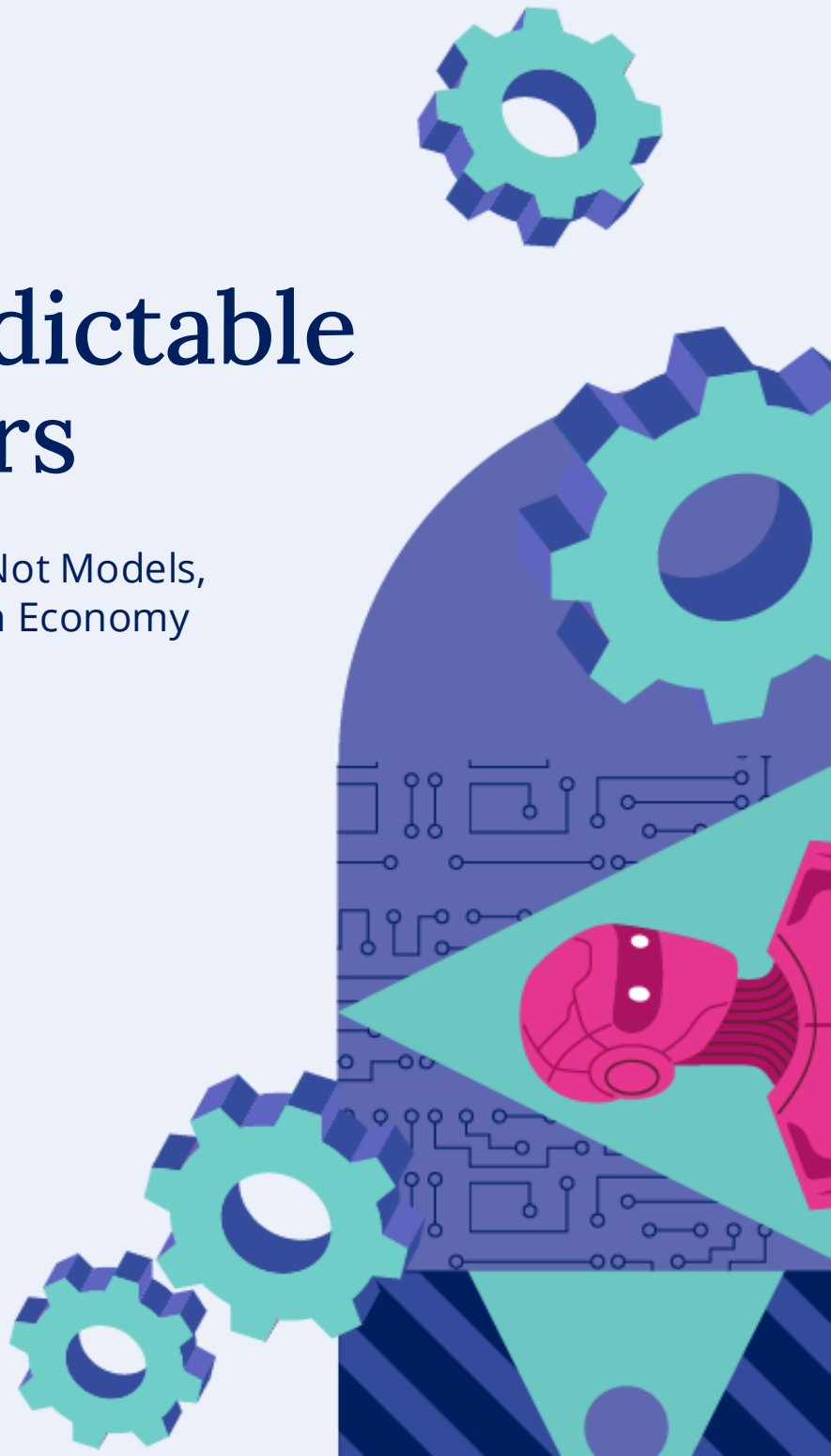
PEGA

Why Predictable AI Matters

Governing Decisions, Not Models, in the Age of the Token Economy



Pega's Structural Advantage

Generative AI introduces a new operational reality for enterprises: intelligence now carries a measurable, variable cost. Every prompt, retrieval, reasoning step, and agent loop consumes tokens, which in turn consume GPU time, energy, and capital. Unlike traditional software, this cost does not flatten with scale; it grows with usage and complexity, often non-linearly.

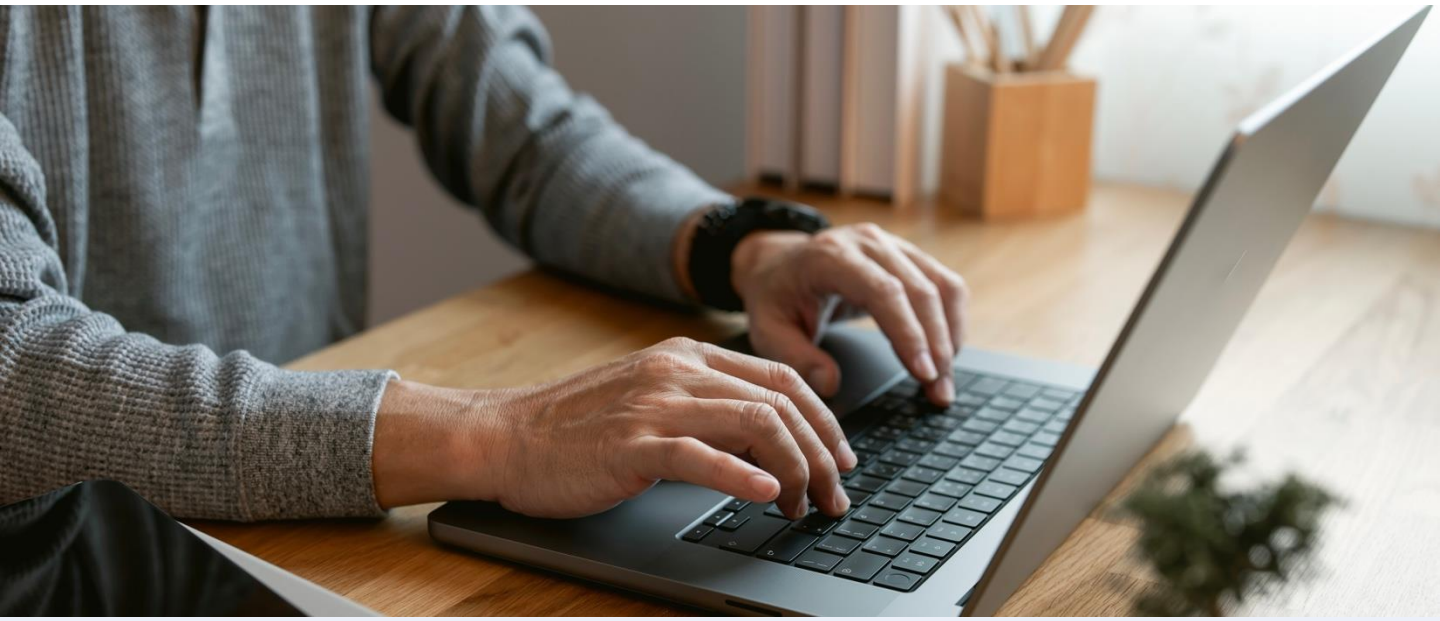
Once tokens are spent, cost and carbon are already incurred. Governance after the fact is ineffective.

This is where Pega is structurally different. Pega already operates at the point where enterprises decide what happens next: evaluating options, applying policy and constraints, and executing decisions in real time. That same decision layer determines whether an entire execution path should be delegated to probabilistic intelligence at all, and under what constraints. Once delegated, the system must already know the economic, latency, and risk boundaries within which that execution is allowed to run.

Predictable AI is therefore not a model feature or a prompt framework. It is an architectural position: Pega sits at the control plane where GenAI economics can be governed in real time.

This reframes Predictable AI as token-level economic governance, evaluated at design-time and enforced at runtime, when decisions are made, rather than delegated to individual applications.

Design time evaluation defines economic intent and policy; runtime enforcement is where cost is controlled.



The token economy breaks the digital cost curve

For decades, enterprise software followed a predictable economic pattern: marginal cost declined with scale.

GenAI reverses that curve. Tokens represent a direct mapping to compute, energy, and infrastructure cost. As organisations scale AI usage, token consumption grows with richer context, longer histories, retrieval augmented generation, and agentic workflows, because improvements in quality are typically achieved by adding context, retrieval, and reasoning steps, each of which increases tokens per interaction.

Crucially, these costs are structural, not temporary. The infrastructure that underpins GenAI – dense GPU clusters, energy intensive datacentres, and accelerated hardware depreciation – creates an inflationary cost base that enterprises cannot offset through efficiency alone.

Industry analysts now broadly acknowledge that AI infrastructure costs behave fundamentally differently from traditional software economics, with sustained investment patterns that reflect persistently rising, not declining, marginal cost.

This makes GenAI not just a technical challenge but an economic and architectural one.

Why governance must move from models to decisions

Most GenAI discussions focus on model selection: which LLM is more accurate, faster, or cheaper per token. While important, this misses the core issue. Cost and risk are driven less by model choice than by invocation frequency and decision context.

In this context, a “decision” is not a single branching point in a flow. It is the commitment to execute an end-to-end path that may involve prompt construction, retrieval, reasoning loops, tool use, and agentic iteration. Once that execution begins, token consumption becomes continuous and largely opaque. The economic risk, therefore, is not confined to individual decision nodes, but to the act of delegating whole workflows to a probabilistic system.

Without governance, enterprises often drift toward ‘LLM everywhere’ architectures. Large models are invoked even when rules, small language models, or deterministic logic would suffice. Token consumption becomes unpredictable, budgets spiral, and sustainability impact compounds.

Because tokens are consumed dynamically at runtime, governance cannot be applied retrospectively through budgeting or procurement. It must be enforced when a decision is made.

That shifts governance from the model layer to the decision layer.

Decision tiering: predictable AI in action

A practical response to the token economy is decision tiering. Instead of asking "Which model should we use?", organisations classify decisions by complexity, risk, and business value.

Rules handle deterministic, low variability decisions at near zero token cost

Small, task-specific, and predictive models address moderate complexity efficiently

LLMs are reserved for ambiguity, synthesis, and reasoning

Agentic workflows are used only where continuous reasoning justifies the cost

Routing decisions through progressively more expensive tiers ensures that enterprises consistently apply the lowest cost viable intelligence at runtime.

This is where Predictable AI becomes operational. It is not about avoiding LLMs; it is about making escalation explicit, governed, and accountable. As this routing is enforced centrally rather than embedded in individual applications, it can be standardised, reused, and governed consistently across the enterprise rather than re-implemented case by case.

A common counter argument is that economic governance can itself be delegated to AI. On the surface, this appears to solve the problem by making cost control proactive rather than retrospective. In practice, it simply pushes the most critical economic decision – how to trade quality, completeness, and correctness against cost – into the hands of an opaque, non-deterministic system.

The organisation no longer governs outcomes; it governs intent and hopes the model interprets that intent appropriately.

Business architecture and technical architecture must work together

This is not only a technical architecture concern.

From a business architecture perspective, enterprises must define:

- Which decision classes justify high-cost inference
- Where business value outweighs token and energy consumption
- Who owns approval for expensive or agentic AI patterns
- How cost per decision aligns to outcomes and KPIs

From a technical architecture perspective, platforms must be able to:

- Enforce routing and policy centrally
- Measure token usage per decision and per workflow
- Apply thresholds for cost, latency, and risk
- Provide deterministic fallback paths when escalation is not justified

Pega's advantage is that these two architectures converge in the same place. The platform already operationalises policy, governance, and execution at scale. The token economy simply makes that convergence unavoidable.

This convergence also demands clear ownership. Governing AI economics requires teams that can translate architectural choices into financial impact – linking decision design, cost per decision, and business value rather than treating AI consumption as a purely technical concern.

The coming reality: From economic governance to regulatory pressure

Today, most enterprises have little meaningful governance over GenAI usage. LLMs are embedded directly into workflows, applications, and products with minimal runtime control over when they are invoked, how often, or at what cumulative cost. As long as spend remains experimental, this lack of discipline is tolerated. That tolerance will not last.

The first unavoidable phase of the coming reality is not regulation, but economic governance. As GenAI moves from pilots to scaled operations, leadership teams will demand answers to fundamental questions:

Why is AI spend growing faster than usage?

Which decisions genuinely justify high cost inference?

What is our cost per decision – and how does it scale as volume increases?

How predictable is our AI cost base as adoption grows?

Because token consumption is variable, non-linear, and incurred at runtime, traditional financial controls cannot answer these questions – the primary controllable moment is before invocation. Budgeting, procurement, and post hoc FinOps reporting all arrive too late – after tokens have already been spent.

As a result, enterprises will be forced to introduce economic governance first: mechanisms that control AI usage before execution. This means routing decisions to the lowest cost viable form of intelligence, enforcing thresholds, and making escalation to high-cost models deliberate rather than implicit.

Cost per decision (the total inference, retrieval, and orchestration cost required to produce an outcome) becomes the unit economics of Predictable AI, replacing abstract debates about model pricing with operational reality.

Over time, this economic discipline will become the foundation for regulatory pressure. As AI energy consumption becomes material at enterprise and national scale and pressure mounts to adapt public revenue models as automation reshapes workforces and erodes traditional labour-based tax bases – a concern already visible in policy discussions around the future of public revenue systems – it is increasingly plausible that governments and regulators are likely to build on cost transparency with formal requirements. These responses may include carbon-linked reporting, energy-based levies on compute intensive workloads, or audit obligations requiring organisations to explain where and why GenAI is being used.

When that happens, organisations without runtime economic governance will struggle. Without decision-level controls, it becomes impossible to demonstrate that expensive or energy intensive inference was justified, necessary, or proportionate. Compliance will require more than dashboards; it will require architectural proof that alternatives were evaluated and enforced.

Seen in this light, Predictable AI is not a reaction to future regulation. It is the precondition that allows enterprises to reach that future without disruption. Economic governance comes first. Regulatory resilience follows.

Why predictable AI matters more over time, not less

Many of today's GenAI differentiators – accuracy gains, hallucination reduction, prompt frameworks – will inevitably converge. Improvements in models and tooling will narrow those gaps.

What will not converge is the economics of AI.

Token cost, infrastructure pressure, energy usage, and regulatory scrutiny will persist and intensify.

This creates a narrow strategic window to definePredictable AI around governance and economics before these concerns become universal and commoditised.

This mirrors a broader pattern already visible in analyst and advisory research: early advantage in AI accrues to capability, but long-term advantage accrues to operating discipline, economic control, and governance at scale.

In that environment, competitive advantage will belong not to the organisations that deploy the most GenAI, but to those that deploy it most intelligently and economically.

Predictable AI, grounded in decision-level governance, gives enterprises a way to scale GenAI responsibly – financially, operationally, and sustainably.

Conclusion

Predictable AI is not about slowing innovation. It is about ensuring that intelligence is applied deliberately, consistently, and with economic intent.

By governing decisions rather than models – and by aligning business architecture with technical execution – enterprises can harness GenAI at scale without surrendering control of cost, carbon, or compliance.

This is where Pega's architecture is not just relevant, but essential.

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