MONDERMAN

INSTITUTIONAL PERFORMANCE RESEARCH

After the First Lap

How Token Economics Will Define the Next Phase of Enterprise AI

A structural dependency on subsidized foundation model pricing is forming now. The companies that recognize the exposure early — and procure the engineering discipline that mitigates it — will define the next decade of enterprise AI.

Institutional Performance Research • May 2026
connect@monderman.com • www.monderman.com

ABOUT THIS PAPER

This is a synthesized insight paper, not original empirical research. It draws on publicly available analyst forecasts, industry surveys, and corporate reporting from sources including IDC, Gartner, Deloitte, Citi, Andreessen Horowitz, AnalyticsWeek, NVIDIA, the FinOps Foundation, the International Energy Agency, and others, listed in full in the references at the back. Numerical figures in this paper are forecasts, estimates, or modeled projections from those sources, with their attribution stated in-line where they appear. The paper synthesizes those inputs into a structural argument about where enterprise AI economics are heading and what an emerging intermediary tier — Deterministic AI Infrastructure — looks like in practice. Examples and walked-through scenarios (such as the hospital diagnostic comparison in Part III) are illustrative rather than measured outcomes; specific token reductions and cost gaps will vary by workflow complexity and deployment context.

EXECUTIVE SUMMARY

IDC forecasts worldwide AI spending of $301 billion in 2026, with broader market definitions placing the figure higher. The investment is real, the productivity gains are real, and the strategic urgency driving the spend is justified. What is also real — and increasingly visible to the CFOs whose budgets fund it — is that the cost of running AI in production is now growing faster than the cost of buying it. According to industry reporting collected by Andreessen Horowitz, some Fortune 500 companies are now reporting monthly AI inference bills in the tens of millions of dollars. Analyst estimates of inference's share of enterprise AI compute spend now range from roughly two-thirds (Deloitte) to as high as eighty-five percent (AnalyticsWeek). Forty-two percent of enterprises name workflow optimization as their top AI spending priority for 2026 — overtaking expansion for the first time.

$301B
IDC-forecast worldwide AI spending in 2026; broader definitions (Gartner) place 2026 spending higher.

⅔ to 85%
Range of analyst estimates for inference share of enterprise AI compute spend.

31% to 63%
Rise in organizations reporting AI as an active FinOps concern, 2024 to 2025.

This paper argues that a structural reckoning is now underway. The first lap of enterprise AI was defined by three conditions running together: urgency to adopt, real and measurable productivity gains, and relative affordability. Pricing was calibrated to drive usage velocity, and that affordability is what made the urgency rational. The second lap — already beginning — will be different. Affordability is ending, urgency persists, and the productivity gains are now embedded in operations enterprises cannot easily wind down. Boards will start asking what every dollar of AI spend produces. The companies that cannot answer that question will face a forced choice between rationing AI use, raising prices, cutting headcount further, or absorbing costs they cannot sustain.

Behind this reckoning is a more specific dynamic: a structural dependency on subsidized foundation model pricing is forming inside enterprise operations. Current per-token prices are calibrated for adoption velocity, not for sustainable margin. As workflows become load-bearing for production operations, switching costs rise sharply. Foundation model labs face their own real economics — capital expenditure on data centers, energy costs, talent, and eventually investor expectations of returns. Most analysts therefore expect pricing to drift toward sustainable levels over time. This is not a critique of foundation labs. It is the normal arc of infrastructure-sector maturation. The exposure it creates for end-user enterprises, however, is real and worth planning for now rather than later.

The companies best positioned for the second wave are not the foundation model labs and not the end-user enterprises trying to build AI capability internally. They are a third category — intermediaries whose proprietary engineering does the structured cognitive work offline, then invokes large language models only for the linguistic articulation that LLMs are uniquely good at. This category needs a name. We propose one: **Deterministic AI Infrastructure**.

1.

The Numbers Are Becoming Impossible to Ignore

The first lap of enterprise AI ran on three conditions in alignment: adoption pressure that made urgency rational, productivity gains that made spending justifiable, and pricing that made the economics tolerable. The first two conditions remain. The third is changing. The numbers in this section establish the scale and direction of that change.

Adoption has crossed the tipping point

Enterprise AI is no longer experimental. Sixty-four percent of organizations report actively using AI in operations — seventy percent in North America. Eighty-six percent of enterprises increased their AI budgets in 2026. The number of organizations with at least forty percent of AI projects in production is set to double within six months. The pilots have become deployments.

Spending follows. IDC's worldwide AI spending guide forecasts $301 billion in 2026; broader market definitions place the figure higher — Gartner's 2026 worldwide AI spending forecast reaches $2.5 trillion, and Citi projects the total AI market above $4.2 trillion by 2030 with enterprise AI roughly half of that. Across all definitions, the trajectory is sharply up. According to NVIDIA's State of AI Report 2026, the hyperscalers are collectively forecast to exceed $600 billion in capital expenditure in 2026, with roughly $450 billion tied directly to AI infrastructure. Gartner projects that data center systems spending will jump 55.8 percent in 2026 — the largest single growth category in enterprise IT.

Inference has replaced training as the cost driver

Two activities consume AI compute. Training is the one-time work of building a model in the first place; inference is the recurring work of running that model every time someone uses it. Public attention has focused on training because the eye-watering costs of frontier model development make headlines. The shift in enterprise AI economics, however, is happening on the inference side, where the cost compounds with every query, every workflow, every deployment. Estimates vary on the exact share — Deloitte expects inference to account for roughly two-thirds of all AI compute in 2026, while AnalyticsWeek puts inference as high as eighty-five percent of the enterprise AI budget. The direction is the same across estimates: the cost story has moved from training to inference. By 2026, many enterprise AI cost models had crossed into inference-dominant economics — a shift some analysts have described as the Inference Flip.

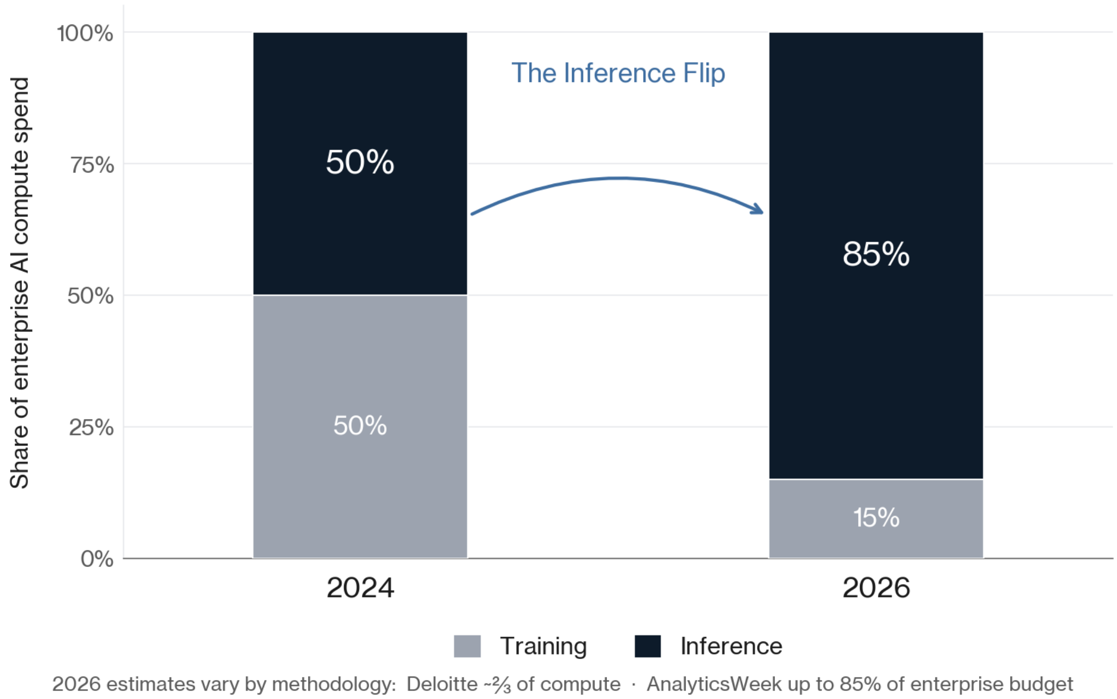

Figure 1. *The Inference Flip. Training-dominant compute economics in 2024 became inference-dominant by 2026, driven by the move from pilot deployments to production-scale agentic workflows. Sources: AnalyticsWeek 2026 Inference Economics Report; Deloitte 2026 AI Infrastructure analysis.*

The cause is structural. Foundation model pricing is denominated in tokens — the units of text the model processes, roughly equivalent to fragments of words. A simple chatbot query triggers one inference call and consumes a predictable number of tokens. An agentic AI workflow — where an autonomous agent reasons through a task, retrieves context, calls tools, verifies outputs, and self-corrects — may trigger ten to twenty inference calls per single user-initiated task. According to Gartner, agentic models require between five and thirty times more tokens per task than a standard generative AI chatbot. A workflow priced as a $0.001 chatbot interaction during the pilot phase becomes a $0.10 to $1.00 transaction in agentic production — a hundred to thousand-fold multiplier on unit cost.

Most enterprises did not model this multiplier when they committed to scaled deployment. The pilot economics, calculated on single-query API calls, bore no relationship to the production economics of multi-step agentic loops running thousands of times per day. The bills arrive after the architecture has been built.

The cost discipline conversation has started

The leading indicators of a coming spending discipline are already in the data. According to the FinOps Foundation's 2026 State of FinOps Report, organizations reporting AI as an active concern for FinOps — the discipline of governing cloud and infrastructure spending against business value — jumped from thirty-one percent in 2024 to sixty-three percent in 2025. CloudZero's 2024–2025 FinOps Adoption Survey finds that AI and ML workloads now represent twenty-two percent of total cloud costs at SaaS and IT companies. Roughly seventy percent of large enterprises maintain a dedicated FinOps or cloud economics team. Forty-two percent of enterprises name workflow optimization as their top spending priority for 2026 — overtaking expansion for the first time.

Token leaderboards and token budgets are becoming standard management tools across enterprise AI teams. The average enterprise AI budget grew from $1.2 million per year in 2024 to $7 million in 2026 — a roughly 480 percent increase in two years. IDC's FutureScape 2026 warns that organizations with dedicated FinOps resources still underestimate AI infrastructure costs by up to thirty percent.

These numbers are not signs of crisis. They are signs of maturity. Enterprises are doing what they should do when a new technology category moves from experimental to operational: they are starting to count what it costs.

The energy footprint is real

AI's compute footprint is also a literal infrastructure footprint. The International Energy Agency projects that global data center electricity demand could double between 2022 and 2026, fueled significantly by AI adoption. By 2030, global data center electricity demand is projected to reach approximately 945 terawatt-hours — slightly more than Japan's entire current electricity consumption. AI-related electricity consumption is projected to reach 4.5 percent of total US electricity demand by 2027.

Residential electricity prices in the US jumped 7.1 percent in 2025 — more than double the inflation rate, and over twenty percent in some states. AI data centers are not the only driver, but they are a substantial one. The cost of every token generated is ultimately a cost in megawatts and grid capacity, and that cost is increasingly visible to ratepayers, regulators, and CFOs alike.

AI is not free. The marginal cost of every token generated is real money, real megawatts, and real strain on grids that were not designed for this workload. Enterprises that have been operating as if AI were a free good — which most still effectively are — will be confronted with that reality through their cloud bills, their utility bills, and increasingly through public pressure on the electricity costs their AI consumption is helping to drive.

2.

Why This Reckoning Is Structurally Unavoidable

The two curves running in opposite directions

The most confusing aspect of the current moment, for enterprise finance teams, is that two opposing trends are running at the same time. Per-token inference prices have fallen sharply — by roughly nine to nine-hundred-fold per year across various capability benchmarks tracked by Epoch AI — and Gartner forecasts a further ninety percent reduction by 2030. At the same time, total enterprise AI bills are rising rapidly. Both can be true because token prices and total token consumption are not the same thing. Costs per unit are collapsing. Units consumed are exploding. The product of the two is what shows up on the bill.

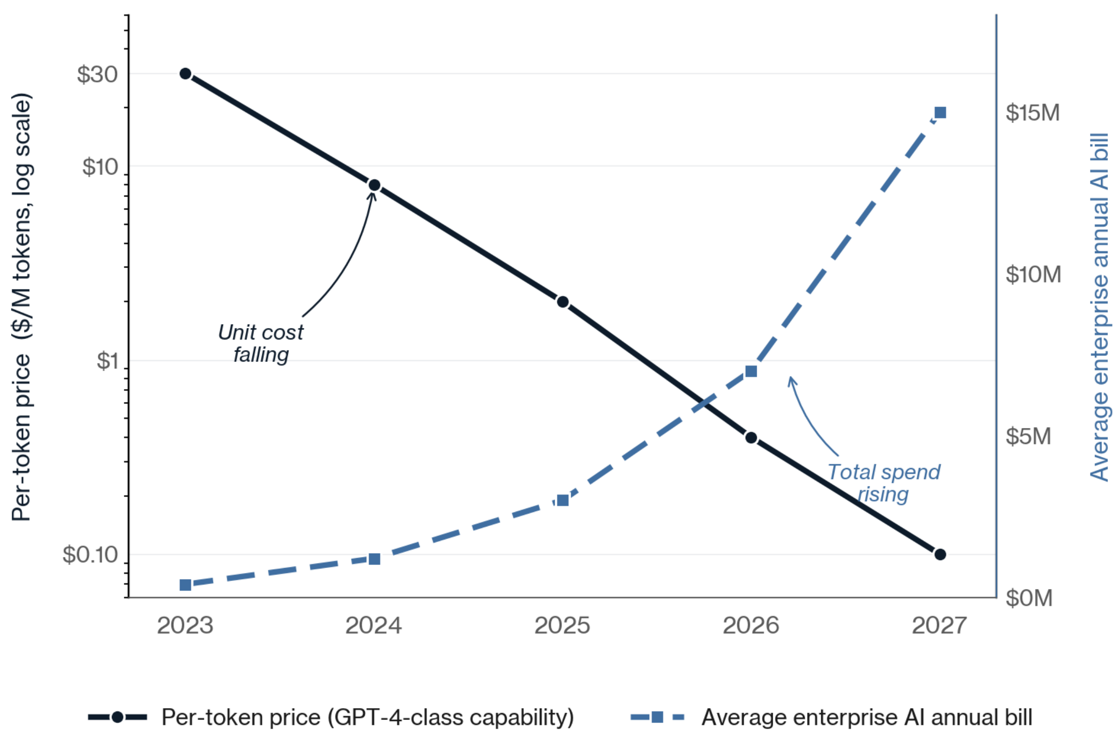

Figure 2. *Two curves running in opposite directions. Per-token API prices for GPT-4-class capability fell roughly 75-fold between 2023 and 2026, while average enterprise annual AI bills rose nearly 18-fold over the same period. Sources: Zylos Research; AnalyticsWeek; Oplexa; Epoch AI.*

This pattern is not new. The same dynamic played out in cloud computing a decade ago: per-instance prices fell consistently, while total cloud bills grew steadily because consumption grew faster than unit prices fell. The pattern produced an entire industry — cloud FinOps — whose function is to discipline consumption against business value. AI is on the same trajectory, only steeper, because the consumption multiplier from agentic workflows is structurally larger than anything cloud computing produced.

Why end users cannot solve this themselves at scale

When a new cost category appears, enterprises typically follow a predictable pattern: first they ignore it, then they try to manage it through procurement, then they try to build internal capability to address it directly. AI is now in the first stages of the second phase, and a meaningful number of enterprises are starting to consider the third — building internal AI infrastructure, training their own models, running inference on their own GPUs.

For most enterprises, that is the wrong answer. Building an internal AI capability that matches what a foundation model lab provides requires an order of investment that is rarely justifiable for any single enterprise. The talent is scarce, the infrastructure is expensive, the iteration speed is slower than what specialist labs can sustain, and the gap between an internal model and a frontier model widens every quarter as the labs continue to push the capability frontier. Eli Lilly building a $200 million purpose-built supercomputer for pharmaceutical AI is the exception, not the model. For most companies, doing this internally either underdelivers on capability, overdelivers on cost, or both.

What enterprises actually need is not their own foundation model. They need their existing AI consumption to produce more value per token. That is a different problem, and a different category of solution.

Why the cost-governance layer must sit elsewhere

Current foundation model pricing reflects a specific moment in market formation: subsidized adoption pricing. Per-token prices are calibrated to drive usage velocity, not to recover the full cost of operating frontier-scale infrastructure. The arrangement is sustained by venture capital, hyperscaler cross-subsidies, and growth-stage investor tolerance for losses. None of this is malpractice; it is the normal early-market pricing pattern for capital-intensive infrastructure categories. Cloud computing, telecom, and electric utilities all passed through equivalent phases.

What follows next is also normal: pricing tends to drift toward sustainable margin over time. Foundation labs face their own real economics — capital expenditure on data center buildout, energy costs that fluctuate with global commodity markets, talent costs that compound with capability competition, and eventually investor expectations of returns commensurate with the capital deployed. None of these costs trend downward in a coordinated way, and none of them are negotiable by foundation labs in any structural sense. Most analysts therefore expect costs to be passed through to customers over time, as they typically are in capital-intensive infrastructure categories. Even if per-token prices continue falling, enterprises remain exposed if production token volumes and workflow dependence grow faster than unit-cost reductions — which the consumption multipliers from agentic workflows make likely.

The exposure this creates for end-user enterprises is the specific combination that makes early-market infrastructure pricing dangerous: real urgency to adopt (competitive pressure), real value (productivity gains are genuine), and engineered affordability (current pricing is below true operating cost). Each factor alone is manageable. Together, they produce structural dependency. By the time pricing normalizes, AI workflows will be load-bearing for operations enterprises cannot easily wind down. Switching to a different foundation lab solves nothing if all foundation labs face the same eventual cost normalization. Even modest per-token price increases will land hard on enterprises whose token volumes are now measured in production scale rather than pilot scale. The risk is not that AI becomes unaffordable overnight; it is that enterprises allow AI workflows to become operationally indispensable before their unit economics are governable.

Foundation labs are not structurally positioned to resolve this exposure for their own customers. Their primary incentive is for total token consumption to grow, not for any given customer's consumption to be disciplined against business value. Every saved token at the application layer is revenue they do not capture. They also cannot credibly stand between a customer and their own pricing model — they are the entity the customer is paying. The disinterested party, the one whose financial incentive aligns directly with the customer getting more value per dollar, has to be a third entity.

End-user enterprises can attempt to build that disinterested function internally, and many will try. The economics of doing so eventually fail. A hospital's core competency is healthcare delivery, not token economics. Internal AI cost-governance teams are real expenses that compete with mission-critical investment. The first dollar spent on internal token FinOps produces meaningful returns; the tenth dollar produces less; the hundredth often produces a team whose ongoing cost exceeds the savings it generates. Enterprises that let internal cost-governance scale without bound end up with an administrative function that is itself a form of bureaucratic drag — capacity absorbed by managing AI consumption rather than delivering on the mission AI was deployed to support.

Deterministic AI Infrastructure resolves the structural trap by amortizing the engineering work that ends-user enterprises cannot amortize alone. The intermediary tier specializes in disciplined consumption as its core competency rather than as a distraction from it. Its costs spread across many customers; its incentives align with each customer's value per token; its engineering depth compounds in ways internal teams cannot easily replicate. The category is not a luxury feature for cost optimization. It is the risk-mitigation layer the market is forming because the alternative is structural exposure that end users cannot fully hedge themselves.

> **"Anyone can call a foundation model API. Few teams can architect a workflow where the structured logic does eighty percent of the cognitive lift and the LLM does only what is irreducibly linguistic."**

The result is not simply a cost problem. It is a market-structure problem: the function enterprises need is economically misaligned with the two categories they currently rely on, which is why a third category is forming.

The three-bucket market

The enterprise AI market is resolving into three structurally distinct buckets, each with different incentives and different roles in the value chain.

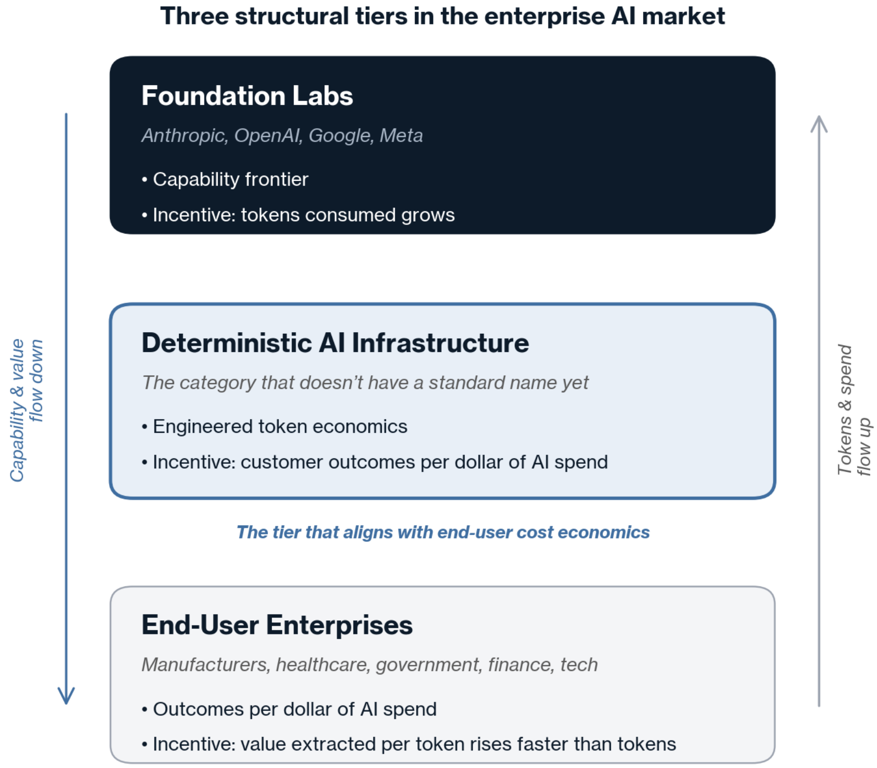

Figure 3. *Three structural tiers in the enterprise AI market. Foundation labs sit at the capability frontier, end-user enterprises consume capability, and a middle tier — Deterministic AI Infrastructure — holds the engineering discipline that aligns token consumption with business outcomes. The middle tier has the clearest structural incentive to align AI consumption with end-user cost economics.*

Of these three buckets, the third has the clearest structural incentive to make the end-user enterprise's AI spend produce more value per dollar. That alignment is the moat. It is also the reason this category will exist as a distinct tier, not as a feature of either of the other two.

3.
Deterministic AI Infrastructure

Defining the category

Deterministic AI Infrastructure is the engineering discipline of making AI workflows produce consistent, predictable, high-quality outputs at substantially lower token cost than naive prompt engineering achieves. The discipline rests on a single architectural principle: structured logic does what structured logic is good at, and large language models do what large language models are uniquely good at — and the work is divided accordingly.

A naive AI workflow asks the LLM to do everything: reason about the data, compute the answer, structure the response, and articulate the result in natural language. Each of those four functions costs tokens. Three of them — reasoning, computation, and structure — can usually be done deterministically in code, with the answer fixed before the LLM is invoked. The LLM is then asked only to articulate, which is the function it is genuinely best at and which is irreducibly linguistic.

The architectural pattern — separating deterministic logic from linguistic articulation — has antecedents in compound AI systems, structured outputs, AI workflows, and validation pipelines. What is new is the packaging: the consolidation of these patterns into a procurement-recognizable infrastructure tier defined by cost-per-decision economics rather than by any single technical primitive.

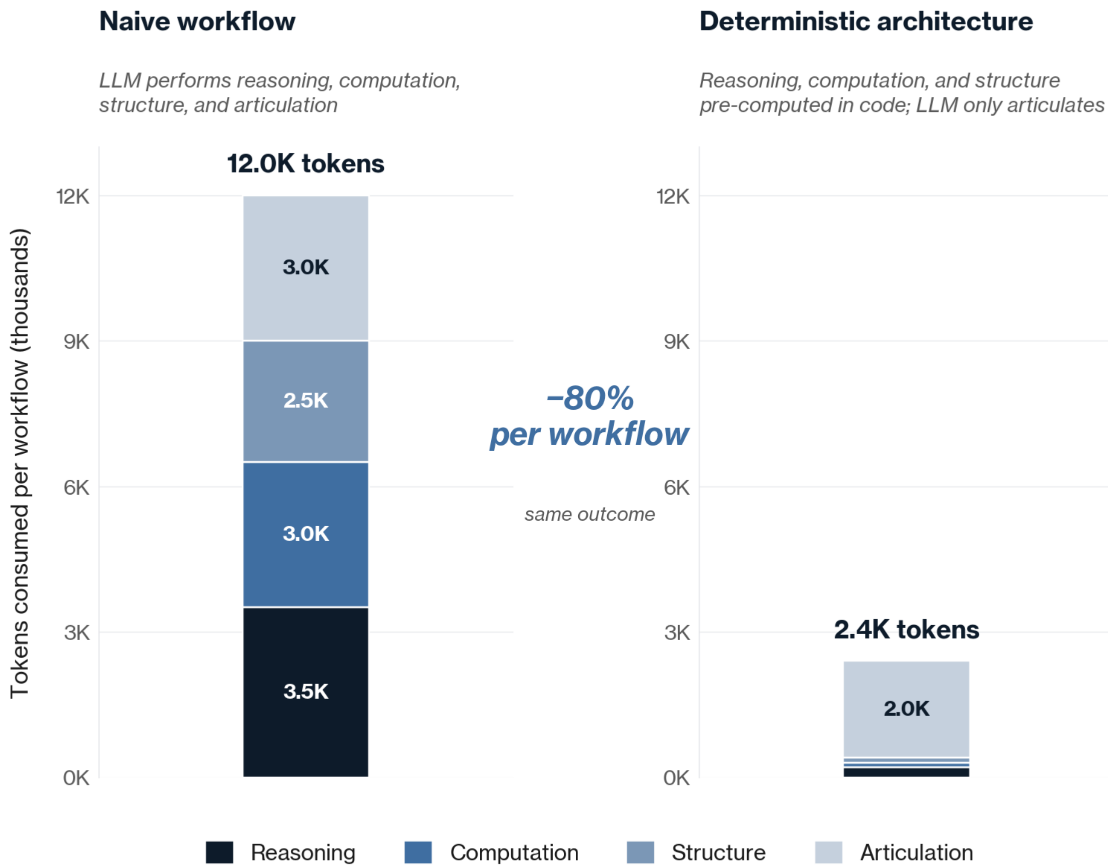

Figure 4. *Cost per workflow: naive vs deterministic architecture. The same business outcome can be produced with substantially lower token consumption when reasoning, computation, and structure are pre-computed in code, leaving the LLM to perform only the irreducibly linguistic work of articulation. Illustrative; reductions in production deployments vary by workflow complexity.*

The result is dramatic. The same business outcome — the same diagnostic read, the same analysis, the same recommendation — can be produced for a fraction of the tokens a naive workflow would consume. Output quality is higher because the deterministic layer does not hallucinate. Output is consistent across runs because the answer was computed before the LLM saw the data. Cost is predictable because the token consumption per workflow is a known engineering constant rather than a function of how much reasoning the LLM happens to do on a given call.

Why it cannot be easily replicated

Anyone can call a foundation model API. Few teams can architect a workflow where the structured logic does eighty percent of the cognitive lift and the LLM does only what is irreducibly linguistic. That is craft, and it compounds with team experience. It is not a moat that can be reproduced from a press release or a quarterly engineering hire. The companies operating at this level today will widen the gap with companies that begin trying to develop the discipline in 2027 — and the gap will compound as token economics tighten.

There are several specific engineering patterns that distinguish a Deterministic AI Infrastructure tier from a generic AI consultancy or prompt-engineering shop:

Pre-computed structure. The diagnostic, score, classification, or analysis is computed in deterministic code before the LLM is called. The LLM is told the answer; it only writes about it.

Locked-facts injection. The prompt explicitly tells the model what is true and what it must not contradict. Tokens are spent once on rails rather than retried after every contradiction.

Bounded output schemas. Required JSON keys force a fixed output shape. No preamble, no rambling synthesis, no rebuilt context. Every output token serves a specific report field.

Validation-first pipelines. Output is validated against the canonical structure before it reaches the customer. Inconsistencies are caught and corrected, not shipped.

Token telemetry as a first-class metric. Input and output tokens per workflow are tracked, optimized, and reported alongside output quality.

The approach is not without trade-offs. Building a Deterministic AI Infrastructure tier requires substantial upfront investment in structured logic, canonical descriptors, and validation pipelines, and the resulting workflows are less flexible for rapidly evolving or genuinely open-ended use cases than naive LLM invocation. Frontier model improvements in reasoning capability may also narrow the token-efficiency gap over time for simpler tasks. For the high-volume, structured enterprise processes that dominate production AI spend, however, the advantages are decisive: the value is not token reduction alone but lower variance, stronger validation, domain-correct logic, predictable per-decision cost, and reduced exposure to foundation-pricing volatility. These compound as token volumes grow.

The domain-expertise nexus

There is a second moat that the engineering-patterns frame alone obscures. The structured logic doing the cognitive lift has to encode correct domain knowledge. A diagnostic engine for HR cases needs to encode employment law and the realities of how HR organizations actually function. A diagnostic engine for clinical operations needs to encode clinical workflow norms. A diagnostic engine for federal program performance needs to encode contracting regulations and program management practice. Software engineers without that subject-matter depth can build a beautifully architected system that produces structurally elegant nonsense, because the rules embedded in the deterministic layer are wrong or incomplete.

This is a different and more dangerous failure mode than LLM hallucination. A naive LLM workflow that hallucinates may at least reveal instability through variance across runs; the variance itself is a signal that something may be off, and a domain expert reviewing two or three outputs can usually spot the inconsistency. A deterministic system encoding incorrect domain logic can be wrong consistently, silently, and at scale. The same flawed answer gets produced reliably across every workflow, looks credible because the architecture around it is sound, and only fails the moment a domain expert audits the underlying rules — which usually happens long after decisions have been made on the strength of its output.

There is a tempting counter to this concern: developers can use LLMs to fill their domain knowledge gaps when building the structured layer. LLMs can legitimately assist in drafting code, taxonomies, and rules. The problem is what happens next. If the code's logic is derived from LLM output without independent validation by domain experts, the deterministic layer is deterministic only in execution, not in epistemic authority. The same flawed reasoning will produce the same flawed answer reliably, with the appearance of engineered rigor and the substance of LLM reasoning, and with worse traceability than either. The LLM-as-knowledge-source approach is valid for narrow or well-documented domains; in any domain where practical expertise matters, it requires a domain expert in the loop to validate the rules before they are encoded. There is no shortcut.

Deterministic AI Infrastructure done well is therefore a meaningfully different kind of organization than a prompt-engineering shop or a general AI consultancy. It requires both engineering depth and lived subject-matter depth, and the integration of the two is itself a form of craft. It is harder to staff, harder to imitate, and a substantially stronger moat than engineering competence alone.

How the architecture changes the economics

Consider, as an illustrative example, a hospital deploying an AI workflow to surface operational patterns from internal data. A naive implementation passes the data to a foundation model with a system prompt and asks it to identify issues, prioritize them, write recommendations, and produce a report. That workflow may consume 8,000 to 15,000 output tokens per run. At enterprise volumes, this scales into meaningful monthly bills, with output quality varying significantly across runs because the LLM is doing both the analysis and the writing simultaneously.

The same hospital, working with a Deterministic AI Infrastructure provider, would receive the same operational read produced through a different architecture. A structured diagnostic engine — built on the provider's proprietary logic — performs the analysis offline. The output is a canonical descriptor of findings, severities, priorities, and recommended actions. The LLM is then invoked only to write a sector-aware narrative around facts that have already been determined. Token consumption drops from roughly 12,000 to 2,400 tokens per run — the eighty percent reduction shown in Figure 4. Output quality is more consistent, because the analysis itself is deterministic and only the language varies. Per-run cost becomes a known engineering constant the provider can quote in advance.

The hospital does not see this architecture. The hospital sees a faster, more reliable, and more cost-predictable diagnostic. The provider sees a sustainable margin structure that does not erode as foundation model pricing pressure works through the market.

Where SaaS fits

Deterministic AI Infrastructure is most naturally a SaaS layer, sitting between an end-user enterprise and the foundation model APIs the workflow ultimately consumes. The SaaS provider holds the proprietary logic, the structured diagnostic engine, the pre-computed canonical descriptors, and the validation pipelines. The customer pays for outcomes — reports produced, decisions surfaced, diagnostics completed — at a price the SaaS provider can offer profitably because the underlying token consumption is engineered to be a known and disciplined constant.

This pricing model has a property that pure foundation model APIs cannot match: the customer's per-decision cost is predictable. A hospital can quote a cost-per-diagnostic to its CFO and have it remain accurate over time, even as foundation model pricing changes underneath. The SaaS provider absorbs the volatility in exchange for the engineering investment it has made in disciplined consumption. The value to the customer is not token reduction alone; it is lower variance, stronger validation, domain-correct logic, and reduced exposure to foundation-pricing volatility.

4.

What Happens to Enterprises That Don't Adapt

The forced choice

Enterprises currently absorbing AI costs as a strategic premium will eventually face the same financial discipline that every other expense category faces. Boards do not tolerate indefinite spending growth without measurable return. CFOs will start asking what every dollar of AI spend produces. The Deloitte 2026 enterprise AI report finds that twenty percent of organizations are already growing revenue through their AI initiatives, while seventy-four percent hope to in the future. That gap will not stay open indefinitely. The companies in the seventy-four percent will be asked to demonstrate their hopes are converging on outcomes.

When that discipline arrives — and the FinOps adoption data suggests it has already started arriving for the majority of large enterprises — companies that have not engineered their AI consumption will face a forced choice among four options, all of them painful:

Ration AI use. Restrict the workflows employees can run, throttle agent loops, cap monthly token budgets per team. This works financially but reverses the productivity gains AI was deployed to deliver.

Pass costs to customers. Raise prices to cover unsustainable AI bills, eroding competitive position and inviting market share loss to competitors with more disciplined cost structures.

Cut headcount further. The layoffs already justified by AI deployment turn into a second wave of layoffs to fund AI deployment. This compounds rather than resolves the cost problem.

Spend through it. Continue absorbing AI costs as strategic, until the absorption is no longer sustainable. This is currently the most common posture and the least defensible.

The alternative

The alternative is to govern AI consumption so that it produces more value per token before the financial discipline arrives. For a small number of enterprises operating at hyperscale, this can be done with internal engineering capability. For most enterprises, it cannot. The patterns are specialized, the testing infrastructure is expensive to build, the canonical structures must be calibrated to the domain, and the ongoing optimization requires continuous familiarity with foundation model behavior that most internal teams do not have time to acquire. What practical action looks like for most enterprises, therefore, is procurement: identifying vendors whose products already embody this engineering discipline, and integrating them in place of, or alongside, naive foundation model invocation.

Enterprises that take this path will not need to ration use, raise prices, or cut headcount to fund AI. They will be paying meaningfully less per workflow than competitors who deferred the engineering work, and producing comparable or better output. That cost gap compounds quarterly as AI consumption volumes grow. This is the entry point for the Deterministic AI Infrastructure tier.

5.

What Enterprise Buyers Should Measure Now

Most enterprise AI procurement conversations are still demo-driven. The shift to unit-economics conversations is happening, but it has not become standard. Buyers who start asking unit-economics questions now will be ahead of their peers in eighteen months when those questions become the standard procurement frame. The procurement question is no longer simply whether an AI product works; it is whether the product's economics remain governable when usage becomes operationally routine. Five questions to ask any AI vendor in 2026, whether they sell foundation model access, an AI application, or infrastructure between the two:

1. What is the per-transaction cost?

Not the API price per million tokens. The all-in cost per workflow the customer actually runs. A vendor that cannot answer this in concrete dollars is a vendor that has not engineered the workflow to a known cost structure.

2. What is the variance distribution?

Average cost is not enough. The standard deviation matters as much as the mean. A workflow that averages $0.50 per run but occasionally spikes to $50 because of agent retry loops or context bloat is not a workflow with predictable economics. Ask for the p50, p95, and p99 cost per workflow.

3. What happens to your margin if foundation model prices rise twenty percent?

Foundation model pricing is currently subsidized by venture capital and hyperscaler cross-subsidies. The leading frontier labs have publicly reported losses that substantially exceed their revenue, even as revenue grows rapidly. That dynamic is unlikely to continue indefinitely. Vendors whose margin is destroyed by a moderate price increase have built their pricing on someone else's unsustainable economics. Vendors with engineered consumption will absorb a price increase with margin compression but not with margin collapse.

4. What percentage of the work is done by your structured logic versus by the LLM?

This is the diagnostic question that distinguishes Deterministic AI Infrastructure from prompt engineering. Vendors who answer with confidence — who can describe specifically which parts of the workflow are deterministic and which parts are linguistic — are operating at the engineering tier.

Vendors who answer vaguely, or who treat the question as unimportant, have not done the work.

5. Show me the validation pipeline.

Output that has not been validated against a canonical structure has not been engineered for consistency. A vendor whose output is whatever the LLM happens to return is selling demos, not infrastructure. The presence of a validation layer — and the reporting of validation warnings to customers — is a marker of a mature AI product.

6.

A Note on Naming

Categories that are unnamed are categories that are hard to procure. Enterprise buyers cannot create line items for things they do not have language for. Analyst firms cannot create market quadrants for categories that do not yet exist in their taxonomies. Investors cannot allocate capital to a thesis that has not been articulated.

We have proposed Deterministic AI Infrastructure as the name for this tier. The reasoning:

“Deterministic” signals the technical reality that distinguishes this tier from generative AI — outputs are consistent because the structured logic does the heavy cognitive lift before the model is invoked.

“Infrastructure” signals scale and pricing tier appropriate to the actual investment required to build at this level — capital infrastructure, not consulting hours.

Together they distinguish cleanly from “Generative AI” without dismissing it. Generative AI is the foundation. Deterministic AI Infrastructure is the engineering layer that makes Generative AI economical for enterprise production.

It is hard for prompt engineering shops to claim credibly. The discipline cannot be performed without the underlying structured logic, and the logic is what takes years to build.

The category will exist with or without this specific name. Other terms — AI Efficiency Layer, Structured AI, Token Economics Infrastructure, Inference FinOps — are circulating in adjacent conversations. The name that ultimately sticks will be the one that buyers, sellers, and analysts converge on. The companies that operate at this tier today have a small window to influence which name does.

7.

Conclusion

The first lap of enterprise AI rewarded urgency. The companies that moved first captured early productivity gains, built the institutional muscle for AI-augmented work, and absorbed compute costs at prices calibrated for adoption velocity rather than sustainable margin. That lap is now ending, and the conditions of the next one will not be the same.

Most analysts expect foundation model pricing to drift toward sustainable economics over time — not as a hostile act but as the normal arc of infrastructure-sector maturation. Real capital expenditure, real energy costs, and real investor expectations of returns make it likely. As that pricing normalization arrives, AI workflows will already be load-bearing for enterprise operations that cannot easily be wound down. Even modest per-token price increases would land hard on companies whose token volumes have grown into production scale.

The companies that come through the transition strongest will not be the ones with the largest AI budgets. They will be the ones whose AI consumption is governed by the highest engineering discipline — whether built in-house at the rare scale that justifies it, or, far more commonly, procured from intermediaries who specialize in it. They will operate with the highest output value per token and the lowest exposure to foundation pricing volatility. They will produce comparable or better business outcomes for a fraction of the cost, and that cost-and-exposure gap will compound as AI volumes grow.

Most enterprises are not in a position to build that combined discipline — engineering depth, validation pipelines, domain-aware structured logic, and continuous token economics — on their own. The category that will provide it, Deterministic AI Infrastructure, is starting to take shape now, in advance of the demand wave that will require it. The window for foresight is open. It will not stay open indefinitely.

> **"AI is not free. The companies that internalize that fact first — and engineer accordingly — will define the next decade of enterprise AI."**

REFERENCES

Andreessen Horowitz. *How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025.* February 2026.

AnalyticsWeek. *2026 Inference Economics Report.* 2026.

CloudZero. *FinOps Adoption Survey, 2024–2025.*

Consumer Reports. *AI Data Centers and Their Impact on Electric Bills.* March 2026.

Deloitte. *State of AI in the Enterprise — 2026 AI Report.*

Deloitte Insights. *The AI Infrastructure Reckoning: Optimizing Compute Strategy in the Age of Inference Economics.* February 2026.

Ecosystm. *The Emerging Economics of Enterprise AI: A Practical Guide for 2026.* February 2026.

Epoch AI. *State of Frontier AI Inference Economics.* 2026.

FinOps Foundation. *2026 State of FinOps Report.*

Gartner. *AI Spending Forecast 2026 / Data Center Systems Spending Forecast 2026.*

IDC. *Worldwide Artificial Intelligence Spending Guide (2026)* and *FutureScape 2026.*

International Energy Agency. *Data Center Electricity Demand Projections.*

Lawrence Berkeley National Laboratory. *Data Center Electricity Demand Report.* 2024.

MIT Sloan Management Review. *AI Has High Data Center Energy Costs — But There Are Solutions.* February 2026.

NVIDIA. *State of AI Report 2026.*

Oxford Economics. *AI Is Reshaping the Composition of Enterprise Tech Spending in the US.*

TechCrunch. *VCs Predict Enterprises Will Spend More on AI in 2026 — Through Fewer Vendors.* December 2025.

TheStreet. *The Next Phase of AI Spending Is Already Underway.* May 2026.

WalkMe. *2026 Enterprise AI Budget Research.*

Zylos Research. *Inference Economics: AI Agent Compute Markets in 2026.* April 2026.

ABOUT MONDERMAN

Monderman is an institutional performance research company building Deterministic AI Infrastructure for organizational diagnostics. Its diagnostic platform produces structured operational reads for enterprises across sectors, including defense, healthcare, government, financial services, technology, manufacturing, and higher education.

MONDERMAN

Organizations deliver at the speed of their administrative reality.

connect@monderman.com • www.monderman.com