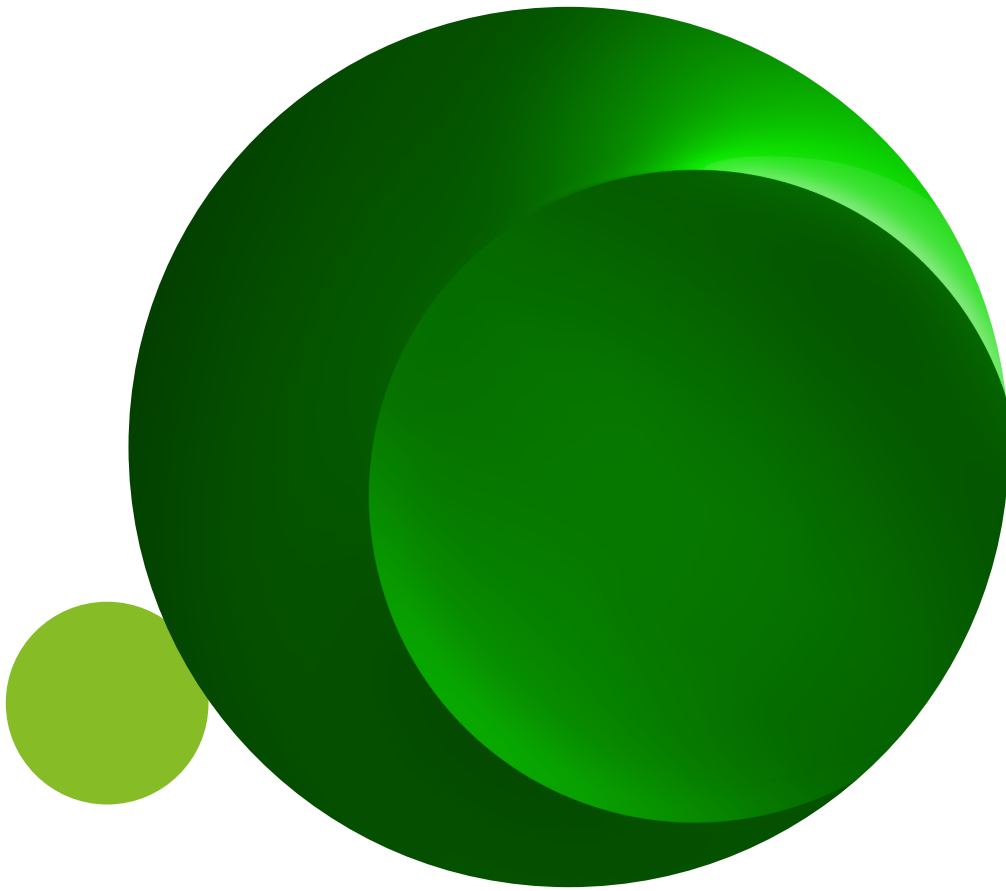




The lifecycle economics of AI

Managing AI as an asset across its lifecycle

AI value needs managing across its lifecycle like any other technology asset. But the spend does not stop changing once a contract is signed. Architecture, model choice and the way people use it can all change what an organisation ultimately pays.

EXECUTIVE PERSPECTIVE

Why AI cost governance matters now

An artificial intelligence (AI) service can cost more than planned even when the negotiated rate is competitive. An application may process longer inputs, retry failed tasks or use a different model while the contract stays unchanged.

AI cost is not settled at signature. It is set and reset across the life of the service it contributes to, at classification of the use case, when contracting the technology, in running the service and when retiring the technology, and each of those stages has different contributors. Managing one stage well does not help if the others are left unmanaged. Adoption has run ahead of governance, and many organisations are still working out how to forecast AI spend, allocate it and show what they get in return. Deloitte UK's GenAI Workforce Survey, the largest study of workplace generative AI (GenAI) use conducted in a single country,

asked 25,000 UK workers about their use in May and June 2026. Of those using GenAI at work, 31% did so without their employer's knowledge.¹ The market is moving just as quickly, with models changing, prices shifting and usage data often fragmented.

Governance across those stages depends on a governed record, and that record should sit within the organisation's existing asset management, extended for AI. IT Asset Management (ITAM) functions should hold the inventory of AI applications, models, suppliers and contracts and carry it from classification through to retirement, as they already do for software and hardware. Without it, a price change, a depreciation notice or an entitlement question cannot be traced to the applications and owners affected, and the lifecycle breaks into separate decisions that no one can connect.

Key Takeaways

01 Cost is set by design and use, more than the negotiated rate.

Model choice, context length, output length, retries and agent runs change what an organisation pays while the contract stays the same.

02 The bill is affected by every lifecycle stage.

Service classification, technology contracting, service operation and retirement each affect the cost base. No single function can manage the AI bill on its own. A cross-stage asset lifecycle governance model is needed.

03 Shift the controls left, before a service scales.

Set usage budgets, spending thresholds and decision rights while workloads remain small and, most importantly, buy the right technology for the right purpose. Retrofitting controls later isn't as effective.

04 The Business Owner carries the value decision but needs relevant inputs.

Procurement, ITAM and FinOps (including Finance) should supply the cost, asset, dependency and consumption data, as well as the commercial context. Chief Technology Officers (CTOs) set technology direction so the Business Owner can make an informed value decision.

THE COST MODEL

First things first: What determines the bill

This article covers generative AI services, whether accessed through application programming interfaces (APIs), embedded in software or run by agents across several steps and tools, and self-hosting as a sourcing option. **Tokenomics** here means how usage becomes cost, and how that cost connects to business value.

AI expenditure spans software-as-a-service (SaaS) subscriptions, managed model services and owned infrastructure. Each requires different controls. Seat-based services call for scrutiny of purchased licences and active use. Credit packages require buyers to understand how actions consume the allowance and how unused credits or overages are treated. Usage-based APIs depend on the work actually performed. Reserved or self-hosted capacity makes utilisation particularly important.

Seat-based SaaS spend can often be estimated from user numbers. AI API and agent spend is less forgiving. For text models, a token is a unit of text, such as a word fragment or punctuation mark. Model choice, context length, output length and the number of calls needed to finish a task all affect consumption.

Some reasoning models also generate tokens that are absent from the final answer but billed as output. The treatment depends on the supplier and model, so a short response can cost more than its visible length suggests.²

Consumption becomes a charge through **input and output rates, caching rules, tool fees** and other service terms. The total invoice may also include **subscriptions** and **minimum commitments**, regardless of actual use. For agents, one business task can span several models and tools, so a forecast has to follow the complete task path, including failures and retries.

Build a base forecast and a higher-consumption case using task volumes, tokens and calls per task, retries, cache reuse, tool charges and contractual commitments. Revise the forecast when application design, model selection, pricing or demand changes, and explain deviations through those drivers.

Spending is already significant. Gartner forecasts worldwide end-user spending on AI models and platforms of approximately **US\$64 billion in 2026**, an increase of **approximately 63%** on 2025.³

01 **Model** which model handles the request

02 **Context** how much input the model receives

03 **Output** how much text the model generates

04 **Model calls** how many calls a task needs

05 **Reasoning** working tokens billed but never shown

US\$ **64**bn

estimated global AI models and platforms spend in 2026³

63%

increase on 2025³

Up to **50%**

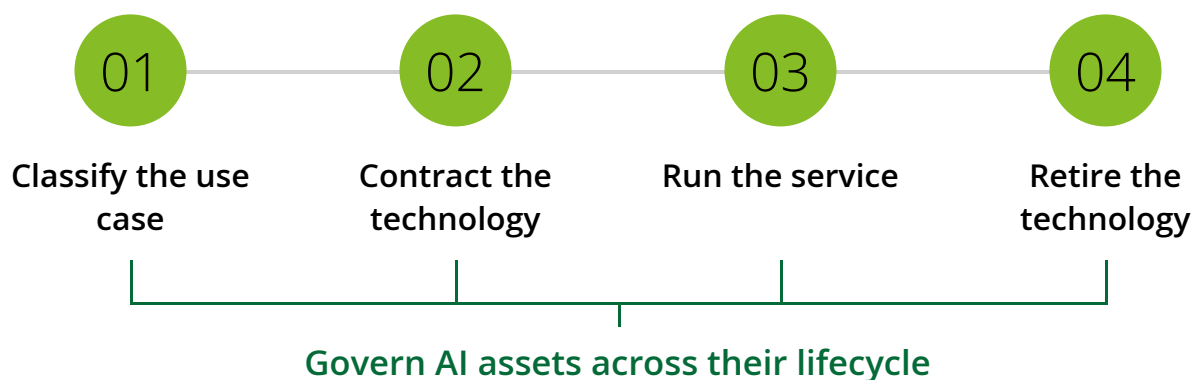
of digital initiative budgets allocated to AI at some surveyed firms⁴

LIFECYCLE CONTROL

Manage AI through its lifecycle, and shift control left

By the time an unexpectedly high invoice arrives, the architecture and the model are usually already in place. That is why control needs to be in place earlier, before multimillion commitments are signed. Practitioners call this **shifting left**. Designing **tokenomics** controls only once the service is

running (Stage 3) is too late. Controls need to be embedded when the use case is first classified (Stage 1), ahead of the commercial deals (Stage 2). It is also important to build control in around retirement (Stage 4) and avoid a new era of technical debt.



Any technology asset governance needs to run the **full length of the lifecycle**. ITAM, as a discipline, is well placed to define this governance framework. It should already maintain the record of models, model versions, suppliers, contracts, renewal dates and dependent applications, and keep it current. AI assets need all of this, but they do not sit still. Suppliers retire models on their own schedule, and a version can change behind a name that stays the same. Future agent access could also introduce new licensing or metering requirements.⁵

Asset record-keeping processes built for slow-moving stock have to adapt to this pace of change.

As part of this governance, the Business Owner remains accountable for the value decision associated with the use case, but the CTO must take the lead in determining technical suitability and value for money. This direction must be **informed by ITAM, Procurement and FinOps, which sit across the lifecycle**. FinOps also leads on cost guardrails.

LIFECYCLE CONTROL

The suggested mapping of actions below draws on the FinOps Foundation personas⁶ and other IT Asset Management best practice. Organisations should map it to their own delegated powers.

The Goal	Action	Suggested Lead
Identify the right AI tools	Regularly assess your AI tool options for value for money, and communicate direction to your Business Owners early	CTO
Use AI to generate value	Define the business outcome and success criteria for each use case, and use CTO direction to select the right technology	Business Owner
Maintain commercial resilience	Review suppliers, commitments, concentration risk, data residency and exit rights and communicate this to the CTO	Procurement
Set cost guardrails	Set guardrails for token budgets, spending thresholds and approval rights	FinOps
Provide AI visibility	Build and maintain the inventory of AI applications, models, suppliers, contracts and dependencies	ITAM
Measure AI consumption	Meter consumption by application, team and use case	Engineering
Manage the AI's exit	Review realised value and decide to continue, migrate or retire	Business Owner

STAGE 1, CLASSIFY THE USE CASE

Classify the use case and define value

01 Classify the use case

02 Contract the technology

03 Run the service

04 Retire the technology

Start with the outcome

(Business Owner, CTO, ITAM and Procurement)

Every use case requires a named Business Owner to define the intended outcome, the existing process baseline and the conditions for success. Specify acceptable quality, response time and risk alongside expected demand. These requirements establish what the organisation is buying and how its performance will be judged.

Check what the organisation already owns. AI features often sit inside software it already licenses, and an existing entitlement may cover the use case without a new purchase. ITAM usually holds this answer, and asking early avoids paying twice for the same capability.

Procurement belongs in this stage as well. Different AI applications, such as customer-service assistants, code-review tools or research services, may necessitate distinct models, controls and buying routes. Shaping these choices is more effective before requirements are finalised.

Set the measure of value

(CTO, Business Owner, ITAM, FinOps and Procurement)

Measuring value remains difficult, but that does not require false precision. Unit economics expresses cost per defined unit of work or business result. A useful measure is cost per customer case resolved or document accepted to the required quality standard. The Tokenomics Foundation's value classification makes this explicit. Only outputs that clear an agreed quality floor count, and the review, triage and rework the AI adds are subtracted from the benefit.⁷ The CTO should define the value-for-money measures for each category of outcome the Business Owners are pursuing, and ITAM, FinOps and Procurement supply the data behind this.

Value assessment also needs to separate time released from financial benefits realised. Faster work may create useful capacity without reducing expenditure. Deloitte UK's survey puts the average saving at 70 minutes a week and finds that most of it goes into more work for the same employer.¹ The Tokenomics Foundation's classification treats freed capacity that is not redeployed as zero realised benefit until it produces an outcome.⁷ The Business Owner should explain how that capacity will be used and review the evidence as adoption develops.

Tech Tip: In setting value measures, CTOs should apply a consistent cost boundary and time horizon across AI options, including integration, data services, evaluation, operations and human review. For self-hosting options, account for accelerator capacity and its utilisation, energy, specialist skills and maintenance. Its relative cost depends on sustained demand, service requirements and operating efficiency.

STAGE 2, CONTRACT THE TECHNOLOGY

Contract for movement and changing demand



To compare suppliers, use the same representative tasks and acceptance criteria, evaluating them against the organisation's own expected demand and quality threshold rather than just headline rates. The commercial assessment should then test how an offer behaves under expected demand and under a higher-consumption scenario, because a discount tied to an unrealistic minimum can be costly.

Longer commitments may still be justified by steady demand, capacity assurance or support requirements. A standing charge plus a consumption rate can provide a middle ground, and reserved throughput may be worth negotiating where demand is genuinely steady.

Procurement should work with ITAM, Business Owners, Engineering, FinOps, Legal and Risk to settle six questions before signing.

- 01 **What is billable?** Define the charging units and the treatment of reasoning, failed requests, retries, tools and automated access to other software.
- 02 **What is committed?** Check minimum purchases, credit expiry, transfer rights, overage rates and the scope for resizing commitments.
- 03 **What performance is required?** Agree the relevant quality, latency, throughput and availability requirements, and how they will be assessed.
- 04 **What evidence is available?** Establish access to sufficiently detailed usage data, billing reconciliation and a process for resolving discrepancies.
- 05 **What can change?** Address price changes, model retirement, notice periods, transition support, data residency, data use and retention, and relevant usage rights.
- 06 **What would exit involve?** Examine data export, integrations, access controls, migration support and remaining financial obligations.

What the contract records decides what can be checked later. Entitlements, renewal dates, usage rights and reporting duties belong in the ITAM record at signature, rather than when the first question about them arrives.

Supplier dependency can extend beyond the model. Deloitte's *Four futures for technology infrastructure* explores possible futures in which control over identity, data access, orchestration and governance creates further dependencies.⁵ Because a relatively small number of suppliers account for much of the market, concentration risk, data residency and switching costs should be considered alongside price.

A practical exit assessment runs representative tasks with an alternative supplier and examines the changes needed around them. Record the effort required to migrate data, tools, permissions and monitoring.

STAGE 3, RUN THE SERVICE

Run on measurement and limits

01 Classify the use case

02 Contract the technology

03 Run the service

04 Retire the technology

Attribute the spend

(ITAM, FinOps and Engineering)

FinOps unites Finance, Engineering and business teams to manage technology costs and value. For AI, that collaboration should commence when design and commercial choices are still open and flexible. Assign an owner to each service identity, link usage to the application, team, use case and environment, and for agents connect model and tool calls through a common task or run identifier. Reconcile this telemetry with supplier invoices. Cost and value also need the same boundary. A value claim and the total cost of AI only combine into a return on investment when both describe the same workload.⁷

Some usage never reaches a central platform.

Shadow AI scatters cost and compliance exposure across expense claims, departmental subscriptions and cloud invoices, necessitating discovery across these sources as well.

Embedded AI needs its own check. ITAM reconciles purchased entitlements with assigned and active use, identifies overlapping add-ons and confirms which consumption sits inside the licence and which is billed separately.⁸ The rights required for agent access to existing software belong in the same review.

Where direct attribution is incomplete, use an agreed allocation policy. Separate unresolved gaps from shared costs that are deliberately held in a central budget. Clear ownership matters even when individual applications do not receive a separate charge.¹⁰

Set limits that hold

(FinOps, CTO and Engineering)

Budget alerts and enforceable limits serve different purposes. Microsoft, for example, states that Azure budget thresholds do not stop resource consumption.⁹ Agent controls should cover the whole run, including delegated tasks and tools, with limits on cost, duration, calls and retries. Define who can approve an exception and whether the service pauses, stops or transfers to an alternative process. Implement these controls in the application or platform, because a written instruction to an agent is not enough.

Engineering directly influences both cost and performance. Avoid defaulting to the newest or most capable model. Route each request to the least expensive approved model that meets the required quality, latency and risk thresholds. Deloitte UK's survey found the most common uses to be searching for information, drafting emails and creating summaries, each reported by around a third to just over four in ten users.¹ Everyday work of that kind rarely needs the most capable model available. Trim unnecessary context, assess prompt caching where repeated inputs justify its cost, and keep responses no longer than the use case requires. A broader model portfolio may improve efficiency, but its evaluation, routing and maintenance costs belong in the assessment.

STAGE 4, RETIRE THE TECHNOLOGY

Review, migrate and retire on purpose

01 Classify the use case

02 Contract the technology

03 Run the service

04 Retire the technology

Review value and commitments

(Business Owner, FinOps and Procurement)

The Business Owner should review realised benefits against the original baseline and targets, alongside cost and service performance, before deciding whether to continue or expand the service. Since no single cross-supplier commercial measure combines consumption, quality and business outcome, supplier efficiency claims must be validated against the organisation's own evaluation set and expected use-case outcome.

Review supplier commitments and sourcing assumptions at least quarterly. Monitor consumption more frequently where volatility warrants it, and reassess after material application or supplier changes. At renewal, compare the incumbent with current requirements, alternatives and switching costs, so that a renewal is a fresh competitive decision rather than an automatic extension.

Govern continually and retire on purpose

(ITAM, Business Owner, CTO, Engineering, Procurement and FinOps)

This is where the asset record earns its keep. A depreciation notice is only useful if the organisation knows which applications depend on the affected model, and who owns them. A model's retirement does not necessarily end the business use case, and a valuable application may need migration, regression testing and temporary parallel operation.

Where AI is self-hosted, the inventory should also cover graphics processing units (GPUs) or other accelerators, utilisation, platform dependencies, maintenance and end-of-life costs.

When a use case no longer delivers enough value to justify its consumption, the Business Owner decides whether it should be scaled back, migrated or retired. Engineering then confirms decommissioning, including scheduled runs and access paths. ITAM updates assets and entitlements, Procurement closes or adjusts the associated contract and FinOps checks residual charges. Data export, deletion and required retention follow the agreed process.

CONCLUSION

Govern the economics, and not just the contract

Lower token rates will not remove the need for this work. Less expensive units often encourage organisations to use more of the service, while more capable models and agents can increase the amount processed per task. Overall **spend can rise even as unit prices fall**.⁴

Many commercial controls still revolve around the contract and the renewal date. AI spend does not. Model choice, context length, application changes and agent runs can move the bill in normal operation, long before Procurement gets anywhere near a renewal.

Before consumption scales, organisations need a named owner, workload-level measurement, scenario-based forecasting, a benefit baseline and regular supplier review. The Business Owner holds the value decision, and the CTO sets the technology direction. Engineering shapes the workload. Finance and FinOps forecast and allocate the cost. ITAM carries the asset record across the lifecycle, while Procurement, Legal and Risk manage the commercial exposure around it.

When cost, usage and realised outcomes are managed together, the **AI bill stops being a surprise and becomes an investment decision**.

Contact us with any questions or to discuss your organisation's AI costs, supplier commitments and lifecycle controls!

- 01 Deloitte UK, "GenAI Workforce Survey", September 2026, deloitte.com.
- 02 OpenAI, "Reasoning models", API documentation, developers.openai.com, accessed September 2026.
- 03 Gartner, "Gartner Forecasts Worldwide AI Platforms and Models Market to Grow 63% in 2026", press release, 20 July 2026, gartner.com.
- 04 Merizzi, Smith, Kearns-Manolatos and others, "The pivot to tokenomics: Navigating AI's new spend dynamics", Deloitte, January 2026, deloitte.com.
- 05 Thomas, Patwari, Srinivasan and others, "Four futures for technology infrastructure", Deloitte Insights, July 2026, deloitte.com.

- 06 FinOps Foundation, "FinOps Framework: Personas", finops.org, accessed September 2026.
- 07 Tokenomics Foundation, "Classifying and Measuring Realized Value from AI", Release Candidate 2, September 2026, tokenomics.com.
- 08 FinOps Foundation, "FinOps Framework: Licensing & SaaS", finops.org, accessed September 2026.
- 09 Microsoft, "Create and manage budgets", Microsoft Learn, learn.microsoft.com, accessed September 2026.
- 10 FinOps Foundation, "FinOps Framework: Allocation", finops.org, accessed September 2026.

ABOUT THIS ARTICLE

Authors



Christoph Alkofer

Manager

AI Economics and FinOps Lead



Tom Lamplough

Associate Director

ITAM and Tech Value for Money, Secure Government Lead

KEY CONTACTS

Get in touch



Peter Kunorubwe

Partner

ITAM Practice Lead

+44 20 7007 9530

pekunorubwe@deloitte.co.uk



Anant Jain

Director

ITAM and Technology Cost
Optimisation

+44 20 7007 7236

anantjain@deloitte.co.uk



Tom Lamplough

Associate Director

ITAM and Tech Value for Money,
Secure Government Lead

+44 20 7303 5388

tmlamplough@deloitte.co.uk



Christoph Alkofer

Manager

AI Economics and FinOps Lead

+44 20 8039 8543

christophalkofer@deloitte.co.uk

With thanks to contributors

Charlotte Warburton

Partner

T&T AI Officer

+44 20 7007 8542

cwarburton@deloitte.co.uk

Komal Mestry

Director

AI Value Optimisation, Financial Services

+44 20 7007 9904

komalmestry@deloitte.co.uk

Lewis Keating

Director

AI Governance Lead

+44 113 292 1193

lkeating@deloitte.co.uk

Stephen McMaster

Partner

CTO and AI Engineering

+44 2890 53 1368

smcmaster@deloitte.co.uk

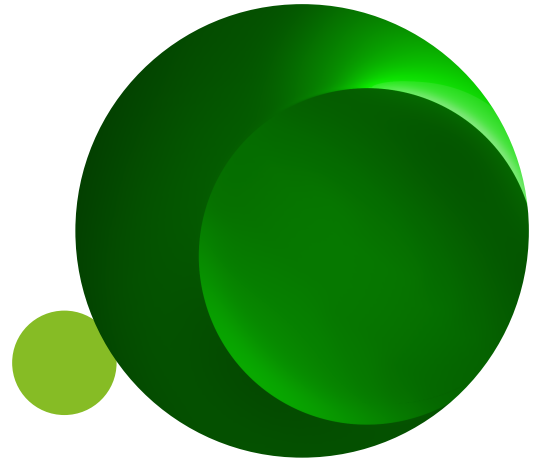
Lorna Weir

Director

AI Strategy and Governance,
Government and Public Sector

+44 117 984 2762

lweir@deloitte.co.uk



This publication has been written in general terms and we recommend that you obtain professional advice before acting or refraining from action on any of the contents of this publication. Deloitte LLP accepts no liability for any loss occasioned to any person acting or refraining from action as a result of any material in this publication.

Deloitte LLP is a limited liability partnership registered in England and Wales with registered number OC303675 and its registered office at 1 New Street Square, London EC4A 3HQ, United Kingdom.

Deloitte LLP's affiliate (Deloitte Management Services Limited) is a shareholder in Deloitte EMEA BV (Deloitte EMEA). Deloitte EMEA is a member firm of Deloitte Touche Tohmatsu Limited (DTTL). DTTL and Deloitte EMEA do not provide services to clients. Services may be provided by other Deloitte entities within the global network of member firms, each of which are separate and independent legal entities. Please see www.deloitte.com/about to learn more about our global network of member firms.

© 2026 Deloitte LLP. All rights reserved.