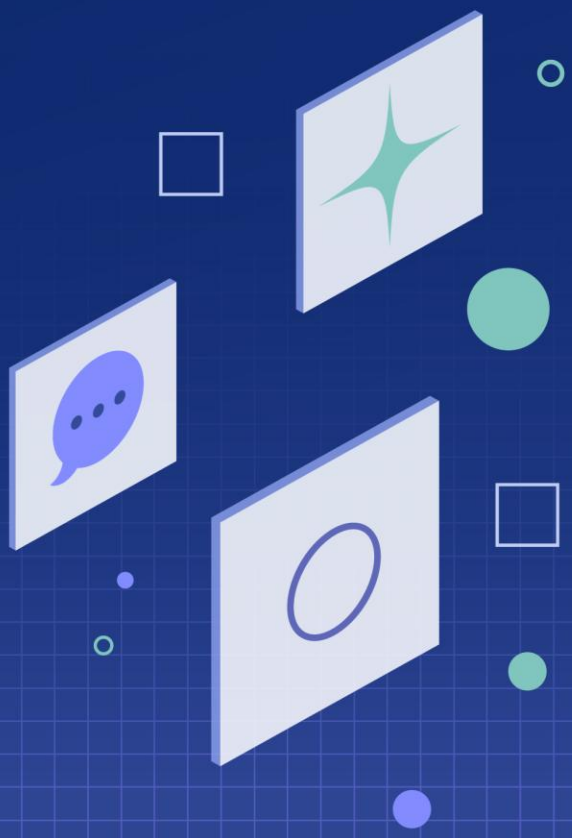
The hidden cost of AI isn't the model — it's how you run it

The case for predictable AI

June 2026 | Don Schuerman, CTO & Head of Marketing



The next phase of AI isn't about possibility. It's about accountability.

Over the past two years, enterprise AI has moved from curiosity to capability. What once required specialized teams and long development cycles can now be done in a matter of days. Teams can generate code, spin up agents, and automate workflows faster than ever before.

That shift has changed the conversation at the executive level.

A year ago, leaders were asking: *What could AI do for us?*
Now the question is more direct: *What is it costing us and what value are we actually seeing?*

As organizations move beyond pilots and begin scaling AI across real operations, expectations are rising. AI is no longer confined to innovation budgets. It is becoming part of how the business runs. And that means it has to stand up to scrutiny around reliability, governance, and cost.

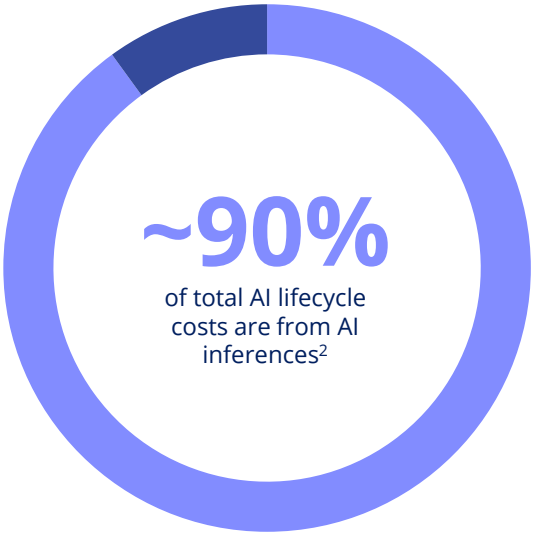
Across industries, there's a growing realization: The AI promise is clear. The economics are not. Recent Gartner research[1] shows most AI initiatives still struggle to deliver meaningful return, with many organizations seeing little to no measurable ROI from early investment.

That isn't a failure of AI. It's a signal that we're entering a new phase: one where success will depend less on what you can build, and more on how effectively you can operate it at scale.

[1] *Gartner research on AI value (gartner.com/en/articles/ai-value)*

AI is getting easier to build, but harder to operate efficiently

The first wave of AI adoption was driven by access. As powerful models became widely available, organizations rushed to experiment. Innovation accelerated. But as those experiments move into production, a different reality is emerging.



Costs that seemed manageable at pilot scale begin to climb. Outputs vary in quality. And in many cases, the link between AI investment and business impact becomes difficult to measure. At the core of this challenge is a structural issue: Most AI systems today are designed for **experimentation, not sustained operation**.

Many rely heavily on *runtime reasoning*, where the model evaluates each interaction from scratch. While that flexibility can be useful, it comes at a price. Every interaction consumes compute. Every escalation in complexity increases that consumption. And as usage grows, those costs compound.

Every time AI has to reevaluate a task, interpret context, or decide what to do next, it incurs additional cost. At small scale, this is manageable. At enterprise scale, it compounds because the system isn't just executing work, it's repeatedly doing the same thinking over and over again. In effect, many organizations are now experiencing what amounts to a kind of "token tax," where every additional step of reasoning carries a marginal cost that compounds at scale.

And at scale, that begins to change the cost equation entirely. In many enterprise deployments, the cost of inference – running models continuously in production – quickly outweighs the initial cost of building them, turning AI from a one-time investment into an ongoing operational expense.

The real cost of AI isn't what you create. It's what you have to keep running continuously. Organizations are paying continuously to "figure out" work that could have been structured in advance. This is why solutions that perform well in a controlled pilot often struggle to scale effectively in the real world. **The constraint isn't potential. It's efficiency.**

[2] *MIT Technology Review, "AI's energy usage and climate footprint" (2025)*

The real cost of AI lives in how work gets done

When leaders look to manage AI spend, the instinct is often to focus on models – comparing providers, capabilities, and pricing structures.

But there's a more important question to ask: **Where in your operation is AI doing the work? And how is that work structured?**

In many organizations, AI has grown organically, with different teams deploying different models, agents, and workflows with little coordination. The result isn't just complexity. It's duplication. The same decisions are made multiple times, in multiple places, often with different logic.

Forecasts now point to enterprise AI workloads driving more than a 20-fold increase in token usage over the next several years[3]. As agents scale across the business, that growth doesn't just expand usage – it expands cost. As a result, AI is often layered onto existing processes without fundamentally changing how those processes function.

This leads to incremental improvement, but not transformation, which creates two systemic issues:

1. It preserves inefficiency. If the underlying process is fragmented or overly complex, introducing AI doesn't eliminate that complexity. It simply accelerates it.
2. It shifts too much responsibility to runtime. Instead of defining how work should happen ahead of time, organizations rely on AI to determine the next step dynamically, interaction by interaction.

That flexibility can feel powerful. But it introduces variability, increases cost, and makes outcomes harder to control. This is ultimately an architectural issue:

- Well-designed systems reduce effort before work begins.
- Poorly structured systems incur cost every time they run.

When AI is applied without rethinking how work flows across systems and teams, cost becomes something that grows with usage rather than stabilizes with scale.

That's why many organizations see spending rise faster than results.

[3] *Goldman Sachs, "AI agents forecast to boost tech cash flow as usage soars"*

Adding AI to broken processes doesn't fix them – it scales their problems

A frequently-seen pattern is what might be described as a "bolt-on" approach to AI. A business identifies a workflow – like customer service, onboarding, claims – and inserts an AI layer to improve it. The experience becomes more conversational. The system becomes more responsive.

But the underlying process remains unchanged. Multiple systems are still involved. Decisions are still fragmented. Work still moves through a series of disconnected steps. To compensate, the AI takes on more responsibility. It interprets, routes, and adapts in real time, effectively trying to hold the process together.

That approach can improve the experience in the short term – but it doesn't change the cost structure. In fact, it often increases it. Because the system is now performing both the original work and the ongoing interpretation required to hold that work together.

Each time the system has to reassess or reconstruct the flow of work, it consumes additional resources. Multiply that across thousands or millions of interactions, and you quickly see the impact. AI, in this model, becomes a layer of continuous interpretation sitting on top of structural complexity.

That's not sustainable at scale. If the goal is to change the cost equation, the starting point has to be the process itself.

Cost control starts with design, not runtime optimization.

The organizations making meaningful progress with AI are approaching the problem differently. They are shifting to an orchestrated model – where AI is coordinated, structured, and embedded within how work is designed and executed.

Rather than relying on AI to make decisions in the moment, they are using it earlier – during the design of how work should be done. At this stage, AI can be incredibly powerful. It can help analyze existing processes, identify inefficiencies, and propose better ways to structure work across systems and teams.

Once that structure is defined, execution becomes far more efficient. At runtime, the system no longer needs to "figure out" each step. It follows a defined path, using AI selectively where it adds the most value – such as interpreting intent or handling unstructured inputs.

This shift from *runtime reasoning* to *design-time intelligence* changes both the cost profile and the consistency of outcomes. It allows organizations to:

- Reduce unnecessary computation
- Eliminate repeated decision-making
- Ensure that work is executed in a repeatable, governed way

AI, in this model, becomes part of a system of execution – not a substitute for it.

Predictability enables enterprise-scale AI

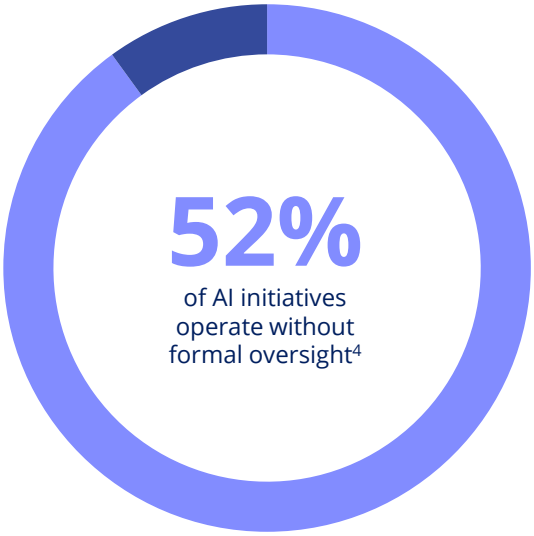
For AI to become foundational in the enterprise, it has to operate with a level of discipline that matches other core systems.

That starts with predictability. Leaders need confidence in three areas:

1. **What it will cost**
2. **What result it will produce**
3. **How it will behave under real-world conditions**

Without that, scaling becomes risky.

Recent EY research shows more than half of AI initiatives are operating without formal oversight – exposing organizations to cost overruns, compliance gaps, and inconsistent outcomes as adoption scales.



This is where a more structured approach to AI begins to take hold. When AI operates within clearly defined workflows, supported by orchestration and governance, something important happens:

- Outcomes become more consistent.
- Costs become easier to anticipate and increasingly tied to decisions, transactions, and real business outcomes.
- Operations become easier to manage.

And importantly, this predictability doesn't just apply to token usage. It extends across the full cost of ownership – operations, governance, compliance, and the ongoing effort required to evolve systems over time. This is where total cost of ownership becomes clearer, extending beyond tokens to include the cost of operating, governing, maintaining, and scaling AI over time.

Instead of treating AI as an open-ended capability, organizations begin to treat it as a controlled, measurable part of how work is done. That shift allows them to move from tracking consumption to tracking outcomes. And that's a critical step toward making the economics of AI work.

[4] EY, "Autonomous AI adoption surges as oversight falls behind" (2026)

The organizations that succeed will change how they think about AI

Many organizations are still operating under a false choice: Move fast with AI or maintain control. In reality, the constraint isn't speed – it's architecture.

The companies that realize the full value of AI won't simply be those that adopt it quickly.

The companies that get this right rethink how their business operates. In practical terms, that means focusing on a different set of priorities:

Shift the focus from tools to systems

The key question is not what AI can do in isolation, but how it integrates into the flow of work across the enterprise.

Redesign before you automate

Applying AI to a broken process rarely fixes it. Reimagining the process first creates the foundation for meaningful impact.

Measure outcomes, not activity

Success should be tied to business results – completed work, improved experiences, reduced friction – not raw usage.

Treat AI as operational infrastructure

As AI becomes more embedded in core processes, it needs to meet the same standards for reliability and governance as any other enterprise system.

This is less about adopting new technology and more about applying discipline to how that technology is used.

From experimentation to execution

We are at a turning point in enterprise AI. The early phase demonstrated what was possible. The next phase will define what is practical and what scales.

As organizations expand their use of AI, the challenge is no longer generating insights or automating individual tasks. It is operating those capabilities in a way that is stable, efficient, and aligned to business outcomes.

That's the difference between isolated success and enterprise-wide impact.

The organizations that close that gap will be able to move faster, operate more efficiently, and adapt more effectively to change.

But they won't achieve that by simply deploying more AI.

They will achieve it by building the right foundation for how AI operates within their business.

The bottom line

AI is already reshaping what organizations can do.

Now the focus shifts to how it performs under real conditions at scale, across complex systems, and within the constraints of cost and control.

The largest hidden cost in AI isn't tied to the model itself. It comes from how the work is structured, executed, and repeated over time.

The organizations that solve for that won't just reduce cost. They'll build a system they can actually run their business on.

And in the next era of enterprise AI, those are the organizations that pull ahead – because they've built AI they can control, measure, and operate as part of the business.

About Pega

Pega provides the leading AI-powered platform for enterprise transformation. The world's most influential organizations trust our technology to reimagine how work gets done by automating workflows, personalizing customer experiences, and modernizing legacy systems. Since 1983, our scalable, flexible architecture has fueled continuous innovation, helping clients accelerate their path to the autonomous enterprise. **PEGA.COM**