STRATEGY BRIEFING · 60 MINUTES

Inference Economics & the Strategy for AI Adoption

What AI actually costs to run, what it returns, and the operating model required to capture the value.

Prepared for the Executive Committee · Board-ready business case · September 2026

Structured on the Deloitte storyline method: Situation – Complication – Resolution – Business case – Activation

AGENDA

How this hour works

1	**Situation**	The economics of inference — what a token really costs and why the bill grows as prices fall	8 min
2	**Complication**	Where value leaks between the pilot and the P&L	6 min
3	**Resolution**	Seven levers on cost per outcome, and the sourcing decision	9 min
4	**Business case**	Build cost, run cost, maintenance, ROI, break-even, hiring	14 min
5	**Future strategy**	Three horizons and the no-regret moves	4 min
6	**AI Pioneers**	Steering committee, decision rights, upskilling and reskilling	7 min
7	**Decisions**	The ask — four approvals required today	2 min

10 minutes reserved for questions and the decision round at the close.

METHOD

The framework this deck follows

Deloitte's storyline discipline: answer first, structure MECE, land every page on a 'so what'. Each page below carries an explicit takeaway strip.

S

Situation

The facts everyone in the room already accepts. No argument, no defensiveness. Establishes the shared baseline.

C

Complication

What has changed or is breaking. Creates the tension that makes a decision necessary rather than optional.

Q

Question

The single question the executive committee must answer. Everything downstream serves it.

A

Answer

The recommendation, stated up front on page four and then defended with evidence.

Assessed across six dimensions on every use case

- Strategy & value
- Data & technology
- Talent & culture
- Governance & risk
- Operating model
- Unit economics

SO WHAT: We are not presenting technology. We are presenting a capital allocation decision with a cost curve, a return, and an operating model attached.

ANSWER FIRST

Executive summary

Recommendation: fund a six-use-case production portfolio with a hard cost-per-outcome ceiling, governed by a standing steering committee, and staffed with 12 net-new roles.

$4.1M	$5.3M	$14.2M	17 mo
One-time build across platform, data and six use cases	Annual run cost at steady state, including maintenance	Annual benefit at full ramp, risk-adjusted	Payback; 3-year NPV of $8.9M at a 12% discount rate

1. **Inference is not your biggest cost.**
   Compute is 25% of run cost. People, platform and assurance are 75%. Optimising tokens while over-staffing the run team is the classic error.

2. **Unit price falls, total spend rises.**
   Cheaper tokens unlock more use cases and longer agent chains. Budget for volume growth of 3-5x per year, not for price relief.

3. **The controllable metric is cost per resolved outcome.**
   Not cost per token. Routing, caching and context discipline move it by 5-10x without touching quality.

4. **Governance is the throughput constraint.**
   Pioneers do not have better models. They have faster decisions, clearer kill criteria and a workforce that can actually use the tools.

SO WHAT: If you read one page, read this one. Everything that follows is evidence for these four claims.

01

SECTION

Situation

The economics of running models have replaced the economics of building them. Most enterprise budgets have not caught up.

1. The cost centre has moved from training to inference
2. What is actually inside a unit of inference cost
3. Unit prices are collapsing at roughly 10x a year
4. Total enterprise spend is rising anyway

8 minutes

SITUATION

The centre of gravity has moved from training to inference

Training economics
2021-2023

- One-time, capitalised, concentrated in a handful of labs
- Cost is someone else's balance sheet
- Enterprises were spectators
- Budget question: 'should we build a model?'

Inference economics
2024 onward

- Recurring, variable, scales with every user and every workflow
- Sits in your operating budget, forever
- Enterprises are the buyers
- Budget question: 'what does each outcome cost to produce?'

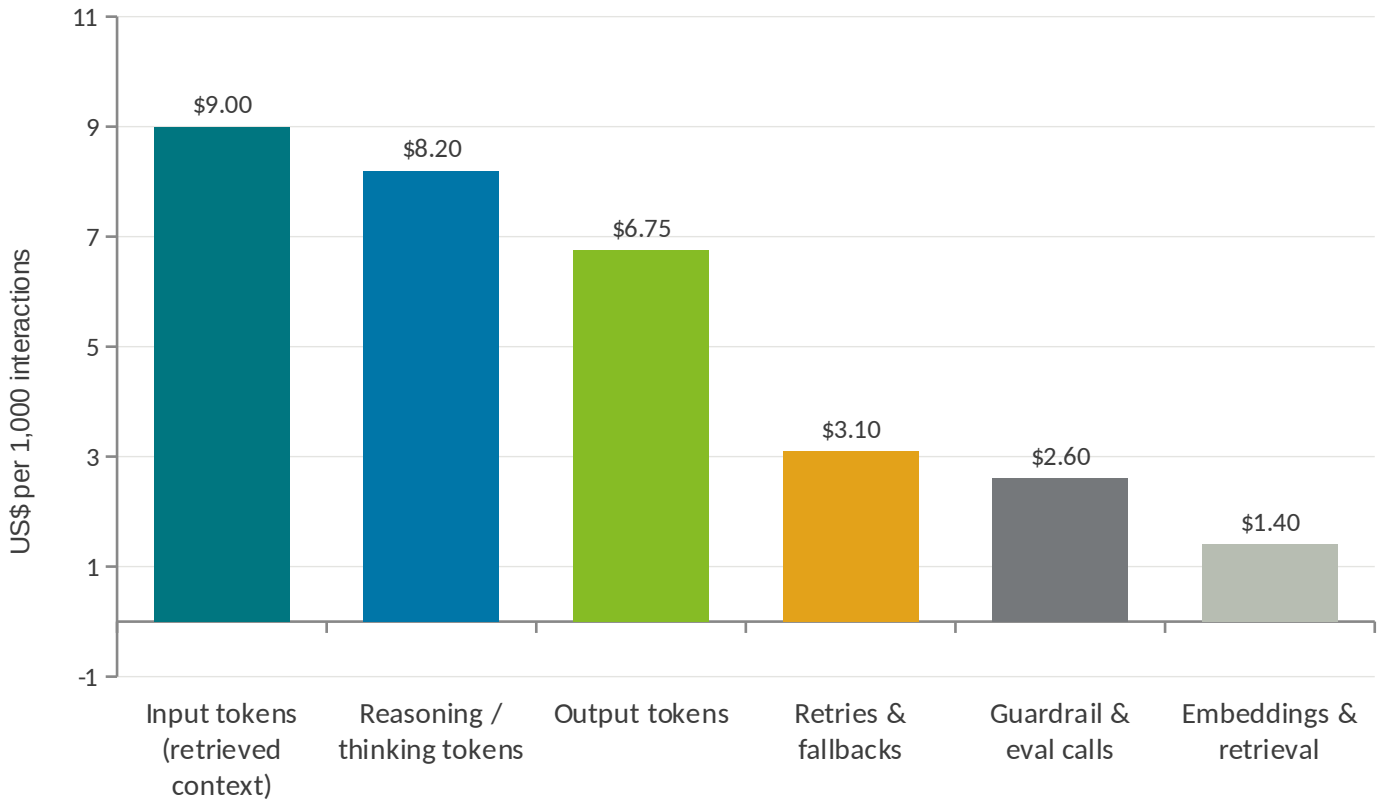
Why this matters for the way we budget

- Inference is a cost of goods sold, not a capital project. It belongs in gross margin, not in an innovation line item.
- It scales with adoption — the more successful the deployment, the larger the bill. Success is the risk event.
- It is elastic and controllable. Unlike headcount, a 60% cost reduction can be engineered in weeks, without renegotiating anything.

SO WHAT: Treat AI spend like cloud spend with a multiplier, not like a software licence. It is variable, it is forecastable, and it must be owned by someone.

SITUATION

Anatomy of an inference bill — what you are actually paying for



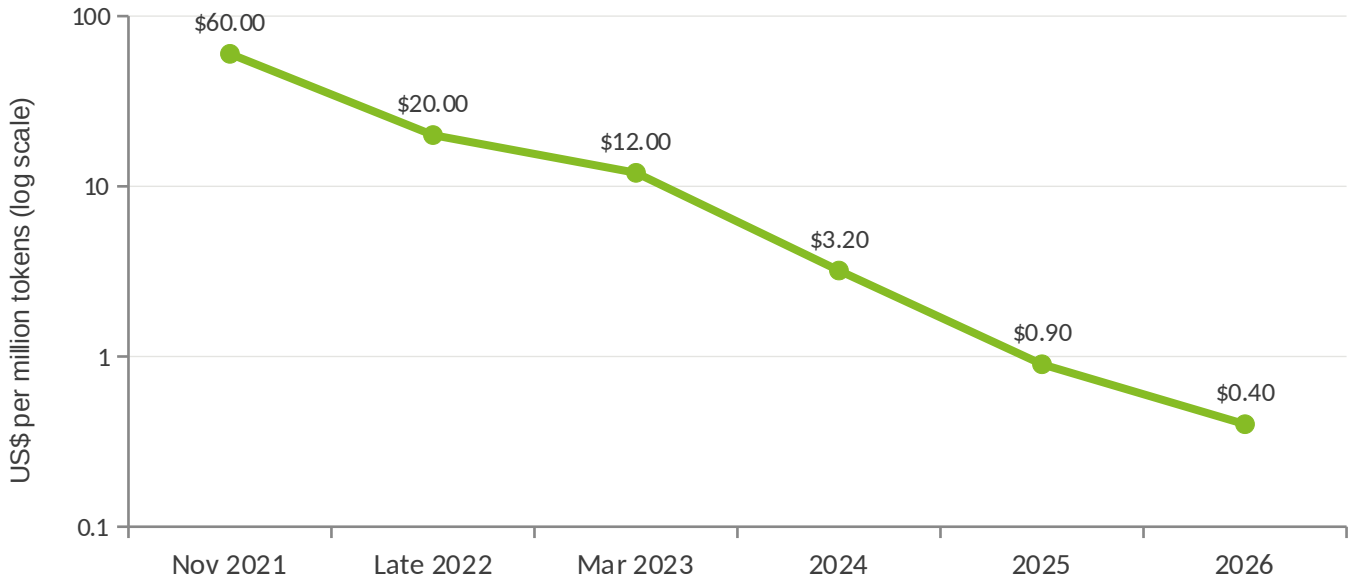
Four facts most business cases miss

- Output and reasoning tokens cost 3-5x more than input tokens. Verbose answers are expensive answers.
- Retrieved context dominates volume. A 40-page document stuffed into every prompt is paid for on every call.
- Retries, guardrails and evaluation calls add 15-25% that no vendor price list shows.
- Agentic workflows multiply all of the above by the number of steps in the chain.

SO WHAT: Roughly $0.031 per interaction here — trivial in isolation, $3.1M at 100 million interactions a year. Volume, not unit price, is the variable to govern.

SITUATION

The unit price of intelligence is collapsing



What is driving it

- Smaller, better-trained models reaching last year's frontier quality
- Quantisation from 16-bit toward 4-bit inference
- Serving software: batching, paged attention, speculative decoding
- Open-weight competition compressing margin across the stack

Analyses attribute most of the decline to software and algorithmic efficiency rather than cheaper hardware.

~1,000x

Fall in cost for a fixed capability benchmark, 2021 to 2024

~10x / yr

Typical decline for frontier-class capability

3-5x / yr

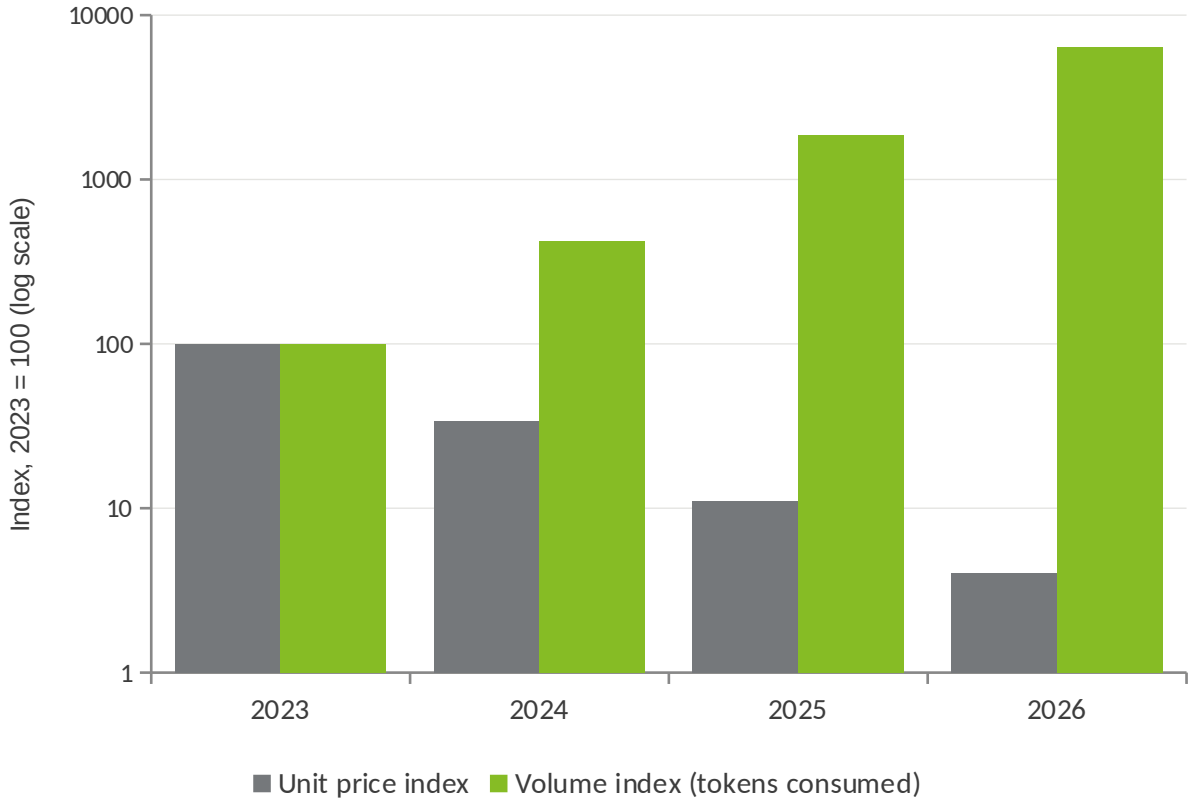
Realistic forward expectation through 2027, then tapering

4-5x

Output token premium over input tokens across providers

SO WHAT: Never sign a three-year business case that assumes today's token price. Assume the price you pay per unit of capability falls at least 3x a year — and assume competitors are modelling the same thing.

So why is the bill going up?



Jevons paradox, applied to tokens

When a resource becomes cheaper, we consume dramatically more of it. Every price cut makes a previously uneconomic use case viable, and each new use case is more token-hungry than the last.

Single-shot chat	~1K tokens per task
Retrieval-augmented answers	~10K tokens per task
Extended reasoning	~40K tokens per task
Multi-step agents with tools	**~250K+ tokens per task**

SO WHAT: Plan for the bill to grow 3-5x a year even as prices fall. The discipline is not to buy cheaper tokens — it is to decide which tokens are worth buying.

SITUATION

Four drivers determine your token bill — all four are controllable

1

Volume

Interactions per user per day × eligible users. Grows with adoption, which is the outcome you want. Govern by use case, not globally.

2

Tokens per interaction

Context length, retrieval strategy, reasoning depth, output verbosity. The most over-provisioned variable in most deployments.

3

Model tier

Frontier versus mid-tier versus small. A 30x price spread for tasks where quality often differs by a few percentage points.

4

Service level

Latency targets, redundancy, uptime and region. Sub-second interactive responses cost multiples of an overnight batch job.

The governing equation

Annual cost = interactions × tokens per interaction × blended price per token × (1 + retry & guardrail overhead)

Three of the four terms are engineering decisions we control. Only the price per token is set by the market — and it is falling. Any team that reports 'AI costs are out of control' is reporting an architecture problem, not a market problem.

SO WHAT: Every one of these four terms must have a named owner and a monthly number. Today, none of them do.

SITUATION

Where enterprises actually are

66%

Report productivity and efficiency gains from enterprise AI adoption

20%

Are already growing revenue from AI, against 74% who expect to

25%

Of leaders say generative AI is already transforming their business

6-12 mo

Typical time from pilot to defensible ROI, often longer

What the leaders have in common

- A small number of high-impact use cases, not a wide portfolio of experiments
- Generative AI layered onto existing processes rather than new processes invented around it
- Centralised governance to drive adoption and reuse
- Investment above 5% of IT budget — correlated with materially higher project success rates

What the laggards have in common

- Dozens of proofs of concept, none with a production owner or a budget line
- Benefits claimed in slideware, never traced to a cost centre or a headcount plan
- No unit economics — nobody can state the cost per resolved outcome
- Enablement treated as a training video rather than a workforce transition

SO WHAT: The gap between leader and laggard is not model access — everyone buys from the same four vendors. It is portfolio discipline, governance and workforce readiness.

02

SECTION

Complication

The technology works. The economics leak. Between a successful pilot and a P&L impact sit five specific failure modes.

1. The pilot-to-production gap
2. Five ways AI economics break
3. The costs nobody budgets for
4. The quality, latency and cost triangle

6 minutes

COMPLICATION

The pilot-to-production gap

100	30	9	3	1
Ideas raised	Pilots built	Reach production	Measured benefit	Scaled enterprise-wide

Why attrition happens at each gate

Idea → pilot	No value case. Enthusiasm is not a hurdle rate.
Pilot → production	No owner, no integration budget, no security sign-off. The pilot ran outside the systems that matter.
Production → measured benefit	Nobody defined the baseline before launch, so the saving cannot be proved and finance will not book it.
Measured → scaled	The economics worked at 200 users and broke at 5,000 because cost per outcome was never engineered.

SO WHAT: The failure is rarely the model. It is the absence of an owner, a baseline and a cost ceiling before the pilot starts.

COMPLICATION

Five ways AI economics break

1	**Frontier-model default**	Every call routed to the most capable model regardless of task difficulty.	**3-10x overspend**
2	**Context stuffing**	Whole documents pushed into every prompt because retrieval was never tuned.	**2-5x overspend**
3	**Unbounded agent loops**	No step budget, no cost ceiling, no circuit breaker on retries.	**Unbounded tail risk**
4	**Benefit leakage**	Time is saved but never converted into capacity, revenue or avoided cost.	**100% of the case**
5	**Shadow AI**	Uncontrolled consumer-tool use — invisible spend, invisible data exposure.	**Unmeasured, uninsured**

SO WHAT: Four of these five are engineering defaults, fixable in weeks. The fifth — benefit leakage — is a management failure and takes a quarter to fix.

COMPLICATION

The costs nobody puts in the business case

Data readiness

- Access controls and entitlement mapping
- Document cleanup and chunking strategy
- Retrieval index refresh and drift

Assurance

- Evaluation harnesses per use case
- Red-teaming and adversarial testing
- Human review of sampled outputs

Change & adoption

- Workflow redesign, not tool rollout
- Role-based enablement at scale
- Manager coaching and incentives

Vendor churn

- Model deprecation on 6-12 month cycles
- Re-testing prompts and agents after upgrades
- Contract and pricing renegotiation

Rule of thumb for a credible case

Inference / compute	**20-30% of run cost**	People (run team)	**35-45% of run cost**
Platform & licences	**10-15% of run cost**	Assurance, maintenance & governance	**20-25% of run cost**

If a business case shows compute as the dominant cost, it is incomplete.

SO WHAT: Budget maintenance from day one. Models deprecate, prompts rot, retrieval indexes drift — an unmaintained AI system degrades faster than any software we run.

COMPLICATION

Quality, latency and cost — pick the two the use case actually needs



No configuration optimises all three. The design job is to state which one the use case is allowed to sacrifice — explicitly, in writing, before build.

Customer-facing chat
Optimise: Latency + quality
Accept higher cost per interaction

Overnight document processing
Optimise: Cost + quality
Accept 8-hour turnaround, batch everything

Internal drafting assistant
Optimise: Cost + latency
Accept a good-enough draft with human edit

03

SECTION

Resolution

Cost per outcome is an engineering variable. Seven levers move it by an order of magnitude without a quality trade-off.

1. The seven levers, ranked by payback
2. Routing, right-sizing and distillation
3. Caching, batching and context discipline
4. Buy, rent or own — the sourcing crossover
5. FinOps for AI and the metric that matters

5 minutes

RESOLUTION

Seven levers on cost per outcome

Lever	What it does	Typical cost reduction	Time to value
Model routing	Send each request to the cheapest model that clears the quality bar	**40-70%**	Weeks
Right-sizing & distillation	Fine-tune or distil a small model for high-volume narrow tasks	**50-90%**	1-2 quarters
Prompt & semantic caching	Reuse repeated context and near-duplicate answers	**20-50%**	Weeks
Batching & scheduling	Move non-interactive work to batch windows and higher concurrency	**30-60%**	Weeks
Context discipline	Tighter retrieval, aggressive chunk limits, compressed history	**20-45%**	Weeks
Agent step budgets	Hard ceilings on steps, tool calls and retries per task	**15-40%**	Days
Output shaping	Structured, bounded outputs instead of free-form prose	**10-30%**	Days

Stacked realistically, these levers deliver a 5-10x reduction in cost per resolved outcome. They compound: routing to a smaller model also shortens latency, which lowers the cost of the retry budget.

SO WHAT: Levers one to three deliver most of the value. Do them before considering any infrastructure investment.

RESOLUTION

Levers 1-3: route, right-size, distil

1

Model routing

- Classify each request by difficulty before it is served
- Small model handles the 70-80% that are routine
- Escalate only on low confidence or explicit complexity
- Requires an evaluation set to prove no quality loss

2

Right-sizing

- Default to the mid-tier model, escalate by exception
- Reserve frontier models for reasoning-heavy work
- Re-benchmark quarterly — last year's frontier is this year's mid-tier
- A 30x price spread rewards getting this right

3

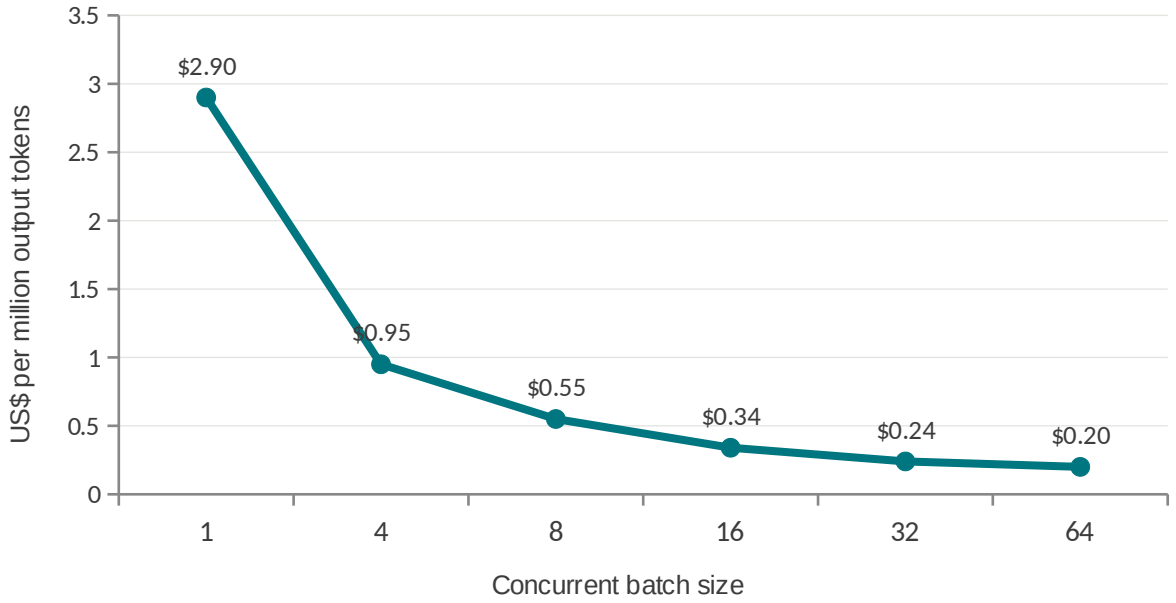
Distillation

- Use a large model to generate training data for a small one
- Viable once a workflow is stable and high-volume
- Delivers frontier-adjacent quality on a narrow task at a fraction of the cost
- Adds a maintenance obligation — budget for it

Worked example — routing a 40 million call per year workflow

Before: 100% frontier	**$0.042 per call**	**$1.68M / year**
After: 78% small, 22% frontier	**$0.013 per call**	**$0.52M / year**
Quality delta on the evaluation set	**−0.4 points**	**Within tolerance**

Levers 4-5: cache aggressively, batch everything that can wait



Utilisation is the whole game on owned infrastructure

A GPU costs the same whether it is 8% or 80% utilised. Batching converts idle capacity into throughput, which is why the same hardware can produce tokens at a 14x cost difference.

Three caches worth building

- Prompt cache — reuse the fixed system and policy preamble across every call
- Retrieval cache — stop re-embedding documents that have not changed
- Semantic cache — serve near-duplicate questions from a stored answer with a freshness rule

SO WHAT: Roughly 60% of enterprise traffic is repetitive or deferrable. Serving it interactively at frontier prices is the most common avoidable cost in the estate.

RESOLUTION

Levers 6-7: context discipline and agent budgets

6. Context discipline

- Retrieve the top 5 relevant chunks, not the top 50
- Summarise conversation history rather than replaying it in full
- Strip boilerplate, headers and navigation from indexed documents
- Set a hard token ceiling per request and alert on breaches

7. Agent step budgets

- Maximum steps, maximum tool calls and maximum retries per task
- A hard dollar ceiling per task, enforced at the gateway
- Circuit breakers that escalate to a human instead of looping
- Log cost per completed task, not cost per API call

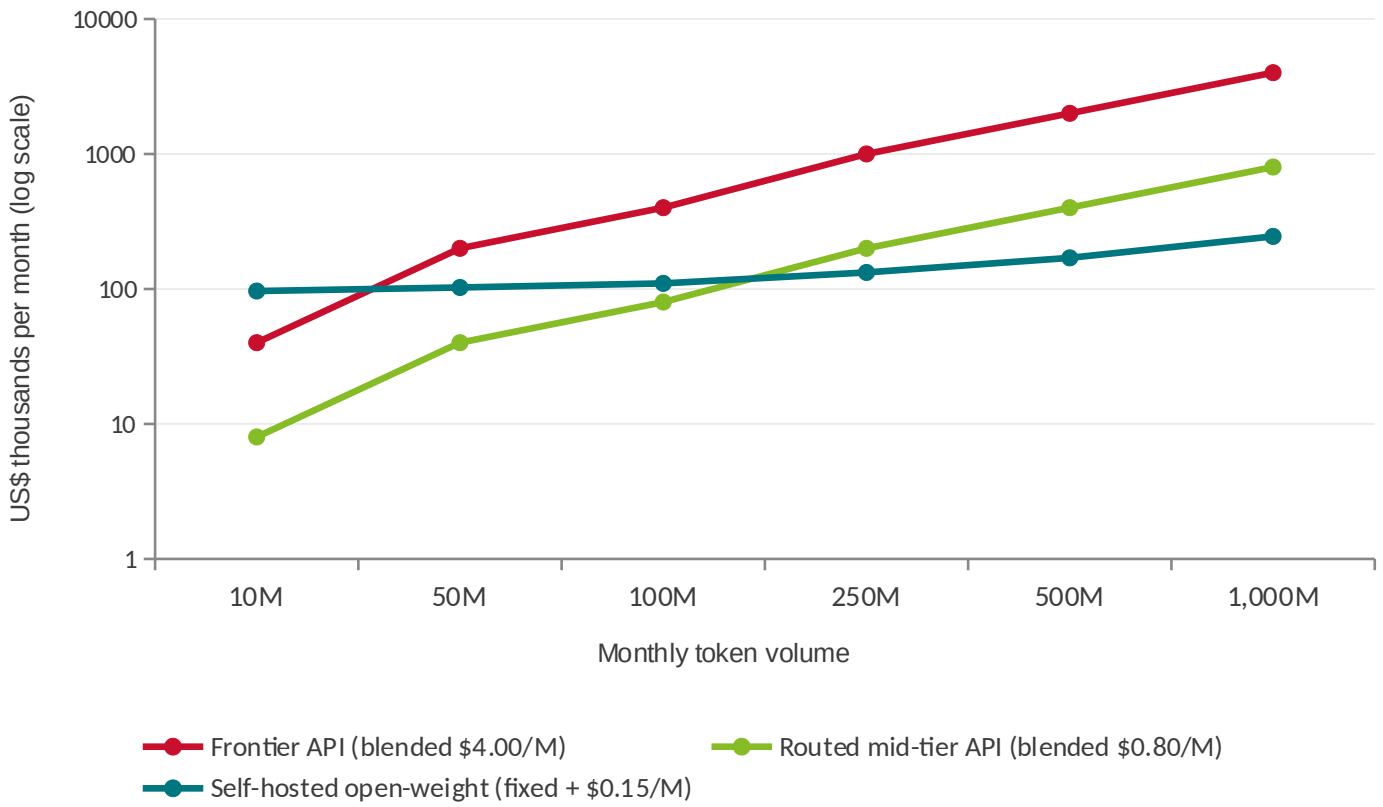
The control point: a single AI gateway

Every one of the seven levers is enforceable in one place if all traffic passes through a common gateway. Without it, each team re-implements policy, nobody can attribute spend, and shadow usage is invisible. This is the single highest-leverage platform investment in the programme.

Routing policy | Cost attribution by team | Rate & spend limits | Guardrails & logging | Model version pinning

RESOLUTION

Buy, rent or own — where the crossover actually sits



Our position and the decision rule

Below ~100M tokens / month

Managed API. Owning infrastructure cannot beat it once staffing is counted.

100-400M tokens / month

Hybrid. Route the stable, high-volume tail to owned or reserved capacity.

Above ~400M tokens / month

Owned or reserved capacity for the commodity tier; keep frontier work on API.

Current run rate: ~34M tokens / month. Forecast at full ramp: ~180M.

SO WHAT: We do not build infrastructure this year. Routing to mid-tier models beats self-hosting at our volume, and it carries none of the staffing or obsolescence risk.

RESOLUTION

FinOps for AI: instrument before you optimise

1

Attribute

Every call tagged to a use case, team and cost centre at the gateway. No tag, no traffic.

2

Forecast

Volume model per use case, refreshed monthly against actuals. Variance over 15% triggers review.

3

Optimise

Quarterly efficiency target on cost per outcome, owned by the platform team and reported to the steering committee.

4

Govern

Budgets, alerts and hard spend ceilings per use case. Automatic throttle, not a monthly surprise.

Six numbers the steering committee should see every month

- Cost per resolved outcome, by use case
- Total spend against forecast
- Token volume growth rate
- Cache hit rate and routing mix
- Evaluation pass rate
- Benefit realised against plan

RESOLUTION

The metric that matters: cost per resolved outcome

Cost per token is an input price. Cost per resolved outcome is a business metric — and it is the only figure that can be compared to the human cost it replaces.

Use case	AI cost per outcome	Incumbent cost	Margin	Volume p.a.	Annual value
Tier-1 service resolution	$0.31	$6.80	96%	396,000	$2.57M
Proposal first draft	$1.90	$140	99%	18,500	$2.55M
Contract clause review	$0.85	$34	98%	62,000	$2.05M
Claims document extraction	$0.12	$3.10	96%	740,000	$2.21M
Code review assistance	$0.44	$28	98%	104,000	$2.87M

Read this table carefully: the margin column is not the value column.

A 96% cost advantage per unit is meaningless unless the freed cost is actually removed from the plan. The annual value column only becomes real when a manager commits to a specific headcount, contractor or throughput change. That commitment is a condition of funding, not a hope.

SO WHAT: Every funded use case carries a cost-per-outcome ceiling and a named benefit owner. Breach the ceiling or lose the owner, and the use case goes back through the gate.

04

SECTION

The business case

What it costs to build, what it costs to keep running, what it returns, and who we have to hire to make it real.

1. How the case is framed and what is excluded
2. One-time build cost and annual run cost
3. Ongoing maintenance — the line item that kills ROI
4. Return, break-even and sensitivity
5. The hiring plan and how we source it

nutes

BUSINESS CASE

How the case is framed

Scope

- Six production use cases across service, sales, legal, operations and engineering
- One shared platform, one gateway, one evaluation harness
- Three-year horizon, 12% discount rate

Included

- All build, run, maintenance and assurance costs
- 12 net-new FTE, fully loaded
- Change management and enablement
- Compute growth of 3x per year

Excluded

- Speculative revenue upside from new AI products
- Second-order productivity claims that cannot be booked
- Existing IT costs that would be incurred anyway
- Any assumption of falling token prices

Three deliberately conservative choices

Flat token prices	Prices have fallen roughly 10x a year. We model zero decline. Any deflation is upside.
Benefit realisation at 75%	Every gross benefit is haircut by 25% before it enters the case.
No revenue growth assumed	Only cost, capacity and avoided-loss benefits are counted, despite revenue being the larger long-term prize.

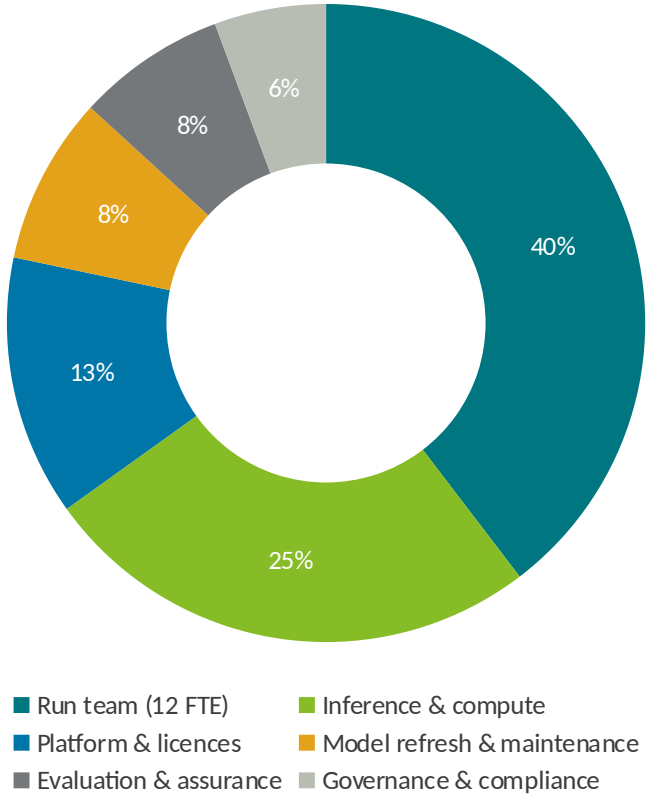
SO WHAT: The case is built to survive scrutiny, not to win an argument. If it clears the hurdle on these assumptions, it clears it in reality.

What it costs to build — one-time investment

Workstream	Scope	Cost ($M)	Timing
Discovery & value engineering	Use-case selection, baselining, cost-per-outcome targets	0.30	Q1
Data readiness	Pipelines, quality, entitlements, retrieval indexes	0.75	Q1-Q2
Platform build	AI gateway, RAG services, guardrails, observability, evaluation harness	1.10	Q1-Q3
Use-case build & integration	Six use cases at ~$0.22M each, including systems integration	1.30	Q2-Q4
Security, legal & model risk	Threat modelling, DPIAs, third-party review, model risk framework	0.25	Q1-Q3
Change management & enablement	Workflow redesign, role-based training, adoption campaign	0.40	Q2-Q4
Total one-time build		**4.10**	**Year 1**

Note the shape: only 32% of the build is use-case delivery. The other 68% is the platform, data and control environment that makes use cases seven through thirty cost a fraction of the first six. If the committee funds only the use cases and not the platform, unit costs will not fall and the programme will not scale.

What it costs to run — annual, at steady state



Line item	Annual ($M)	Share
Run team — 12 FTE, fully loaded	2.10	40%
Inference & compute	1.35	25%
Platform, tooling & vendor licences	0.70	13%
Model refresh, prompt & agent maintenance	0.45	8%
Evaluation, red-teaming & assurance	0.40	8%
Governance, audit & compliance reporting	0.30	6%
Total annual run cost	**5.30**	**100%**

Inference is 25% of run cost. The dominant cost is people.

Which means the leverage on total cost is a productive, well-tooled run team — not a marginal saving on tokens.

Ongoing maintenance — the line item that kills ROI

AI systems degrade without intervention. This is not optional spend; it is the cost of the asset continuing to work.

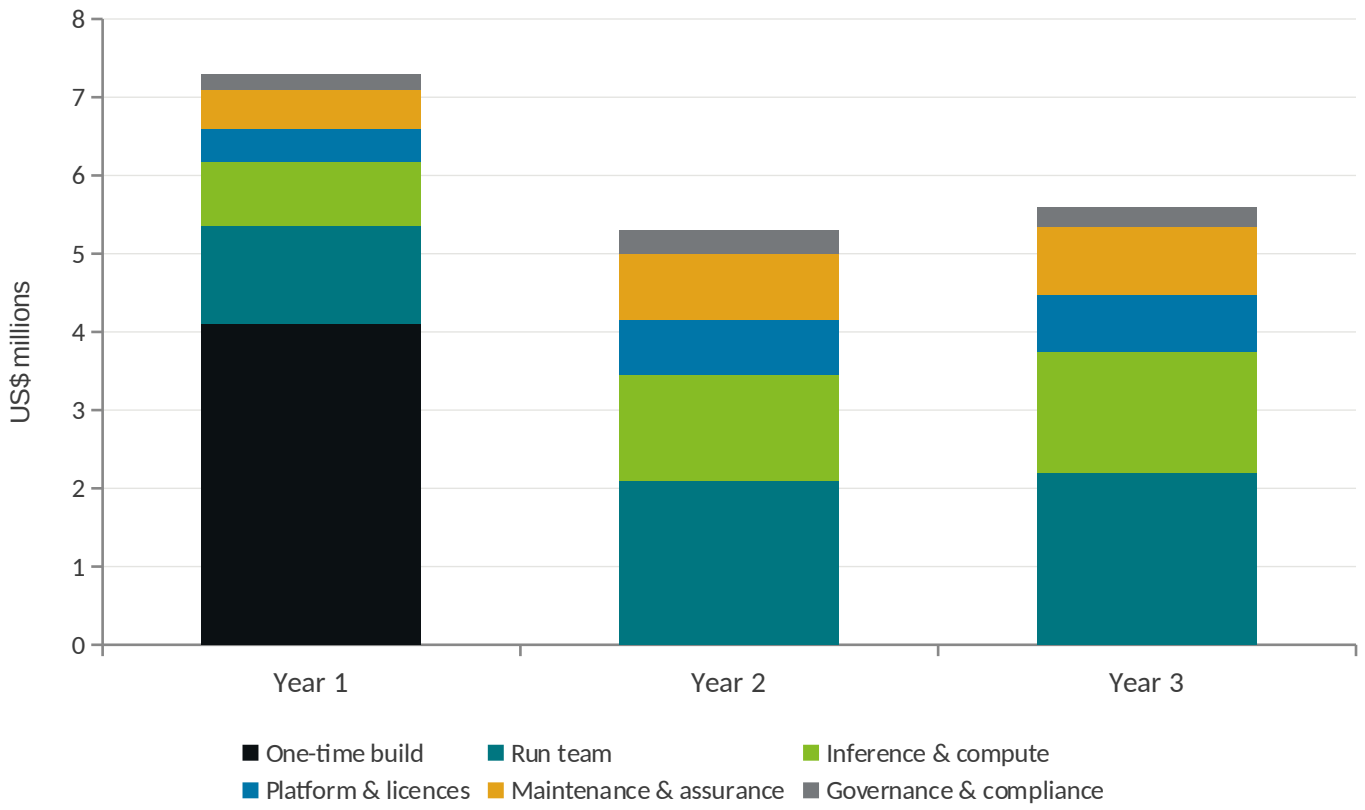
Maintenance obligation	Why it is unavoidable	Annual cost	Cadence
Model deprecation & upgrades	Providers retire models on 6-12 month cycles. Every upgrade requires re-testing prompts, agents and evaluation sets.	$0.18M	Continuous
Prompt & agent drift	Business rules, products and policies change. Prompts encode them and silently go stale.	$0.12M	Monthly
Retrieval index maintenance	Documents change, permissions change, embeddings need refreshing. Stale retrieval is the top cause of wrong answers.	$0.09M	Weekly
Evaluation set curation	Test sets rot as the business evolves; without refresh, quality gates measure last year's problem.	$0.06M	Quarterly
Regression & incident response	Behaviour changes after a vendor update; triage, rollback and root cause.	Within run team	On demand

The comparison that lands with a technology committee

A traditional enterprise application is re-platformed every five to seven years. The model layer underneath these systems will be replaced roughly every nine months. Maintenance is therefore a permanent 8-10% of annual run cost, and the first programme to under-budget it will be the first to report that 'AI stopped working'.

SO WHAT: Maintenance is budgeted at $0.45M a year and is ring-fenced. It is not a source of savings in a tight quarter.

Three-year total cost of ownership



$18.2M — Three-year total cost of ownership

$4.10M — One-time build, incurred in Year 1

$5.30M — Annual run cost at steady state

23% — Of three-year cost is build; 77% is run — this is an operating commitment, not a project

SO WHAT: Approving this is approving a permanent $5.3M annual operating line, not a $4.1M project. Fund it that way or do not fund it.

Where the value comes from

Benefit stream	Mechanism	How it is booked	Gross ($M)	Risk-adj. ($M)
Service operations	396K tier-1 contacts resolved without an agent; 22% handle-time reduction on the rest	Contractor reduction and attrition non-replacement	4.80	3.60
Sales & proposals	Faster first drafts, higher bid coverage, more qualified pipeline per rep	Revenue per rep in the sales plan	5.47	4.10
Engineering velocity	Code review, test generation and documentation assistance	Feature throughput per squad, deferred contractor spend	3.87	2.90
Document processing	740K documents extracted and validated without manual keying	Operations headcount plan and outsourcing spend	2.93	2.20
Risk & error avoidance	Earlier detection of contract, compliance and quality exceptions	Loss provision and rework budget	1.87	1.40
Total annual benefit at full ramp			**18.94**	**14.20**

Every stream has a named benefit owner in the business, a pre-agreed baseline measured before go-live, and a booking route into a specific budget line. Benefits with no booking route are excluded from this case entirely — they are reported as capacity, not as savings.

Expected return on investment



68% Three-year return on invested cost

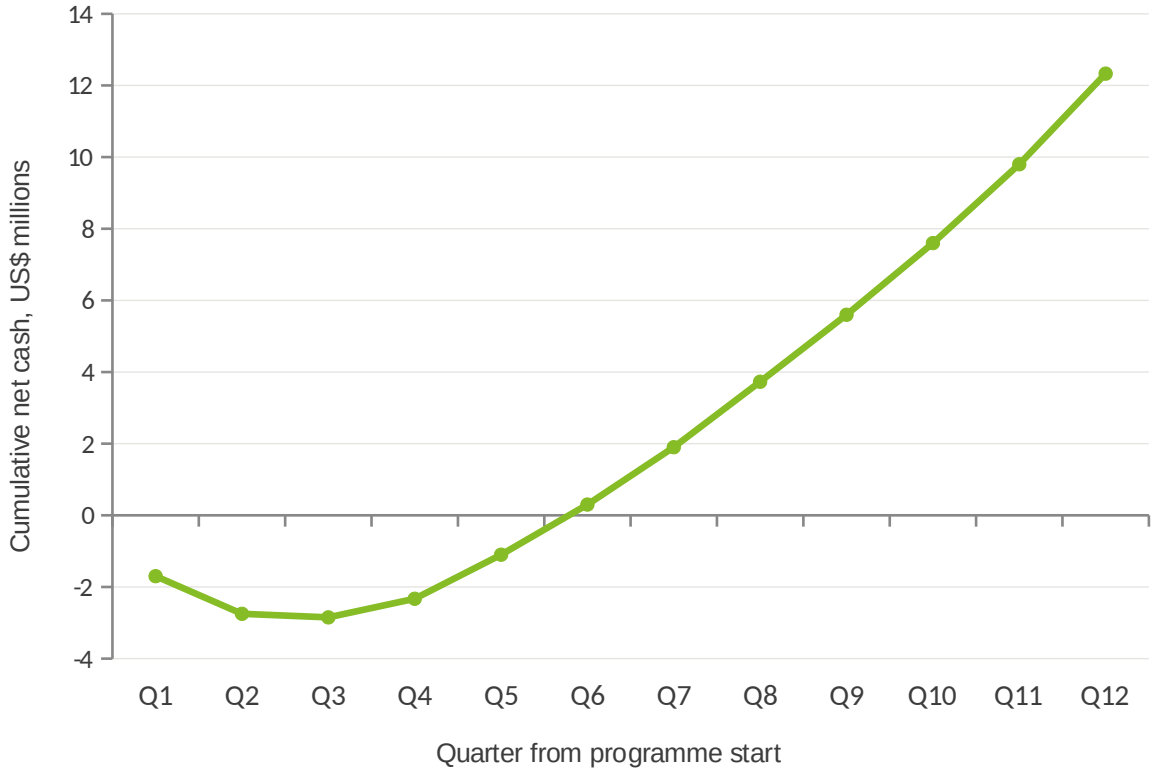
$8.9M Net present value at a 12% discount rate

17 months Payback from programme start

2.7x Benefit-to-run-cost ratio at steady state

SO WHAT: Three-year cost of $18.2M against risk-adjusted benefit of $30.5M. The case clears the hurdle with flat token prices and a 25% benefit haircut already applied.

Break-even and sensitivity



Break-even in Q6 — month 17

Sensitivity: impact on three-year NPV

Adoption 25% below plan	$4.1M	Still positive
Benefit realisation at 60%, not 75%	$5.6M	Still positive
Run team 3 FTE larger	$7.5M	Absorbed
Token prices rise 50%	$7.9M	Absorbed
Token prices fall 50%	$9.8M	Upside
Adoption 25% above plan	$13.6M	Upside

The case only fails if adoption collapses and benefits are never booked — both governance failures, not technology ones.

Will other hiring be necessary? Yes — twelve roles

Role	FTE	Why it is needed	Start	Loaded cost
AI product owner	2	Owns the use case, the baseline and the benefit booking	Immediate	$0.34M
ML / LLM engineer	3	Retrieval, routing, agent design, evaluation harnesses	Immediate	$0.60M
Data engineer	2	Pipelines, entitlements, index freshness	Immediate	$0.34M
Platform / MLOps engineer (incl. FinOps)	2	Gateway, observability, cost attribution and controls	Quarter 2	$0.36M
Evaluation & quality lead	1	Owns test sets, quality gates and regression after model upgrades	Quarter 2	$0.16M
AI risk & model governance	1	Model risk framework, audit evidence, regulatory engagement	Quarter 3	$0.18M
Change & enablement lead	1	Workflow redesign, curriculum, adoption campaign	Immediate	$0.14M
Total net-new headcount	**12**			**$2.12M**

Roles we are deliberately not creating

No standing 'prompt engineer' function and no separate AI centre of excellence detached from delivery. Both create dependency instead of capability.

Net workforce position

12 roles added; ~40 roles materially redefined; no involuntary reductions planned in the three-year horizon. Benefits are taken through attrition and contractor reduction.

How we source the twelve — build, borrow, buy

B

Build (6 of 12)

- Internal engineers and analysts redeployed with a 12-week conversion programme
- Cheapest, fastest to productive, retains domain knowledge
- Requires backfill planning in the source teams

B

Borrow (3 of 12)

- Partner or contractor capacity for platform build and evaluation tooling
- Deliberately time-boxed with a knowledge-transfer exit gate
- Avoids hiring for a peak that will not persist

B

Buy (3 of 12)

- External hires for scarce roles: LLM engineering, AI risk, product ownership
- Expect a 20-30% market premium and a 3-4 month time to hire
- Start the search before approval closes, not after

The retention risk is real and it is cheaper to pre-empt than to replace

People we train on this technology become immediately more marketable. The mitigations are unglamorous and effective: interesting problems, published internal career paths for AI roles, salary bands reviewed at six months rather than twelve, and no single points of knowledge failure. Budget replacement cost at 1.5x salary and design the team so no one departure stops a use case.

05

SECTION

Future strategy

Where this goes after the first six use cases — and which moves are no-regret regardless of how the technology develops.

1. Three horizons of AI value
2. Strategic options and the portfolio
3. No-regret moves we can start now

4 minutes

FUTURE STRATEGY

Three horizons — what the strategy could include

H1 Now — 12 months

Assist

- Copilots inside existing workflows
- Six use cases in production
- Cost-per-outcome discipline established
- Value: efficiency and capacity

70% of investment

H2 12-30 months

Automate

- Agentic workflows that complete tasks end to end
- Humans supervise exceptions rather than execute steps
- Process redesign, not process acceleration
- Value: structural cost and cycle time

25% of investment

H3 30 months +

Reinvent

- AI-native products and services sold to customers
- Pricing on outcomes rather than effort
- New operating models and possibly new business lines
- Value: revenue and competitive position

5% of investment

SO WHAT: Fund all three horizons every year, at a fixed ratio. A portfolio that is 100% Horizon 1 is efficient today and irrelevant in three years.

FUTURE STRATEGY

Strategic options and no-regret moves

Options to keep open

Vertical depth

Push AI deep into one function until it becomes structurally cheaper than any competitor's version of it.

Horizontal platform

Make the platform good enough that any team can ship a use case in weeks, and let demand find the value.

Data advantage

Convert proprietary process and outcome data into fine-tuned models competitors cannot replicate by buying an API.

Product monetisation

Package an internal capability as a customer-facing product with outcome-based pricing.

No-regret moves

Correct under every scenario, so start them now:

- Get data entitlements and quality right — the constraint under every future model
- Build the gateway and evaluation harness — model-agnostic by design
- Avoid architectural lock-in to any single provider
- Raise workforce AI fluency broadly, not narrowly
- Instrument cost per outcome before scale, not after

06

SECTION

AI Pioneers

The organisations pulling ahead are not the ones with better models. They govern faster and their people can actually use the tools.

1. What pioneers do differently
2. The AI steering committee — mandate and membership
3. Decision rights and operating cadence
4. Stage gates: funding and killing use cases
5. The upskilling and reskilling plan

Attributes

What the pioneers do differently

1

Concentrate

A small number of high-impact use cases in proven areas, not a wide field of experiments

2

Layer, don't rebuild

Generative AI applied on top of existing processes first; redesign follows evidence

3

Centralise governance

One committee, one standard, one platform — which is what makes local adoption fast

4

Invest above the threshold

Above 5% of IT budget, project success rates rise materially; token investment fails quietly

5

Redesign roles

Treat AI as a workforce transition, not a tool rollout, and fund enablement accordingly

6

Measure honestly

A defined baseline before launch and a named benefit owner after it

SO WHAT: None of these six is a technology capability. All six are governance and workforce decisions this committee controls directly.

Yes — a standing AI steering committee, chartered

Mandate

- Own the AI portfolio and allocate capital across the three horizons
- Approve, defer and kill use cases at defined stage gates
- Set enterprise standards: acceptable use, guardrails, data handling, model selection
- Unblock data access, legal and procurement obstacles within one cycle
- Own the workforce transition — reskilling, role redesign and communication
- Review cost per outcome and benefit realisation monthly

Membership and cadence

Chair	COO or Chief AI Officer — accountable for value, not for technology
Core	CFO, CIO/CTO, Chief Data Officer, CHRO, General Counsel / CRO
Business	Two rotating business unit leaders with live use cases
Standing	AI platform lead and AI risk lead, present but non-voting

Cadence

Monthly, 90 minutes	Portfolio, gates, spend and benefit review
Quarterly, half day	Capital reallocation and horizon rebalancing
Model risk council	Fortnightly, delegated authority

Decision rights — who decides what

Decision	Steering cttee	Business owner	Platform team	Risk & legal
Fund or kill a use case	**Decide**	Recommend	Inform	Consult
Set cost-per-outcome ceiling	**Approve**	Propose	Model	Inform
Choose model and vendor	Inform	Consult	**Decide**	Consult
Approve production release	Inform	**Approve**	Recommend	**Approve**
Accept a residual model risk	**Approve**	Consult	Inform	Recommend
Reallocate budget between use cases	**Decide**	Consult	Inform	Inform
Change workforce structure	**Decide**	Recommend	Inform	Consult

The single most important row is the first. Most organisations have a mechanism to start AI work and no mechanism to stop it, so the portfolio silently fills with projects nobody will defend but nobody will cancel. The gate on the next page is the stopping mechanism.

Stage gates: how use cases get funded and killed

G0 — Idea

Problem stated, sponsor named

1 week

›

G1 — Value case

Baseline measured, benefit owner signed, cost-per-outcome target set

3 weeks

›

G2 — Pilot

Evaluation thresholds met, risk assessment cleared, unit cost inside target

6-8 weeks

›

G3 — Production

Integrated, monitored, supported, adoption plan live

8-12 weeks

›

G4 — Scale & realise

Benefit booked into a budget line, unit cost trending down

Ongoing

Kill criteria — applied without negotiation

- Cost per outcome above ceiling for two consecutive months
- Quality below the agreed evaluation threshold with no path to close
- Benefit owner unwilling to book the saving into a budget line
- Adoption below 30% of eligible users six months after launch

Why this speeds things up rather than slowing them down

Clear gates with pre-published criteria mean teams stop lobbying and start evidencing. G0 to G2 should take under twelve weeks. Anything that cannot clear G1 in three weeks does not have a real business problem behind it, and finding that out in three weeks rather than nine months is the point.

The upskilling and reskilling plan

Persona	Population	Curriculum	Hours	Outcome
All employees	100%	AI fluency, acceptable use, data handling, what the tools cannot do	4	Safe, confident everyday use
Power users	~20%	Prompt patterns, workflow redesign, verification habits, cost awareness	16	Redesign their own workflow
Managers & leaders	~5%	Value cases, portfolio choices, risk appetite, leading redesigned teams	8	Sponsor and govern credibly
Builders	~3%	Retrieval, agents, evaluation design, cost-aware architecture	40+	Ship production use cases
Deep specialists	<1%	Certification, fine-tuning, serving optimisation, model risk	80+	Own the platform

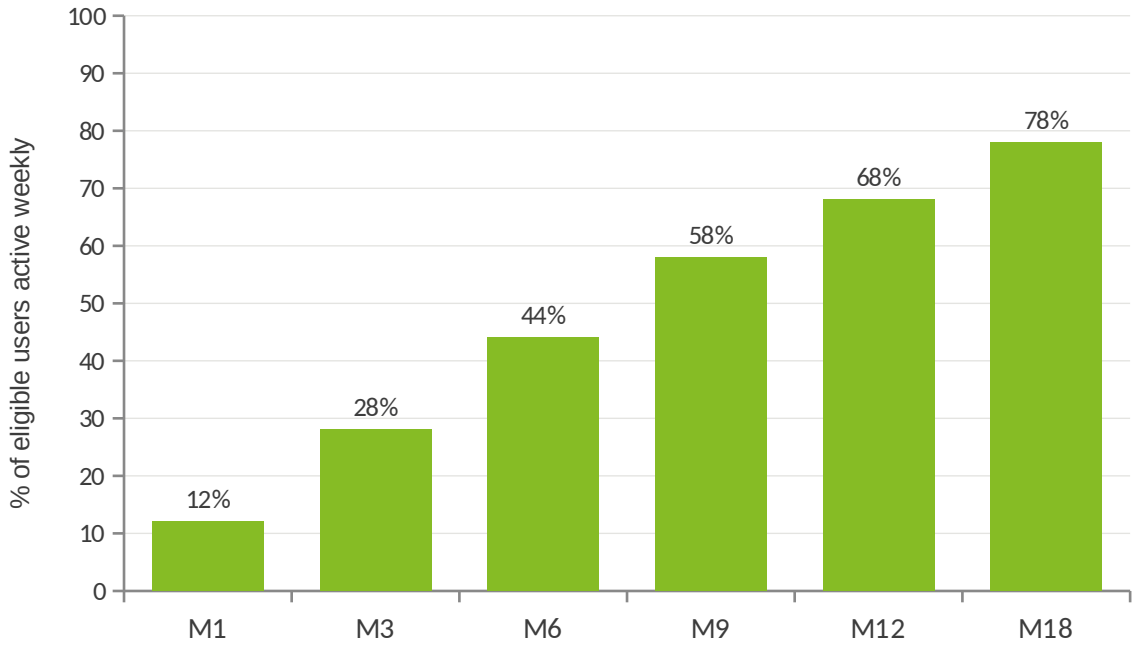
Reskilling — for roles with over 30% task exposure

Approximately 40 roles change materially. Service agents move to exception handling and AI supervision; analysts move from producing outputs to orchestrating and validating them. Each affected person gets a named pathway, a funded curriculum and a six-month transition window — published before deployment, not after.

Funding and accountability

$0.40M in Year 1, roughly 10% of build cost, held by the change and enablement lead and protected from reallocation. Completion is tracked per persona and reported monthly. Managers are accountable for their team's fluency rate, not the learning function.

Adoption is the variable the whole case rests on



What actually moves adoption

Embed in the tool people already use

Adoption collapses when the AI lives in a separate application

Manager modelling

Teams whose manager uses it weekly adopt at roughly twice the rate

Visible wins, named people

Peer proof beats a corporate campaign, every time

Remove the old path

Where quality allows, retire the manual route rather than leaving it as a comfortable default

SO WHAT: Adoption 25% below plan is the single largest downside in the sensitivity analysis. It is also the variable with the least technology in it.

ACTIVATION

The first 90 days

Days 1-30

- Charter and stand up the steering committee
- Confirm the six use cases and name benefit owners
- Measure baselines before anything is deployed
- Open the three external searches

Days 31-60

- Deploy the AI gateway with cost attribution from day one
- Publish acceptable-use policy and close down shadow usage
- Launch fluency training for all employees
- First two use cases through gate G1

Days 61-90

- First use case into production with a cost-per-outcome ceiling
- Monthly reporting pack live with the six FinOps metrics
- Reskilling pathways published for affected roles
- Q2 portfolio review and first kill decisions

Sequencing rule: measure the baseline before the first deployment. A benefit that cannot be evidenced against a pre-agreed baseline will not be booked, and an unbooked benefit is not a benefit.

ACTIVATION

Principal risks and mitigations

Risk	Severity	Mitigation
Benefit leakage — savings never booked	High	Named benefit owner per stream; booking route agreed at G1; monthly realisation reporting
Adoption below plan	High	Embedded in existing tools; manager accountability; adoption is a kill criterion at six months
Vendor price or policy change	Medium	Model-agnostic gateway; two qualified providers per tier; annual repricing built into the plan
Model deprecation breaking production	Medium	Version pinning, regression suites, ring-fenced maintenance budget
Data leakage or inappropriate output	Medium	Guardrails at the gateway, entitlement-aware retrieval, human review on sensitive paths, model risk council
Key-person dependency and attrition	Medium	Build-borrow-buy mix, documented runbooks, no single owner per use case
Regulatory change	Medium	Model inventory and audit evidence maintained continuously, not assembled on request

SO WHAT: The two high-severity risks are both organisational. Neither is mitigated by better technology, and both are owned by people in this room.

DECISIONS REQUIRED TODAY

The ask

1. **Approve the investment**
   $4.1M one-time build and $5.3M annual run cost, placed in the operating plan rather than the change portfolio.

2. **Approve the headcount**
   12 net-new roles, sourced six internal, three partner, three external, with searches opening this month.

3. **Charter the steering committee**
   Chaired outside technology, membership confirmed within two weeks, first meeting inside 30 days.

4. **Adopt the gates and kill criteria**
   Including the cost-per-outcome ceiling and the commitment that unbooked benefits are not counted.

Decision requested at the close of this session. Deferral has a cost: every quarter of delay pushes break-even out by a quarter and compounds the capability gap with competitors already at scale.

APPENDIX

Assumptions, definitions and sources

Key assumptions

- 5,000 employees; six use cases in production by end of Year 2
- Discount rate 12%; three-year evaluation horizon
- Token prices held flat — no deflation banked
- Benefit realisation haircut of 25% applied to all gross benefits
- Token volume growth of 3x per year through the horizon
- Fully loaded FTE cost of $170K; contractor premium 40%
- Run team of 12 FTE reaching full strength in Quarter 3
- No revenue growth or new-product benefit included

Definitions

Inference

Running a trained model to produce an output; the recurring cost of AI, distinct from one-time training

Token

The unit of text a model processes and the unit providers price against

Cost per resolved outcome

Total cost of producing one completed business result, including retries, guardrails and retrieval

Routing

Directing each request to the cheapest model that meets the quality threshold

Distillation

Training a small model on a large model's outputs to serve a narrow task cheaply

Sources: Deloitte, State of AI in the Enterprise (2026 edition, survey of 3,235 leaders across 24 countries); Deloitte, State of Generative AI in the Enterprise series; a16z analysis of inference price decline; 2026 published inference pricing benchmarks. Cost and benefit figures are an illustrative model for this organisation and require validation against actual baselines before approval.