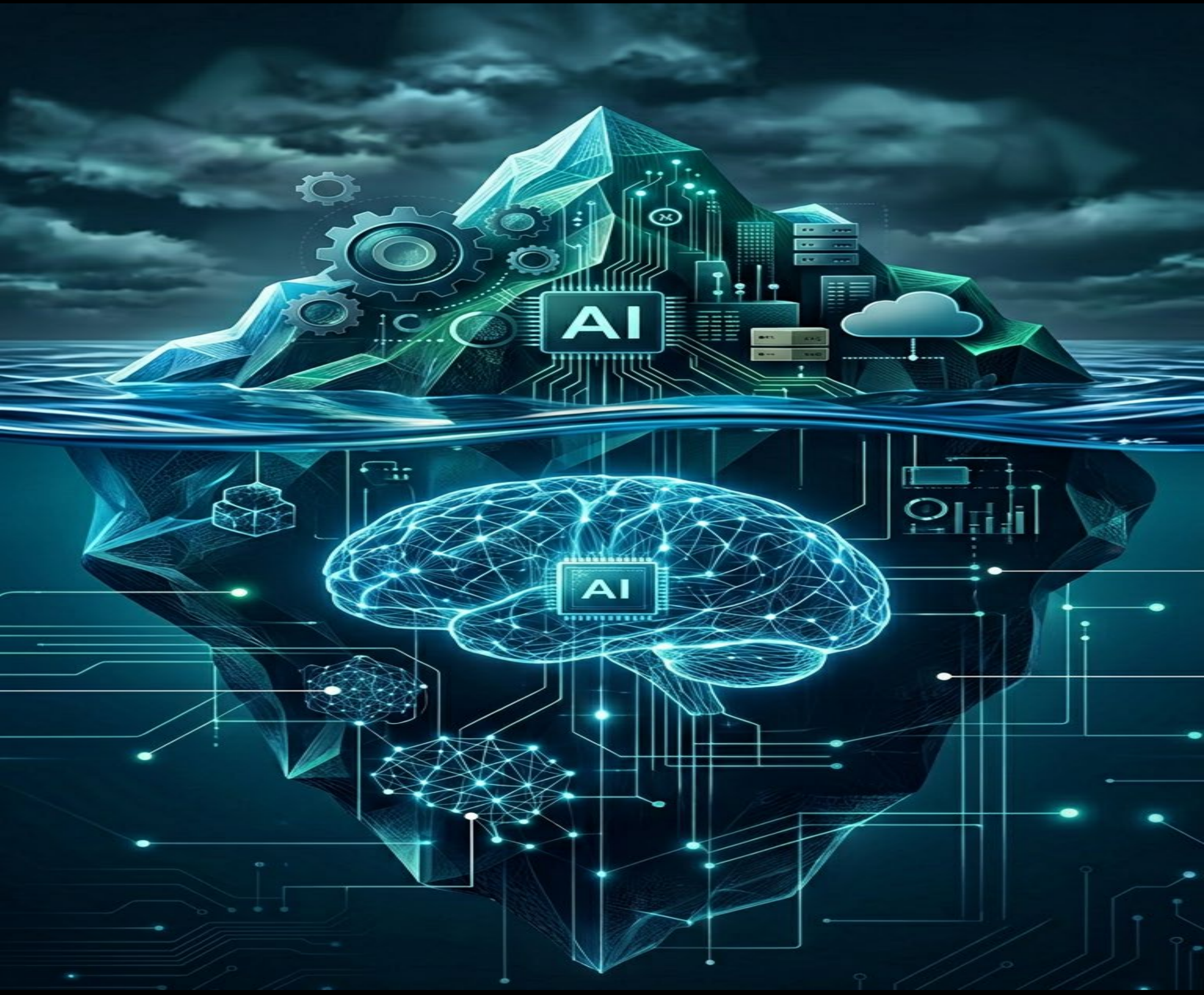


What Agentic AI Really Costs and What It Actually Returns

Deloitte Regulatory, Risk & Forensic

July 2026

Contents

This document presents a Deloitte point of view on the economics of agentic AI. Figures in the legal contract-review example are for illustrative purposes only and demonstrate the measurement method, not a specific client engagement.

Introduction

What agentic AI costs, and why the question matters

Introduction

Many artificial intelligence (AI) programs are failing to meet return on investment (ROI) expectations. Oftentimes, this is because organizations are asking the wrong question about agentic AI. The focus tends to be model pricing, but a more appropriate question is what the full workflow costs and what value it creates over time once agents run in production at scale.

What "agentic AI" means in this point of view

A prompt answers a question. Agentic workflows go beyond prompts and traditional automation: they interpret a goal, execute multistep work using tools and data, and trigger human review when judgment or control is required. In this point of view (POV), agentic AI refers to systems that plan and execute multistep work using tools, data sources, and workflows, with defined points of human oversight.

Why the timing matters

Many organizations are moving from experiments into live workflows. As agents move beyond prompt assistance and begin using tools, calling systems, and handling exceptions, the cost base shifts. Spend migrates from tokens to system interactions, integration work, governance, exception handling, and operational support. Meanwhile, value is often tracked as activity or "productivity," not outcomes.

The leadership question has changed

The issue is not adoption. It is whether the work has been designed to scale economically. Productivity is not ROI unless it converts to one of five value types with evidence: hard savings, captured capacity release, cost avoidance, business performance, or quality and risk outcomes.

Hard savings	Capacity release	Cost avoidance	Business performance	Quality and risk
Reduced external spend or budgeted cost reductions, validated by Finance	Time freed and redeployed to higher-value work	Prevented losses or avoided future spend	Revenue uplift, conversion, service expansion, output gains	Lower error rates, reduced rework, stronger controls

This POV explores what agentic AI really costs and what it actually returns over time. The real issue is not model pricing in isolation. It is whether the enterprise can deploy, operate, and govern agentic AI in a way that scales economically.

The full cost stack of agentic AI

Where spend really accumulates once agents run in production

The full cost stack of agentic AI

The model bill is only one part of the real cost and often not the largest part. Based on Deloitte's experience across enterprise deployments, when organizations first evaluate agentic AI, the conversation almost always gravitates toward model pricing: cost per token, per application programming interface (API) call, per seat. That framing is understandable, but it is incomplete. As agents become more capable, execute multi-step workflows, call on external tools, retrieve data dynamically, and escalate to humans when uncertain, the cost center shifts. Tokens become a smaller share of the total picture, while system interactions, infrastructure, oversight, and operational support grow to become dominant. Without a grip on the full economics, leaders risk making deployment decisions with only half the picture.

To close that gap, organizations need a consistent definition of scope, specifically what is included when calculating the total cost. The full cost stack spans four layers, all of which should be counted to build a defensible ROI case:

Build	Design, development, integration, and testing
Run	Inference, orchestration, tooling, retrieval, and licensing
Change	Training, process redesign, and adoption
Govern	Compliance, human oversight, audit, and ongoing support

Together, these layers represent true total cost of ownership. Model inference fees, which is the number most vendors advertise, sit in just one line of one layer. Everything else helps determine whether agentic AI can scale economically or creates costs that outweigh realized value.

Without this scope, organizations may struggle to make apples-to-apples comparisons across vendors, build vs. buy decisions, or deployment scenarios. According to the International Data Corporation (IDC), survey data indicated that 96% of organizations deploying generative AI and 92% implementing agentic AI reported actual costs were higher or much higher than initially expected. [1]

96%
of GenAI deployers reported costs higher than expected

92%
of agentic AI adopters reported the same

[1] IDC, Insights from Early Enterprise Deployments: A Strategic Approach to Scaling GenAI and Agentic AI, November 2025.

The full cost stack of agentic AI (cont.)

Within those four layers, costs interact and compound in ways that make agentic AI fundamentally different from traditional software to budget. Model inference, which is the most visible cost, grows quickly because agents do not respond to a single prompt. These agentic agents reason through multi-step problems, call tools, retrieve documents, evaluate their own outputs, and loop back when results fall short, with each step generating additional model calls. As a multi-turn research agent works through a task, token consumption can grow substantially with each step and, at enterprise scale, this compounds quickly across users and workflows.

Data and retrieval costs scale alongside this. Because most enterprise agents rely on retrieval-augmented generation (RAG), the embeddings, vector databases, storage, and search operations grow quickly with data volume. Poor data quality compounds the problem, with bad embeddings driving retrieval retries and higher token spend.

Integration costs are similarly underestimated, as traditional enterprise systems are rarely designed for real-time autonomous agent interactions. Finally, risk, compliance, and oversight costs are recurring as agents become more autonomous, governance requirements intensify. In regulated industries such as financial services, health care, and legal, compliance alone requires audit trails, guardrails, output validation, drift monitoring, and human-in-the-loop review infrastructure.

Five multipliers that inflate cost at scale

What makes agentic systems uniquely difficult to budget, however, is that these costs scale non-linearly once agents move from pilot to production. Five multipliers often drive inflation beyond early projections:

Agentic systems can drive several-fold higher token consumption per task than a standard chat interaction

1. **Retry loops**
   Agents are designed to persist, but when a tool call fails or an output is uncertain, they retry, often with additional context. In some implementations, retry loops can consume many multiples of the tokens required for a single linear pass. At scale, retry behavior can increase cost without a proportionate increase in business output.

2. **Tool-call volume**
   Compared with standard chat, agentic systems usually incur a significantly higher token footprint per task. In multi-agent architectures, a single agent request can chain multiple sequential API calls to complete one task, compounding overhead further.

3. **Exception rates**
   When agents escalate, humans should review, decide, and return context. Exception rates are not static, as they tend to rise as agents are deployed against more complex workflows and as quality thresholds increase.

4. **Context length**
   Each step in an agent loop sends the entire accumulated context to the model. Teams are reluctant to prune context because doing so feels risky, so every request gets heavier and more expensive over time.

5. **Retrieval depth**
   As agents are applied to broader knowledge bases, retrieval operations deepen. Tool metadata, document chunks, and re-ranking operations all consume context capacity, and retrieval costs scale non-linearly as the scope and ambition of the use case expands.

Five multipliers that inflate cost at scale (cont.)

The cost stack of agentic AI is therefore not a fixed number to be quoted at procurement; it is a dynamic system shaped by agent design choices, workflow complexity, exception rates, and governance requirements. The practical implication is straightforward: organizations should measure and manage cost at the workflow level, not the model level. What does it cost to complete a process end-to-end, across all four layers and accounting for the multipliers that inflate them at scale? That question leads directly to the other side of the equation: what value does the workflow actually create?

Why firms undercount cost and overclaim value

The gap between pilot economics and production economics

Why firms undercount cost and overclaim value

Many AI business cases are directionally aligned but financially incomplete. Firms often evaluate agentic AI through the lens of visible technology spend, including model usage, licenses, and pilot development effort, and then project value from speed, task completion, or early user activity. That creates a distorted picture: costs look lighter than they will be in production, while benefits look larger than they are likely to be once workflow complexity, enterprise controls, adoption realities, and actual value capture come into view. More critically, the gap between pilot economics and production economics often widens when adoption succeeds. This is due to scaled usage exposing exception rates, control requirements, and support demands that do not surface in controlled environments with small, expert user groups.

On the cost side

Organizations systematically undercount what it takes to run AI-enabled workflows reliably at scale because they focus on visible spend and miss the embedded enterprise cost stack around it. The build, governance, and licensing layers all compound as usage grows, and none of them appear on a vendor's pricing sheet.

On the value side

Firms often overclaim returns by labeling them as productivity without proving whether that productivity became hard savings, real capacity release, lower external spend, improved business performance, or measurable quality and risk improvement. That is where many business cases break down.

Why firms undercount cost and overclaim value

Enterprise AI Bill of Materials: The Cost-Value Gap

Enterprise AI Bill of Materials: The *Cost-Value* Gap

Why AI economics are harder to close than they appear

THE ACCUMULATED COST
Systematically undercounted

1 Embedded costs

The model bill is one component. Integration, orchestration, workflow redesign, testing, and the operating environment often represent a larger share of true production cost.

2 Understated post-pilot costs

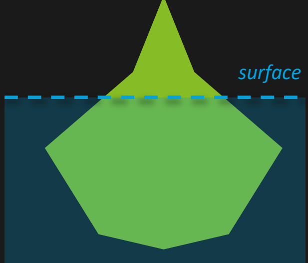
Business cases understate post-pilot reality: governance, security, compliance, control requirements, and ongoing maintenance obligations all add to the run rate.

3 Pilot assumptions don't scale

Retries rise, exception handling expands, tool calls multiply, and usage grows across teams. Scaled economics rarely match the original pilot assumptions.

4 Hidden overheads

Shared platform and licensing costs absorbed elsewhere leave workflow economics looking cleaner than they really are.



VISIBLE COST

Model bill

HIDDEN COST STACK

Integration · orchestration

Governance · compliance

Licensing · maintenance

TOTAL UNDERCOUNTED

Hidden cost stack

THE RECONCILED VALUE
Frequently overclaimed

A Productivity gains without proof

Time saved is counted without showing whether labor demand changed, throughput increased or spend declined, projecting efficiency rather than demonstrating it.

B Activity metrics ≠ value

Usage, tasks completed, and engagement are treated as evidence of value even when they reflect only activity, not impact on labor, spend, or business performance.

C Unfit process design

The workflow hasn't been redesigned to convert AI output into end-to-end improvement. Technology speeds up part of the process without changing its overall economics.

D Siloed use case building

Teams build redundant agents and workflows in silos, duplicating investment, multiplying governance overhead, and inflating licensing spend across redundant vendor relationships.

E Premature benefit claims

Benefits are claimed before accuracy, exception rates, and control performance have stabilized for scaled deployment.

F Review burden underestimated

AI outputs still require substantial human checking. Quickly assembled solutions can create rework, debugging, and control risk rather than reduce effort.

VALUE ADJUSTMENT

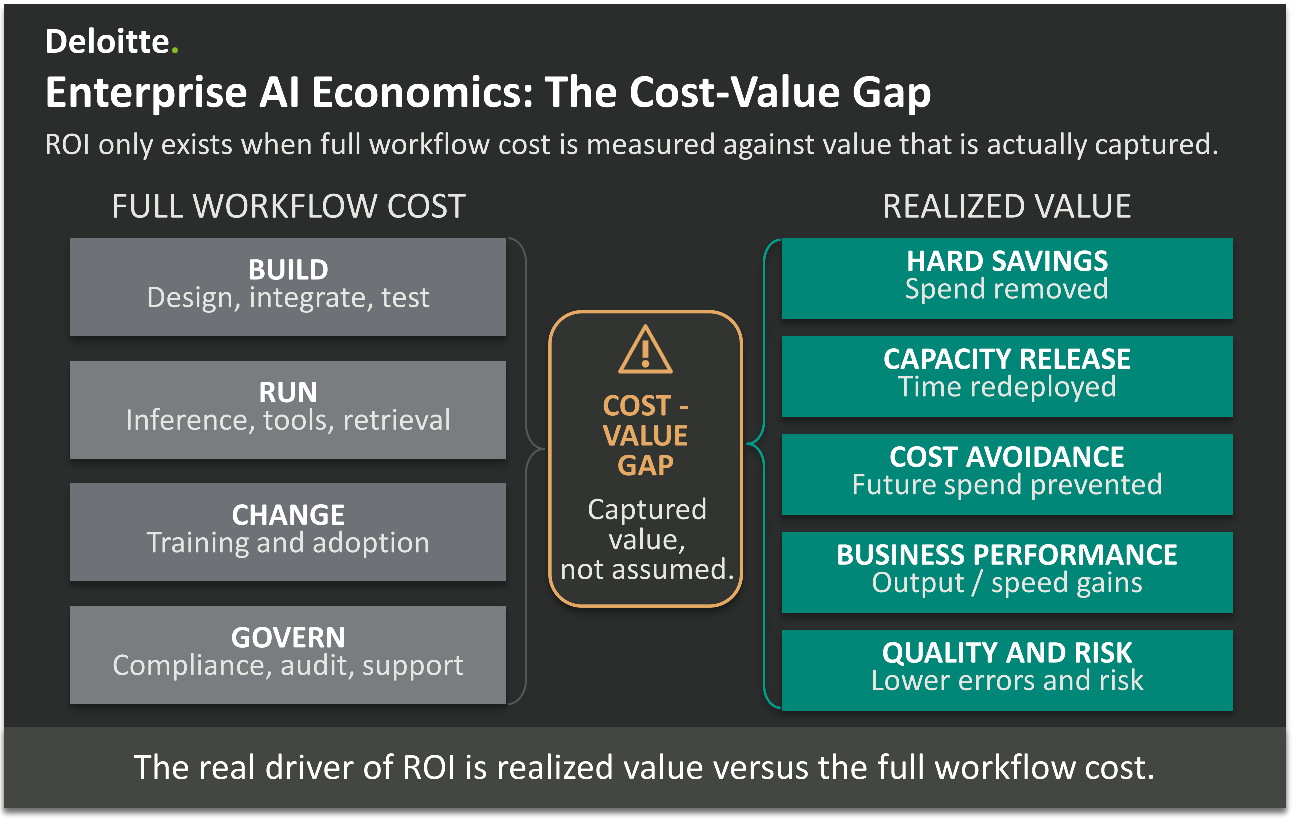
Overclaimed returns

The central challenge: *not whether AI can generate productivity, but whether cost and value can be measured, validated, and governed to reflect full workflow economics from pilot through scaled deployment.*

A disciplined ROI approach means accounting for hidden costs outside the model bill, testing how economics change as usage scales, and measuring realized value rather than projected efficiency alone.

Why firms undercount cost and overclaim value (cont.)

Put simply, the entire challenge reduces to one picture: the full cost of running a workflow on one side and the value it captures on the other and the gap between them.



A practical framework for managing return

The workflow ROI discipline, the equation, and a repeatable method

A practical framework for managing return

Return on AI should be managed at the workflow level, not the model level, which is what we call the Workflow ROI Discipline. Managing return requires more than building an upfront business case. Organizations need a repeatable mechanism to define value, measure realized outcomes, and continuously optimize investments over time. The Workflow ROI Discipline provides that structure: it anchors value in end-to-end workflows, not isolated AI capabilities, and establishes a closed-loop system for tracking what delivers returns once the technology meets operational reality.

The ROI equation and what to track

Many organizations do not struggle to identify AI use cases; they struggle to prove return. The reason is straightforward: ROI is too often estimated at the investment level, while value is created, captured, or lost at the level of the business workflow. That is where the ROI equation should start.

For AI-enabled transformation, the relevant unit of analysis is the workflow itself: its start and end points, transaction volumes, service levels, roles, handoffs, and systems touched. This is what allows leaders to move beyond abstract productivity claims and assess whether AI is materially improving business performance.

At its core, the equation remains simple:

ROI = (Value realized – Total cost to achieve) / Total cost to achieve

What changes in an AI context is the discipline required to measure both sides of that equation properly. Organizations that do this well start with a consistent measurement rubric for every workflow. On the cost side, that means tracking the full operating cost required to run the workflow reliably and at scale. On the value side, it means defining in advance how return will be evidenced, validated, and governed over time.

What to track: cost and value

On the cost side, organizations should track the full operating cost of the workflow across the categories that determine whether it can scale economically. On the value side, return should be tracked across five categories that matter to the business.

Cost to track (nine categories)

Cost category	What to measure
Model and inference	Token, compute, and routing costs by model, task, and volume
Tool and API usage	Charges from search, enterprise systems, third-party tools, and downstream actions the agent triggers
Retrieval and data infrastructure	Data preparation, storage, indexing, retrieval, and access management needed to make outputs usable
Orchestration and platform runtime	Workflow engines, agent frameworks, observability, logging, and runtime services
Licenses and vendor spend	Platform fees, seat licenses, and shared services attributed to the workflow
Integration and workflow redesign	System integration, testing, process changes, and the work to embed AI into production operations
Security, risk, and compliance controls	Identity, access, monitoring, governance, validation, and control requirements
Human oversight and exception handling	Review time, escalation paths, rework, and fallback handling when the workflow does not complete cleanly
Support and maintenance	Model updates, prompt and rule tuning, issue resolution, and ongoing operational support

Value to track (five categories)

Value category	What counts as evidence
Hard savings	Reduced external spend or budgeted cost reductions validated by Finance
Capacity release	Time freed and actually redeployed to higher-value work, evidenced by throughput, backlog reduction, or cycle-time improvement
Cost avoidance	Prevented losses or avoided future spend, tracked through agreed assumptions and periodic validation
Business performance	Revenue uplift, conversion improvement, service expansion, or measurable output gains
Quality and risk outcomes	Lower error rates, reduced rework, stronger SLA performance, better control execution, and lower risk exposure

What to track: cost and value (cont.)

The practical point is this: these categories are not just for reporting. They are how leaders make better deployment decisions. If a workflow cannot show its full operating cost, it is likely not ready to scale. If value cannot be tied to one or more of these outcome categories with evidence, it generally should not be counted as ROI.

That is why productivity alone is typically not enough. Time saved creates return only when it is converted into lower cost, higher throughput, better service, stronger revenue performance, or reduced risk. If that conversion cannot be shown, the value remains potential rather than realized.

A practical method helps make that conversion real. Start by baselining the current workflow: cost, cycle time, quality, volume, and exception patterns. Then design the AI-enabled workflow clearly: where the agent acts, where tools are called, and where human review or approval remains necessary. Instrument it so teams can see the real cost drivers and operating behavior, including tool calls, exception rates, human touches, latency, and rework loops. Then convert those measured changes into business outcomes using the value categories above, and use those results to govern where to scale, where to refine, and where to stop.

The tracking model should be equally practical. In pilot, teams should review weekly. During scale, monthly is usually sufficient. In steady state, quarterly reviews help confirm that value is holding and controls remain effective. Ownership should also be explicit, and the good news is that these roles already exist in most organizations. The product owner should own workflow performance, the finance partner should validate value conversion, and the risk owner should ensure controls remain effective and residual risk stays within tolerance. No new organizational structure is typically required; what is required is clarity about who is accountable for each dimension and a cadence that keeps them engaged.

The broader point is this: ROI is better managed as an operating discipline for improving workflow economics over time. Organizations that do this well will not just deploy more AI. They will know where it creates measurable value, where the economics do not hold, and what it takes to improve performance before scaling further.

A repeatable five-step method for managing return

Measuring the return on agentic AI is not a reporting exercise; rather, it is an operational discipline. For agentic AI specifically, measurement is likely to focus on cost savings, process redesign, risk management, and longer-term transformation, not the simple productivity proxies that worked for earlier generations of AI tooling. Many AI projects fail to demonstrate ROI not because the technology fails, but because of measurement failures: no baseline before deployment, the wrong KPIs, one-time measurement instead of continued tracking, and isolated measurement that misses workflow redesign benefits. The method below is structured around five steps that any team can apply consistently across workflows, and that any leader can use to make scaling decisions on defensible grounds.[2]

1

Baseline the workflow

Without a documented baseline, any improvement claim is an assertion, not evidence. Before deploying an agent, measure what the current process actually costs and how it performs across: cost per workflow; cycle time and throughput; quality levels and error rates; volume and variability; and exception patterns and escalation rates.[3]

2

Design the AI-enabled workflow

Specify upfront where agents act, which tools they use, where human review is required, and what triggers escalation.

Workflows where agents handle ~80% of cases with exception-based human oversight have fundamentally different cost and value profiles than those requiring full sign-off. Organizations that intentionally design human–machine integration are more likely to exceed ROI expectations than those taking a technology-first approach.

The design stage defines how agents, systems, and humans interact, directly shaping cost and value. It should clearly specify:

- Agent roles and tasks within the workflow
- Tools, systems, and data sources invoked
- Points of human review, approval, and escalation
- Control points for quality, compliance, and risk

Getting this right is critical—poorly defined roles and controls increase cost and limit ROI.

[2] Deloitte, AI ROI: The Paradox of Rising Investment and Elusive Returns, October 2025.
[3] Deloitte Canada, AI Awareness and Access Have Skyrocketed, Yet Real Enterprise Value and ROI Are Rare, February 2026.

A repeatable five-step method for managing return (cont.)

3

Instrument cost and performance

Once deployed, the workflow should be instrumented to capture the metrics that drive both cost and value. Without comprehensive observability, teams struggle to understand how agents make decisions, diagnose failures, and maintain quality at scale. Key signals: model usage, tool calls, and infrastructure; number of tool interactions and retries; exception and escalation rates; human intervention frequency; and latency, rework, and workflow completion time.

4

Convert to dollars and outcomes

Operational improvements must be mapped to business value:

- **Hard savings**: Direct cost
- **Capacity release**: Time redeployed to higher-value work
- **Cost avoidance**: Future spend prevented
- **Business performance**: Gains in revenue, speed, or quality
- **Quality and risk**: Reduced errors and stronger compliance

Without this conversion, improvements should not be counted as ROI.

5

Govern scaling decisions

Return is not realized at deployment, but at scale that is governed. Leading organizations require a material ROI to be signed off by the finance partner and the business unit head before moving from experiment to production. Each expansion should be treated as a new ROI test, not an assumption of proportional return. This includes rationalizing licenses and vendors as the portfolio grows, consolidating where possible to reduce pricing complexity and remove redundant tooling.

Governance does not end at the scaling decision. ROI should be tracked over time with a defined cadence and accountable owners. In practice, this means weekly tracking during pilots, monthly reviews at scale, and quarterly assessment in steady state. Accountability should span product (workflow performance), finance (value realization against assumptions), and risk (exception rates, compliance, and oversight levels). Without named owners and a review cadence, even well-instrumented workflows drift: costs creep, exception rates rise, and the ROI case that justified deployment quietly erodes. The goal is not the cheapest AI stack, but the best business outcome per dollar invested, which can only be managed if it is consistently measured.

Putting the method to work

An illustrative legal contract review example

Worked example: legal contract review

Legal makes agentic AI economics easier to see. The work is high-value, structured enough to break down, and review-intensive enough that both the savings and the hidden costs show up quickly. That is why legal is a useful illustrative test case for this POV. It shows that the question is not whether AI can do part of the work. It can. The real question is whether the workflow is redesigned so the work scales economically, with clear controls and measurable value.

Consider a common contract review process for commercial agreements. A typical enterprise legal team may handle 8,000 to 12,000 contracts each year. Cycle time often runs from three to seven days per contract, depending on contract type, queue length, and how often non-standard terms trigger escalation. The work usually starts with junior legal staff or outside counsel conducting an initial review, comparing terms against playbooks, and marking deviations for follow-up. Roughly 25 to 40 percent of contracts may require senior lawyer intervention, particularly when indemnities, liability caps, data rights, or regulatory terms fall outside approved language. The cost base is spread across internal legal labor, outside counsel used for overflow or specialized review, and a set of legal technology tools such as contract lifecycle management platforms, document repositories, and clause libraries.

Illustrative key figures of a legal workflow

8,000–12,000	3–7 days	25–40%
contracts reviewed per year	typical cycle time per contract	require senior lawyer intervention

The tables in the following pages use one illustrative workflow so the method can be read in numbers, not just described in prose.

Step 1 – Baseline the workflow

Without a documented baseline, any improvement claim is an assertion, not evidence.

Before deploying any agent, the current workflow should be measured in operational terms, not estimated from sentiment or self-reported time savings. The table pairs the current workflow with a target-state reference point so the reader can see, in one view, what is expected to change and by how much.

Workflow baseline and target-state reference

Metric	Current workflow	AI-enabled workflow	Impact
Contracts reviewed per year	10,000	10,000	Volume held constant
Human review time per contract	120 min	65 min	-55 min (-46%)
Annual human review hours	20,000 hrs	10,833 hrs	9,167 hours released
Senior escalation rate	32%	18%	-14 pts (-44%)
Senior escalations per year	3,200	1,800	1,400 fewer escalations
Average cycle time	5.0 days	2.5 days	-2.5 days (-50%)
Economic cost per contract	$450	$320	-$130 (-29%)
Annual workflow cost	**$4.50M**	**$3.20M**	**$1.30M lower run-rate cost**

The baseline clarifies where the economics are breaking today. Most legal teams track total legal spend, but few track cost per contract, cost by contract type, or the true cost of escalation and rework. That makes it easy to overstate AI's upside and hard to prove improvement later. In this example, the biggest issues are not just labor hours. They are queue delay, broad escalation, and routine work priced too close to complex work. While agentic AI capabilities would help accelerate this process and reduce inefficiencies that come from manual baseline calculations, it is not a replacement for humans.

Step 2 – Design the AI-enabled workflow

A workflow where the agent handles 80% of cases autonomously has a fundamentally different cost and value profile than one where every output requires sign-off.

An agentic contract workflow is not AI as a drafting assistant at the margin. It is a redesign of the operating model. Agents perform the repeatable first-pass work within defined controls, while lawyers stay focused on exceptions, judgment, and negotiation. Agents act on clause extraction and classification, comparison against approved playbooks, detection of deviations from standard language, suggested redlines, and intake triage by contract type, counterparty, and risk level. Humans stay in the loop for exception handling and escalation, final approval on material risk, and negotiation of non-standard issues.

Workflow redesign and control points

Workflow stage	Current	AI-enabled	Effect	Control point
Intake and triage	Manual queueing and prioritization	Agent classifies by contract type, risk, and urgency	5-10 min saved per contract	Human reviews flagged exceptions
Clause extraction and classification	Junior staff or outside counsel reads and tags manually	Agent extracts clauses and classifies terms	10-15 min saved per contract	Flagged clauses checked by reviewer
Playbook comparison and deviation detection	Manual comparison against standard terms	Agent compares against playbook and identifies deviations	15-20 min saved per contract	Non-standard terms escalate automatically
First-pass redlining	Reviewer drafts fallback language manually	Agent proposes first-pass redlines	10-15 min saved per contract	Human approves material edits
Escalation and approval routing	Broad escalation driven by caution and queue delay	Rule-based escalation on defined triggers	Escalation rate falls from 32% to 18%	Senior lawyer reviews material risk only
Complex negotiation	Lawyer-led	Lawyer-led	No direct time saving	Human judgment remains required

This is the operating model change that drives the deltas shown in the baseline table. The time savings roll up into the lower review minutes, lower escalation rate, and shorter cycle time. If every output still receives full manual review, the workflow may move faster, but cost per contract will not fall enough to sustain ROI. A single contract in this model may require 20 to 50 tool calls across retrieval, reasoning, redlining, validation, and policy checks. Those interactions create new cost drivers that need to be measured explicitly.

Step 3 – Instrument cost and performance

Without comprehensive observability, teams struggle to understand how agents make decisions, diagnose failures, and maintain quality at scale.

Once the workflow is redesigned, the next job is to make the cost stack visible. Manual review time may decline, but cost does not vanish. It moves. In the AI-enabled model, cost shifts across a broader system stack that includes models, tools, retrieval, orchestration, controls, and support.

AI-enabled cost stack to instrument

Cost category	Current $/contract	AI run-rate $/contract	Change	Year-1 note
Model and inference	$0	$12	$12	New run-rate cost
Tool and API usage	$0	$10	$10	New run-rate cost
Retrieval and data infrastructure	$0	$8	$8	New run-rate cost
Orchestration and platform runtime	$0	$7	$7	New run-rate cost
Licenses and vendor spend	$50	$55	$5	Slight increase in shared tooling cost
Integration and workflow redesign	$0	$0	n/a	$300k one-time in Year 1
Security, risk, and compliance controls	$0	$8	$8	New run-rate cost
Human oversight and exception handling	$400	$210	($190)	Largest cost reduction
Support and maintenance	$0	$10	$10	New run-rate cost
Total run-rate cost per contract	$450	$320	($130)	Excludes one-time build cost
Annual run-rate workflow cost	**$4.50M**	**$3.20M**	**$1.30M lower**	**Excludes one-time build cost**

If this shift in cost composition is not measured explicitly, the business case will look better on paper than it does in production. That is why model price alone is the wrong metric to focus on. The real question is whether the full run-rate cost per contract stays below the cost it replaces, while control and review requirements remain intact.

Step 4 – Convert to dollars and outcomes

Organizations that track only direct cost reduction may systematically miss return, particularly from capacity release and risk reduction.

In a well-instrumented contract workflow, raw operational metrics should be translated into business value across five categories.

Value conversion and realization treatment

Value category	What changed	Illustrative impact	Annual value	Realization treatment
Hard savings	Outside counsel on routine contracts reduced	$1.8M to $0.8M	$1.0M	Count when spend is removed
New AI operating cost	New run-rate cost for models, tools, data, runtime, controls, support, and license uplift	New cost to run the workflow	($0.6M)	Count immediately
Net hard savings, steady state	External savings less new AI run-rate cost	Steady-state cash impact	$0.4M	Base realized ROI
Less one-time integration and redesign	Build and embed the workflow into production	Year 1 implementation cost	($0.3M)	Include in Year 1 ROI
Net Year 1 hard savings	Steady-state hard savings less one-time implementation cost	First-year cash impact	$0.1M	Use for Year 1 ROI
Capacity release	Internal review time drops from 20,000 to 10,833 hours	9,167 hours released	~$0.9M eq.	Count only if redeployed and evidenced
Cost avoidance	20% volume growth absorbed without new headcount or outside counsel	2,000 more contracts handled	$0.4-0.7M	Conditional. Validate with Finance
Business performance	Faster contract turnaround	2.5 days faster	–	Case-specific KPI unless tied to revenue timing
Quality and risk outcomes	More consistent playbook application, fewer missed deviations	Lower rework, stronger adherence	–	Case-specific KPI unless loss history supports it

Step 5 – Govern scaling decisions

Each expansion should be treated as a new ROI test, not an assumption of proportional return.

Exception rates in a mature contract workflow may fall to 10 to 25 percent, but they do not disappear. Human review levels, exception thresholds, and cost per contract should be governed as the portfolio grows and should not be assumed to hold.

Scale governance thresholds

Governance area	Target threshold	Current state	Decision rule	Owner and cadence
Cost per contract	<= $320	$320	Scale only if cost holds or improves	Product + Finance, monthly
Senior escalation rate	<= 20%	18%	Pause expansion if rate rises above threshold	Product + Risk, monthly
Human review model	Exception-led, not universal sign-off	Exception-led	Do not scale if full-review behavior returns	Legal Ops, monthly
Cycle time	<= 3.0 days	2.5 days	Expand only if SLA holds at higher volume	Product, weekly then monthly
Tool-call efficiency	Within expected range by contract type	20-50 calls per contract	Tune before scale if retries or latency spike	Product + Engineering, monthly
Quality and policy adherence	No material control degradation	Stable	Stop or redesign if quality drops	Risk + Legal, monthly
Vendor footprint	Rationalized and non-duplicative	Controlled	Consolidate before broader rollout	Product + Procurement, quarterly
ROI review cadence	Weekly pilot, monthly scale, quarterly steady	Defined	No further scaling without named owners and rhythm	Product + Finance + Risk

Scaling should follow evidence, not enthusiasm. If cost per contract drifts up, exception rates rise, or human review creeps back toward the baseline, the workflow should be refined before scope expands. That is how leaders protect both the economics and the controls as adoption grows.

Bottom line: the organizations that win will not be the ones with the cheapest model bill. They will be the ones that deliver the best contract outcome per dollar invested, with controls strong enough to scale. That is what this method is designed to produce. The legal example makes the larger argument concrete: agentic AI does not create value just because human effort goes down. It creates value when work is redesigned so that lower-touch processing, tighter escalation rules, and faster first-pass review translate into lower external spend, higher throughput, better turnaround times, or stronger control outcomes.

Closing

Closing

The mistake is not underestimating the model bill. It is mistaking activity for return.

In practice, the economics of agentic AI are shaped by workflow design, tool use, data readiness, exception rates, and the level of human oversight required to produce a reliable result. That is why the right question is not how fast to deploy, but which workflows can deliver better outcomes at an acceptable cost, under real operating conditions, and with controls that hold at scale.

Agentic AI is not a technology cost problem. It is an enterprise system-economics and value-realization challenge.

Leaders should start by assessing where AI is already deployed, mapping its position within full workflows; measuring actual productivity impact (not just usage); and building feedback loops that connect investment decisions to realized outcomes. The work is not building more pilots. It is building the discipline to know what is working, why, and at what cost.

Contacts

For more information, contact:

Donald Williams

Managing Director
Deloitte Transactions and Business Analytics LLP
dowilliams@deloitte.com

Jennifer MacMillan

Senior Manager
Deloitte & Touche LLP
jemcmillan@deloitte.com

Manmohan Sharma

Manager
Deloitte & Touche Assurance and Enterprise Risk Services India Private Limited
manmsharma@deloitte.com

Minh Tra

Manager
Deloitte Transactions and Business Analytics LLP
minhtra7@deloitte.com

Deloitte.