference demand grows, but it is only one part of the efficiency picture. Metrics such as cost per million tokens, utilization, throughput, latency, cost per outcome, and tokens per watt can help leaders understand how efficiently AI value is being produced.



A key economic determinant is whether demand is steady enough to optimize



FUZZ HUSSAIN

Portfolio marketer at Dell Technologies

UTILIZATION DETERMINES WHETHER CONTROLLED INFRASTRUCTURE PAYS OFF

Controlled or self-hosted infrastructure becomes more attractive when demand is predictable enough to keep compute resources productively used. Hussain said a key economic determinant is whether demand is “steady enough to optimize.”

Utilization still matters, but it should not be treated as the only variable. A high-volume workload does not always run at perfect utilization. Even so, predictable, governed workloads are often easier to improve over time because teams can plan capacity, tune infrastructure, adjust batching, refine model routing, and spread demand more efficiently.

Tokenomics planning should look at demand patterns, expected concurrency, opportunities to batch work, and likely growth.



If infrastructure is underused, the organization may pay for idle capacity. If it is well used, the cost of each completed task can improve. Tokenomics planning should therefore look at demand patterns, expected concurrency, opportunities to batch work, and likely growth.

OPTIMIZATION LEVERS IMPROVE VALUE PER TOKEN

Several optimization levers can reduce waste before costs rise:

-  **Model right-sizing** means using the smallest model that can reliably complete the task. Routine summarization, classification, or extraction may not need a frontier model. More complex reasoning may justify a larger model.
-  **Context management** means sending the model only the information the workflow needs. Larger context windows can help agents complete complex tasks, but unnecessary context adds token load and can slow performance. The Dell AI Data Platform with NVIDIA helps teams prepare and orchestrate enterprise data so agents can retrieve more relevant context with less waste.
-  **Retrieval quality** affects both cost and accuracy. If retrieval brings back irrelevant or duplicative material, the agent may process more tokens and produce weaker results. The Dell AI Data Platform with NVIDIA supports accelerated vector indexing, automated data orchestration, and a composable architecture built on open formats such as Iceberg, Delta Lake, and Parquet.
-  **Batching** groups non-urgent work so infrastructure can process it more efficiently. This can work well for document analysis, reporting, embedding generation, or other tasks that do not require immediate response.
-  **Concurrency tuning** helps teams manage how many users, agents, or workflows can run at the same time without creating unnecessary latency, infrastructure strain, or token waste.
-  **Quantization** can reduce model memory and compute requirements for suitable workloads. For leaders, the point is the ability to lower resource demand while preserving acceptable quality.
-  **Continuous batching and KV cache optimization** can improve throughput, concurrency, and long-context performance. These methods are especially useful when many users or agents are active at once, or when workflows depend on long-running context. For agentic AI, KV cache optimization can help reduce repeated computation and improve the economics of high-volume inference. KV cache offload can take that a step further by moving cache data from GPU memory to Dell G4 storage engines, helping reduce memory pressure and improve \$/token for long-context workloads.

These techniques should not be applied blindly. Optimization choices need to match the workflow and its risk tolerance. A high-risk legal or clinical workflow may have different tolerance for model compression or routing than a routine internal summarization task.

COST INTELLIGENCE MAKES OPTIMIZATION VISIBLE

Organizations need to attribute token usage to the applications, workflows, and teams driving it, then review workload patterns and pricing regularly. Without that visibility, they may know total AI spend but not which workflows are driving it or what needs to change.

Cost intelligence also helps teams understand whether a workflow is improving over time.

A TOKEN-SMART AI PROGRAM SHOULD BE ABLE TO ANSWER QUESTIONS SUCH AS:



Which workflows consume the most tokens?



Which data sources produce the most irrelevant retrieval?



Which models are being used for tasks that smaller models could handle?



Which agents retry most often?



Which workloads have the highest cost per completed outcome?



Which workloads should move to a different environment?

AI economics do not stay fixed once workloads move into production. Regular review helps teams adjust before outdated assumptions become budget, performance, or governance problems.

Governance, Guardrails, and a Token-Smart Operating Model

Agentic AI needs governance because agents do not simply generate text. They can make tool calls, interact with enterprise systems, and trigger downstream workflows. Tokenomics therefore becomes part of enterprise risk and operational control.

As Hussain put it, agents should be treated as “governed, observable systems — not black boxes.” That distinction matters. A chatbot that gives a weak answer may create rework. An agent with access to tools, data, and business systems can create cost, security, compliance, and operational consequences if its behavior is not visible and controlled.

Telemetry is the operational backbone of a token-smart AI environment. Continuous instrumentation of each agent step turns token usage, tool calls, latency, anomalies, and outcomes into measurable signals. Without that visibility, organizations cannot govern agent behavior or optimize token economics at scale.

MINIMUM CONTROLS BEFORE AGENTS ACT

Before agents act across enterprise systems, organizations need controls that limit unintended behavior and runaway consumption. Secure agent runtimes, policy controls, and confidential computing can help organizations govern autonomous agents across hybrid environments, especially as agents gain access to sensitive data, proprietary models, and enterprise systems.

AT A MINIMUM, TEAMS SHOULD DEFINE:

- A named business owner and clear business objective for the workflow
- A risk classification that determines where human approval is required
- Least-privilege identity and access controls that specify which agents can operate and under whose authority
- Tool and data permissions that limit which systems, APIs, and datasets agents can use
- Policy boundaries, token budgets, and usage caps that prevent runaway consumption from retries or loops
- Isolated execution where needed for sensitive workloads
- Action logging, traceability, and auditability so teams can reconstruct what an agent did and why
- Observability dashboards that track cost, latency, anomalies, retries, token burn, and outcome quality

Agentic AI can create security exposure and economic exposure quickly when agents have too much freedom to act or retry. Budget controls matter as much as access controls. Usage caps can force escalation before an agent burns through too many tokens or repeats an action too many times.

Local agentic AI can also change the economics of high-volume inference. Third-party analysis shows that running AI on the Dell AI Factory with NVIDIA can deliver up to 87% lower token spend over two years compared with public cloud APIs, with a breakeven point achievable in as little as 3 months.

Governance should not be treated as a brake on AI adoption. Done well, it gives teams the confidence to move more workflows into production because operating boundaries are already defined.

A token-smart operating model should guide agentic workflows from design through deployment and ongoing optimization. A practical model includes five steps.

THE FIVE STEPS OF A PRACTICAL TOKEN-SMART OPERATING MODEL

1

Classify the workflow

Teams should assess the workflow's business value, risk level, data sensitivity, and expected usage before deployment. A low-risk internal summarization tool does not require the same controls as an agent that can update customer records or trigger financial actions.

2

Design the workflow economics

Before production, teams should estimate the workflow's expected cost drivers and define a baseline that can be tested against real usage.

3

Set governance and budgets

Teams must define the permissions, spending limits, and escalation paths that govern agent behavior across systems. Budget controls should be linked to the workflow's business priority and risk level.

4

Deploy with observability

Teams need visibility into how agents consume tokens, use tools, affect latency, and perform against expected outcomes. Without observability, organizations cannot distinguish a productive agent from one that is looping, over-retrieving, or using the wrong model.

5

Optimize continuously

Teams should use production data to refine prompts, retrieval, model routing, batching, concurrency, infrastructure utilization, context management, and policies as usage patterns change. This becomes especially important as organizations adopt recursive or self-improving agentic workflows.



The right operating model will not look the same for every workload. Some agentic workflows may belong in API-based services or cloud infrastructure. Others may be better suited to colocation, self-hosted environments, or desktide agentic AI — especially when data sensitivity, cost predictability, space, power, or edge-server constraints shape the decision.



AI factories are becoming the production systems that turn enterprise data, models, and workflows into usable intelligence at scale

FUZZ HUSSAIN

Portfolio marketer at Dell Technologies

Dell's approach helps organizations operationalize this model by pairing infrastructure choices with services that support use case prioritization, data readiness, governance requirements, workload placement, and KPI tracking.

The Dell Automation Platform can also help reduce the operational work involved in deploying AI software. Dell cites an average reduction of one week in software deployment time, which matters when teams are trying to move agentic workflows from pilot environments into production. Modular Dell PowerRack and AI Factory foundation infrastructure gives organizations a path to scale capacity as AI demand grows, rather than rebuilding the stack for each new workload.

That repeatable approach reflects Dell's broader AI Factory experience across more than 5,000 customers. For tokenomics, the value is operational consistency: teams can apply shared patterns for workload placement, data readiness, governance, and cost tracking as agentic AI adoption expands.

This operating model also helps keep AI adoption consistent as more teams build agents, choose models, define controls, and measure success. Shared standards make token costs easier to manage and governance easier to enforce as agentic AI expands.

Scaling Agentic AI With Predictable Economics

Scaling agentic AI requires more than lower token prices. As production usage grows, leaders need to understand what each workflow costs, what value it creates, and whether the completed outcome justifies the spend.

A token-smart foundation connects AI economics to the way workflows are designed, placed, governed, and improved over time. That discipline helps organizations move from isolated pilots to production environments with greater control over cost, performance, and operational risk.

As agentic AI matures, enterprise infrastructure is becoming more than a place to run models. "AI factories are becoming the production systems that turn enterprise data, models, and workflows into usable intelligence at scale," Hussain said.

In that model, tokens are the output of the AI factory: the measurable work produced when models interpret information, reason through tasks, retrieve context, use tools, and generate useful outcomes. The objective is not simply to produce more tokens or cheaper tokens. It is to produce the right intelligence for the right workflow as efficiently, securely, and predictably as possible.

That requires more than compute. A token-smart AI factory needs an operational foundation for inference, orchestration, data access, security, observability, and lifecycle management.

Dell Technologies and NVIDIA help organizations move beyond disconnected experiments toward governed, token-smart agentic AI deployments. The Dell AI Factory with NVIDIA brings together infrastructure, software, services, and validated approaches to help teams operationalize AI across deskside, edge, data center, and cloud-connected environments. That gives organizations a more repeatable way to tune model placement, performance, governance, and scale instead of stitching together custom components for every deployment.



Learn how the **Dell AI Factory with NVIDIA** can help your organization build a token-smart foundation for scaling agentic AI.