THE CFO BOARD

10 YEARS

THE TOTAL COST OF OWNERSHIP AND RETURN ON INVESTMENT OF ENTERPRISE AI



October 2026

TABLE OF CONTENT

Message from the CEO.....3

1. Executive Summary 4

Findings at a glance 4

What this paper asks the CFO to do .....'...5

2. Purpose, Scope, and the Case for a Version 2.0 6

2.1 Purpose 6

2.2 Scope.....6

2.3 Why this document is versioned 6

3. Why AI Total Cost of Ownership Behaves Differently 7

3.1 Five structural differences 7

3.2 Deployment context changes the cost and risk profile 8

3.3 The unit economics shift 8

4. The Architecture Is the Cost Model: 9

4.1 The foundation layer — how AI connects, reasons and stays accurate 9

4.2 The operating layer — how AI collaborates, remembers and is governed 10

4.3 What these blocks mean financially 10

4.4 Reading the blocks as a forecast 11

5. The AI Cost Taxonomy 12

5.1 The three categories most often mispriced 13

5.2 Attribution: the precondition for control 13

6. The Visibility Gap 14

6.1 Four failure patterns in AI business cases 14

6.2 The pilot-to-production cliff.....15

6.3 Shadow AI: the cost you have not been told about.....15

7. From Cost to Value: An ROI Framework 16

7.1 The first principle: value must be booked somewhere.....16

7.2 The measurement set: twenty practical KPIs 16

7.3 Three rules that make the numbers defensible 18

7.4 What the evidence says separates high performers 18

8. The Risk Ledger: Pricing What Can Go Wrong..... 19

8.1 Risk-adjusted TCO 20

8.2 Three risks with distinctive financial behaviour..... 20

TABLE OF CONTENT

9. Governance as a Cost Discipline 22

9.1 The AI FinOps operating model 23

9.2 Build, buy or outsource the FinOps capability 23

9.3 Build versus buy versus orchestrate 23

10. The CFO Question Bank 24

10.1 At investment approval 24

10.2 At the quarterly business review 25

10.3 On architecture and vendors 26

10.4 On risk, compliance and the board 26

11. What the CFO Should Be Tracking 27

11.1 Reporting cadence 28

12. A Twelve-Month Implementation Path 29

12.1 First ninety days 29

12.2 Months four to six 30

12.3Months seven to twelve 30

13. Version 2.0: What This Series Will Test Next 31

13.1 Hypotheses to be tested 31

13.2 Open questions carried forward 31

13.3 The Claims verification list 32

13.4 What readers can contribute 32

14. Conclusion 33

Appendix A · AI TCO Worksheet Template 34

Appendix B · Roundtable Discussion Questions 35

Appendix C · Glossary 36

Appendix D · Further Reading 37

MESSAGE FROM THE CFO BOARD

Enterprise AI is moving rapidly from experimentation into the operating fabric of organisations. For finance leaders, that shift changes the question. The issue is no longer simply how much an AI solution costs to acquire, but how its architecture, usage, governance and operating model shape the full cost of ownership and the value ultimately realised.

This paper brings these dimensions together through a CFO lens. It looks beyond licences and visible technology spend to examine inference, compute, data engineering, human oversight, compliance, technical debt, FinOps and the other costs that can determine whether an AI investment delivers sustainable value.

It also places equal emphasis on measurement. AI benefits need a baseline, a clear owner and a method for verification. Productivity claims are useful signals, but they become financially meaningful only when they can be connected to cost reduction, redeployed capacity, revenue or risk outcomes. The paper is deliberately practical. It provides a cost taxonomy, an ROI measurement framework, a risk ledger, a CFO question bank and a twelve-month implementation path. These are intended to support better investment conversations today while creating a consistent basis for measuring progress over time.

Enterprise AI economics will continue to evolve. The objective, therefore, is not to predict every change, but to establish the financial discipline needed to understand those changes as they occur. A common language for cost, value and risk can help finance leaders govern AI as a business capability rather than as a collection of technology projects. Given the pace at which enterprise AI is evolving, The CFO Board intends to revisit this topic with a follow-on report to help finance leaders assess emerging developments, risks and opportunities as they unfold.



Nitin Madan
Member, The CFO Board

1. EXECUTIVE SUMMARY

Three years into enterprise generative AI, the boardroom question has changed. It is no longer whether AI improves productivity; that is largely settled and largely visible. It is whether AI improves profitability, and whether the organisation can demonstrate that improvement with the same rigour it applies to every other capital and operating decision.

The evidence suggests it usually cannot. Published research indicates that while roughly 88% of organisations report regular AI use, only about 39% report any measurable enterprise-wide EBIT impact. That gap is not primarily a technology gap. It is a financial governance gap: costs are under-scoped, benefits are un-baselined, and the architecture that drives both is treated as an engineering detail rather than a financial one.

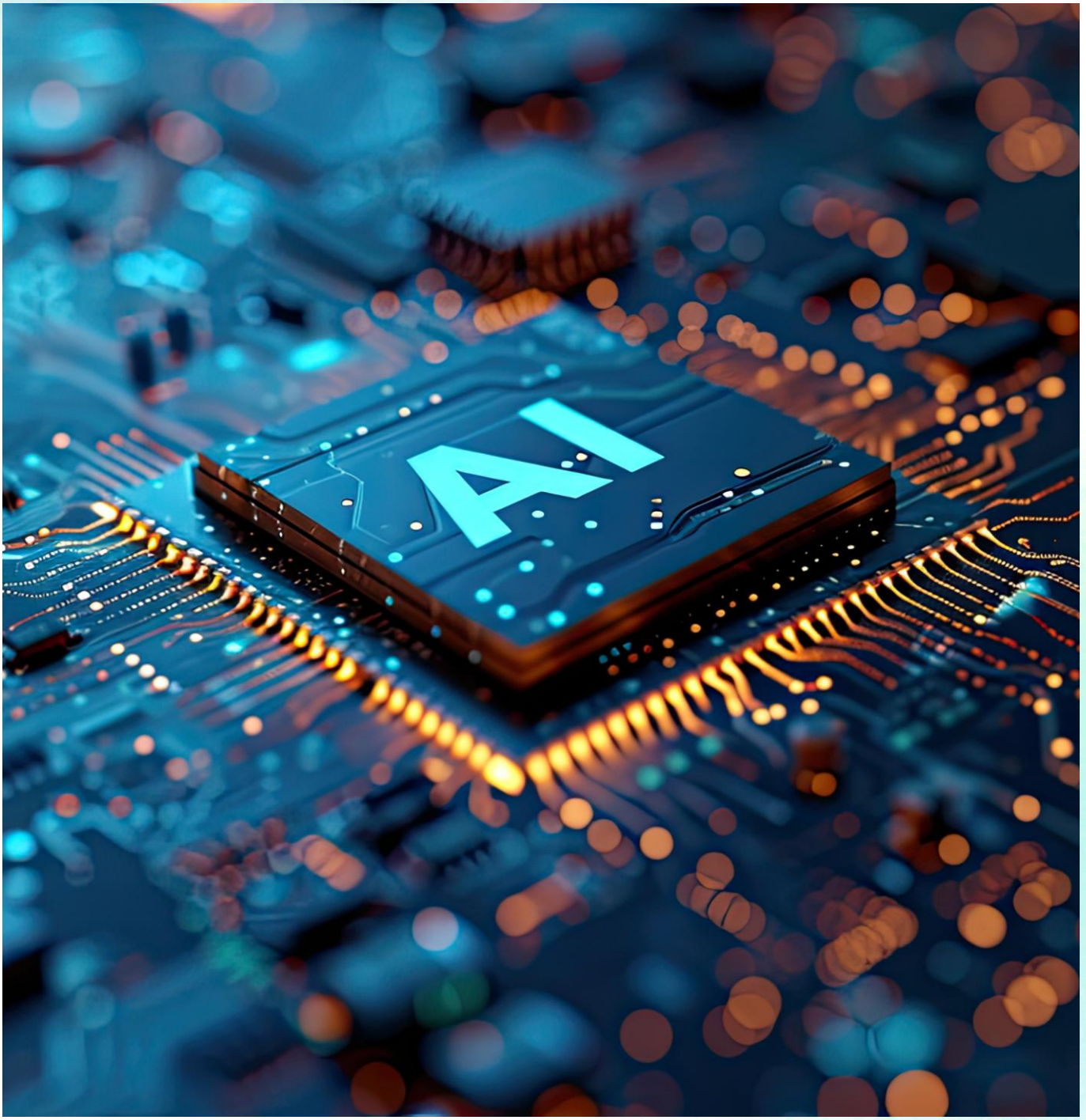
> The central argument of this paper: AI cost is not a procurement outcome, it is an architectural outcome. A CFO who can read the architecture can forecast the cost curve.

Findings at a glance

1. **AI TCO is structurally different from software TCO**. Traditional software has fixed licences, planned upgrades, linear infrastructure and finite maintenance. AI systems have variable inference costs, continuous retraining, auto-scaling compute, agentic workflows that compound API calls, and a permanent governance overhead. The economic model has shifted from software ownership to continuous AI operations management.
2. **Technical building blocks now determine the cost model.** Six of the twelve reduce total cost of ownership; six increase it. Every one of them is the outcome of a decision that someone in the organisation is making - often without the CFO in the room.
3. **Organisations underestimate AI cost by 10–40%, and hidden costs commonly account for 40–60% of true TCO.** Data engineering, human-in-the-loop oversight, legal and regulatory compliance, model drift monitoring, shadow AI and change management are the categories most consistently missing from business cases.
4. **The pilot-to-production cliff is where cost curves change materially.** Pilots run at small volume with relaxed governance and look inexpensive. Production adds reliability engineering, security review, observability, enterprise support and scaling infrastructure. A business case built on pilot economics is not a business case.
5. **Model routing is the single largest controllable cost lever.** Sending every task to a frontier model regardless of complexity is the AI equivalent of flying every parcel first class. Routing converts AI spend from uncontrolled variable cost into managed variable cost.
6. **Governance is now a designed cost centre, not an overhead to be minimised.** Evaluation harnesses and guardrails add real, recurring spend. They are also the only mechanism that makes AI auditable, defensible and board-reportable — and the cheapest form of incident insurance available.
7. **Return must be proven, not asserted.** Productivity anecdotes and licence counts are not ROI. Value has to be booked somewhere: cost removed from the P&L, capacity redeployed to revenue-generating work, revenue uplifted, or risk cost avoided. Capacity released that is never redeployed is not a saving.
8. **Deployment context changes the provisioning standard.** Inward-facing AI, software delivery, knowledge management, HR, back office, carries contained risk. Outward-facing AI, customers, regulators, public interfaces, introduces uncontrolled query volumes, regulatory scrutiny and brand exposure, and warrants a materially higher governance and legal provision.

What this paper asks the CFO to do

1. Own the AI cost taxonomy — all categories, not the three that appear on an invoice.
2. Require that the architecture be explained in financial terms before funding is released.
3. Insist on a measured baseline before deployment; without one, benefit claims are unfalsifiable.
4. Stand up an AI FinOps capability with cost attribution by project, model and business unit.
5. Price risk explicitly and provision for it, rather than discovering it as an incident.
6. Track a fixed metric set on a fixed cadence, so that the next twelve months are comparable to the last.



2. PURPOSE, SCOPE, AND THE CASE FOR A VERSION 2.0

2.1 Purpose

This paper consolidates a series of published articles on enterprise AI economics and governance into a single reference document written for the office of the CFO. It is intended to be used in four ways:

a) as a briefing for the finance leadership team;

b) as a checklist during AI investment approval;

c) as an agenda for a CFO roundtable; and

d) as a baseline against which the same organisation can measure its own progress twelve months from now.

2.2 Scope

The scope is enterprise AI deployed to run the business, both inward-facing systems (software and application delivery acceleration, internal knowledge management, HR, finance and back-office automation) and outwardfacing systems (customer service, advisory interfaces, regulated disclosures, public-facing agents). AI embedded in products sold to customers raises additional questions of product costing, pricing and margin attribution; those are noted where relevant but are not the focus of this edition.

2.3 Why this document is versioned

Enterprise AI economics is not a settled field, and a white paper that pretends otherwise ages badly. Three things will change materially within twelve months: unit inference prices and the capability of open-weight models; regulatory obligations, particularly as AI Act-style requirements move from readiness to enforcement; and the internal maturity of the organisations reading this. A framework that is correct today may be mispriced by a factor of two next year.

The response is to version the paper rather than rewrite it. Version 1.0 establishes the taxonomy, the metric definitions and the tracking cadence. Version 2.0, planned for six to twelve months after publication, will hold those definitions constant and report movement against them, which is the only way an organisation can distinguish genuine progress from re-labelled spend. Section 13 sets out, in advance, exactly what Version 2.0 will test. Committing to those hypotheses now is what makes the comparison credible later.



3. WHY AI TOTAL COST OF OWNERSHIP BEHAVES DIFFERENTLY

The first error in most AI business cases is a category error. AI is procured through software channels, approved through software governance and budgeted on software templates but it does not behave like software. Understanding precisely where the analogy breaks is the foundation for everything that follows.

Table 2 · Traditional software economics versus ai economics

Traditional software — predictable economics	AI systems — the new economic reality
Fixed licence costs negotiated upfront	Variable inference and API consumption costs
Periodic, planned upgrade cycles	Continuous model maintenance and retraining
Stable, largely linear infrastructure	Dynamic, auto-scaling compute workloads
Human-controlled, auditable workflows	Agentic workflows that compound API calls
Finite maintenance and support scope	Permanent governance, observability and compliance overhead
Combination of Capex and Opex	Mostly Opex given the evolution and short life span

3.1 Five structural differences

Consumption pricing replaces seat pricing

Software cost scales with people. AI cost scales with use, and use is set by prompt length, output length, retry behaviour, retrieval volume and the number of model calls per task. Two teams with identical licence counts can differ by an order of magnitude in spend. The consequence for the CFO is that headcount is no longer a reliable proxy for cost, and budget cannot be set by seat count alone.

Non-determinism creates a quality cost

The same input can produce different outputs. Where the workflow requires a defensible answer, that variability has to be absorbed somewhere: by evaluation infrastructure, by retrieval grounding, or by a human reviewer. The cheapest of these is engineering; the most expensive, and the most commonly used by default, is the human. This variability may vary from region to region and model pricing. Human oversight cost is therefore a direct function of model quality and of how well the trust threshold is calibrated.

Strengthen Tax Analytics

Project Anveshan, a nationwide effort aimed at uncovering weaknesses in GST collection, has exposed numerous cases of tax avoidance and fraud. Tamil Nadu could leverage on the experience thus far and adopt AI-powered compliance and analytics systems, similar to those implemented by Karnataka and Maharashtra, to detect tax evasion and improve GST collection.

Agentic workflows compound cost non-linearly

A single question answered by a single model call is a predictable unit of cost. An agent that plans, retrieves, calls three tools, critiques its own output and retries on failure may consume twenty or more model calls for the same user request.

As agentic patterns spread, the relationship between user activity and token consumption becomes non-linear and unbounded loops become a genuine financial risk, not just a technical one. Spend caps, loop limits and pertask cost ceilings belong in the architecture.

Governance is a running cost, not a project cost

Security review, privacy control, audit trail, model evaluation and regulatory compliance are not one-off implementation tasks. They recur for every new use case, every model change and every regulatory update. Treating them as a project line item that ends at go-live is the most common structural error in AI budgeting.

3.2 Deployment context changes the cost and risk profile

The underlying cost drivers are broadly consistent across deployments, but the context materially changes the risk and provisioning profile.

Table 3 · Inward-facing versus outward-facing deployment

Dimension	Inward-facing AI	Outward-facing AI
Typical use cases	Software and application delivery acceleration, internal knowledge management, HR, finance and back-office automation	Customer service and advice, public-facing agents, regulated disclosures, partner and regulator interfaces
Volume behaviour	Bounded by employee population and working patterns; forecastable	Uncontrolled query volumes; spikes driven by campaigns, incidents and external events
Failure cost	Largely operational e.g. rework, delay, internal frustration	Operational, reputational and regulatory simultaneously
Governance standard	Proportionate controls; risk contained inside the organisation	Higher provisioning standard: output validation, audit trail, escalation paths, legal review, incident response
CFO implication	Model on productivity and cycle-time value; watch consumption growth	Provision explicitly for legal, compliance and incident cost; treat volume as a risk variable, not just a cost variable

> The practical rule: an outward-facing use case is not the same business case as an inward-facing one with a bigger user number. It is a different risk class and should carry a different provision.

3.3 The unit economics shift

The most useful mental adjustment a finance team can make is to stop asking “what does this tool cost per user per month?” and start asking **“what does this cost per task, and what is that task worth?”** Cost per AI-assisted transaction is the metric that survives across pilots, production, model changes and vendor switches. It is also the metric that makes AI comparable to the human process it is replacing or augmenting, which is the comparison the board actually wants.

4. THE ARCHITECTURE IS THE COST MODEL: TWELVE BUILDING BLOCKS

Between 2023 and 2026, enterprise AI moved from custom, brittle, single-model implementations to standardised, routed, retrieval-grounded and governed systems. That transition can be described precisely as twelve technical building blocks. Each block has its own cost curve. Each is the outcome of a decision. Six of the twelve reduce total cost of ownership; six increase it — and the six that increase it are, in most cases, the ones that make the other six safe to rely on.

> AI cost is not one line item. It is the sum of twelve building blocks, each with its own cost curve, and most of them are the outcome of a decision. Some blocks raise spend now to cut it later. Some cut spend immediately. A CFO who understands the architecture can forecast the trajectory of AI cost.

4.1 The foundation layer — how AI connects, reasons and stays accurate

Table 4 · Building blocks 1–6, with 2023 baseline, 2026 architecture and TCO impact

Category	Block	2023 baseline	2026 architecture	TCO impact for the CFO
Models	1. Model Routing	Every task sent to one premium model, regardless of complexity.	A router sends each task to the right-sized model by cost and quality.	↓ Converts AI spend from uncontrolled variable cost to managed variable cost.
	2. Local LLM	Not viable in production — open-source and local models lagged far behind; used only for R&D or narrow, low-risk tasks.	Open-weight models run on-premise or in private cloud, competitive in quality for many tasks, with full data control.	↓↓ Per-token cost, data egress risk and vendor lockin at scale.
	3. Frontier Model	The default choice for every task — the only model reliable enough to use, regardless of cost.	Reserved for highcomplexity, high-stakes work; called only when the task justifies the premium.	↑ Cost per call stays the highest in the stack. ↓ Overall spend as frontier use narrows to fewer, higher-value calls.
Customisation	4. Fine-Tuning	Default approach for customising model behaviour.	Reserved for specialised domains — medicine, banking, legal — as RAG, memory and tools now outperform it.	↑ Training investment, maintenance cost, retraining cost.
	5. RAG	Model answered from memory alone — prone to guessing.	Model retrieves enterprise documents first, then answers.	↑ Engineering cost upfront ↓ Hallucination cost, compliance and risk exposure.
Capabilities	6. MCP	Custom point-to-point connector built per system (ERP, CRM, document stores).	One standard protocol connects AI to any system — the USB-C of enterprise AI.	↓ Upfront build cost, then ↓↓ integration, maintenance and vendor lock-in.

Green = net TCO decrease · Amber = mixed · Red = net TCO increase. Arrows indicate direction of cost movement, not magnitude.

4.2 The operating layer — how AI collaborates, remembers and is governed

Table 5 · Building blocks 7–12, with 2023 baseline, 2026 architecture and TCO impact

Category	Block	2023 baseline	2026 architecture	TCO impact for the CFO
Capabilities	7. Tool Use	Models only answered questions in text.	Models execute APIs, run queries, write code and update enterprise systems.	↓ Manual effort, application switching, cycle time, labour cost.
Agentic	8. A2A	No standard for AI-to-AI communication.	Agents communicate across functions via shared protocols (finance ↔ tax ↔ legal ↔ treasury).	↓ Workflow integration cost, custom development, future migration effort.
	9. Agent Memory	Every prompt started from zero context.	AI retains preferences, projects, context, rules and history.	↓ Repeated prompting, token cost, training effort, user friction.
	10. Multi-Agent	One model attempted every task alone — does not scale.	Specialised agents (planner, finance, legal, reviewer) collaborate.	↑ Engineering cost upfront. ↓ Labour cost, cycle time, manual coordination.
Controls	11. Evals	Ad hoc, subjective "does this look right?" review.	Continuous measurement of accuracy, latency, outcomes and regression.	↑ Governance cost. ↓ Operational failures, compliance issues, business risk.
	12. Guardrails	Minimal, inconsistent input and output controls.	Full stack: safety, policy, compliance, output validation, audit trail.	↑ Governance investment. ↓ Legal exposure, security incidents, reputational risk.

4.3 What these blocks mean financially

1. **Cost shifts from build to run**
   MCP, multi-agent orchestration and agent-to-agent protocols raise engineering spend early and then cut integration and coordination cost for years. This is a classic capitalisable-investment profile inside what most organisations treat as an operating budget. The CFO question is not "why is this more expensive than last year?" but "over what horizon does this build convert into a lower run-rate, and who is accountable for that conversion?"

2. **Variable cost becomes managed cost**
   Model routing is the single biggest lever available. It stops every task from paying premium-model prices and turns a demand-driven, unpredictable spend line into a policy-driven one. Routing is also the mechanism through which local and open-weight models deliver savings: without a router, an open-weight deployment is a parallel stack rather than a substitution.

3. **Governance is now a cost centre by design**
   Evaluations and guardrails add genuine, recurring spend. They are what makes AI auditable, defensible and boardready. The correct framing for the board is not that governance is overhead to be minimised, but that it is the cost of making the other ten blocks safe to scale — and that it is materially cheaper than the incidents it prevents.

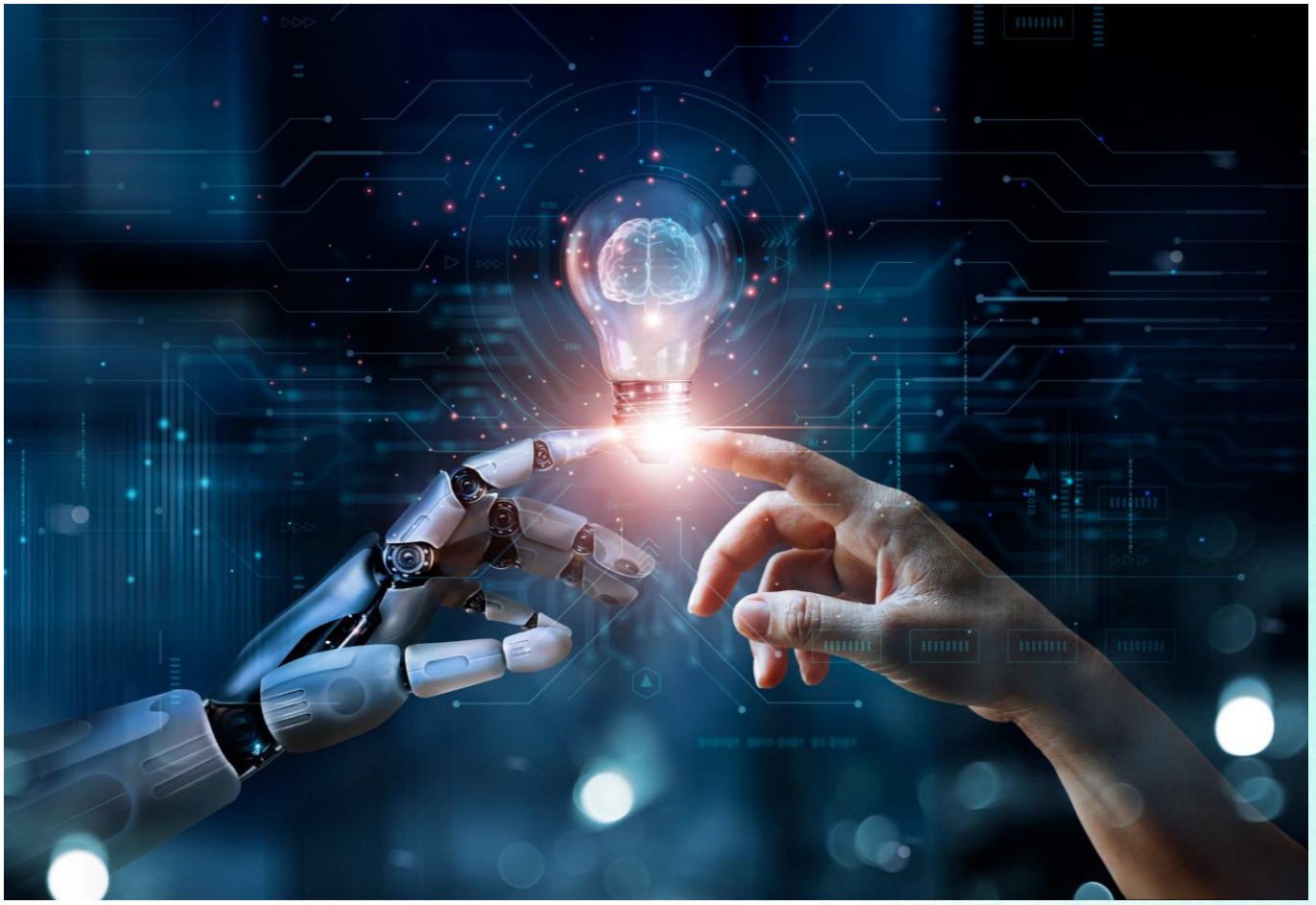
4. **Fine-tuning is no longer the default**
   Retrieval, memory and tool use now outperform fine-tuning for most enterprise use cases at a fraction of the lifecycle cost. Fine-tuning retains a place in specialised domains — medicine, banking, legal — where domain language and reasoning patterns genuinely differ. Elsewhere, a fine-tuning proposal should be challenged: it carries training investment, maintenance obligation and a retraining cost every time the base model moves.

4.4 Reading the blocks as a forecast

Taken together, the twelve blocks produce a predictable cost shape over an AI programme's life — a J-curve. Spend rises first as integration standards, orchestration and controls are built, then falls as routing, retrieval and reuse take effect, and then flattens into a managed run-rate. The failure mode is stopping halfway: organisations that build agentic capability without building routing and evaluation get the cost increase without the cost reduction.

Table 6 · The cost j-curve by programme phase

Phase	Blocks being built	Cost signature	What the CFO should require
Phase 1 — Ground (0–6 months)	RAG, MCP, basic guardrails	Engineering spend up; inference spend low but rising	Integration standard chosen once, not per system; baseline metrics captured before go-live
Phase 2 — Scale (6–18 months)	Tool use, agent memory, multi-agent, A2A	Inference spend accelerates non-linearly; labour cost begins to fall	Per-task cost ceilings, loop limits, cost attribution by project before agentic rollout
Phase 3 — Optimise (12–24 months)	Model routing, local and openweight models, evals	Unit cost falls even as volume rises; governance run-rate stabilises	Unit-cost trend reported monthly; frontier share of tokens managed to a target
Phase 4 — Govern (ongoing)	Full evaluation and guardrail stack, FinOps	Managed run-rate; cost growth tracks value growth	Cost per assisted transaction and benefit realisation reported together, every quarter



5. THE AI COST TAXONOMY

The table below is the core working artefact of this paper: a structured taxonomy of AI cost across fixed, variable and semi-variable categories. For each driver it identifies what increases cost, what reduces it, and the point most commonly underestimated in enterprise business cases. A business case that does not address all twelve lines is not complete — it is a licence estimate.

Table 7 · AI TCO cost sheet — key cost drivers by category

Cost category	Type	Examples	What increases cost	What reduces cost	CFO remarks
AI licences & subscriptions	FIXED	Enterprise AI assistant and copilot subscriptions, platform seats	More users, premium tiers, multi-vendor sprawl	Enterprise contracts, shared licence pooling, renegotiation at scale	Anchor point of the budget; negotiate upfront at enterprise scale rather than department by department
Inference / API usage	VARIABLE	Token consumption, model calls, agent loops	Large prompts, excessive retries, agentic chains, frontier models	Prompt optimisation, model routing, caching, output compression	Fastest-growing cost lever; usage explodes in agentic workflows and must be capped by design
Cloud compute & GPU	VARIABLE	GPU hosting, vector databases, embedding workloads	Always-on infrastructure, peak workloads, no autoscaling	Spot instances, serverless, reserved capacity, auto-scaling	Often double-counted with inference; needs dedicated tagging to be managed at all
Data engineering & governance	SEMI-VARIABLE	Cleaning, RAG pipelines, vector database maintenance	Poor data quality, siloed systems, manual ingestion	Standardised pipelines, reusable data products, master data investment	Underestimated in the large majority of enterprise deployments; rises with data scope, not user count
Security & privacy reviews	SEMI-VARIABLE	PII controls, threat modelling, audit trails	Regulated industries, multi-cloud, sensitive data	Built-in governance, privacy-by-design, reusable frameworks, data categorisation	Non-negotiable for production; scales with deployment breadth rather than volume
Legal & regulatory compliance	SEMI-VARIABLE	AI Act readiness, contractual IP review, vendor agreements, explainability requirements	Regulated sectors, cross-border deployments, new AI regulation	Reusable compliance frameworks, regulatory precertification, proactive legal review	Often omitted at pilot stage; becomes material at enterprise scale — IP infringement, contractual liability and AI Act obligations are real exposures
Model maintenance & retraining	SEMI-VARIABLE	Fine-tuning, drift monitoring, regression testing	Frequent workflow changes, unstable data	Stable workflows, modular architecture, retraining pipelines	Ongoing operational cost most commonly missing from the business case entirely
Integration & technical debt	FIXED / SEMI	ERP and API integrations, AIgenerated code remediation	Legacy architectures, poor coding standards	Modern API-first architecture, coding governance	AI-generated code accelerates technical debt if unreviewed — a deferred cost, not an avoided one
Human oversight & review	SEMI-VARIABLE	Human-in-the-loop escalations, validation workflows	Low model accuracy, high-risk use cases	Improved model quality, calibrated trust thresholds	Identified in published research as a top differentiator for AI high performers; usually the largest hidden line
Observability & FinOps tooling	SEMI-VARIABLE	Monitoring, tracing, dashboards, cost attribution	Multiple AI vendors, fragmented tooling	Centralised observability, an AI FinOps platform	Emerging as a standalone category; it is what enables cost control at scale
Cost category	Type	Examples	What increases cost	What reduces cost	CFO remarks
Shadow AI & duplicate spend	VARIABLE	Unapproved subscriptions, fragmented procurement	Decentralised adoption, weak governance	Centralised procurement, an AI tool registry	Invisible until it becomes a compliance or cost shock; assume it exists and go looking for it
Change management & training	FIXED / SEMI	Workforce enablement, adoption programmes	Low readiness, complex tools, high turnover	Structured programmes, rolebased learning journeys	Frequently omitted from TCO; without it, benefit realisation stalls regardless of technical quality

Legal & regulatory compliance is included deliberately. It is frequently absent from pilot-stage business cases and becomes increasingly material as AI Act obligations, IP liability and cross-border data requirements take effect.

5.1 The three categories most often mispriced

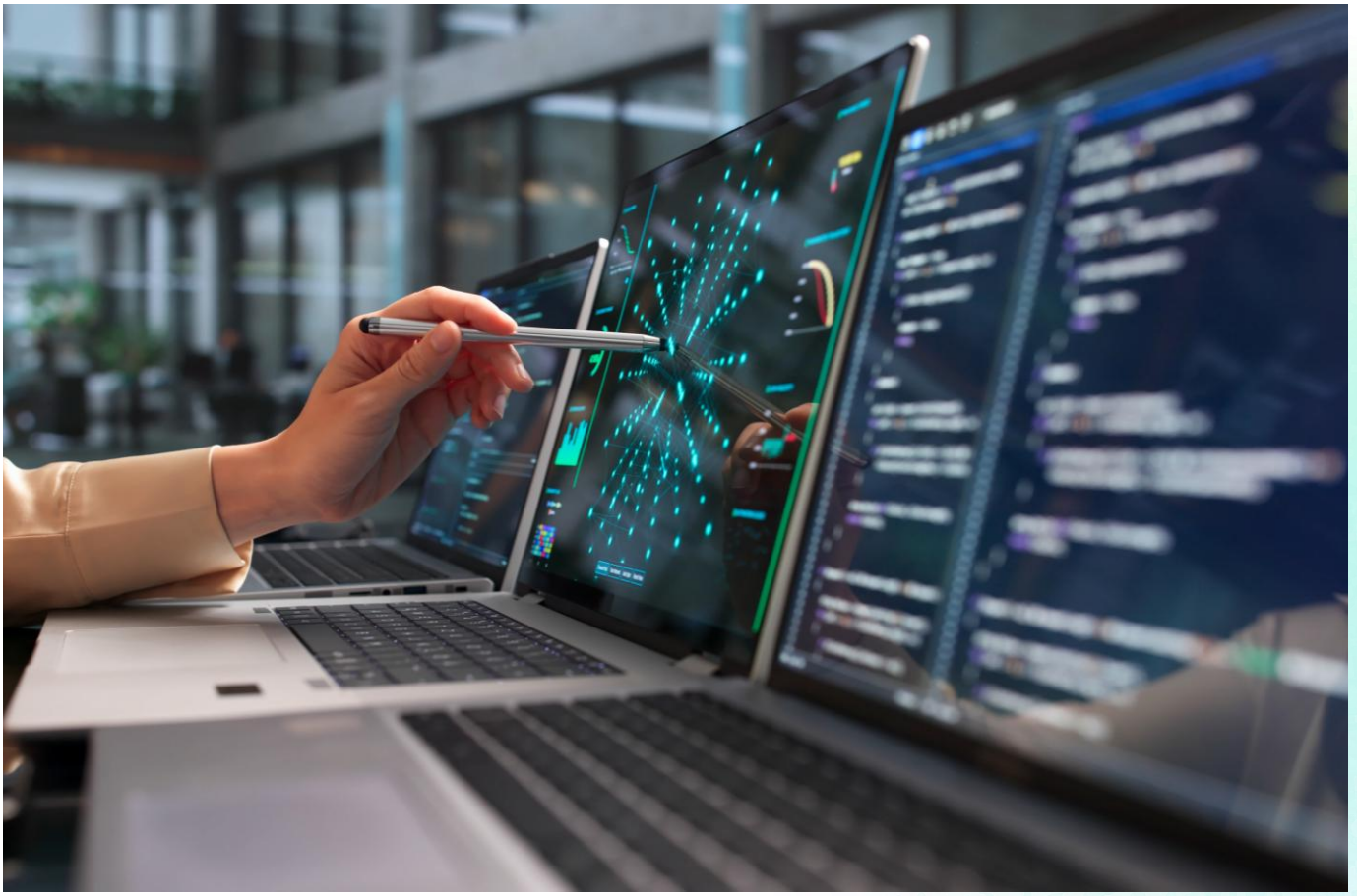
1. **Data engineering and governance**
AI quality is a function of data quality, and data work does not scale with users — it scales with the number of sources, the messiness of each, and the rate at which they change. Organisations routinely fund the model and underfund the pipeline that feeds it, then attribute the resulting poor performance to the model. The financial signature of this error is a rising human-review cost that nobody can explain.

2. **Human oversight**
Every percentage point of model inaccuracy in a high-stakes workflow converts directly into human review hours. This line is rarely modelled because it appears in operational headcount rather than in the AI budget. It is also the line most responsive to investment elsewhere: better retrieval, better evaluation and calibrated trust thresholds all reduce it. Model the trade explicitly — spending on evaluation to reduce review cost is usually a favourable exchange.

3. **Legal and regulatory compliance**
At pilot stage this is close to zero. At enterprise scale, in a regulated sector, with cross-border data movement and outward-facing deployment, it becomes a standing cost with a standing team. The prudent approach is to build the compliance framework once, reusably, and to amortise it across use cases rather than paying legal review per project in perpetuity.

5.2 Attribution: the precondition for control

None of the above can be managed without attribution. Before an organisation can optimise AI cost it must be able to answer three questions on demand: what does this project cost, what does this model cost, and what does this business unit consume? Most enterprises today cannot answer any of the three. Tagging discipline by project, model, environment, business unit and use case is unglamorous and is the highest-return control available in year one.



6. THE VISIBILITY GAP

Industry studies indicate that organisations consistently underestimate AI project costs by 10–40%. The mechanism is not carelessness; it is structural. The costs that are easy to see are billed by a vendor. The costs that are hard to see are absorbed inside existing teams, existing infrastructure budgets and existing operational headcount — which is precisely why they never appear in the business case.

Table 8 · The visibility gap

Visible costs — typically budgeted	Hidden or underestimated costs — often 40–60% of true TCO
• AI platform and licence subscriptions • API token usage, where directly billed • Cloud compute and GPU hosting • Integration development, initial build only • Storage and vector database licences	• Data cleaning, structuring and RAG pipeline maintenance • Prompt engineering, tuning and optimisation cycles • Legal: IP review, AI Act compliance, vendor contracts • Model drift monitoring and retraining • Security, PII controls and privacy reviews Human-in-the-loop oversight and escalation workflows • Shadow AI subscriptions across business units • Technical debt from unreviewed AI-generated code • Organisational change management and workforce reskilling

6.1 Four failure patterns in AI business cases

1. **The pilot extrapolation.**
Pilot unit economics are multiplied by production volume. This ignores that production adds reliability engineering, security review, monitoring, enterprise support tiers and scaling infrastructure — and that pilot users are self-selected enthusiasts whose usage patterns are not representative.

2. **The absorbed-cost omission.**
Work done by existing data, security, legal and operations teams is treated as free because no invoice is raised. It is not free; it is capacity diverted from other work. Charge it in at a loaded rate or the comparison against the status quo is invalid.

3. **The unbaselined benefit.**
A benefit is claimed against a “before” state that was never measured. Once the system is live, the counterfactual is unrecoverable. This single omission accounts for the majority of AI ROI disputes.

4. **The one-off governance line.**
Compliance, evaluation and oversight are budgeted as implementation tasks that end at go-live, when in fact they recur for every model change, every new use case and every regulatory update.

6.2 The pilot-to-production cliff

Pilot-stage AI appears inexpensive because it runs at small volume under relaxed governance. Production deployment introduces reliability requirements, security reviews, monitoring layers, enterprise support and scaling infrastructure. This is the point at which cost curves change materially — and it is also, not coincidentally, the point at which most AI programmes lose executive confidence, because the number presented at approval bears no relation to the number that arrives.

> A practical control: require every AI business case to present two sets of economics — pilot and production-atscale — with the bridge between them itemised. If the team cannot construct the bridge, the programme is not ready for production funding.

6.3 Shadow AI: the cost you have not been told about

Decentralised adoption is now the norm rather than the exception. Published research indicates that a substantial minority of staff — on the order of **29%** in one large vendor study — use AI agents that no one has approved, while a large majority of major enterprises already run agents in production. The financial exposure is threefold: duplicate subscriptions across business units, data governance risk from tools with no contractual protection, and a compliance position that cannot be evidenced to a regulator.

Shadow AI is best treated as a symptom rather than a violation. It appears where the sanctioned path is slower than the unsanctioned one. The remedy is a central registry, a fast approval route and consolidated procurement — in that order.



7. FROM COST TO VALUE: AN ROI FRAMEWORK

Cost discipline without value measurement produces a different failure: a well-governed programme that nobody can justify. This section sets out how to prove return in terms a finance function can defend.

7.1 The first principle: value must be booked somewhere

Every claimed AI benefit falls into one of four classes, and each has a different standard of proof. Confusing them is the most common source of inflated ROI claims.

Table 9 · Four classes of ai value and their evidence standard

Value class	What it means	Evidence standard	Booking treatment
Cost-out	A cost line falls: fewer vendor hours, lower cost-to-serve, reduced rework, avoided contractor spend	Visible in the ledger against a measured baseline	Bookable. Include at full value in the benefit-cost ratio
Capacity release	Hours are freed but headcount is unchanged	Hours saved × fully loaded rate, verified by sampling not self-report	Bookable only if the capacity is redeployed to identified work or a planned hire is not made. Otherwise apply a haircut
Revenue and retention	Conversion, retention, satisfaction or first-contact resolution improve	Controlled comparison or staged rollout with a holdout group	Bookable at the margin contribution, not at gross revenue
Risk cost avoided	Fewer errors, better compliance evidence, lower audit findings	Reduction in observed incident frequency or severity against baseline	Bookable as an expectedvalue reduction; state assumptions explicitly

A practical control: require every AI business case to present two sets of economics — pilot and production-atscale — with the bridge between them itemised. If the team cannot construct the bridge, the programme is not ready for production funding.

7.2 The measurement set: twenty practical KPIs

The following twenty metrics cover the financial, operational, customer, scaling, workforce and adoption value of AI. Few organisations should track all twenty at once. The recommendation is to select four to six as the standing board set at minimum one from each of financial, operational and adoption and to hold their definitions constant across reporting periods.

Table 10 · Twenty ways to measure ai impact

#	Metric	Calculation	What it proves	Lens / cadence
1	Revenue uplift from AI	Post-AI revenue − baseline revenue	Extra revenue created above the baseline trend	Financial · Quarterly
2	Cost savings realised	(Baseline cost − post-AI cost) − AI operating cost	Net saving after the cost of running AI itself	Financial · Quarterly
3	AI investment payback period	Total AI investment ÷ monthly net benefit	Time until benefits repay the investment	Financial · Quarterly
4	AI benefit–cost ratio	Verified AI benefits ÷ total AI costs	Return generated per unit invested	Financial · Quarterly
#	Metric	Calculation	What it proves	Lens / cadence
5	Cost per AI-assisted transaction	Total AI operating cost ÷ AI-assisted transactions	Unit economics; the metric that survives model and vendor change	Financial · Monthly
6	Cycle time reduction	Baseline cycle time − post-AI cycle time	How much faster the process runs	Operational · Monthly
7	Straight-through processing value	Auto-processed transactions × manual cost per transaction	Value of work completed without human handoff	Operational · Monthly
8	Labour capacity released	Hours saved × average fully loaded hourly cost	Monetary value of freed human hours	Operational · Quarterly
9	Sales conversion uplift	Post-AI conversion rate − baseline conversion rate	Lift in prospects converting	Customer · Quarterly
10	Customer retention uplift	Post-AI retention rate − baseline retention rate	Lift in customers retained	Customer · Quarterly
11	Customer satisfaction uplift	Post-AI CSAT − baseline CSAT	Improvement in customer experience	Customer · Quarterly
12	First-contact resolution uplift	Post-AI FCR rate − baseline FCR rate	Lift in issues resolved at first contact	Customer · Monthly
13	Collection efficiency improvement	Post-AI collection rate − baseline collection rate	Working-capital impact of AI in the cash cycle	Customer · Quarterly
14	Time to value	First value date − project start date	How quickly AI delivers measurable benefit	Scaling · Per project
15	Pilot-to-production value rate	Production AI use cases with value ÷ total AI pilots × 100	Share of pilots that scale with proven value	Scaling · Quarterly
16	AI use case value coverage	Value pools covered by AI ÷ total priority value pools × 100	Whether AI is aimed at the value that matters	Scaling · Half-yearly
17	Employee experience uplift	Post-AI eNPS − baseline eNPS	Whether AI improves or degrades the work itself	Workforce · Half-yearly
18	Adoption rate	Active AI users ÷ total eligible users × 100	Share of the eligible workforce actually using AI	Adoption · Monthly
19	Business unit engagement rate	Business units active in AI ÷ total business units × 100	Breadth of adoption across the enterprise	Adoption · Quarterly
20	AI literacy score	Sum of assessment scores ÷ number assessed	Workforce capability, measured rather than assumed	Adoption · Half-yearly

7.3 Three rules that make the numbers defensible

01

Baseline before you build

Cycle time, error rate, cost to serve, resolution rate and satisfaction must be measured before deployment. This is a two-week discipline that determines whether the next two years of benefit claims are credible. No baseline, no benefit — it should be a funding condition.

02

Verify benefits before they enter the ratio

The benefit–cost ratio should use verified benefits only. Verification means a finance-owned check on a sample: did the hours actually disappear from the process, did the cost line actually move, is the improvement distinguishable from seasonality and from other concurrent initiatives? Self-reported time savings from users are an input to verification, never a substitute for it.

03

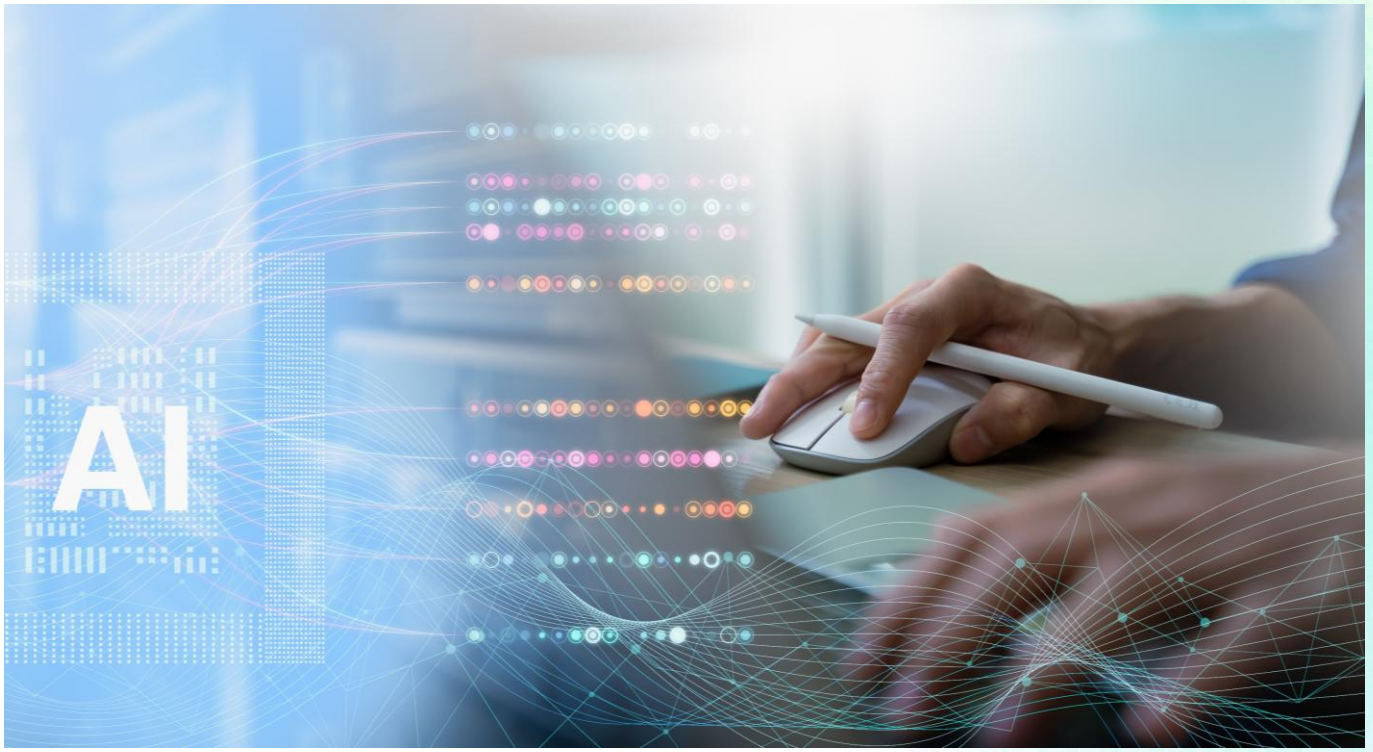
Charge AI's own cost against its benefit

Net cost savings are gross savings minus the cost of running AI, including the hidden categories in Section 5. A programme reporting savings against licence cost alone is reporting a number that will not survive scrutiny.

7.4 What the evidence says separates high performers

Published research on AI high performers — those reporting greater than 5% EBIT impact — finds they are approximately **2.8 times more likely to have fundamentally redesigned workflows** and **three times more likely to have senior leadership actively championing AI governance**. Related work indicates that organisations investing at scale in responsible AI report EBIT impact above 5% more frequently than those that do not.

The significance for the CFO is that neither differentiator is technological. Both are organisational and financial disciplines: redesigning the work rather than bolting AI onto it, and funding governance seriously rather than treating it as a compliance formality. This is a direct rebuttal to the common internal argument that AI value is gated by model capability. It is gated by workflow redesign and governance investment and both of which sit squarely inside the CFO's sphere of influence.



8. THE RISK LEDGER: PRICING WHAT CAN GO WRONG

Risk is a cost line that has not been invoiced yet. Enterprise AI risks fall into five families: data, model, agent, security and operations, and governance and each family maps to identifiable financial exposure and to a control with a known cost. The purpose of this section is not to catalogue risk for its own sake, but to give the CFO a basis for provisioning.

Every claimed AI benefit falls into one of four classes, and each has a different standard of proof. Confusing them is the most common source of inflated ROI claims.

Table 11 · The five ai risk families, their financial exposure and controls

Risk family	Representative risks	Financial exposure	Primary control	Where it lands in TCO
Data risks	Bias, privacy exposure, data quality, unclear lineage, data drift, leakage, missing consent, synthetic-data false confidence, embedded secrets, PII exposure	Regulatory penalty, remediation programmes, customer redress, invalidated outputs requiring rework	Data governance, classification, lineage tracking, consent management	Data engineering & governance; security & privacy
Model risks	Hallucination, poor explainability, overfitting and underfitting, weak robustness, evaluation failure, validation gaps, model drift, prompt sensitivity, metric blindness	Wrong decisions taken at scale, human re-verification cost, loss of user trust and consequent abandonment of the investment	Retrieval grounding, continuous evaluation, regression testing, drift monitoring	Model maintenance; human oversight; observability
Agent risks	Excess autonomy, tool misuse, unintended actions, permission escalation, prompt injection, social manipulation, unsafe delegation, memory exposure, loop failure, inability to roll back	Unauthorised transactions, runaway consumption from unbounded loops, operational disruption, irreversible actions in systems of record	Least privilege, named owner per agent, escalation path and kill switch, spend and loop caps, rollback design	Inference/API usage; integration; observability
Security & operations	Exploitable vulnerabilities, cyber threats, API abuse, token theft, malicious use, data exfiltration, unsafe deployment, integration failure, latency, cost overrun	Incident response, breach notification, downtime, direct financial loss, insurance and audit consequences	Threat modelling, secrets management, rate limiting, monitoring, staged deployment	Security & privacy; cloud & compute; observability
Governance risks	Monitoring gaps, unclear decision ownership, insufficient human oversight, poor traceability, compliance breach, audit gaps, legal exposure, IP and copyright, ethics, regulatory breach	Regulatory action, contractual liability, IP claims, inability to evidence oversight to a board or regulator, reputational damage	Named accountability, audit trail, policy framework, evidenceof-oversight reporting	Legal & regulatory; observability; change management

Derived from the fifty-risk enterprise AI risk taxonomy published in the series. Not every risk applies to every deployment; the families are what should appear on every risk register.

8.1 Risk-adjusted TCO

My learning from being in a bank to assign risk weights to financial numbers and for material AI investments, the business case should present a risk-adjusted view rather than a point estimate:

> Risk-adjusted TCO = Direct Cost + Control cost + Σ (probability × impact) of the residual risks that remain after controls.

The value of this construction is not precision, the probabilities are estimates. Its value is that it forces two conversations that would otherwise not happen: which risks are being accepted rather than controlled, and whether the control is cheaper than the exposure. In most cases evaluation and guardrail investment prices out favourably against a single significant incident, which is the argument that persuades boards.

8.2 Three risks with distinctive financial behaviour

1. **Cost overrun from unbounded agent loops**
   This is the only risk in the catalogue that is simultaneously a technical failure and a direct, immediate cost event. An agent that retries indefinitely, or a multi-agent system in which two agents call each other repeatedly, can consume a meaningful share of a monthly budget in hours. The control is architectural and cheap: per-task token ceilings, maximum iteration limits, and automated alerting on consumption anomalies. This should be a mandatory pre-production checklist item, and the CFO is entitled to ask for evidence that it exists.

2. **IP and copyright exposure**
   Ownership questions around AI-generated material: code, content, designs remain unsettled in several jurisdictions and are moving. The exposure is contractual as much as legal: indemnity terms vary considerably between vendors, and the difference in provision required between a vendor that indemnifies and one that does not can be material. This belongs in vendor evaluation, priced, not treated as boilerplate.

3. **Technical debt from unreviewed AI-generated code**
   AI-accelerated software delivery produces a genuine productivity gain and a genuine deferred liability. Code generated faster than it can be reviewed accumulates as maintenance burden that surfaces twelve to twenty-four months later. The financial treatment should mirror any other deferred maintenance: recognise it, size it, and require that the productivity claim be net of it.



9. GOVERNANCE AS A COST DISCIPLINE

Say “AI governance” to six executives and you will get six different concerns. That is not a problem to be resolved, it is the structure of the accountability. The CFO’s task is not to hold all six concerns, but to know who holds each and to fund the mechanism that connects them.

Table 12 · AI governance by role — what each leader owns and what the CFO needs from them

Role	Owns	Their core concern	What the CFO needs from this role
CEO	Direction. Owns the call.	Governance is what allows the organisation to say yes to AI, not what slows it down. Research indicates a large majority of CEOs now identify themselves as the primary AI decision maker — which means one bad AI decision is theirs to explain to the board.	A stated risk appetite and a clear position on which use cases are in scope, so that funding decisions are not made case by case in a vacuum.
CFO	Financial stewardship. Funds control.	Control is cheaper than cleanup: fund oversight now or fund the incident later. Clear AI accountability correlates with higher maturity, and organisations investing at scale in responsible AI report stronger EBIT impact.	Self-owned: the cost taxonomy, the attribution model, the benefit verification standard, and the tracking cadence in Section 11.
COO	Operations. Owns the impact.	Agents act, they do not merely answer — the risk has moved from saying the wrong thing to doing it. Every agent needs an escalation path and a kill switch, and accountability must sit with a named person rather than a committee.	A named owner per agent, and confirmation that escalation and shutdown controls are tested rather than merely documented.
CIO	Technology and security.	Shadow AI is already present; a substantial minority of staff use agents nobody approved, while most large enterprises already run agents in production. Every agent should have only the access it actually needs.	A complete inventory of what is running, with least-privilege access confirmed — the precondition for both cost attribution and risk control.
CDO	The data foundation.	You cannot put rules on data you cannot find, define or trust. AI governance fails wherever data governance has failed. The standard is “enough, not perfect”: make the data slice the use case needs trustworthy.	A realistic view of data readiness per use case, so that data remediation cost is in the business case rather than discovered midbuild.
Risk & Legal	Proof of oversight.	Can the organisation demonstrate oversight, accountability and a clear audit trail? Standards and regulation are still forming, so systems must be built for what is coming. Risk should be an early input at design, not a late veto.	An early, costed view of regulatory obligations per use case — and a reusable compliance framework rather than perproject legal review.

9.1 The AI FinOps operating model

The CFO's mandate in AI is expanding from financial stewardship to enterprise AI governance. In practice this requires a capability that most finance functions do not yet have: centralised financial management of AI infrastructure, model consumption, team-level attribution and budget enforcement. Three functions define it.

Table 13 · The three functions of AI finops

Function	What it does	Minimum viable version
Visibility	Attribute every unit of AI spend to a project, model, environment, business unit and use case; surface shadow AI	A monthly consumption report by project and model, plus a registry of every approved AI tool and subscription
Accountability	Set budgets and thresholds by team and use case; enforce them; require benefit reporting against the same identifier as cost	A named budget owner per use case and an alert when consumption exceeds a defined threshold in-month
Optimisation	Reduce unit cost without reducing value: routing policy, caching, prompt efficiency, model selection, reserved capacity	A quarterly review of the highest-spend workloads against cheaper alternatives, with the saving quantified

9.2 Build, buy or outsource the FinOps capability

Most organisations should start by building visibility with existing cloud and observability tooling, and buy a dedicated AI FinOps platform only once multi-vendor sprawl makes consolidated reporting genuinely difficult, typically at the point where three or more model providers are in production use. Outsourcing is appropriate for the periodic optimisation review, but the visibility and accountability functions should stay in-house: they depend on internal cost allocation logic that no external party can maintain.

9.3 Build versus buy versus orchestrate

At the architecture level, the winning system in most enterprises is not the most powerful one but the most operationally governable one. Three questions drive the decision, and all three are financial:

Are frontier models necessary for this task, or is this a routing decision? Reserve the premium tier for work that justifies it.

Where does open-weight or local deployment suffice? The saving is not only per-token; it includes data egress control and reduced vendor concentration.

What is the long-term cost of vendor lock-in? Price switching cost explicitly: how much of the integration, prompt and evaluation work would have to be redone if the provider changed? Standardised integration protocols exist precisely to reduce this number.

10. THE CFO QUESTION BANK

The questions below are grouped by the moment at which they should be asked. They are deliberately answerable: each has a right kind of answer, and the absence of one is itself informative.

10.1 At investment approval

Question	What a good answer contains
1. What is the total cost of ownership across all twelve cost categories, not just licences and inference?	A completed cost sheet with an explicit number or explicit zero for every category, including absorbed internal effort
2. What is the baseline we are measuring against, and when was it measured?	Cycle time, error rate, cost to serve and volume, measured before deployment, with the measurement method stated
3. What is the bridge from pilot economics to production-at-scale economics?	An itemised list of what is added at production: reliability, security, monitoring, support, scaling
4. What is the bridge from pilot economics to production-at-scale economics?	An explicit higher provision for legal, compliance and incident response if outward-facing
5. What is the bridge from pilot economics to production-at-scale economics?	A distinction between reusable platform investment and use-case-specific build
6. What is the bridge from pilot economics to production-at-scale economics?	A unit cost that falls or stays flat with volume; if it rises, an explanation
7. What is the bridge from pilot economics to production-at-scale economics?	A named redeployment or an avoided hire; “we will see” means the benefit is not bookable
8. What is the bridge from pilot economics to production-at-scale economics?	A quantified switching cost covering integration, prompts, evaluations and retraining
9. What is the bridge from pilot economics to production-at-scale economics?	A single name, with escalation authority and budget accountability
10. What is the bridge from pilot economics to production-at-scale economics?	A tested procedure with a date, not a documented intention

10.2 At the quarterly business review

Question	What a good answer contains
11. What is the trend in cost per AI-assisted transaction?	A declining or flat unit cost, with the drivers of any increase named
12. What share of total token spend went to frontier models, and is that share falling?	A managed share against a target, evidencing that routing is working
13. Which benefits claimed last quarter have since been verified by finance?	A verification rate; a low rate means the benefit pipeline is not real
14. What proportion of pilots reached production with proven value?	A pilot-to-production value rate, with reasons for the failures
15. What did we discover about shadow AI this quarter?	New tools found, consolidated or shut down — and the spend recovered
16. What is the human-in-the-loop review rate, and is it falling as model quality improves?	A falling review rate, or an explanation of why the risk profile requires it to stay
17. Has any consumption anomaly or runaway loop occurred, and what did it cost?	Incidents disclosed with cost attached, and the control change made
18. Is adoption broad or concentrated in a few enthusiastic teams?	Adoption rate and business unit engagement rate, not just total active users



10.3 On architecture and vendors

Question	What a good answer contains
19. Do we have a routing policy, and who sets it?	A documented policy with an owner, reviewed as model prices change
20. Where are we using fine-tuning, and why is retrieval not sufficient there?	A domain-specific justification; anything else should be challenged
21. Are we integrating through a standard protocol or building point-topoint connectors?	A standard, chosen once; point-to-point work should be the exception with a reason
22. What is our vendor concentration, and what is the contingency if one provider changes terms?	A named alternative that has actually been tested, not one assumed to work
23. What indemnity do our AI vendor contracts provide on IP and output liability?	Specific contractual terms, with the residual exposure quantified
24. What is the incremental cost of adding the next use case to the existing platform?	A falling marginal cost; a flat one means the platform is not being reused
25. How much of our AI spend is committed versus on-demand?	A deliberate mix, with reserved capacity where volume is predictable
26. What would we have to rebuild if we moved to a different model family tomorrow?	A short list, if the architecture is well abstracted; a long one is a lock-in warning

10.4 On risk, compliance and the board

Question	What a good answer contains
27. Can we evidence oversight, accountability and an audit trail to a regulator today?	A demonstration, not an assertion — ideally a sample audit trail produced on request
28. Which AI systems would be classified as high-risk under emerging regulation, and are they provisioned accordingly?	A classification per use case, with the compliance cost attached
29. What is our residual risk after controls, and what have we consciously accepted?	An explicit accepted-risk list signed by a named executive
30. Are our evaluation results trending better or worse over time?	A trend line from a continuous evaluation harness, not a point-in-time assessment
31. What is the single largest uncontrolled AI exposure in the enterprise right now?	A specific answer; “nothing significant” usually means nobody has looked
32. What are we reporting to the board on AI, and is it the same set of metrics as last quarter?	A stable metric set — changing definitions between periods destroys comparability

11. WHAT THE CFO SHOULD BE TRACKING

A question bank works once a quarter. A tracking discipline works continuously. The set below is deliberately small: twelve indicators, each with an owner, a cadence and a trigger threshold that prompts action. The thresholds are illustrative starting points — the discipline is having them at all, and holding their definitions constant so that Version 2.0 of this paper can be compared with Version 1.0.

Table 14 · The CFO AI tracking dashboard

Indicator	Definition	Owner	Cadence	Trigger for action
Total AI run-rate	All twelve cost categories, normalised to a monthly figure	Finance / FinOps	Monthly	Growth outpaces assistedtransaction growth for two consecutive months
Cost per AI-assisted transaction	Total AI operating cost ÷ assisted transactions	Finance / FinOps	Monthly	Unit cost rises quarter on quarter
Frontier share of token spend	Spend on premium-tier models ÷ total inference spend	Engineering / FinOps	Monthly	Share exceeds the routing policy target
Hidden-cost ratio	Non-invoiced cost categories ÷ total AI cost	Finance	Quarterly	Ratio unknown, or below 20% — which usually means it is unmeasured, not absent
Consumption anomalies	Count and cost of runaway loops, retry storms and unexpected spikes	Engineering	Monthly	Any occurrence without a corresponding control change
Benefit verification rate	Claimed benefits verified by finance ÷ claimed benefits	Finance	Quarterly	Below 70%
Benefit–cost ratio	Verified benefits ÷ total AI cost, cumulative	Finance	Quarterly	Below 1.0 after four quarters of production operation
Pilot-to-production value rate	Pilots reaching production with proven value ÷ total pilots	Transformation / PMO	Quarterly	Below 25%
Human-in-the-loop review rate	Outputs requiring human review ÷ total outputs, in the target workflow	Operations	Monthly	Flat or rising while volume grows
Evaluation pass rate	Share of evaluation cases passed by the production system	Engineering / Risk	Monthly	Any regression against the previous release
Shadow AI count	Unregistered AI tools and subscriptions discovered	CIO / Procurement	Quarterly	Any increase, or no discovery activity in the period
Adoption and engagement	Active users ÷ eligible users; business units active ÷ total	HR / Transformation	Monthly	Adoption flat while spend grows — the classic valueleakage signature

11.1 Reporting cadence

Table 15 · Suggested reporting rhythm

Cadence	What is reviewed	Audience
Monthly	Run-rate, unit cost, frontier share, anomalies, review rate, adoption	Finance leadership and the AI FinOps function
Quarterly	Benefit verification, benefit–cost ratio, pilot-toproduction rate, shadow AI, evaluation trend, risk register movement	Executive committee
Half-yearly	Use case value coverage, workforce literacy and experience, vendor concentration and lock-in position, architecture review against the twelve blocks	Board or board committee
Annually	Full TCO restatement, business case reconciliation against actuals, risk-adjusted TCO refresh, strategy reset	Board

One discipline underpins all of this: report the same metrics, defined the same way, every period. The most common reason organisations cannot demonstrate AI progress is not that progress is absent, it is that the measurement changed.



12. A TWELVE-MONTH IMPLEMENTATION PATH

The framework in this paper is not a programme in itself. It is installed alongside whatever AI work is already underway. The sequence below assumes an organisation with AI in production in at least one function and no consolidated financial view of it, which describes most enterprises at the time of writing.

Table 16 · Maturity stages and their financial signature

Stage	Description	Financial signature	The move to the next stage
0 — Ad hoc	Scattered pilots, individual subscriptions, no central view	Cost unknown; benefits anecdotal; shadow AI unmeasured	Build an inventory and a registry. Nothing else is possible first
1 — Visible	Central inventory, consolidated procurement, spend attributed by project	Cost known in aggregate; benefits still self-reported	Introduce baselines and a financeowned benefit verification standard
2 — Accountable	Named owners, budgets and thresholds per use case; verified benefits	Cost and benefit reported against the same identifier	Introduce routing, evaluation and unit-cost management
3 — Optimised	Routing policy, model mix managed, continuous evaluation, reuse across use cases	Unit cost falls as volume rises; marginal cost of the next use case is low	Extend into risk-adjusted planning and portfolio-level allocation
4 — Governed	AI FinOps as a standing function; risk-adjusted TCO in every case; board-level reporting	AI managed as a portfolio with hurdle rates, like any other capital allocation	Sustain, and re-baseline as prices and regulation move

12.1 First ninety days

13 **Inventory everything.** Every AI tool, subscription, model endpoint and agent in production or pilot, with an owner against each. Expect the list to be longer than anticipated.

14 **Build the cost sheet.** Populate all twelve categories for the current portfolio, with explicit estimates for absorbed internal effort. An imperfect number beats an absent one.

15 **Impose tagging.** Project, model, environment, business unit and use case, on every workload. This is the precondition for everything in Section 11.

16 **Set the baseline standard.** No new AI funding is released without a measured pre-deployment baseline. Publish this as policy.

17 **Check the agent controls.** Named owner, escalation path, tested kill switch, per-task spend cap, loop limit. Any agent without all five is an open exposure.

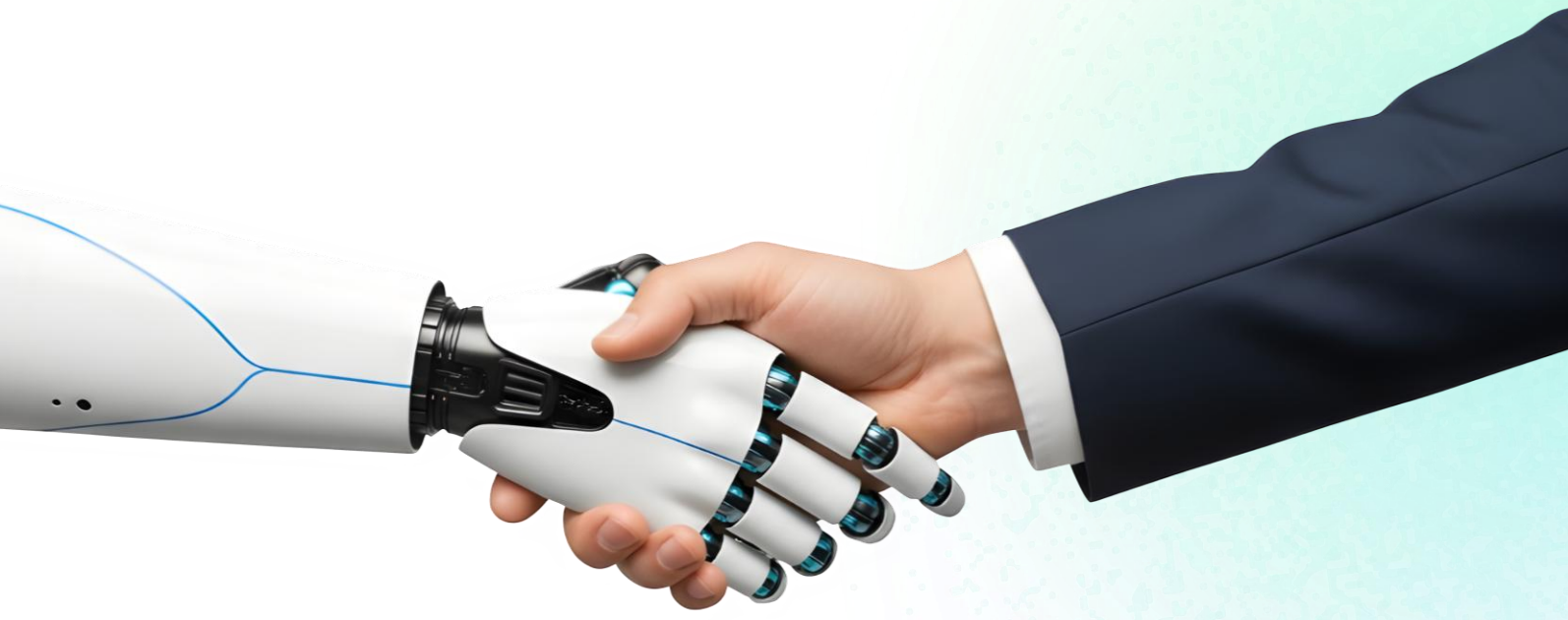
12.2 Months four to six

18. **Stand up AI FinOps reporting.** Monthly consumption by project and model; the first unit-cost trend line.
19. **Introduce benefit verification.** Finance samples and verifies claimed benefits before they enter any reported ratio.
20. **Establish routing policy.** Define which task classes are permitted to call premium-tier models, and who approves exceptions.
21. **Consolidate shadow AI.** Move discovered tools onto enterprise agreements or retire them; record the spend recovered.

12.3 Months seven to twelve

22. **Stand up AI FinOps reporting.** Monthly consumption by project and model; the first unit-cost trend line.
23. **Introduce benefit verification.** Finance samples and verifies claimed benefits before they enter any reported ratio.
24. **Establish routing policy.** Define which task classes are permitted to call premium-tier models, and who approves exceptions.
25. **Consolidate shadow AI.** Move discovered tools onto enterprise agreements or retire them; record the spend recovered.

> The fourth item is the one most organisations skip and the one that changes behaviour most. A published variance analysis between AI business cases and AI actuals does more for forecasting discipline than any framework.



13. VERSION 2.0: WHAT THIS SERIES WILL TEST NEXT

This paper is the first edition of a recurring analysis. Version 2.0 will be published six to twelve months from now, holding the definitions in Sections 5, 7 and 11 constant and reporting movement against them. Stating the hypotheses in advance is what makes the follow-up credible rather than retrospective.

13.1 Hypotheses to be tested

Table 17 · Version 2.0 hypotheses and the evidence that will settle them

#	Hypothesis	Evidence required to confirm or refute
H1	Unit cost per AI-assisted transaction falls materially year on year, even as total AI spend rises	Like-for-like unit cost across the same workflows in the same organisations, over four consecutive quarters
H2	The share of enterprise workloads served by open-weight or locally hosted models rises significantly, driven by cost and data control rather than capability	Token share by model class, and the stated reason for each migration decision
H3	Governance spend stabilises as a predictable share of AI run-rate rather than continuing to grow	Evaluation, guardrail, legal and oversight cost as a percentage of total AI cost, tracked quarterly
H4	Agentic deployment increases total consumption faster than it reduces labour cost in the first year, and reverses in the second	Paired consumption and labour-cost data from organisations at least eighteen months into agentic production use
H5	Organisations that baseline before deployment report materially higher benefit verification rates than those that do not	Benefit verification rate, segmented by whether a predeployment baseline existed
H6	Legal and regulatory compliance moves from an immaterial to a material cost category as AI regulation moves from readiness to enforcement	Compliance cost as a share of TCO, before and after the relevant obligations take effect in each jurisdiction
H7	The hidden-cost ratio falls as attribution improves — not because hidden costs disappear, but because they become visible and therefore managed	The proportion of total AI cost that organisations can attribute to a project, tracked over time

13.2 Open questions carried forward

Does workflow redesign remain the dominant differentiator? The current evidence is strong but was gathered when agentic deployment was still early. It may weaken as agents absorb more of the workflow, or strengthen as redesign becomes the bottleneck.

How should released capacity be valued in a period of hiring restraint? If organisations are not hiring anyway, the avoided-hire argument weakens and the capacity-release class of benefit needs a stricter treatment.

What is the real cost of vendor switching in practice? Switching cost is widely estimated and rarely tested. Version 2.0 will seek cases where a provider was actually changed and the cost was recorded.

Does the twelve-block architecture hold? New blocks may emerge, and one or more current blocks may consolidate into infrastructure and cease to be a decision at all. Model routing in particular may become a default platform capability rather than a build decision.

Where does the technical debt from AI-generated code actually land? The liability has been widely predicted and not yet widely measured. Twelve to twenty-four months of production history should begin to show it.

13.3 The Claims verification list

Claims currently in circulation that this edition has deliberately excluded, and that Version 2.0 will attempt to verify or formally retire:

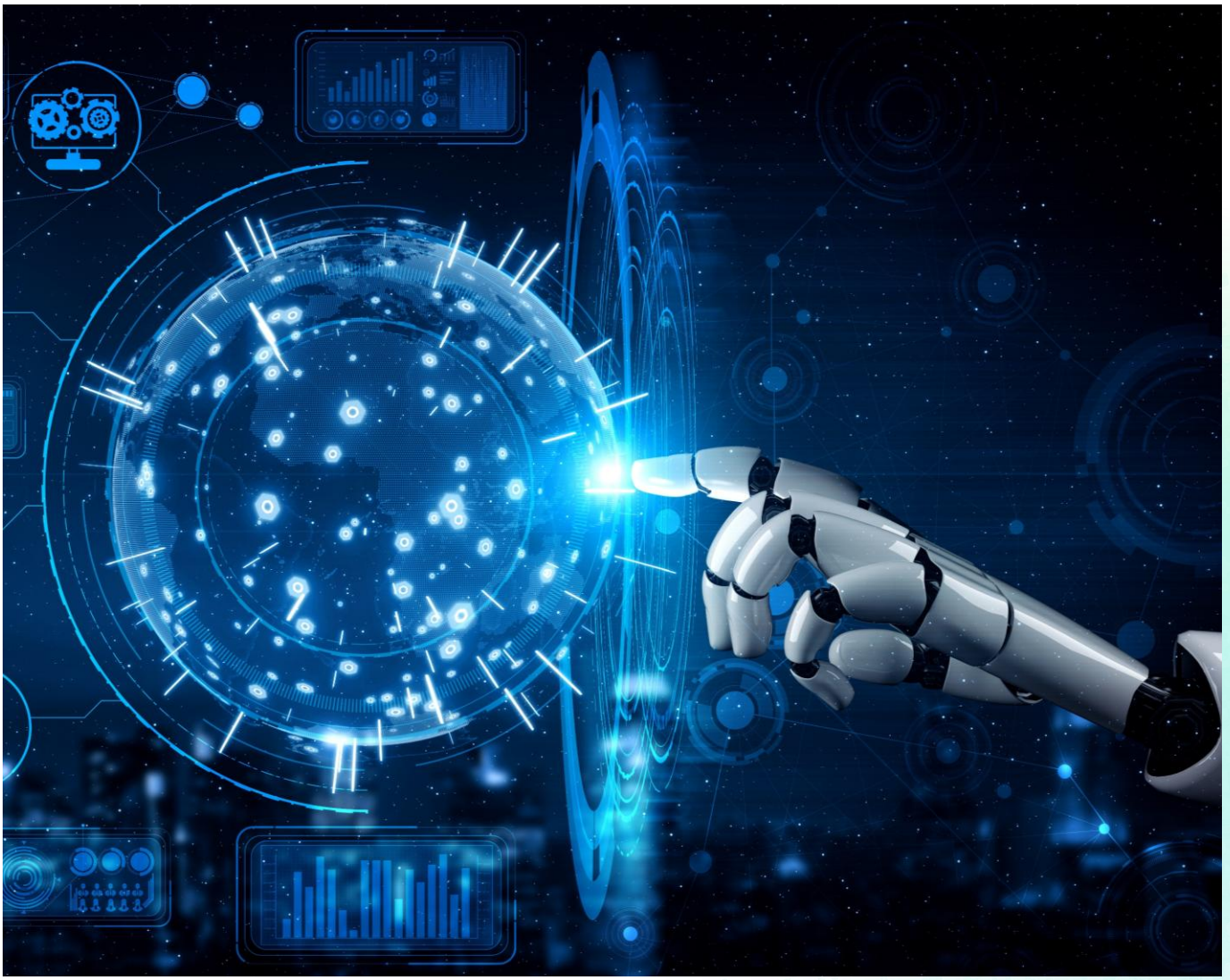
That named large technology companies have consumed an entire annual AI budget within the first four to five months of a financial year.

That AI-driven productivity gains in software delivery are being substantially offset by downstream review and remediation cost — plausible, widely asserted, not yet quantified.

That enterprise AI spend is materially concentrated in a small number of use cases while budgets are allocated broadly — an observed pattern awaiting systematic data.

13.4 What readers can contribute

The value of Version 2.0 depends on comparable data. Organisations willing to contribute anonymised figures against the definitions in Section 11 — particularly unit cost trend, benefit verification rate and hidden-cost ratio will materially improve the next edition, and will receive the comparative benchmark in return.



14. CONCLUSION

The strategic question for the CFO is no longer whether AI creates productivity. That is visible in almost every enterprise that has deployed it. The question is whether AI creates sustainable profitability and answering it requires treating AI as what it actually is: a continuously running operational utility whose cost is determined by architecture, whose value must be measured against a baseline that no longer exists once deployment begins, and whose risk is a cost line that has not yet been invoiced.

Three commitments follow from this paper, and they can be made this quarter without waiting for any technology decision.

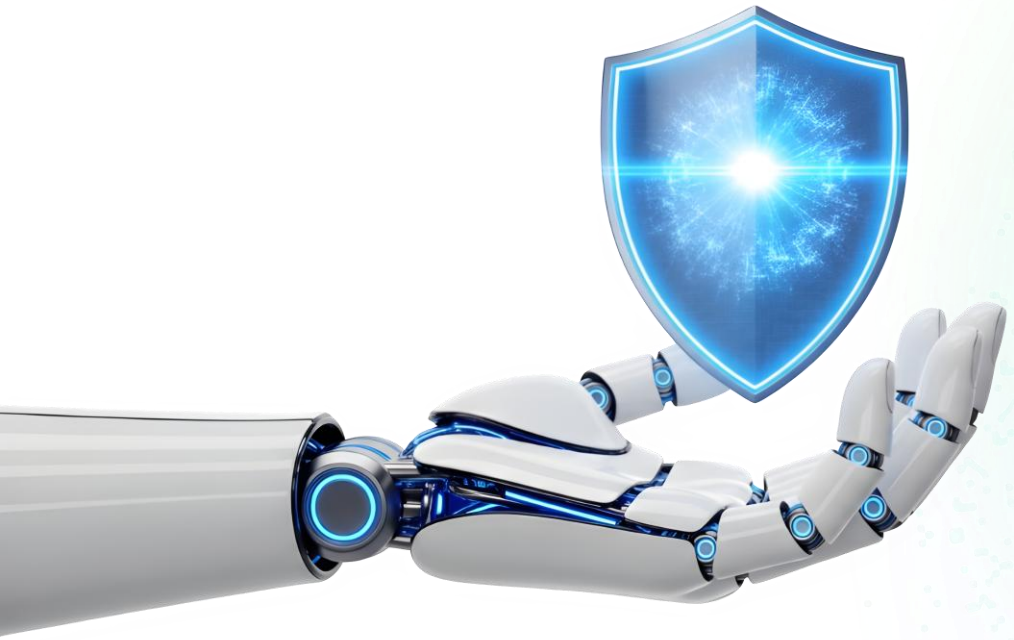
26 **Know the whole cost.** Twelve categories, attributed by project and model, with the hidden categories estimated rather than omitted. A programme whose cost is partly unknown cannot be optimised, only tolerated.

27 **Prove the value.** Baseline before deployment, verify before reporting, and distinguish released capacity from realised saving. The rigour applied to an AI benefit claim should be the rigour applied to any other capital request.

28 **Fund the control.** Evaluation, guardrails, oversight and compliance are what convert AI from a promising capability into a defensible one. Control is cheaper than cleanup — and it is the mechanism by which the other ten building blocks can be scaled with confidence.

None of this requires the organisation to be certain about where the technology is heading. It requires only that the financial discipline keep pace with the deployment. That is the CFO's contribution, and it is the differentiator the evidence keeps pointing to.

> Version 2.0 of this paper will be published six to twelve months from now, measuring the same things in the same way. What matters is not that the numbers in this edition are permanent — they are not — but that the organisation reading it can say, twelve months from now, exactly how much it has moved and why.



APPENDIX A · AI TCO WORKSHEET TEMPLATE

A one-page working template for a single AI use case. Complete every row: an explicit zero is a decision, a blank is an omission.

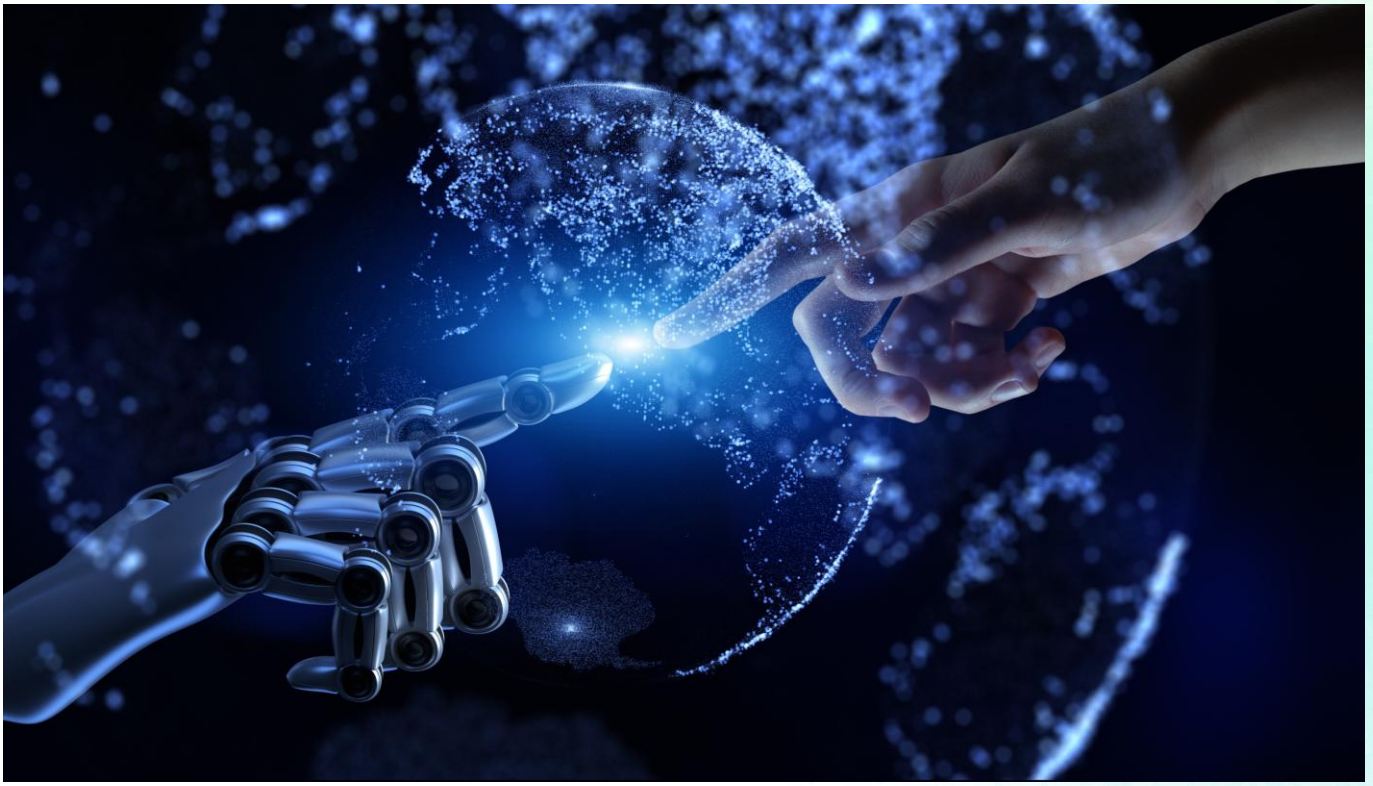
Cost line	Type	Year 1	Year 2–3 p.a.	Basis of estimate / owner
Licences and subscriptions	Fixed			
Inference and API usage	Variable			
Cloud compute, GPU, vector DB	Variable			
Data engineering and pipeline maintenance	Semi-variable			
Security and privacy review	Semi-variable			
Legal and regulatory compliance	Semi-variable			
Model maintenance, drift monitoring, retraining	Semi-variable			
Integration build and technical debt remediation	Fixed / semi			
Human oversight and review	Semi-variable			
Observability and FinOps tooling	Semi-variable			
Consolidated shadow AI / duplicate spend	Variable			
Change management and training	Fixed / semi			
A. Direct TCO (sum of above)				
B. Residual risk provision (Σ probability × impact after controls)				
C. Risk-adjusted TCO (A + B)				
D. Verified benefits (cost-out + redeployed capacity + margin uplift + risk cost avoided)				
E. Benefit–cost ratio (D ÷ C)				

Attach to the worksheet: the pre-deployment baseline measurement, the pilot-to-production bridge, the named system owner, and the accepted-risk list.

APPENDIX B · ROUNDTABLE DISCUSSION QUESTIONS

Eight questions for a CFO peer discussion. They are intentionally open: the value is in the divergence between answers.

Q1	How should enterprises measure AI ROI beyond productivity claims and what financial KPIs should CFOs hold AI investments accountable to?
Q2	What percentage of your organisation's AI cost is currently visible versus hidden — and what governance structure would close that gap?
Q3	How should AI costs be structured: centralised enterprise-wide, or allocated by business unit and project? What are the trade-offs?
Q4	How are you governing shadow AI usage — and what is the financial exposure from unmanaged subscriptions and data governance risk?
Q5	What is the right balance between frontier models and lower-cost open-weight models — and how is that decision being made today?
Q6	How should CFOs evaluate and price long-term vendor lock-in risk into AI investment decisions, given how quickly the market is moving?
Q7	As AI Act obligations and IP liability mature, how are legal and compliance costs being factored into AI business cases?
Q8	Can AI FinOps become a dedicated enterprise governance function — and is it a build, buy or outsource decision?



APPENDIX C - GLOSSARY

Term	Definition for a finance audience
A2A (agent-to-agent)	A shared protocol allowing AI agents in different functions to communicate directly, reducing custom integration work between them.
Agent	An AI system that plans and takes actions — calling tools, updating systems — rather than only producing text. Financially significant because one user request may generate many model calls.
Agent memory	Retained context, preferences and history across sessions, reducing repeated prompting and token consumption.
Drift	Gradual degradation in model performance as the surrounding data, processes or upstream models change. The reason AI maintenance never reaches a stable end state.
Evals (evaluations)	A structured, repeatable test suite measuring accuracy, latency and regression. The governance equivalent of automated testing.
Fine-tuning	Adapting a model's weights on domain data. Carries training, maintenance and retraining cost; now reserved for specialised domains.
Frontier model	The highest-capability, highest-priced model tier. Appropriate for complex, high-stakes tasks; expensive as a default.
Guardrails	Input and output controls covering safety, policy, compliance, output validation and audit trail.
Hallucination	A confident, fabricated output. Its cost appears as human verification effort, rework and, in outwardfacing use, liability.
Inference	The act of running a model to produce an output. The main variable cost line in enterprise AI.
MCP (Model Context Protocol)	A standard protocol connecting AI systems to enterprise applications and data, replacing bespoke point-to-point connectors.
Model routing	Policy-driven selection of the right-sized model for each task by cost and quality. The single largest controllable cost lever.
Multi-agent	Several specialised agents collaborating on a task. Raises engineering cost upfront; reduces labour cost and cycle time thereafter.
Open-weight / local LLM	A model whose weights can be hosted on-premise or in private cloud. Reduces per-token cost, data egress risk and vendor concentration.
RAG (retrieval-augmented generation)	Retrieving enterprise documents before answering, so responses are grounded in source material rather than model memory.
Shadow AI	AI tools and subscriptions in use without approval or central visibility. A cost, compliance and data governance exposure simultaneously.
Token	The billing unit of model usage, roughly a fragment of a word. Cost scales with input length, output length and number of calls.
Tool use	A model's ability to execute actions — call APIs, run queries, update systems — rather than only generate text.

APPENDIX D · SOURCES AND FURTHER READING

Primary source material — the published article series

- **The 12 Building Blocks of Enterprise AI: A CFO's Guide from 2023 to 2026** — architectural comparison and per-block TCO impact.
- **Total Cost of Ownership of AI Solutions: A CFO Perspective** — discussion document prepared for the CFO Board Roundtable, including the twelve-category cost sheet and the visibility gap analysis.
- **AI Governance for Every Leader** — role-by-role governance concerns across CEO, CFO, COO, CIO, CDO and Risk & Legal.
- **20 Ways to Measure AI Impact** — practical KPI set across financial, operational, customer, scaling, workforce and adoption lenses.
- **The Periodic Table of AI Risks** — fifty enterprise AI risks across data, model, agent, security and operations, and governance families.



THIS REPORT HAS BEEN PREPARED WITH GUIDANCE FROM

Mr. Nitin Madan, Mr. Jogendra Singh, Mr. Sugata Sircar and members of The CFO Board

Knowledge Partners:

Sushant Rabra, Partner, KPMG in India

Special Thanks:

Amit Yadav, Co-founder and CTO, Josh Technology Group and Prasoon Shrivastava, Founder and CEO, Zepth

For feedback and queries, write to The CFO Board at secretariat@cfoboard.com

For more information, visit www.cfoboard.com

ASSOCIATION OF CFO WELFARE INDIA

216-241, 2nd Floor, Block-F, International Trade Tower, Nehru Place, New Delhi-110019

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavour to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

For private circulation only.

Some images in this report have been created using artificial intelligence technology.