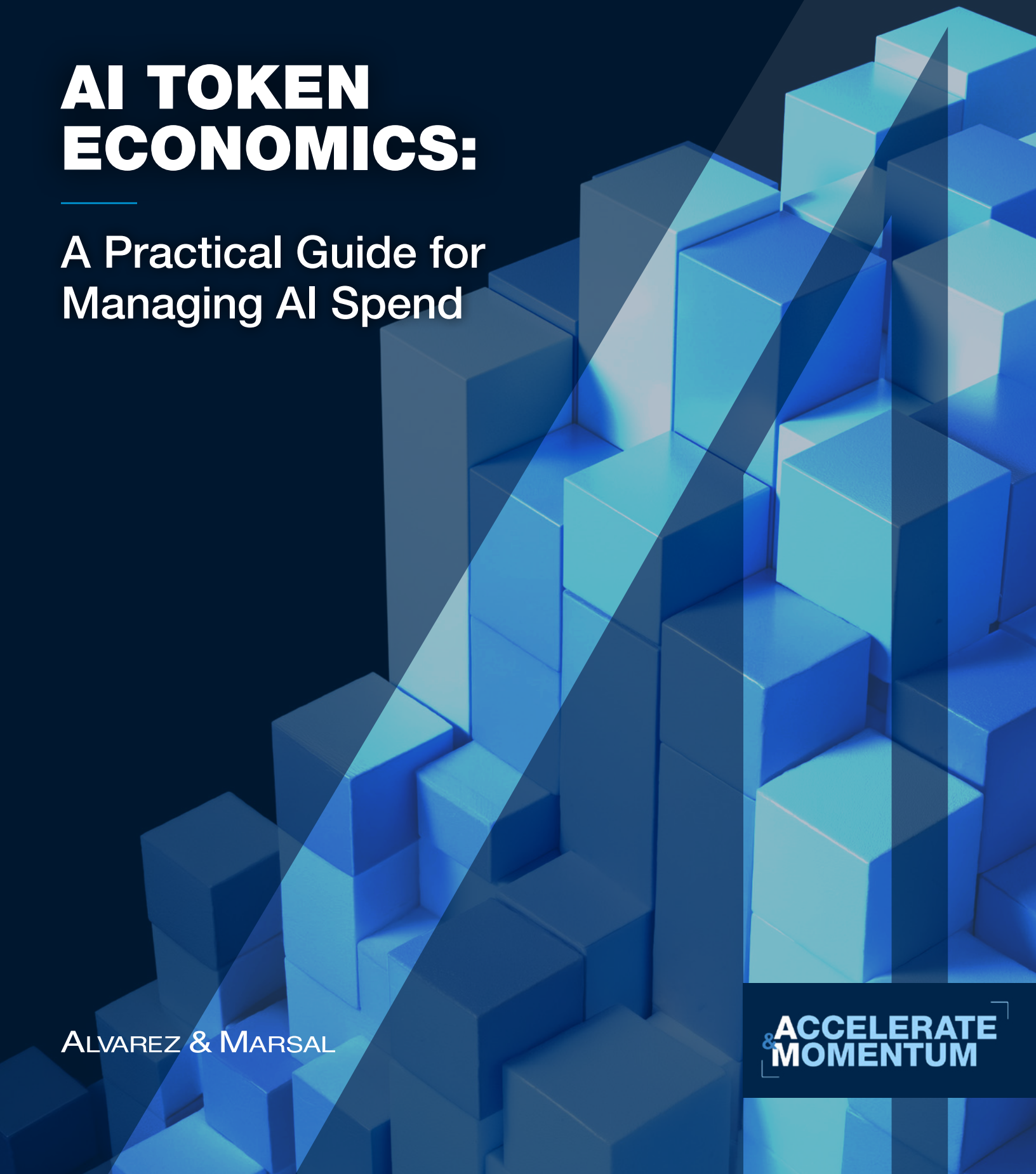
PRIVATE EQUITY
PERFORMANCE IMPROVEMENT

AI TOKEN ECONOMICS:

A Practical Guide for Managing AI Spend

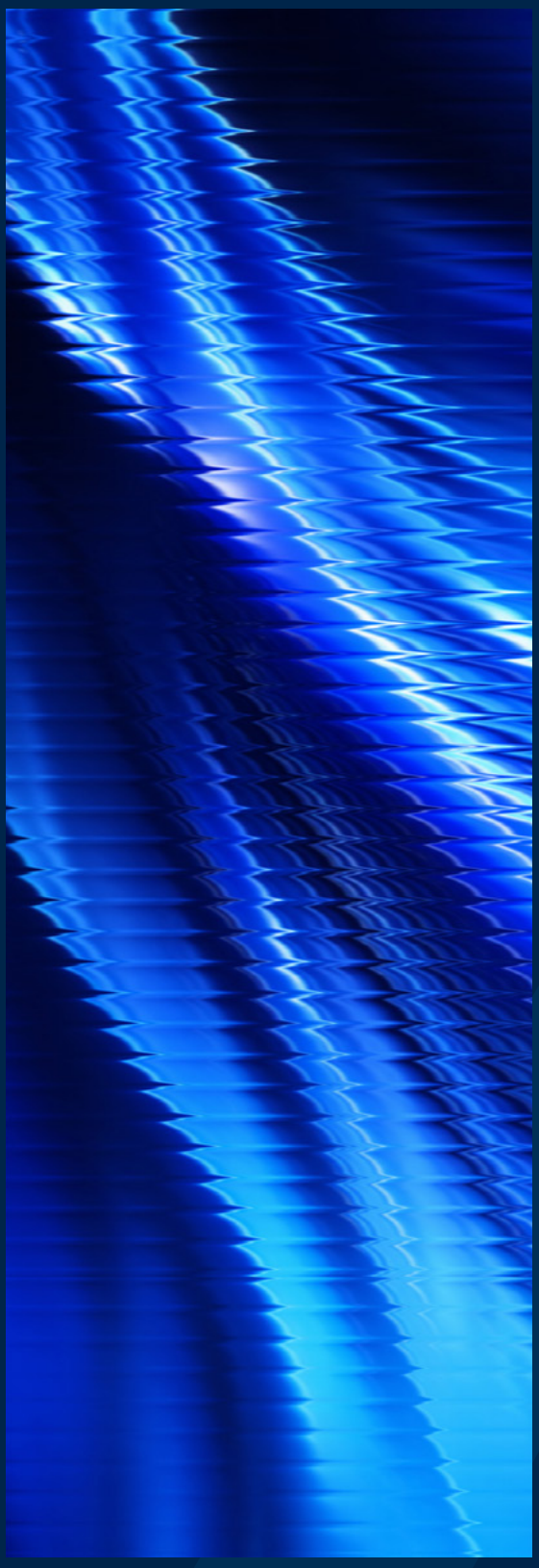


ALVAREZ & MARSAL

ACCELERATE & MOMENTUM

TABLE OF CONTENTS

INTRODUCTION



As artificial intelligence (AI) moves from proof of concept to operational infrastructure, the economics of running it have become a material line item for organizations across industries. Prices are moving quickly and the competitive landscape is shifting on a timescale of months, not years.

Yet most organizations have no systematic way to understand, measure, or manage their AI spend.

At the core of AI pricing are tokens, the fundamental unit of measure that a language model processes. They are at the core of all pricing and usage discussions. As a result, private equity (PE) funds and their portfolio companies need to develop a practical framework to bring the same discipline to token costs that they would apply to any other operational expense as they continue to scale AI.

Several factors—including language, task type, and market competition—influence the cost of AI usage, and structural cost drivers can embed additional expenses into organizations. Companies can manage tokens, and by extension AI costs, by:

- Implementing best practices in design, measurement, configuration, and governance
- Educating users on optimal AI use
- Tracking token efficiency

UNDERSTANDING THE BASICS: TOKENS, PRICING, AND COSTS

In practical terms, one token corresponds to roughly four characters in the English language. The phrase "private equity due diligence" consumes around six tokens, for instance. Code, structured formats such as JSON or XML, and certain languages including Chinese and Arabic are substantially more token-dense than plain English prose.

Every model interaction consumes tokens across two separate pools:

1. Input tokens cover everything sent to the model: the system prompt, any conversation history, retrieved context, and the user message itself.
2. Output tokens cover everything the model generates in response.

Providers price these two pools differently, with output tokens typically costing three to five times more than input tokens at equivalent volume.

Most enterprise AI providers bill on a per-million-token basis. As a working reference across major frontier models in mid-2026:

- **Input tokens**: approximately $1–$15 per million tokens, depending on the model tier[1]
- **Output tokens**: approximately $5–$75 per million tokens[2]
- **Cached input tokens**: 80%–90% cheaper than standard input where caching is supported[3]

The implication is direct: a model configured to produce verbose, fulsome responses is materially more expensive to operate than one tuned for conciseness, even when handling identical tasks. This is also true when performing agentic tasks; as a task adds more discrete steps, the total token cost will be significantly higher.

One of the most frequently overlooked dynamics in AI cost management is that token consumption varies dramatically by the type of work being performed, independent of the volume of text involved.

CONVERSATIONAL CHAT

A user asking a question and receiving a text answer is a relatively low-token interaction. The costs are bound by the length of the conversation and grow predictably with usage. For example, a customer-facing support chatbot has predictable token usage given the fact that the only possible interaction is question and answer, not agentic tool use.

AGENTIC AND COMPUTER-USE TASKS

When AI moves from answering questions to taking actions, browsing the web, operating software, editing files, or executing multi-step workflows, each action generates its own input and output cycle. A workflow requiring 10 steps to complete a single request carries 10 times the base token cost, plus any overhead from tool outputs injected back into context. If a model is asked to edit an Excel file, that model must load the tools to parse, understand, and modify Excel files on top of the requested analysis.

Organizations building agentic systems should budget for this non-linearly: an AI that "does things" rather than "says things" should be expected to consume tokens at an order of magnitude higher rate than a conversational assistant.

EXTENDED THINKING

Many frontier models—industry-leading models that push the limits of AI technology—now offer "extended thinking" or reasoning mode configurations, which enable AI tools to take longer to complete queries to provide more thorough responses. The tradeoff is straightforward: extended thinking uses significantly more tokens through internal chain-of-thought processing before producing an output. It reaches a higher-quality answer at higher total cost and often faster than a typical back-and-forth cycle of conversations.

Operators and product teams need a clear policy regarding when extended thinking should be used. Extended thinking is appropriate for high-stakes decisions, complex analysis, or tasks where answer quality is the dominant concern. Meanwhile, standard conversational chat and agentic modes should be used for high-volume, lower-complexity tasks where speed and cost matter more.

Applying extended thinking indiscriminately across a product is the equivalent of routing every database query through the most expensive compute tier regardless of complexity. The fix is a deliberate mode selection policy, not a blanket default in either direction.

THE SHIFTING COMPETITIVE LANDSCAPE

Token economics is one of the fastest-moving cost curves in enterprise technology, with many factors affecting pricing. When selecting AI models, organizations need to hold several dynamics in view simultaneously.

MARKET COMPETITION

The combination of competition among leading labs, the maturation of open-source models, and the emergence of heavily discounted overseas providers is driving down token prices. Guidance on model selection and pricing tiers from 12 months ago may already be materially wrong.

OPEN-SOURCE MODELS

As of mid-2026, open-source models run roughly six to eight months behind the frontier in capability terms.[4] Within the next 12–18 months, locally-hosted open-source models will likely be capable enough to complete most agentic and workflow tasks, with some experts claiming the technology is already there.[5] An organization that can run inference on its own hardware will pay for computations and operations rather than per-token application programming interface (API) costs, fundamentally changing the unit economics of AI at scale.

OVERSEAS APIS

Providers such as DeepSeek offer frontier-class capabilities at 1%–5% of Anthropic and OpenAI pricing.[6] For organizations without data residency constraints, export control considerations, or regulatory requirements tied to model provenance, overseas providers represent a legitimate cost savings lever. Businesses shouldn't default to tier-one providers by convenience, though. Before any decision is made, the tradeoff across cost, data security, vendor stability, and geopolitical risk must be assessed.

IMPLEMENTING BEST PRACTICES TO CONTROL COSTS

When reviewing an organization's AI cost structure, the following four categories represent the largest sources of addressable waste.

- **System Prompt Bloat –** Every API call includes the full system prompt. A 2,000-token system prompt sent 10,000 times per day generates 20 million input tokens daily before any user interaction occurs. Many organizations build system prompts organically over time and never conduct a structured audit. Condensing a system prompt reduces the ongoing input token cost for that deployment and often does not reduce output quality, especially as models improve.
- **Context Window Carryover –** In conversational chat applications, the full conversation history is typically resent with each new message. The cost of a 30-turn conversation is not 30 times the cost of a single turn; it is the cumulative sum of an ever-growing context window input. Organizations building internal chatbots without conversation truncation or summarization strategies will see costs compound non-linearly as user sessions lengthen and adoption grows.
- **Retrieval-Augmented Generation Overhead –** To handle very large amounts of information in a finite context window, retrieval augmented generation (RAG) is used. RAG is the ability to store and recall data as needed, and RAG pipelines inject retrieved information into the context at every call. A retrieval that pulls five 500-word chunks adds roughly 3,000–4,000 tokens per request. When retrieval recall is poorly calibrated and returns irrelevant content alongside relevant content, the client is paying for noise at scale. Tuning the retrieval parameters is often a high-return optimization that sits entirely within the client's control.
- **Agentic Retry and Validation Loops –** Workflows that call the model multiple times for output validation, self-correction, or structured response parsing—extracting unstructured data into structured responses—multiply token consumption. A workflow making three model calls to complete one user task triples the effective cost, and that is before accounting for the compounding context growth described above. To limit costs, agentic pipelines should be designed with explicit call budgets and fail-fast logic rather than open-ended retry tolerance.

These structural cost drivers can be easy to fall into and difficult to reverse. However, by following best practices and implementing initiatives across four key categories, token usage can be made predictable.

DESIGN – ARCHITECTURAL DECISIONS

- Structure prompts to maximize caching by placing static content first
- Establish a windowing, summarization, or hard truncation context management strategy
- Ensure RAG chunk sizing and retrieval precision with k-value tuning and reranking
- Enhance model tier selection with task-based routing rather than blanket frontier model use
- Implement a policy that outlines what task categories require standard or extended modes

CONFIGURATION – MODEL-LEVEL CONTROLS

- Set maximum output token ceilings per use case
- Enforce model and mode selection by use case, not by user preference

MEASUREMENT – INSTRUMENTATION

- Track token consumption by workflow, model, and user cohort
- Analyze cost per completed task, not cost per API call
- Review agentic workflow step counts, as they serve as a leading indicator of cost escalation
- Assess output-to-input token ratio to identify high ratios that may signal verbose generation needing investigation

GOVERNANCE – OPERATIONAL DISCIPLINE

- Build a prompt versioning and structured review cadence
- Communicate and enforce a model tiering policy across teams
- Create Spend alerts by application and team
- Assess the vendor diversification strategy periodically as capabilities and pricing evolve

Taking cost control measures across any of these four categories can help manage AI spend, but there are specific strategies that have been shown to have the greatest impact.

EDUCATING USERS ON OPTIMAL AI USE: THE HIGHEST-LEVERAGE INTERVENTION

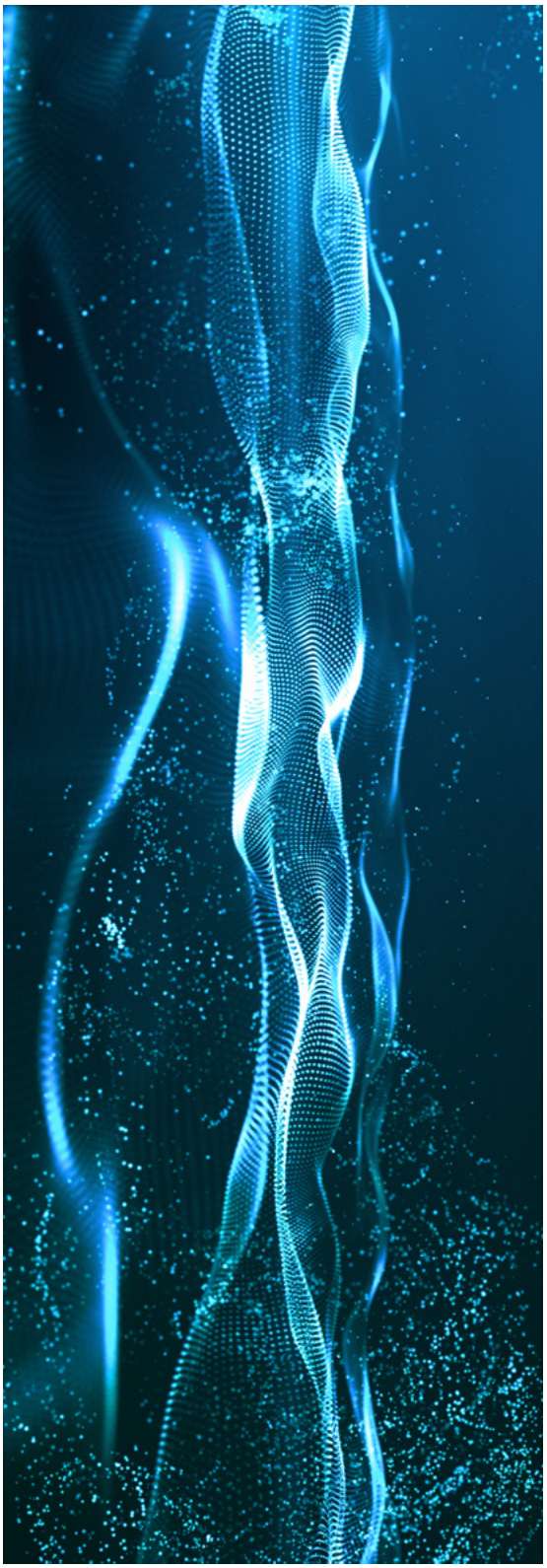
The single highest-leverage intervention in AI cost management is not technical optimization at all; it is teaching users which model to reach for. Frontier models may be powerful, but they are expensive, and the majority of everyday tasks do not need them. A workforce that understands the relative strengths of each model and the cost difference between tiers can route work to the cheapest tool that can do the job well.

This intuition is one users already apply to ordinary software. Nobody opens Excel to add two numbers together, a calculator does the job instantly and a spreadsheet is overkill, even though Excel can technically run the same calculation. Models work the same way: they are specialized variants of an underlying capability, and matching the model to the task is a skill that can be taught. A short internal guide pairing common task types with a recommended model tier, reinforced through onboarding and everyday practice, shifts consumption patterns faster and more durably than any back-end change.

However, education works only when it is paired with guardrails. Hard usage caps and active consumption monitoring are table stakes for deployment. They prevent runaway spend, surface misuse early, and generate the data the organization needs to refine its guidance over time. Caps set the ceiling, monitoring shows where the money is going, and education shifts the default behavior underneath both.

Because this lever operates on user behavior rather than infrastructure, it scales across every workflow at once and compounds as adoption grows. Technical optimizations such as prompt structuring and caching remain worthwhile, but they only tune the cost of work that has already been routed to a given model. Getting that routing decision right in the first place is where the largest and most durable savings come from.

TRACKING TOKEN EFFICIENCY: BOARD-LEVEL FRAMING AND ENGINEERING TEAM BENCHMARKS



To accurately control AI spend, organizations must be able to consistently track usage and costs. For PE clients, the relevant frame is unit economics, not absolute spend. The question is not how much the client is spending on AI, but what the cost per unit of value produced looks like and whether it is trending in the right direction as volume scales.

Token efficiency belongs in the same category as cloud infrastructure utilization or software license harvesting. It is an operational key performance indicator that should be owned, measured, and optimized on a regular cadence. A client spending $50,000 per month on AI that is driving measurable throughput in underwriting, diligence, or operations may be dramatically underinvesting. Meanwhile, a client spending $20,000 per month with no instrumentation, no caching, and no model tiering strategy is almost certainly overspending by a large margin relative to what sound architecture would cost.

The organizations that build this discipline now will be positioned to capture the next wave of cost deflation as open-source and locally hosted models mature. The ones who do not will find themselves with a cost structure that is opaque, unmanaged, and increasingly difficult to justify to investors.

According to proprietary A&M data, for organizations deploying AI tools to software engineering teams, internal benchmarking is approximately 25–30 million tokens per engineer per month.[7] This covers the range of typical engineering use cases: code generation, code review, documentation, and technical ideation.

Raw consumption benchmarks are necessary but not sufficient. Engineers need to develop the judgment to select the appropriate model for each task. Defaulting to a frontier model for every request is one of the most common and most avoidable sources of AI spend waste. A capable mid-tier model like Claude Sonnet or GPT-5.4, for example, handles most code completion, summarization, and drafting tasks at a fraction of the cost. Higher-tier models such as Claude Opus or GPT-5.5-class models should be reserved for strategy, system design, architecture, and other longer-horizon tasks that genuinely require extended reasoning, complex judgment, or novel synthesis.

THE PRICING DYNAMICS OF AI TOKENS

PRICE UPWARD PRESSURE
Capability Leadership, Demand

PRICE DOWNWARD PRESSURE
Supply, Competition, Self-hosting

The Balance of Forces: AI token pricing is the result of competing forces. When supply and self-hosting pressures dominate, prices fall. When capability leadership and demand expansion dominate, prices rise.

1

BARGAINING POWER OF FRONTIER MODEL PROVIDERS

Newer iterations of model capability (GPT 5.6, Fable, etc.) drive pricing up.

For a period of time, leading labs can command premium pricing until others catch up.

Higher quality, more capable models justify higher prices – but only temporarily. Advantage erodes as competition closes the gap.

2

DEMAND: AGENTIC USE DRIVES TOKEN CONSUMPTION UP

Agentic use will drive token consumption up as users rely on models to do more tasks on their machines.

More use cases, longer workflows, more autonomy = more tokens.

More value realized from models drives more usage, expanding the total market.

Improved model intelligence drives higher usage (up)

Lower prices attract more usage over time (up)

Alternatives increase supply and competition (down)

High usage incentivizes alternatives (down)

3

THREAT OF SELF-HOSTING & DISINTERMEDIATION

As highly capable models become available from open-source / overseas providers, there may be a string push towards self-hosting and not paying provider API fees.

Reduces willingness to pay and shifts spend away from APIs.

Increased self-hosting pressure caps prices providers can charge and shifts value to infrastructure.

4

SUPPLY POWER: OVERSEAS & OPEN-SOURCE MODELS DRIVE PRICES DOWN

Overseas APIs and open-source models increase supply and drive token prices down.

More providers, more access, more competition.

Greater supply and lower switching costs put downward pressure on prices.

LOOKING AHEAD: TOKEN BUDGETS AS PART OF THE EMPLOYMENT AGREEMENT

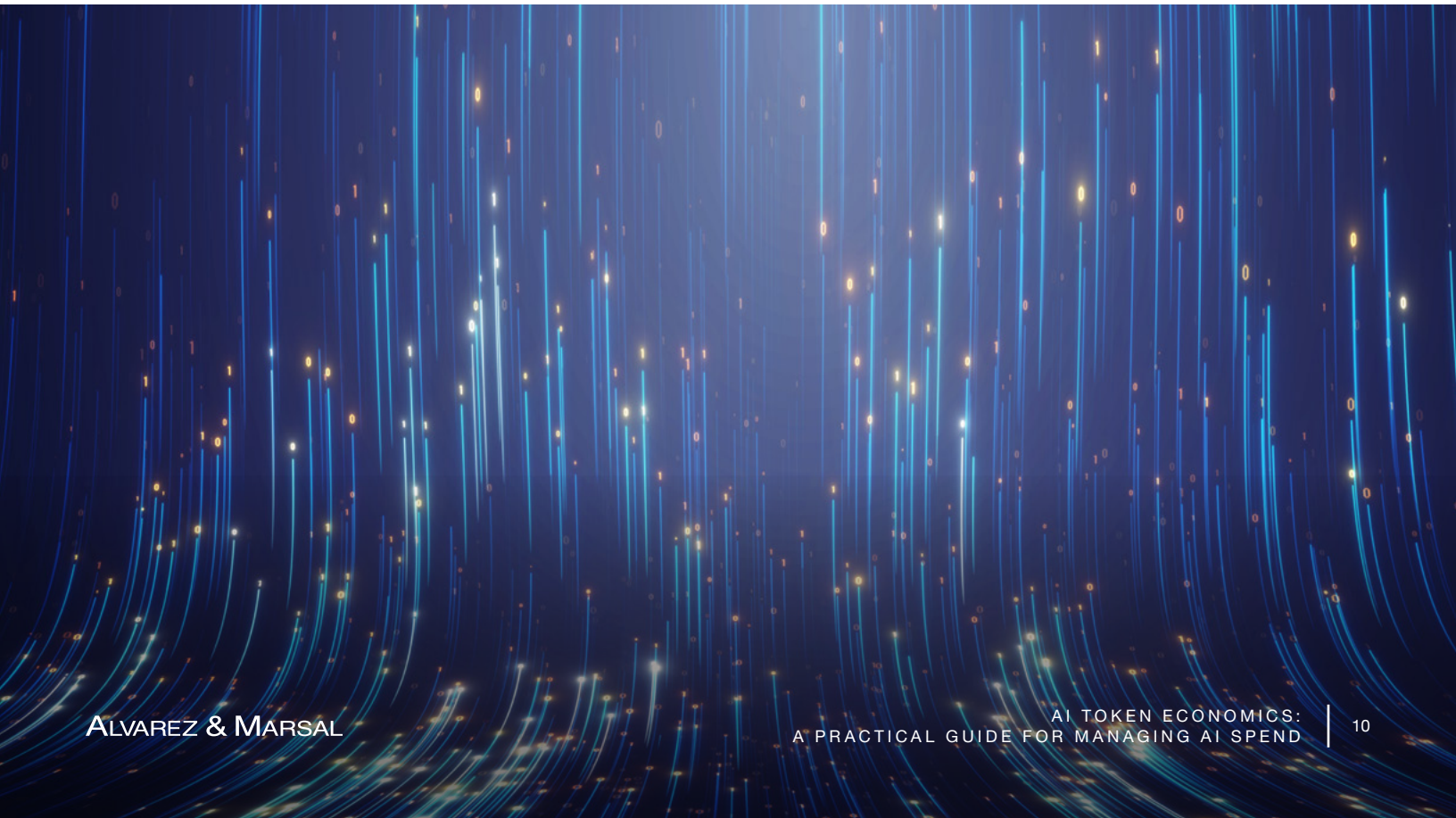
As AI becomes embedded in everyday work, token consumption is likely to become an explicit part of how organizations define and resource roles. Just as employees today are provisioned with a laptop, software licenses, and a budget, it is foreseeable to expect that future employment agreements and role definitions will specify a token budget or monthly allotment calibrated to the work the role performs.

This shift will be driven by the same logic that governs every other operational input: token access will increasingly determine how productive an individual can be. A knowledge worker with a generous allotment and the judgment to deploy it well will out-produce a peer who is rationed or untrained, much as access to better tooling has always separated high and low performers. Organizations will need to decide how to allocate that budget across roles, how to handle overages, and how to treat token efficiency as a dimension of individual and team performance.

For now, the practical implication is to begin instrumenting consumption at the level of teams and individuals, not just applications. The organizations that understand who is using tokens, for what, and to what effect will be the ones positioned to turn token budgets into a deliberate lever for productivity rather than an uncontrolled and opaque cost.

HOW CAN A&M HELP?

A&M's Private Equity Performance Improvement practice helps PE portfolio companies enhance their AI use. As your partner, we specialize in guiding clients through this critical integration to unlock value. Our AI services offer a tried and tested methodology for introducing AI technology and educating users through a hands-on, collaborative approach. With the help of our seasoned professionals, clients can ensure their AI spend and usage is optimized for their specific needs and tasks.



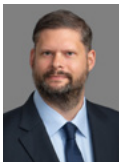
ENDNOTES

1. LLM API Pricing 2026: 20+ Models, Cost Per Token
2. What Is Token-Based Pricing for AI Models | MindStudio
3. Prompt caching | OpenAI API
4. Open Source & the Bifurcated AI Frontier
5. Open-Source vs Closed-Source AI Models: Which Should You Use for Agentic Workflows? | MindStudio
6. Low-cost Chinese AI models like DeepSeek gain traction in the U.S. - Rest of World
7. For reference, this is roughly one million tokens per working day. With many models now offering context windows approaching one million tokens, a single coding session can span an entire day's work, allowing the model to keep the full context of what an engineer is building in view throughout.

CONTACTS

AUTHORS

ROBERT KREFT
Managing Director
rkreft@alvarezandmarsal.com



JONATHAN ROSS
Managing Director
j.ross@alvarezandmarsal.com

AZMAT MOHAMMED
Senior Director
amohammed@alvarezandmarsal.com

VAMSI NADIMPALLI
Senior Director
vnadimpalli@alvarezandmarsal.com

MICHAEL ROTH
Senior Associate
m.roth@alvarezandmarsal.com

ABOUT ALVAREZ & MARSAL

Founded in 1983, Alvarez & Marsal is a leading global professional services firm. Renowned for its leadership, action and results, Alvarez & Marsal provides advisory, business performance improvement and turnaround management services, delivering practical solutions to address clients' unique challenges. With a worldwide network of experienced operators, world-class consultants, former regulators and industry authorities, Alvarez & Marsal helps corporates, boards, private equity firms, law firms and government agencies drive transformation, mitigate risk and unlock value at every stage of growth.

To learn more, visit: **AlvarezandMarsal.com**

Follow A&M on:

WHICH MODEL CLASS TO USE FOR WHICH TASK

A PRACTICAL GUIDE FOR ROUTING AI WORK TO THE RIGHT TIER

SMALL/FAST

Use cases: Lightweight text interaction and routine language tasks

- Email drafting and editing
- Simple Q&A / chat
- Summarization
- Rewriting and tone adjustment
- Short-form content generation
- Basic extraction or classification

Rule of thumb: If the task is mostly words in, words out, and low risk, start here.

Models:

- Anthropic: **Claude Haiku**
- OpenAI: **GPT-5.4 nano**

MID-RANGE GENERAL PURPOSE

Use cases: Everyday knowledge work, coding, and limited tool use

- Code writing and code completion
- Code review and debugging support
- Lower-complexity tool calling
- Document drafting
- Internal assistant workflows
- Structured analysis with moderate nuance

Rule of thumb: Use this tier for most productive work that is more than basic chat, but does not require deep reasoning or complex orchestration.

Models:

- Anthropic: **Claude Sonnet**
- OpenAI: **GPT-5.4**

HIGH-CAPABILITY

Use cases: More demanding analysis, multi-step tasks, and agentic work

- Web browsing and research
- Mathematics and quantitative reasoning
- Excel / spreadsheet tasks
- Agentic engineering workflows
- Multi-step tool use
- Complex code-writing tasks
- Higher-judgment analysis

Rule of thumb: Use this tier when the task involves multiple steps, external tools, or greater precision and complexity.

Models:

- Anthropic: **Claude Opus**
- OpenAI: **GPT-5.5**

FRONTIER

Use cases: Strategy, planning, scoping, and highest-complexity work

- Strategy development
- Planning and prioritization
- Scope definition
- System design
- Architecture
- Complex synthesis
- Novel or ambiguous problem-solving

Rule of thumb: Use the highest tier only when the work truly requires advanced reasoning, broad synthesis, or high-stakes judgment.

Models:

- Anthropic: **Claude Fable**
- OpenAI: **GPT-5.6**

QUICK DECISION GUIDE

Ask three questions before selecting a model:

1. *Is this a simple text task?*
2. *Does it require coding, tools, or multi-step work?*
3. *Does it require deep reasoning, strategy, or high-stakes judgment?*

If the answer is no to all three, use a lower-cost model.

If the answer is yes to the third, use the highest appropriate tier.

COST CONTROL PRINCIPLES

- Route work by task type, not user preference
- Use the lowest-cost model that meets the need
- Reserve frontier models for the few tasks that truly need them
- Monitor usage by workflow, team, and model tier
- Apply explicit guardrails for agentic and multi-step tasks