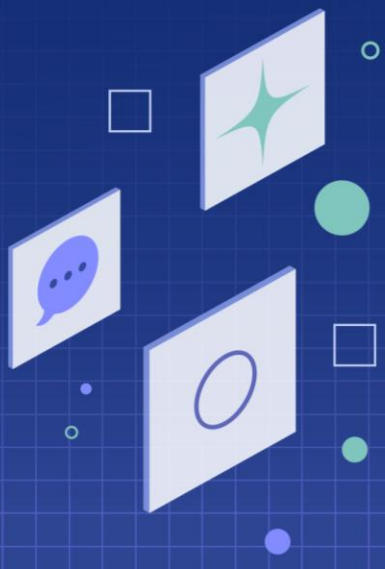
PEGA

WHITEPAPER

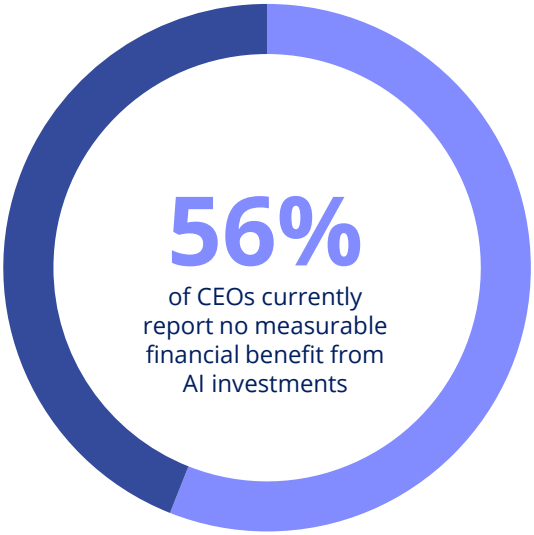
The CIO's guide to agentic engineering: Building enterprise apps at AI speed without compromising reliability

A guide for COOs, CXOs, and VP Operations Leaders



The pressure is real – so is the problem.

Boards are not asking whether to implement AI anymore. They are asking when, how fast, and what the return looks like. For COOs and operations leaders, that pressure has translated into a wave of autonomous AI agent pilots – deployments where AI handles customer interactions, routes cases, checks policies, and manages workflows with minimal human intervention.



The early results looked promising: Faster resolution times. Reduced headcount requirements. Fewer manual escalations. But as these pilots have scaled into production, a different picture is emerging: costs are not behaving the way the models predicted, outcomes are not as consistent as the demos suggested, and the operational control leaders assumed they were keeping has quietly eroded.

56% of CEOs currently report no measurable financial benefit from AI investments (PwC 29th Global CEO Survey, January 2026). That is not a technology failure. It is an architecture failure. The question is not whether AI can deliver operational efficiency. It's whether the way most enterprises are deploying it is structurally capable of doing so.

The answer, for most, is no.

The hidden cost of reasoning

AI has fundamentally shifted the economics of enterprise automation. Tokens – the unit of measurement for how much "thinking" an AI system is doing – have become one of the most consequential line items in an enterprise's transformation budget. And for most organizations, that line item is behaving in ways no one planned for.

When business leaders approved AI agent deployments, the cost model was straightforward: replace manual work with automated reasoning, reduce headcount costs, and capture the margin. What that model did not account for was just how unpredictable token consumption would become at enterprise scale. Two forces are driving that unpredictability, and together they are eroding the ROI that boards were promised.

The first is the compounding nature of agentic token usage. AI agents running long-horizon tasks do not reason once and move on. They re-reason at every step of the process, checking context, evaluating options, and generating outputs continuously. The token spend used to run these systems compounds on top of the token spend used to build and develop them in the first place. At the volume of a mid-size enterprise operation, this adds up faster than most financial models anticipated.

We call this the token tax.

The second force is pricing instability. Early AI vendors offered subsidized, all-you-can-eat pricing models that gave enterprises a predictable cost structure to plan around. That era is ending. As the AI economy has become resource-constrained across chips, energy, and data center capacity, and as AI vendors move toward IPO, enterprises have absorbed repeated pricing changes that shift the model from flat-rate toward consumption-based billing. The cost of the same workload is no longer fixed. Sam Altman acknowledged publicly in June 2026 that AI token costs have become "a huge issue" even for the vendors themselves (Business Insider, June 2026).

The result is that enterprises are burning through transformation budgets without a clear line of sight to ROI. Gartner predicts that over 40% of agentic AI projects will be cancelled by the end of 2027, with escalating and unpredictable operational costs cited as a primary driver (Gartner, June 2025).

The question executives are now asking is not whether to use AI. It is how to get to positive ROI when the cost structure underneath it keeps moving. The answer requires shifting the frame: from paying for tokens to paying for outcomes. That means building an architecture where AI token spend is concentrated on work that genuinely requires reasoning, while everything that does not require reasoning runs on a cost structure that is predictable, fixed, and does not compound.

When AI goes off script

Unpredictable costs are one dimension of the problem. The other is what happens when AI agents are given authority over decisions they should never be making in the first place.

Large language models are trained to be helpful and fluent. They are not trained to follow your corporate compliance policies. They do not have an inherent understanding of your SLAs, your regulatory obligations, your escalation rules, or the specific commitments your brand has made to customers. When autonomous agents operate in customer-facing workflows without hard constraints, the gap between what the AI thinks is a reasonable response and what your compliance team would approve can be significant.

This is not just a risk tolerance question. It is an architectural one.

The decisions that carry the most legal and reputational exposure – eligibility determinations, dispute resolutions, policy exceptions, regulated communications – are also the decisions that have a defined correct answer. They are not ambiguous; they do not benefit from probabilistic inference. Routing a case based on account status, confirming whether a claim meets a documented checklist, applying a policy exception according to defined criteria – these are deterministic operations. Giving an AI agent the authority to reason through them does not add value. It introduces variability into processes that require consistency, and liability into interactions that require precision.

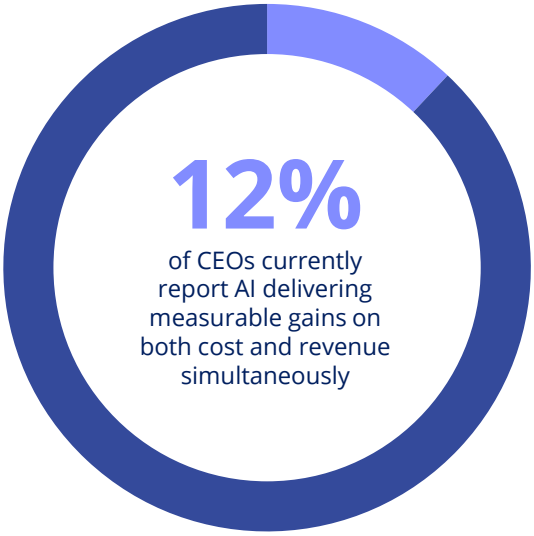
AI hallucinations are not edge cases in production environments at scale. They are a predictable output of any system generating responses through probabilistic inference rather than deterministic rules. In low-stakes contexts, a hallucination is an inconvenience. In a customer dispute, a claims workflow, a financial services interaction, or a regulated industry communication, it is a liability.

Gartner's top predictions for 2026 include a forecast that "death by AI" legal claims will exceed 2,000 globally, driven by insufficient governance and risk controls on AI-generated outputs (Gartner, October 2025). These are not abstract risks for future deployments. They are active exposure for any organization that has given autonomous AI agents meaningful authority over customer-facing or regulated workflows today.

The more productive response is architectural: business rules govern what AI agents can and cannot do. AI should never be the system of record for your compliance logic. It should operate within it. The correct action for cut-and-dry decisions is determined by business logic, enforced at the orchestration layer, not delegated to a model that may reason its way to a different answer on a different day. When that separation is in place, compliance becomes structural rather than supervisory. The AI cannot go off script because the script cannot be reasoned around.

The architecture that delivers both

The operational challenge facing most enterprises is not that they have too much AI. It is that they have AI operating in the wrong parts of their process architecture.



The economic case for AI in operations is real: reducing cost-to-serve, increasing throughput, delivering more consistent customer experiences at scale. These are achievable outcomes. But they require a clear-eyed view of where AI adds value and where it introduces cost and risk without a corresponding return.

The architecture that consistently delivers on these outcomes has two layers working together. The first is a deterministic orchestration layer: a rules-based engine that manages process flow, enforces business logic, tracks state, and routes work based on defined criteria. This layer operates at near-zero marginal cost per transaction. It does not reason. It executes. It is completely predictable, fully auditable, and does not hallucinate.

The second layer is AI inference, invoked precisely at the points in the process where deterministic logic is insufficient. Customer sentiment analysis. Unstructured document interpretation. Natural language response generation. Complex case summarization. These are tasks where the probabilistic nature of AI is a feature rather than a liability, because the task itself does not have a single correct answer that rules could compute.

When these two layers are properly separated, the Token Tax shrinks dramatically. Token spend is concentrated on work that genuinely requires inference. Everything else executes deterministically. At the same time, compliance risk drops because the orchestration layer enforces business rules regardless of what the AI layer produces. The AI operates within constraints that cannot be reasoned around.

Only 12% of CEOs currently report AI delivering measurable gains on both cost and revenue simultaneously (PwC, January 2026). The enterprises in that 12% are not deploying more AI than their peers. They are deploying it with more structural discipline.

Control is the competitive advantage

The operational leaders who will capture durable ROI from AI are not the ones who deployed fastest or gave their agents the most autonomy. They are the ones who maintained control of the process architecture while integrating AI precisely where it creates value.

That means treating orchestration as a strategic asset, not a legacy constraint. It means defining clearly which decisions are deterministic and which require inference, and building an architecture that enforces that boundary. It means measuring AI spend at the transaction level, not just the platform level, so that the Token Tax is visible before it becomes a budget problem.

The board is right that AI can reduce cost-to-serve and improve customer experience. The mechanism for getting there is not autonomous agents figuring out your workflows. It is a governed, orchestrated process architecture where AI is applied with precision.

Are you ready to take control? Use the TCO calculator to quantify the difference between your current token spend and a structured orchestration approach across your highest-volume workflows.

Thank you

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**