ecosystm.

AI Inference Economics: Managing the Cost of Scale

AUTHORED BY
Darian Bird
Principal Advisor, Ecosystm

AUGUST 2026

Table of Contents

Introduction

AI is getting cheaper to run, but enterprises are spending more on it.

The early period of AI investment centred largely on the high upfront costs of training models. The economics are now shifting towards inference. Many enterprises now only need to refine their already developed custom models or can rely on third-party models, such as open-weight variants or those delivered as a service. As a result, recurring inference costs are becoming a larger share of AI spending for many organisations.

3.4x
Expected annual growth in global inference capacity

10x
Recent annualised growth indicated by token-demand proxies

Source: Epoch AI (Is a Compute Crunch Coming, 2026)

Recent advances in silicon and model efficiency have reduced the cost of each unit of AI, but enterprise spending is still rising. Adoption is expanding, while power users such as software developers are consuming AI at much higher volumes. This is a classic example of Jevons Paradox: as a technology becomes cheaper to use, demand can increase enough to drive total spending higher. The challenge for organisations is therefore to capture the productivity gains from greater AI use without allowing consumption – and costs – to run ahead of value.

Innovations such as reasoning models and autonomous agents are adding another dimension. AI can now take on more complex tasks, but doing so can require substantially more inference. Larger context windows can increase consumption when processing longer conversations or inspecting full codebases, while retrieval from external sources adds further processing. Always-on agents can also continue operating longer than intended if they are not properly governed.

As organisations become more experienced with AI, they will develop ways to manage these costs before they escalate. Precise, reusable project context and instructions can reduce unnecessary inference. Differentiating between models makes it possible to match higher-cost AI with higher-value tasks. Over time, organisations will need to optimise and govern inference as systematically as any other technology cost – and link it more closely to the business outcomes it enables.

Why AI Consumption is Rising Faster than Expected

Inference spending is expected to rise as AI is adopted by more people and used more often. However, some enterprises have experienced bill shock, as consumption increases unexpectedly fast due to several factors that can have a multiplier effect on costs.

Usage is expanding

Adoption figures alone understate how much consumption can grow, because usage expands in both breadth and depth. More employees, customers, and applications begin using AI, while existing users apply it to a growing share of their work. Software development is the most mature example of this, where LLMs were initially used for simple tasks, such as documentation creation, code completion, and debugging. Due to advances in coding agents, engineering teams are increasingly relying on them for architectural tasks.

AI is processing richer inputs

AI applications today process images, documents, audio, and video alongside text. These richer inputs can require substantially more inference than a short text interaction. Multimodality also opens new categories of enterprise workload that were previously difficult to automate, from analysing scanned documents to reviewing photographs or recorded calls. It expands both the inference required for some tasks and the volume of enterprise information that AI can process.

Reasoning adds hidden compute

Reasoning models change the relationship between what a user sees and the inference required to produce it. Beyond the visible prompt and response, these models can consume substantial additional reasoning tokens or computation before reaching a final answer, so a short response can represent considerably more inference than its length suggests. This can improve performance on complex tasks, but wider use of reasoning models can substantially increase the inference consumed per completed task.

320x

Year-on-year growth in average reasoning-token consumption per organisation among OpenAI enterprise customers

Source: OpenAI, State of Enterprise AI 2025

Context carries a cost

Enterprise applications increasingly draw on substantial context, including conversation history, codebases, and information retrieved from enterprise systems. Coding assistants and agents may repeatedly process this context as they work. RAG can improve relevance and reduce hallucination, but retrieved information still adds to the volume a model must process, while more elaborate retrieval architectures can trigger additional calls for query generation, reranking, or summarisation.

Agents multiply inference

Agentic applications can turn a single request into a sequence of inference events as the system plans, invokes models, calls tools or external systems, evaluates results, and repeats until the task is complete. A coding agent, for example, may explore a codebase, plan its approach, and make repeated calls from one instruction. Cost reflects the work an agent performs rather than the number of prompts submitted.

1,000x more tokens, with 30x variation

Agentic coding tasks consumed approximately 1,000 times more tokens than coding chat or single-turn reasoning, with repeated attempts at the same task varying by as much as 30x

Source: Bai et al. How Do AI Agents Spend Your Money?, 2026

Reliability requires more processing

Producing an initial output is often only part of the inference a production workload requires. Enterprises may add model-based steps such as self-checking, verification by another model or agent, retries on low-confidence outputs, evaluation and validation, or guardrails applied to inputs and outputs. Each additional model call adds consumption, meaning the cost of producing a dependable result can exceed the cost of generating the first response.

Inference Pricing is Shifting with the Work

AI pricing is starting to reflect the work being done, not simply access to the technology.

AI pricing is moving away from treating every request, user, or seat as equivalent. As AI applications become more capable, the amount of inference required to complete a task can vary dramatically. A query that once triggered one model call may now involve extended reasoning, substantial retrieved context, several coordinated model invocations, tool use, and retries. Two requests that look similar to a user can therefore generate very different costs.

This variability is putting pressure on fixed per-seat subscriptions and uniform request allowances, which can overcharge light users while undercharging intensive users. Providers are responding with usage credits, model-specific rates, premium tiers offering greater access or capacity, charges tied to agent actions, and pricing linked to completed outcomes. These mechanisms increasingly sit alongside subscriptions, allowing providers to segment demand, protect margins, and capture more of the value generated by AI.

Coding assistants are among the clearest examples of this shift

Cursor's USD 20 monthly subscription originally included 500 fast requests using premium models, but in June 2025 it replaced this with USD 20 of usage priced at the model providers' API rates. The vendor explained that its hardest requests cost an order of magnitude more than simple ones. GitHub subsequently replaced Copilot's premium-request allowances and model-specific multipliers with credits based on input, output, and cached tokens consumed. It noted that its previous structure charged users the same for a brief chat prompt and a multi-hour autonomous coding session, leaving GitHub to absorb the cost difference. Both companies retained subscriptions but tied the included allowance more closely to actual inference consumption, making expenditure more variable.

The major LLM providers are moving in the same direction

OpenAI, Anthropic, and Google have retained monthly subscriptions for individual and small-business plans but have established more defined usage limits and the option of additional credits. Enterprise plans go further by combining seats with pooled credits or metering inference separately, shifting more of the cost associated with intensive users and agentic workloads to customers. Consumption pricing is also becoming more differentiated, with rates varying by model, input and output tokens, caching, and delivery requirements. Anthropic, for example, offers a 50% discount for batch processing, allowing customers to trade response speed for lower inference costs.

For autonomous agents, pricing can go one step further: charging for the outcome rather than the AI consumed

Fin, which Salesforce recently agreed to acquire, helped establish pay-per-resolution pricing when it launched its customer service agent, charging USD 0.99 per resolved enquiry rather than for tokens or messages. Zendesk has also adopted automated-resolution pricing and distinguishes between verified resolutions and customer silence or abandonment. Customer service is a natural starting point because each resolution is a discrete unit of work that can be compared with the cost of a human agent. The model charges only for successful resolutions, excluding unsuccessful or escalated interactions. Fin has extended the model beyond support, charging USD 0.99 for completed workflows and disqualified leads, and USD 9.99 for a customer-qualified sales lead. Pricing is therefore beginning to reflect not just how much AI is consumed, but what the AI accomplishes.

Optimising Inference Economics

As AI adoption scales, inference becomes an operating cost rather than simply an adoption cost.

Consumption can vary significantly by model, workload, context, and workflow design, making cost management a question of how AI is built and used, not simply how much it is purchased.

1. MODEL CAPABILITY AND TASK COMPLEXITY

Not every inference task demands the most capable model. Using frontier or reasoning models for routine work means paying for capabilities the task may not require.

- Smaller or specialised models are well suited to bounded tasks such as extraction, classification, formatting, and summarisation.
- Larger or reasoning-capable models become more relevant when a task involves greater complexity, ambiguity, or judgement, rather than across an entire workflow.

2. INFERENCE EFFICIENCY

Optimising inference costs is not only about paying less per token, but reducing the inference required to complete a task. Repeated context, oversized inputs, and unnecessary output can all increase consumption without improving the result.

- Precise, reusable project context and instructions can reduce unnecessary retrieval, irrelevant output, and subsequent rework.
- Large inputs can be broken into manageable chunks and combined or summarised rather than processed as a single task.
- Recurring prompts, context, or intermediate results can be cached to avoid repeated processing.

3. AI WORKFLOW DESIGN

Agentic applications can multiply inference consumption, since a single task may trigger multiple model calls, tool invocations, and retries. Application design therefore directly affects the inference required to complete each task.

- Simple, clearly scoped tools can reduce errors, retries, and wasted calls.
- Complex workflows can be broken into smaller, well-defined tasks so models receive clear objectives and only the information required for each step.
- Monitoring the full workflow cost is important, including external APIs and RAG infrastructure, which can escalate in more complex applications.

4. PRICING AND CONSUMPTION PATTERNS

No single pricing model is cheapest for every organisation. Pooled, per-seat, and metered pricing produce different economics depending on team size and how concentrated usage is among heavy users.

- Pooled consumption can be more efficient across larger teams where usage varies significantly between individuals.
- For startups and small teams with a handful of heavy users, enterprise or metered pricing may be considerably cheaper than fixed-price individual subscriptions.
- The optimal commercial model can change as organisations and their AI usage grow, making periodic reassessment important.

Spotlight

EY: Making Model Selection an Efficiency Lever

EY introduced an internal AI router in April 2026 that directs requests from specialist tools towards models suited to the complexity of each task. Deployed in areas including tax and risk, it routes simpler work away from more expensive frontier models without changing how employees interact with the tools. EY reports that the router, combined with training and governance measures, reduced token consumption by up to 60% in the environments where it was deployed.

METLIFE:
Building an AI FinOps Discipline

While cloud FinOps has matured into a well-established discipline for allocating expenditure and forecasting usage and spending, AI has proven harder to govern. Usage is spread across internal platforms, embedded SaaS features, and external model providers, and billed through a mix of tokens, subscriptions, and platform fees that can be difficult to allocate across business units.

MetLife is developing a more unified approach to financial governance as AI scales. At FinOps X 2026, the company highlighted the difficulty of monitoring consumption distributed across internal platforms, SaaS tools with embedded AI, and external APIs. The challenge is compounded by a rapidly changing technology landscape while budgets are set 18 months in advance. A unified view of spending is therefore becoming important for identifying where usage occurs before it can be optimised or governed.

Greater visibility into costs is also informing model selection. Rather than defaulting to the most capable option available, MetLife weighs cost, performance, accuracy, and workload complexity for each use case, aiming for the least expensive model that reliably delivers the required outcome. The assessment begins with a more fundamental question: is AI necessary at all? Making this judgement at design time reduces the risk of building an application around an unnecessarily expensive model and having to re-architect it once usage grows. This moves cost management upstream, rather than treating it as a retrospective exercise once expenditure has already occurred.

MetLife is also embedding controls directly into its AI platforms, including model-routing logic, workload classification, and spending limits. Access to expensive reasoning models is differentiated by role. Engineers working on complex problems may select the most capable models, while employees in functions such as marketing or legal are offered a narrower set of options tied to the needs of their work. The aim is to match model access to task requirements while preserving autonomy within defined guardrails and reducing unnecessary bureaucracy.

The greater challenge is connecting AI consumption to business value. Usage and outcome data often sit in separate systems, making it difficult to link inference costs to efficiency, productivity, or revenue.

> **“How do you tie token spend to business value? This requires correlation of data across different systems to tie consumption to value realisation. This is not easy.”**
>
> **Sonali Niswander, SVP, MetLife’s Global Technology Office**

Efficiency use cases, which can be benchmarked against an existing process, are easier to measure, while growth-oriented applications often have no historical baseline for comparison. In these cases, MetLife relies on a stated value hypothesis and time-limited pilots to assess prospective returns.

What AI FinOps Needs to Cover

AI FinOps extends traditional cost management beyond model invoices to the full economics of AI applications and agents. This requires visibility into where AI spending occurs, what generates it, and what value it produces.

98% of FinOps teams now manage AI spending, up from 63% in 2025 and 31% in 2024.

Source: The State of FinOps Report, 2026

Full-stack cost visibility captures expenditure across model providers, AI-enabled SaaS products, agents, data and RAG platforms, external APIs, GPU infrastructure, and other cloud resources supporting AI workloads.

Instrumentation and allocation combines provider billing with usage telemetry and workload metadata, attributing model calls, tokens, agent steps, and supporting costs to the applications, teams, features, customers, or business processes that generated them.

Workload unit economics calculates the complete cost of a useful unit of work, such as a resolved enquiry, accepted code change, or processed document. This includes retries, verification, retrieval, tool calls, and orchestration, rather than measuring only the initial model response.

Dynamic forecasting and control accounts for user growth, usage per user, and inference consumed per task. Budgets, alerts, model-access rules, and agent iteration limits can then be applied at the level of teams, applications, or workloads.

Outcome and value measurement connects the cost of completed work to operational measures such as speed, quality, and labour saved, and ultimately to financial value where this can be established reliably.

Priorities for Managing AI Costs

As inference costs become a more significant part of AI spending, AI, technology, engineering, and finance leaders will need to manage consumption alongside performance, reliability, and business value.

MEASURE THE COST OF COMPLETED WORK

Token consumption shows how much inference an application uses, but not what that expenditure produces. Where possible, organisations should calculate AI spending per recognisable unit of work: a resolved customer enquiry, a processed document, an accepted code change. This connects spending to tangible outcomes and allows comparison with the cost, speed, and quality of an existing or comparable process. Attributing expenditure precisely to financial value remains difficult, particularly for new applications, so this measure is a practical starting point for assessing value rather than a complete ROI calculation.

MAKE INFERENCE EFFICIENCY A DESIGN REQUIREMENT

Inference consumption should be managed as an application-performance metric alongside quality, latency, and reliability. Teams should set expected consumption for representative tasks, test it during development, and monitor it in production. If the same work begins consuming substantially more inference, this should be treated as a performance regression. Application design should incorporate reusable context and instructions so models do not repeatedly rediscover information, reducing consumption without compromising output standards.

MATCH MODEL ACCESS TO THE WORK PERFORMED

Organisations should not send every request to the most capable or expensive model by default. Model capability should match the complexity of the work, with routine, bounded tasks directed to smaller or specialised models and frontier or reasoning models reserved for genuine complexity, ambiguity, or judgement. This can be operationalised through automated routing or differentiated access by role, application, or workload, with escalation available where needed. Selection should weigh cost against accuracy, quality, latency, and reliability, and be reassessed as capabilities, prices, and requirements evolve.

GOVERN AND FORECAST CONSUMPTION DYNAMICALLY

Organisations need visibility into AI consumption across providers, applications, and teams, with responsibility shared across engineering, finance, and business leadership. Because consumption often concentrates among intensive users or inference-heavy agents, adjustable budgets, alerts, and spending limits are more effective than uniform allowances. Controls should manage unexpected growth without unnecessarily restricting legitimate use. Forecasts should account for changes in user numbers, usage levels, and inference per task, since falling token prices may not reduce total expenditure.

About the Author

Darian Bird

Principal Advisor, Ecosystm

Darian helps businesses navigate transformation, providing insight into cloud, automation, data management, infrastructure, and telecommunications. He has spent two decades advising leaders on leveraging technology to enter new markets, improve client experiences, and enhance service delivery, with a particular focus on infrastructure and strategic resilience. At Ecosystm, Darian conducts in-depth research on cloud, data, and AI sovereignty regulations across the Asia Pacific, exploring their implications for enterprises and technology providers.

ecosystm.

Ecosystm is a leading technology market analyst and advisory firm that helps stakeholders navigate innovation in the digital economy through data, insights, and expertise. We connect enterprises, technology companies, digital-native founders, investors, and policymakers to enable informed decision-making. With ongoing research and access to top analysts and strategic advisors, we empower business planning, go-to-market activities, thought leadership, and innovation strategy consulting. Visit ecosystm.io