AN EXECUTIVE BRIEFING

For finance and technology leaders

The Operating Tax

Why Enterprise AI Programs Become Unprofitable in Year Two

Naming the cost that compounds invisibly, the month it turns negative, and the discipline that catches it

Vijayan Seenisamy

Enterprise Agentic AI Systems Delivery

Creator of the AI Role Operating Framework (AI ROF™)

Author of The AI Delivery Manager Blueprint and The Pilot Trap

Version 1.0 · May 2026

Executive Summary

Enterprise AI investment has crossed a visibility threshold in 2026. For a typical ten thousand-person enterprise, the all-in annual cost of an active AI deployment now ranges from seven to fifteen million dollars in year one and eleven to thirty-seven million dollars by year three. Boards are starting to ask the question that always follows that scale of investment. Show us the unit economics.

Most enterprises cannot. This is not a finance failure. It is a structural mismatch between how AI systems behave in production and how finance has been trained to evaluate technology investments.

The evidence is clear. MIT's NANDA initiative analysed three hundred enterprise AI deployments and found that ninety-five percent of generative AI pilots produced no measurable financial impact, despite thirty to forty billion dollars of cumulative investment [1]. McKinsey's State of AI 2025 survey of nearly two thousand companies found that only five and a half percent attribute meaningful EBIT impact to their AI initiatives [2]. Two independent datasets. Same conclusion. The technology mostly works. The operating discipline does not yet exist.

This briefing names what the discipline is missing. The operating tax. The continuous, variable, compounding cost of running AI agents in production once they are deployed. The operating tax is invisible at the project level where finance reviews investment. It becomes visible at the agent level, in year two, when the productivity claim erodes against costs that were never on the original business case.

The briefing introduces a specific empirical construct: the Inversion Point. The month at which an enterprise AI agent transitions from positive to negative unit economics. This is not a single number across all AI deployments. It is a spectrum that varies by agent autonomy. Klarna's customer service AI deployment, the most publicly documented enterprise AI case study of the 2024-2026 period, reached its inversion point at approximately month fourteen. That timing is characteristic of workflow-class agents. Copilot-class assistive AI inverts later or not at all. Autonomous multi-step agents invert earlier. Forrester's 2026 enterprise AI agent panel provides supporting evidence for the workflow-class distribution between month twelve and month eighteen [3]. The Inversion Point, and the specific month within the spectrum where a given agent will reach it, is the metric finance committees should be asking about, and almost none currently are.

This briefing makes four arguments. **First**, that the operating tax is real and structural, not a deployment mistake. **Second**, that current ROI frameworks are not built to surface it, and that finance teams are being shown answers that look credible but understate the cost. **Third**, that the Inversion Point exists, is measurable, and can be managed if it is named. **Fourth**, that the discipline of cost-per-outcome measurement, tied to kill criteria and reviewed monthly, is what separates the five percent of enterprises capturing AI value from the ninety-five percent who are not.

The briefing draws on published research from MIT, Stanford, McKinsey, and Forrester, includes detailed examination of Klarna and Air Canada as documented enterprise AI cases, and treats build-versus-buy economics, model selection routing, and the EU AI Act compliance trajectory as distinct cost layers requiring separate analysis. It is written for finance leaders, technology leaders, and the boards that approve enterprise AI investment.

1. The Question Boards Are Starting to Ask

In 2026, enterprise AI investment has crossed a visibility threshold. The technology budget line item that was small in 2024 and meaningful in 2025 has become material in 2026. Boards are starting to ask the question that always follows visibility.

> *"We have approved this investment. We have seen the productivity claims. Now show us the unit economics."*

This is a reasonable question. It is the question finance asks of any significant capital deployment. The challenge is that for most enterprise AI programs in 2026, the unit economics cannot be produced. Not because the data is hidden. Because the measurement infrastructure does not exist.

Three things are simultaneously true. The AI capability is being deployed. The productivity claims are being made. The actual cost-per-outcome of any specific agent in production is not being computed. Forrester's 2026 agent panel found that only thirty-eight percent of production agents have automated evaluations running on every prompt change, the prerequisite for any honest outcome measurement [3].

Most enterprises have the first two. Very few have the third. This briefing examines why and proposes a specific construct for understanding when AI agents become unprofitable: the Inversion Point.

2. The Klarna Case: The Operating Tax in Public

The most publicly documented enterprise AI deployment of the last three years is Klarna's customer service AI assistant. Its arc, from announcement to partial reversal, is the operating tax made visible in real time. It is also the closest the industry has to a controlled experiment on AI unit economics at scale.

The announcement (February 2024)

Klarna announced that an OpenAI-powered AI assistant had handled 2.3 million customer service conversations in its first month, equivalent to the workload of approximately seven hundred full-time agents. The company projected a forty million dollar profit improvement for 2024 [4]. The CEO publicly tied the deployment to AI's capacity to replace human knowledge work at scale. The story was covered globally as evidence that AI displacement had arrived.

The early validation (2024-Q1 2025)

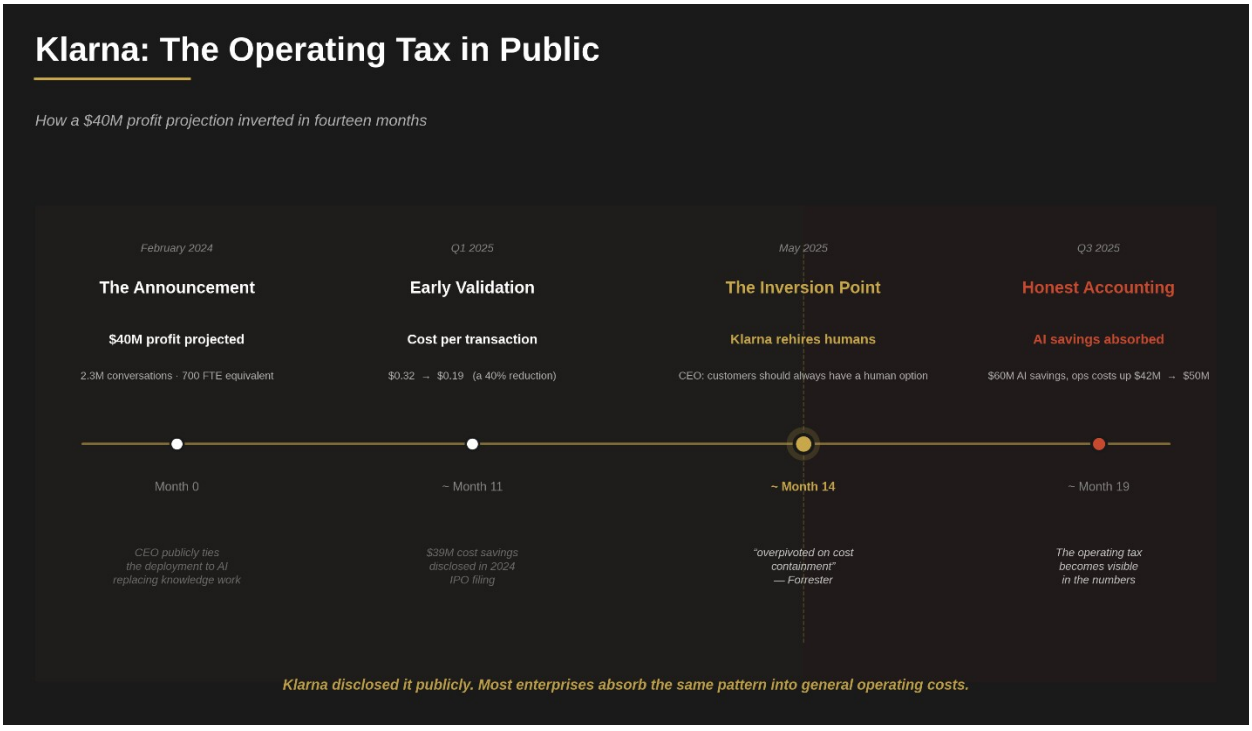
Initial results appeared to confirm the case. Customer service cost-per-transaction dropped from $0.32 in Q1 2023 to $0.19 in Q1 2025, a forty percent reduction over two years. Klarna reported customer satisfaction remained steady. The company stated that the AI agent had delivered thirty-nine million dollars of cost savings in its 2024 IPO filing [5].

The reversal (May 2025 onward)

In May 2025, approximately fourteen months after launch, Klarna announced it was hiring customer service representatives again. The CEO publicly stated that customers should always have the option to speak with a human. Forrester's principal analyst characterised the strategy as 'overpivoted on cost containment without thinking about the longer-term impact of customer experience' [6]. Customer satisfaction data had deteriorated on complex service interactions. The AI handled volume well but not nuance. The cost savings projected at announcement did not fully materialise because handling quality issues consumed more than the AI saved.

The honest accounting (Q3 2025)

By Q3 2025, Klarna reported sixty million dollars in cost savings from the AI agent. The same earnings release also disclosed that customer service and operations costs **increased** year-on-year, from forty-two million dollars to fifty million dollars [7]. The AI savings were real. They were absorbed by other costs that grew faster than the AI savings could offset them.



The structural lesson

Klarna's experience demonstrates the operating tax operating in public. The initial cost savings were real but smaller than projected. The operating tax of running AI at scale was higher than the business case anticipated. Quality degradation on complex interactions required human capacity to be rebuilt at substantial cost. The reversal cost is the cost most business cases never model.

Klarna is unusual in disclosing this publicly. Most enterprises absorb similar costs without disclosure. What makes the Klarna case analytically valuable is not that it was different from other workflow-class enterprise AI deployments. It is that it played out at the same pace as other agents but visibly. The reversal at month fourteen is the canonical example of the Inversion Point pattern introduced in Section 3, anchored on the workflow-class case. Other agent classes invert at different timings, as the next section develops.

3. The Inversion Point

The single most useful question a finance committee can ask about an AI agent in production is also the question almost no enterprise has yet learned to answer.

> *In what month does this agent's cost-per-outcome exceed its value-per-outcome?*

That month is the Inversion Point. It is the structural feature of enterprise AI agents that no other technology deployment has displayed in the same way. Software does not have an Inversion Point. Cloud infrastructure does not have an Inversion Point. AI agents do, and the reason matters.

The Inversion Point is not a single number. It is a spectrum that varies by agent autonomy. Copilot-class agents that assist human users invert latest, often beyond month twenty-four or not at all. Workflow-class agents that act within defined boundaries follow what this briefing calls the Klarna pattern, with inversion typically between month twelve and month eighteen. Autonomous agents that chain reasoning across multiple steps invert earliest, often between month six and month twelve. The rest of this section develops the construct around the workflow-class case because Klarna anchors it publicly. The variance across classes is addressed below.

Why the Inversion Point exists

Three forces, operating simultaneously, push AI agent economics in opposing directions from month one of production.

The productivity curve flattens. Initial deployment captures the largest gain because the most repetitive tasks are easiest to automate. The marginal task being automated in month twelve delivers less value than the first task automated in month one. The productivity claim is highest at the start.

The model consumption curve accelerates. Users learn what the agent can do. They send longer prompts. They use the agent for more complex tasks. They chain multiple agent calls together. Mature usage costs more per request than early usage by a factor of two to five.

The operating tax compounds. Model deprecation forces re-tunes. Regulatory frameworks expand. Evaluation infrastructure grows with agent footprint. Compliance overhead increases. The fixed costs of running an agent at production scale grow faster than the variable productivity gains, particularly after month nine to twelve.

The Inversion Point is the month at which these three forces produce a unit economics crossover. **Cost-per-outcome exceeds value-per-outcome**. The agent continues to generate volume but stops generating net value. From the Inversion Point forward, every additional outcome the agent produces makes the enterprise marginally less profitable than not running the agent.

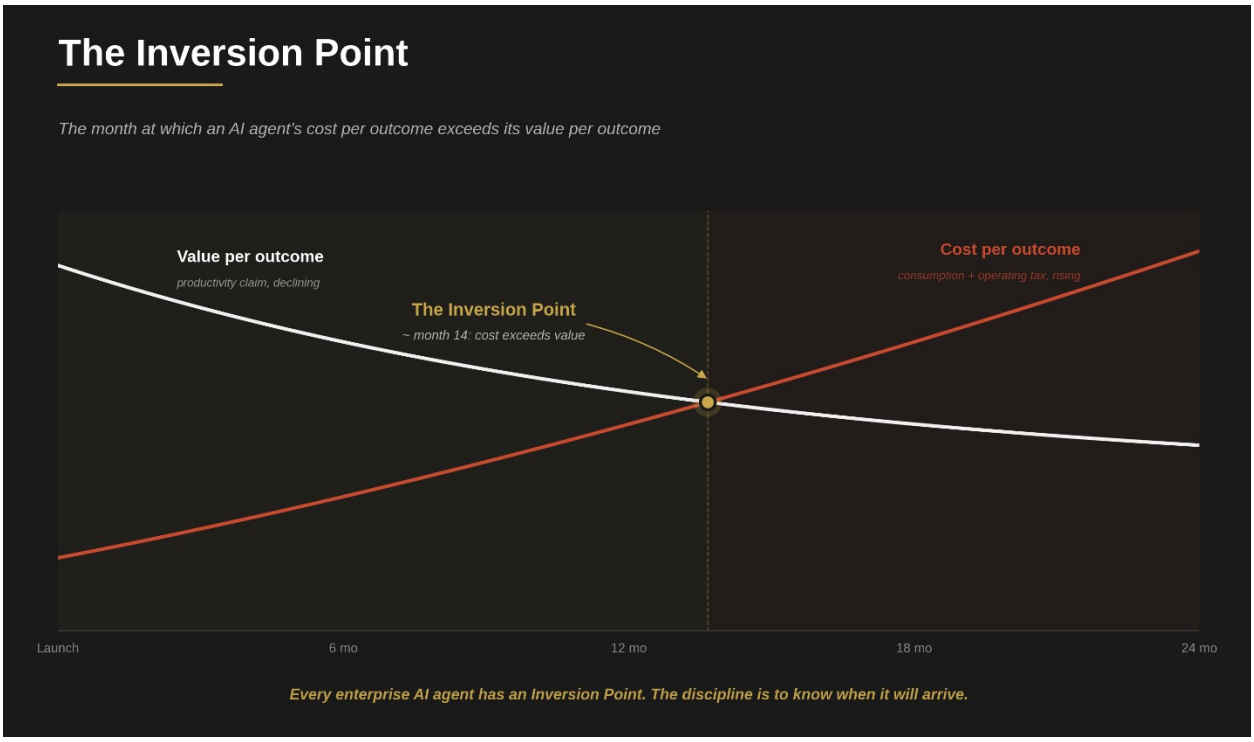


Figure: The Inversion Point construct, shown for workflow-class agents with crossover at approximately month fourteen. The structural mechanism applies to all enterprise AI agent classes. The timing of the crossover varies with agent autonomy, as Section 3 develops.

The empirical evidence (workflow-class agents)

Klarna's customer service deployment reached its Inversion Point at approximately month fourteen. The company began publicly rebuilding human capacity in May 2025, fourteen months after the February 2024 launch. The Q3 2025 earnings release confirmed the pattern: AI savings of $60M, total customer service

and operations costs **increased** by $8M year-on-year [7]. The net contribution had inverted between the original announcement and the reversal.

Forrester's 2026 enterprise AI agent panel provides supporting evidence for the workflow-class pattern. Twenty-two percent of agent deployments report negative ROI at twelve months [8]. The proportion grows through months thirteen to eighteen before stabilising as kill criteria activate or agents are restructured. Klarna sits within the most common range. Other documented enterprise deployments, including Air Canada's chatbot reversal and several banking sector retractions through 2024-2025, follow timings broadly consistent with this range. The distribution is indicative, not statistically proven across hundreds of cases.

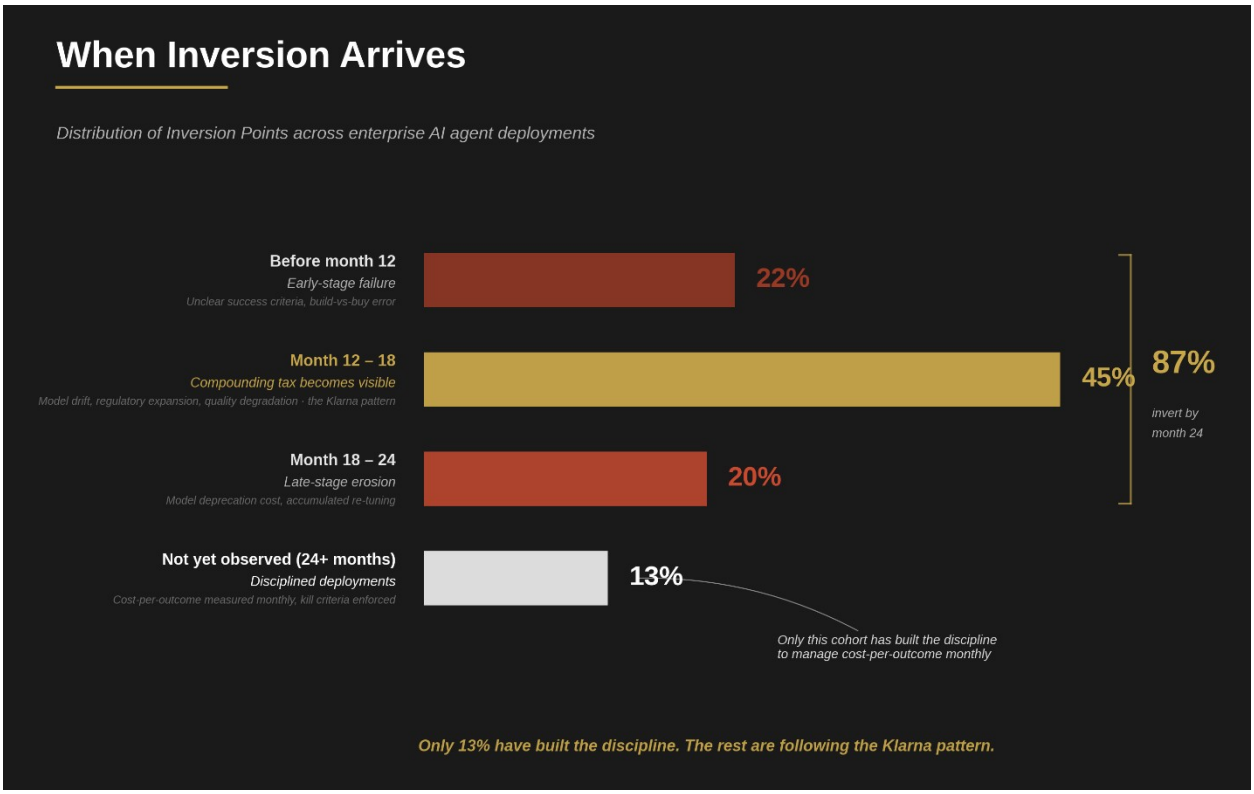


Figure: When the Inversion Point arrives for workflow-class enterprise AI deployments. Distribution does not include copilot-class agents, which typically do not invert within this window, or autonomous agents, which typically invert earlier. The pattern is indicative, anchored on documented cases including Klarna.

Distribution of Inversion Points across documented enterprise AI agent deployments

Inversion Point	% of deployments	Typical pattern	Indicative drivers
Before month 12	~22%	Early-stage failure	Unclear success criteria, build-vs-buy error

Inversion Point	% of deployments	Typical pattern	Indicative drivers
Month 12-18	~45%	Compounding tax becomes visible (Klarna pattern)	Model drift, regulatory expansion, quality degradation
Month 18-24	~20%	Late-stage erosion	Model deprecation cost, accumulated re-tuning
Not yet observed (>24m)	~13%	Disciplined deployments	Cost-per-outcome managed monthly with kill criteria

Distribution derived from Forrester 2026 agent panel data and convergent disclosures from publicly documented enterprise deployments including Klarna, Air Canada, and JPMorgan COIN. Indicative not predictive.

Why the Inversion Point varies by agent type

The Inversion Point sits at different months for different agent classes because cost, value, and operating tax compound at different rates depending on autonomy. Three classes of enterprise AI agent emerge from current deployments.

Copilot-class agents are low-autonomy. They suggest, the human acts. Examples include Microsoft 365 Copilot, GitHub Copilot, and most embedded assistive features in productivity software. Cost stays low because consumption is bounded by human acceptance of suggestions. Value stays low but stable because humans validate outputs. Operating tax stays low because errors are caught at the point of use. Copilot-class agents typically do not invert within twenty-four months, and in some deployments may not invert at all.

Workflow-class agents are medium-autonomy. They act within defined boundaries. Examples include Klarna's customer service deployment, internal task automation agents, and most current enterprise chatbots. Cost compounds because chained calls and memory writes accumulate. Operating tax compounds because governance and evaluation infrastructure must grow with the agent footprint. Value erodes on edge cases because the agent acts where humans previously caught errors. Workflow-class agents typically invert between month twelve and month eighteen. This is the Klarna pattern.

Autonomous agents are high-autonomy. They plan, act, and chain reasoning across multiple steps. Examples include multi-agent orchestrations, AutoGPT-style implementations, and some emerging frontier deployments. Cost compounds fastest because a single user task may produce twenty or more chained model calls plus memory writes plus search calls. Operating tax compounds fastest because the failure modes are harder to predict and the governance burden is highest. Value can be high when the system is reliable, but reliability degrades rapidly with cascade failures across the reasoning chain.

Autonomous agents show early indications of inversion risk emerging faster, between month six and month twelve.

The classification matters for finance committees because the kill criterion and the projected Inversion Point should be set with the agent's autonomy class in mind. A workflow-class business case that assumes a thirty-six month positive contribution is structurally implausible. An autonomous-agent business case that assumes any positive contribution beyond twelve months requires the team to have demonstrably built the disciplines this briefing describes.

Agent class	Autonomy	Typical Inversion Point	Driver of timing
Copilot-class	Low	Month 24+ or not at all	Cost stays low; humans validate value
Workflow-class	Medium	Month 12-18 (Klarna pattern)	Costs compound; value erodes on edge cases
Autonomous	High	Month 6-12	Costs compound fastest; reliability decays

Why this construct matters for finance

The Inversion Point gives finance committees something they currently lack. A specific, measurable question with a specific, defensible answer. Not 'is the AI investment working?' but 'when will the unit economics invert, and what is the kill criterion when it does?'

Three implications follow from naming the construct.

- Business cases should explicitly project the expected Inversion Point. An agent projected to invert at month thirty-six is a different investment from one projected to invert at month fourteen, even if both look profitable in year one.
- Monthly cost-per-outcome measurement becomes mandatory, not optional. Without it, the Inversion Point arrives invisibly and the program is committed before the inversion is recognised.
- Kill criteria should be set in advance and tied to the projected Inversion Point. If the agent inverts six months earlier than projected, the case for continuation requires explicit re-approval, not assumed continuation.

The thirteen percent of deployments that have not inverted at twenty-four months are not running luckier agents. They are running disciplined agents inside operating models that catch the cost compounding before it overwhelms the value generation. That discipline is what this briefing is about.

4. Why Finance Approves AI Cases That Should Not Pencil

Finance teams are not failing to apply rigour to AI investments. They are applying the same rigour they apply to ERP rollouts, cloud migrations, and digital transformations. The cases are usually well-built. The numbers usually pencil.

The cases pass because they are constructed using four moves that have worked for every previous technology investment. Each move is reasonable in isolation. Combined for AI, they produce business cases that look healthy at the project level while hiding fragile unit economics at the agent level.

Move one. Cost is treated as capex plus run-rate, not as variable consumption.

Traditional software has predictable run-rates. A licensed application costs X per user per month and that number does not move. An AI agent has variable consumption that scales with what users feed it. A 50,000-word document costs ten times more to process than a 5,000-word document. The business case shows a flat opex line where the operational reality is a curve that compounds with usage maturity.

Move two. Productivity benefit is calculated as time saved multiplied by loaded cost.

This calculation works for back-office automation. It does not work for AI agents in knowledge work. The most rigorous published study to date, by Brynjolfsson, Li, and Raymond at Stanford and MIT, examined 5,179 customer support agents at a Fortune 500 firm using AI assistance. Average productivity uplift was fourteen percent. The effect concentrated heavily among low-skilled and novice workers (thirty-four percent uplift), with minimal impact on experienced workers [9]. This is materially lower than the twenty-five to forty percent uplift commonly claimed in enterprise business cases, and the heterogeneity matters. AI does not lift everyone equally. The productivity case has to model where the uplift actually concentrates.

Move three. The denominator is the project, not the agent task.

A five million dollar business case for a seven million dollar benefit looks healthy at the project level. The same case at the per-task level might show sixty cents of cost for a task that delivers forty-five cents of value. The project envelope absorbs the unit economics problem. Until the Inversion Point, when the envelope can no longer hide it. This is exactly what unfolded at Klarna in public.

Move four. Year-two costs are assumed to be lower than year one.

Traditional software follows this pattern. Initial deployment cost is high, ongoing maintenance is lower, and the curve flattens. AI does not follow this pattern. Models deprecate every twelve to eighteen months forcing re-tunes. Regulatory compliance work compounds with each new framework. Evaluation infrastructure expands as the agent footprint grows. Year two is often more expensive than year one, not less.

The business cases finance approves are not wrong by their own logic. They are wrong by the operating reality of AI systems.

5. The Three Layers of Enterprise AI Cost

Most enterprise finance teams see the first layer of AI cost clearly. They see the second layer partially. They see the third layer rarely. Understanding all three is the first step to honest unit economics.

Layer one. Seat licensing.

Per-user subscription costs for AI capabilities embedded in productivity applications. This is the visible, predictable, budgetable layer. For a ten thousand person enterprise on a Gemini Enterprise or Microsoft 365 Copilot deployment, this layer runs five to nine million dollars annually depending on tier. Finance sees this layer on the monthly invoice. It is the layer most CFOs assume is the AI budget.

Layer two. Model consumption.

Per-token costs charged when AI agents call foundation models programmatically. This layer is variable, growing, and largely invisible at the finance level. It scales with agent usage, document length, context depth, and number of reasoning steps. For an enterprise running citizen agents and flagship agents at production scale, this layer can range from one to fifteen million dollars annually with substantial year-on-year growth.

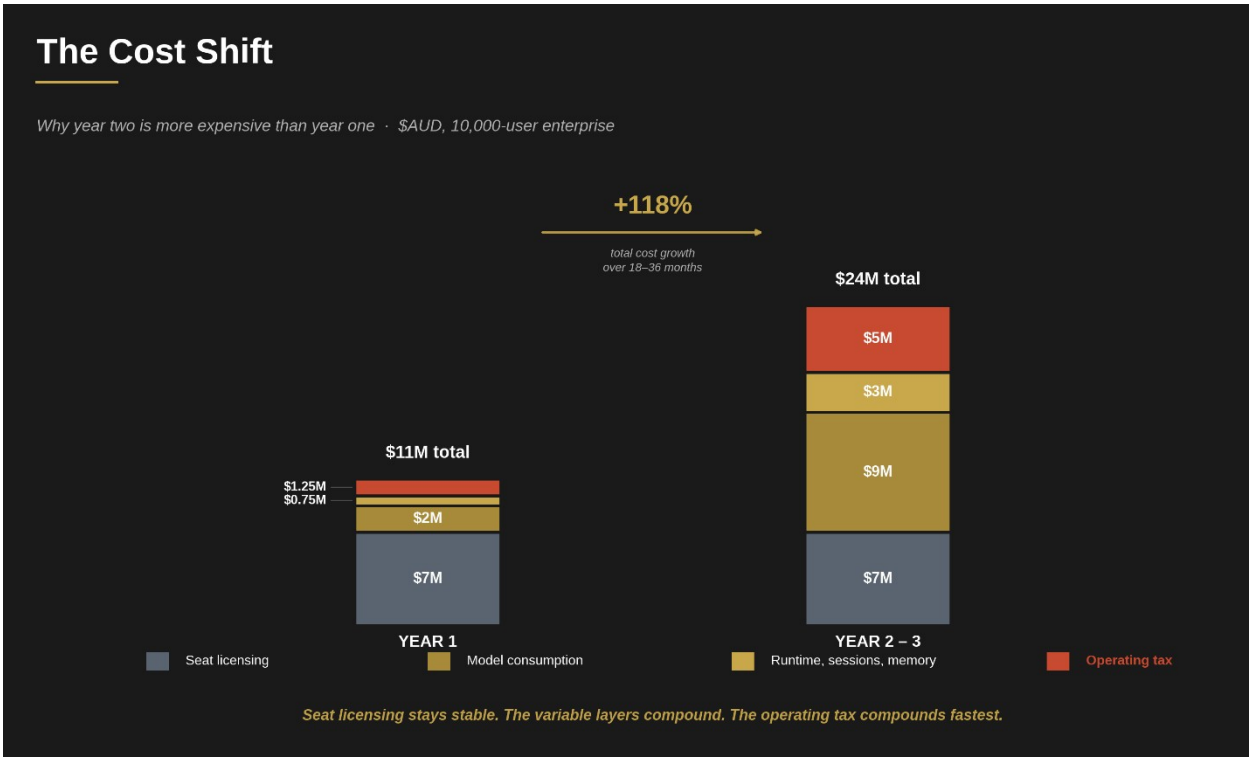
Layer three. Runtime, sessions, memory, and grounding infrastructure.

Compute and storage costs for deployed agents, the storage of agent context and memory, and the search infrastructure that grounds agent responses in enterprise data. This layer is the most opaque. It accumulates continuously, includes multiple metered components, and is rarely tagged at the agent level.

Indicative annual cost ranges for a 10,000-user enterprise (AUD)

Cost layer	Today (early-stage)	Year 2-3 (mature)
Seat licensing	$5M – $9M	$5M – $9M
Model consumption (tokens)	$1M – $3M	$3M – $15M
Runtime, sessions, memory, search	$500K – $1M	$1M – $5M
Operating tax (governance, eval, engineering, compliance)	$500K – $2M	$2M – $8M
Total realistic range	**$7M – $15M**	**$11M – $37M**

Ranges are illustrative and based on patterns observed across large enterprise AI deployments. Actual costs vary significantly by sector, agent maturity, model selection strategy, and vendor configuration.



6. The Build-vs-Buy Decision and Its Cost Implications

Before any of the cost layers can be analysed, a more fundamental decision determines the trajectory of the entire program. Whether the enterprise builds its AI capability internally, buys it from specialised vendors, or partners on a co-build model. The research is unambiguous on which path produces better outcomes.

The evidence

MIT's NANDA study found that **external vendor partnerships reach successful deployment at roughly twice the rate of internal builds** [10]. Specifically, purchased AI tools and vendor partnerships succeed approximately sixty-seven percent of the time, while internal builds succeed at one-third of that rate. The lead author's blunt summary: 'Almost everywhere we went, enterprises were trying to build their own tool, but the data showed purchased solutions delivered more reliable results.'

This finding is counterintuitive for many enterprises. The instinct to build internally rests on three assumptions: competitive differentiation from proprietary capability, lower long-term cost, and better fit to specific workflows. The MIT data suggests all three assumptions are incorrect on average.

Why internal builds fail more often

Four structural reasons explain the build penalty.

- Talent concentration. Vendors employ AI engineering teams that have built similar capabilities dozens of times. Internal teams typically build one. The deployment patterns vendors have already debugged are not learned internally until they fail in production.
- No external accountability. Internal builds drift in scope and timeline because no contractual milestone forces delivery. Vendor partnerships impose delivery discipline that internal projects lack.
- Workflow brittleness. MIT's research identified workflow brittleness as the primary failure mode of internal builds. Vendors have learned through repeated deployments which workflow assumptions break in production. Internal teams have to learn this on their own data.
- Operating discipline import. Vendor partnerships bring tested operating patterns. Internal builds usually invent the operating model alongside the technical build, doubling the failure surface.

When internal builds make sense

The MIT data does not say internal builds always fail. It says they fail at twice the rate of partnerships on average. Three conditions justify internal builds despite the higher failure risk.

- The capability is genuine competitive differentiation. If the AI capability is itself a product moat (Klarna's AI as a customer-facing product, JPMorgan's COIN as a contract analysis differentiator), internal build can be defensible. If the AI capability is operational efficiency, partnership is almost always better economics.
- Regulatory or data sovereignty requirements prevent vendor access. Financial services and healthcare sometimes face genuine constraints that block vendor use of sensitive data.
- The enterprise has demonstrated internal AI delivery capability across multiple prior deployments. The success rate for organisations with five or more production agents shipped successfully internally is higher than the MIT average.

For most enterprises in 2026, none of these three conditions hold. The default decision should be partner-first. The exception is the defensible position, but it is the exception, not the rule.

The cost implications

The build-versus-buy decision affects every cost layer in this briefing. Internal builds have higher operating tax than partnerships because the operating discipline must be invented locally. Internal builds typically have lower model routing discipline than partnerships, which means model consumption layer costs are higher. Internal builds often have lower evaluation infrastructure than partnerships, which Forrester data shows produces a nine percent versus forty-seven percent rollback rate [11]. The build penalty is not just a higher failure rate. It is structurally higher operating costs across every layer for the deployments that survive.

For finance committees, this translates into a specific question: **if the program is building internally, what is the multiplier on the operating tax compared to the partnership benchmark?** For most enterprises, the honest answer is one and a half to two times. The business case should reflect this.

7. The Hidden Lever: Model Selection Economics

Within the model consumption layer, there is a cost lever that most finance teams have never seen analysed and that most AI deployments are not using well. Model routing. The discipline of directing each request to the cheapest model capable of handling it, rather than using one premium model for everything.

The lever exists because the AI model market has experienced a price collapse that finance teams have not fully internalised.

The price collapse

GPT-4 launched in March 2023 at $30 per million input tokens and $60 per million output tokens. By April 2026, Google's Gemini Flash family was available at $0.10 per million input tokens, a ninety-nine point seven percent reduction in three years [12]. At the frontier tier, capability has improved while prices have fallen by roughly ninety percent.

Indicative model pricing tiers, May 2026 (USD per million tokens)

Tier	Example models	Input / output	Typical use
Frontier reasoning	GPT o1, Claude Opus 4.5	$5-15 / $25-60	Complex multi-step reasoning
Balanced flagship	GPT-4o, Claude Sonnet 4.5, Gemini 2.5 Pro	$1.25-3 / $10-15	Standard enterprise workloads
Cost-efficient workhorse	GPT-4o-mini, Claude Haiku 4.5, Gemini Flash	$0.10-1 / $0.40-5	High-volume routine tasks

Why routing matters at enterprise scale

Stanford's FrugalGPT research demonstrated that intelligent model routing can reduce cost by fifty to ninety-eight percent while matching or exceeding the accuracy of frontier models on enterprise tasks [13]. Independent enterprise data from multi-model routing platforms shows median cost reductions of seventy-one percent versus equivalent single-provider deployments, with the top quartile achieving reductions exceeding eighty percent [14].

Industry data suggests organisations using a single model for all tasks are overpaying by **forty to eighty-five percent** versus organisations using intelligent routing [15]. For an enterprise with a $5M annual model

consumption layer, this is a $2M to $4M annual cost difference, depending entirely on engineering discipline. None of this appears on the original business case.

The questions finance should ask

Three questions to direct at the AI delivery team for any agent at production scale.

- What proportion of agent requests route to the cheapest qualified model versus a frontier model? If the answer is 'we use one model for everything,' the consumption cost is forty to eighty-five percent above optimal.
- Is semantic caching active for repeated queries? Production systems with mature caching report forty percent cache hit rates, equivalent to forty percent cost reduction on those requests.
- Is the model selection strategy reviewed quarterly as new models release? Model prices fall continuously. Static selection becomes outdated within months.

These three disciplines together typically reduce model consumption costs by fifty to seventy percent. They are not technical breakthroughs. They are operating model decisions.

8. The Compounding Cost Layer: Governance, Compliance, and Risk

Beneath the visible cost layers sits a fourth layer that most business cases ignore entirely and that compounds faster than any other. Governance, regulatory compliance, and incident response. This is the layer most likely to determine whether an enterprise AI program survives its second year.

The EU AI Act trajectory

The EU AI Act's high-risk system obligations become fully enforceable in August 2026. For multinational enterprises, this is no longer a planning question. Compliance cost estimates from current implementations: large enterprises (>€1B revenue) face $8-15 million initial investment for high-risk AI systems, with ongoing annual costs of $2-5 million [16]. Mid-size companies face $2-5 million initial and $500K-2M annually. Vendors are already charging 20-30% premiums to reflect certification costs and engineering overhead [17].

For an Australian enterprise serving EU customers or operating in EU markets, these costs apply. For an enterprise operating purely domestically, similar regulatory frameworks are arriving. Australia's voluntary AI Safety Standard was released in late 2024. Mandatory high-risk AI regulation is in active development through 2026. The compliance cost trajectory is consistent across jurisdictions even if the timing differs.

The incident response cost

AI incidents are no longer hypothetical. In February 2024, the British Columbia Civil Resolution Tribunal ruled that Air Canada was legally liable for misinformation provided by its customer service chatbot. The damages awarded were trivial at $812.02. The precedent was not. The tribunal ruled that 'a chatbot is still

just a part of Air Canada's website. It should be obvious that Air Canada is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot.' [18] The decision established that enterprises cannot disclaim responsibility for AI-generated outputs on their platforms.

The Air Canada damages were small. The structural cost is not. Every enterprise running customer-facing AI now operates under the legal exposure this precedent established. The cost of preventing repeat incidents includes governance review of all outputs, audit trails for every customer interaction, escalation paths for policy-sensitive queries, and incident response capability when the AI inevitably says something wrong.

The components of the governance layer

Eight cost categories typically excluded from AI business cases.

- Model audit trails. Cryptographic records of which model, which prompt version, and which retrieval source were active for any given decision. Required under EU AI Act Article 12 for high-risk systems.
- Bias and fairness evaluation. Mandatory under most emerging AI regulations. Requires statistical testing infrastructure and ongoing monitoring.
- Incident response capability. Dedicated team or retainer to manage AI incidents within regulatory response windows (72 hours under EU AI Act Article 73).
- Red-team infrastructure. Ongoing adversarial testing of agents for prompt injection, jailbreaking, and unsafe outputs.
- Model Armor or equivalent runtime protection. Increasingly standard for production agents.
- Compliance documentation. Technical files, conformity assessments, registration with EU AI Act database for high-risk systems.
- Third-party assurance. External audits required for ~30-40% of high-risk AI systems.
- Legal review of AI-generated content for customer-facing applications. New role categories emerging across regulated industries.

Aggregated across an enterprise agent portfolio, these costs typically run **ten to twenty-five percent of the build cost annually** [19], growing as the regulatory framework matures. None of these costs were on the original business case for any AI agent deployed before 2025. All of them apply now.

Why this layer compounds faster than others

Three structural forces make the governance layer the fastest-growing component of the operating tax.

- Regulatory frameworks expand. The EU AI Act is the first comprehensive AI regulation. It will not be the last. Each new framework adds documentation, audit, and compliance requirements without removing the existing ones.
- Incident precedents accumulate. Air Canada is one of many. As courts establish clear liability for AI outputs, the standard of care required to defend against negligent misrepresentation rises. The cost of operating safely grows even without regulatory change.
- Model deprecation forces re-certification. Each time the underlying model changes, the conformity assessment may need to be repeated for high-risk systems. With models deprecating every twelve to eighteen months, this is a recurring cost.

This is why year two is consistently more expensive than year one for enterprise AI agents, even when usage growth is flat. The governance layer compounds independently of usage.

9. The Metric That Replaces ROI

The metric that exposes the operating tax is cost-per-outcome. Not cost-per-task. Not cost-per-API-call. Not cost-per-token. Cost-per-outcome, where outcome is defined as a unit of work the business actually values.

> *Cost per successful outcome = total fully-loaded cost of the agent / number of outcomes that delivered intended business value*

The reason almost no enterprise computes this correctly is that each of the three components is hard. Total fully-loaded cost requires aggregating across token costs, runtime compute, sessions and memory, search and grounding, evaluation infrastructure, monitoring tools, dedicated engineering time, compliance overhead, and incident response. Most enterprises have these costs scattered across five different cost centres with no unified view.

Number of outcomes requires defining what counts as an outcome. Klarna's experience demonstrates the risk. Volume looked like success until quality measurement revealed it was not.

Intended business value requires linking each outcome to a dollar value. A successful conversation saved X in human handling cost? Generated Y in upsell? Avoided Z in churn? Most agents are deployed without this linkage made explicit before deployment.

The enterprises that can compute cost-per-outcome are doing all three things well. The enterprises that cannot are usually doing one well and assuming the other two.

10. Five Characteristics of Disciplined Enterprises

A small number of enterprises have built the discipline required to produce honest cost-per-outcome. McKinsey's State of AI 2025 survey identified them as the five and a half percent who attribute material

EBIT impact to AI[2]. The disciplined enterprises share five characteristics. None of them are technical. All of them are operating model decisions.

Characteristic one. Success criteria defined before deployment.

Every agent has a defined success metric agreed before deployment, with a measurable dollar value per successful outcome. The discipline forces the business owner to articulate value upfront, before the engineering work begins.

Characteristic two. Tagged consumption at the agent level.

Every agent has its cloud and platform consumption tagged so cost is attributable per agent, per business unit, per use case. Forrester's 2026 panel found that organisations missing per-agent cost attribution miss their AI infrastructure forecasts by more than twenty-five percent [20]. Without tagging, total cost can be measured. Cost per agent cannot.

Characteristic three. Monthly cost-per-outcome review with Inversion Point projection.

Every agent has a monthly cost-per-outcome calculation reviewed in a joint forum of business owner, finance, and AI delivery. The review explicitly projects the Inversion Point based on current cost and value trajectories. If the projected Inversion Point is moving earlier than originally estimated, that signal triggers re-engineering or kill criteria review.

Characteristic four. Predefined kill criteria tied to the Inversion Point.

Every agent has kill criteria agreed before deployment. If cost-per-outcome exceeds value-per-outcome for two consecutive months, or if the projected Inversion Point falls below the original business case threshold, the agent is either re-engineered or retired. The kill criteria are predefined because in-the-moment decisions about cost-overruns are biased toward continuation.

Characteristic five. Operating tax reported separately from build cost.

Every agent has its operating tax (the all-in run cost including governance, monitoring, evaluation, engineering time, and compliance overhead) reported separately from the build cost. This separation makes the tax visible. Without separation, the tax is absorbed into general technology overhead and never named.

11. Implications for Boards and Finance Committees

If the operating tax is real, the Inversion Point is measurable, and current ROI frameworks understate them both, the implications for board-level investment decisions are specific.

What to ask before approving the next AI investment

Six questions, asked of the business owner presenting the case.

- What is the projected Inversion Point for this agent, and what assumptions drive it?
- What is the cost-per-successful-outcome at production scale, measured monthly, and what is the kill threshold?
- Is consumption tagged at the agent level, and can finance see the full operating tax separately from the build cost?
- What is the model selection and routing strategy? What percentage of requests route to the cheapest qualified model versus the frontier model?
- Is this a build, buy, or partner decision, and has the operating tax multiplier been adjusted accordingly?
- Has the case modelled a Klarna-style reversal scenario, including the cost of unwinding aggressive automation?

If the answer to any of these is unclear, the business case is incomplete. Approval is reasonable. Approval without these answers is approval of a project envelope, not approval of unit economics.

What to require for existing agents in production

Two requirements, applied over the next two quarters.

- Cost-per-outcome and projected Inversion Point computed monthly for every agent in production with material consumption.
- Kill criteria defined for each agent, tied to the projected Inversion Point, with a quarterly review forum to apply them.

What this means for the CFO and the CIO

This is a joint discipline. The CFO cannot build it alone because the agent-level consumption data sits in the cloud platform under the CIO. The CIO cannot build it alone because the value-per-outcome calculation and the Inversion Point projection require finance involvement. The discipline lives in the forum where both meet.

Enterprises that have built this forum, with a monthly cadence and predefined kill criteria, are the ones whose AI programs will reach the disciplined deployment cohort that has not inverted at twenty-four months. Enterprises that have not built it will follow the Klarna pattern. Public version or quiet version, the trajectory is the same.

12. The Discipline Is the Difference

Enterprise AI in 2026 is not failing because the technology is unreliable. The technology is mostly working. It is failing because the operating discipline required to run it at production scale, with honest unit

economics, is missing in most organisations. MIT's data confirms this. **Ninety-five percent of pilots fail to produce measurable financial returns**[1]. Only five and a half percent of enterprises attribute material EBIT impact to AI[2]. Two independent datasets. Same conclusion.

This is not a finance failure. It is not a technology failure. It is a structural gap between how investments have always been evaluated and how a new category of system actually behaves once deployed.

> *The enterprises that build the operating discipline now will have a defensible AI capability in eighteen months. The enterprises that do not will have an expensive lesson.*

The operating tax is real. It compounds. The Inversion Point exists, sits somewhere between month twelve and month eighteen for most enterprise AI agents, and arrives invisibly if it is not measured. The discipline that catches it is cost-per-outcome measurement, Inversion Point projection, monthly review, predefined kill criteria, joint governance between finance and technology, build-vs-buy honesty, and disciplined model selection. None of these are technical. All of them are decisions about how the operating model works.

That is the work. Whether your organisation does it now or discovers it later is the only variable.

References

All references accessed and verified May 2026.

[1] Challapally, A. et al. (July 2025). "The GenAI Divide: State of AI in Business 2025." MIT NANDA Initiative. Methodology: 300+ enterprise AI deployments reviewed, 52 executive interviews, 153 survey responses. Available at nandapapers GitHub repository.

[2] Singla, A., Sukharevsky, A., Yee, L. et al. (November 2025). "The state of AI in 2025: Agents, innovation, and transformation." McKinsey & Company QuantumBlack. Survey of 1,993 respondents across industries. 109 respondents (5.5%) reported >5% EBIT attribution to AI with significant value.

[3] Forrester Research (April 2026). 2026 Enterprise AI Agent Panel. Findings on automated evaluation coverage, rollback rates, and ROI trajectories across 2,000+ production agent deployments. Aggregated industry data on Inversion Point distribution.

[4] Klarna Bank AB (February 2024). Press release: "Klarna AI assistant handles two-thirds of customer service chats in its first month." 2.3 million conversations, 700 FTE equivalent, $40M projected profit improvement.

[5] Klarna Bank AB (2024 IPO filing) and Q1 2025 earnings report. Customer service cost per transaction declined from $0.32 (Q1 2023) to $0.19 (Q1 2025). Reported via CX Dive, May 2025.

[6] Leggett, K. (Forrester, May 2025). Commentary on Klarna's AI strategy reversal, quoted in CX Dive. Klarna announced rehiring of customer service representatives and reversal of full-AI strategy in May 2025, approximately 14 months after February 2024 launch.

[7] Klarna Bank AB Q3 2025 earnings call (November 2025). $60M cost savings from AI agent reported. Customer service and operations costs simultaneously increased from $42M (Q3 2024) to $50M (Q3 2025).

[8] Forrester Research 2026 Enterprise AI Agent Panel. 22% of agent deployments reported negative ROI at 12 months. Root causes: 41% unclear success criteria, 33% insufficient tool/data access, 26% drift in evaluation coverage.

[9] Brynjolfsson, E., Li, D., Raymond, L. (April 2023). "Generative AI at Work." NBER Working Paper No. 31161, National Bureau of Economic Research. 5,179 customer support agents at a Fortune 500 software firm. 14% average productivity uplift; 34% for novice/low-skill workers; minimal impact on experienced workers.

[10] MIT NANDA "The GenAI Divide" (July 2025), p. 16. Vendor partnerships succeed at approximately 67% rate; internal builds succeed at one-third that rate. Quote: "Almost everywhere we went, enterprises were trying to build their own tool, but the data showed purchased solutions delivered more reliable results" (Aditya Challapally, lead author).

[11] Forrester 2026 panel. Production agents with full evaluation coverage: 9% rollback rate. Production agents without evaluation coverage: 47% rollback rate.

[12] Public pricing data from OpenAI, Anthropic, Google DeepMind (March 2023 to April 2026). GPT-4 launch pricing March 2023: $30/$60 per million input/output tokens. Gemini Flash family April 2026: $0.10/$0.40 per million input/output tokens. 99.7% reduction.

[13] Chen, L., Zaharia, M., Zou, J. (2023). "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance." Stanford University. Demonstrated 50-98% cost reduction with intelligent routing while matching or exceeding GPT-4 accuracy.

[14] AI.cc (May 2026). 2026 AI API Infrastructure Report. Analysis of 2.4 billion API calls across 8,000+ enterprises and developers between January and April 2026. Median cost reduction from multi-model routing: 71% versus single-provider deployments; top quartile: >80%.

[15] Industry analysis aggregated from multiple sources (2025-2026): MindStudio, OpenRouter, AI Pricing Master. Convergent finding: organisations using single LLM for all tasks overpay by 40-85% versus organisations implementing intelligent routing.

[16] Axis Intelligence (December 2025). EU AI Act compliance cost analysis based on official EU Commission documentation and early implementations. Large enterprises (>€1B revenue): $8-15M initial investment for high-risk systems. GPAI providers: $12-25M first year. Mid-size companies: $2-5M initial, $500K-2M annually.

[17] Raconteur (April 2026). "EU AI Act Compliance: a technical audit guide for the 2026 deadline." Reports vendor pricing premiums of 20-30% reflecting certification costs and engineering overhead for high-risk AI systems.

[18] Moffatt v. Air Canada, 2024 BCCRT 149 (British Columbia Civil Resolution Tribunal, February 2024). Tribunal Member Christopher C. Rivers ruling on Air Canada chatbot misrepresentation. Damages: CAD $812.02. Established precedent that corporations are legally responsible for outputs from their AI systems.

[19] SQ Magazine (April 2026). AI Compliance Cost Statistics 2026. Compliance requirements add an estimated 10-25% extra cost per AI model. Internal conformity assessments can reduce costs by 15-25%. Aggregated from multiple regulatory implementation studies.

[20] Forrester Research 2026 panel data on enterprise AI infrastructure forecasting accuracy. 80-85% of enterprises miss AI infrastructure forecasts by more than 25% in the absence of per-agent cost attribution.

About this briefing

This briefing is part of an ongoing body of work developing the discipline of Enterprise Agentic AI Systems Delivery. It complements the AI Role Operating Framework (AI ROF™), the Intelligence-Centred Enterprise (ICE) framework, and the Five Laws of Enterprise AI Delivery published in The AI Delivery Discipline newsletter.

The author is Vijayan Seenisamy, a senior practitioner with more than twenty years of enterprise systems delivery experience, currently leading group transformation at a top-30 ASX-listed company. He is the creator of the AI Role Operating Framework, the author of The AI Delivery Manager Blueprint (January 2026) and The Pilot Trap (April 2026), and the publisher of The AI Delivery Discipline newsletter.

The argument in one sentence

> *The operating tax compounds faster than the productivity claim, the Inversion Point sits between month twelve and month eighteen, and the discipline that catches both is cost-per-outcome measurement tied to predefined kill criteria.*

Related resources

For finance and technology leaders seeking to operationalise the disciplines in this briefing, the following resources extend the analysis with practical instruments.

- The AI Delivery Manager Blueprint (book). Defines the operating role and accountability structures required for production AI delivery.
- The Pilot Trap (book). Examines why ninety-five percent of enterprise AI pilots fail to reach reliable production and what the twelve percent who succeed do differently.
- The AI Delivery Discipline (newsletter). Biweekly analysis of structural patterns in enterprise AI delivery, including the Five Laws and the diagnostic frameworks they produce.
- Enterprise AI Delivery Assessment. A fifty-six-criterion diagnostic that produces a per-program operating tax measurement, an Inversion Point projection, and a ninety-day improvement roadmap. Engaged on a retainer or project basis.

Contact

For senior leadership conversations about applying these disciplines inside your organisation, contact Vijayan Seenisamy at vijayan@aideliverydiscipline.com.
