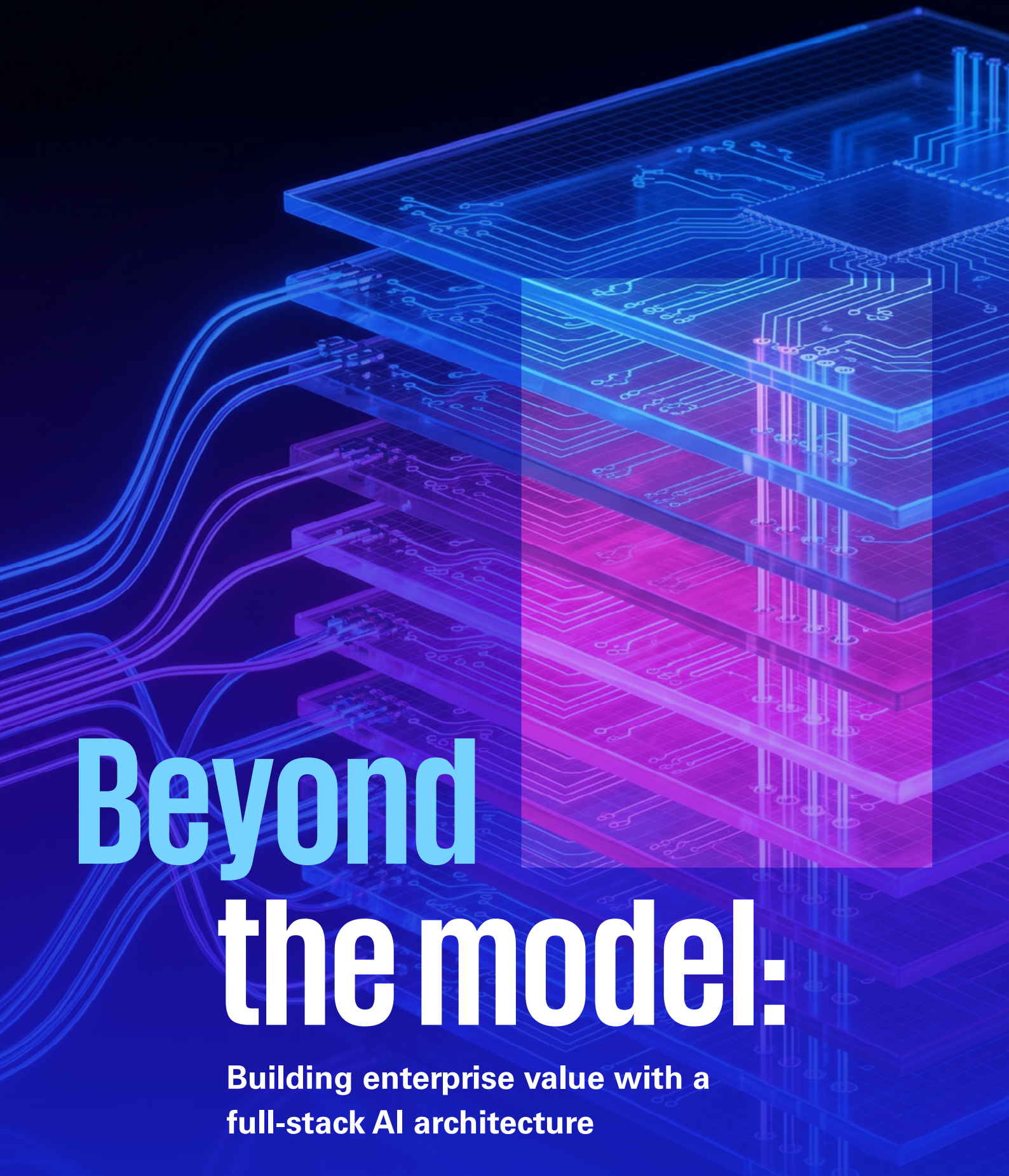
KPMG



Beyond the model:

Building enterprise value with a full-stack AI architecture

July 2026

With AI, progress usually makes headlines when a shinier model emerges or a slicker interface ships. Companies swap one large language model for another, graft an assistant onto an existing application, and declare victory. Pilot projects sprout like weeds, usage climbs, and token consumption accelerates—while value remains elusive.

This tension occurs because models and interfaces are fragments of a larger whole. Yes, there are bright spots: a cleaner month-end close, automated exception handling, faster service resolution. But these impacts don't compound, and when momentum fizzles, transformation stalls.

What if the gateway to value isn't intelligence but architecture design? Enterprise systems are still built on three outdated assumptions: Processes live inside applications, coordination lives in org charts, and judgment lives only in human heads. Add AI to that structure, and work gets faster, but not fundamentally different. What's changing now is the nature of work itself. Work is no longer embedded inside applications or enforced through org charts; it is being planned, executed, and governed by intelligent systems. As AI systems plan, reason, and act, work can be decomposed, executed, and governed by machines. And when work becomes computational, architecture becomes the operating system of the business.

We propose a minimum viable system architecture with 10 distinct layers and no shortcuts. Skip one layer and, instead of leverage, you've gained a liability. At each step, expect to confront uncomfortable questions: Where does genuine competitive advantage emerge? How do you govern it so that autonomy is both trustworthy and compliant? And perhaps most critically, how will your architecture scale to yield meaningful economic returns?

Get the answers right across all 10 layers and AI stops being a bolt-on tool. It becomes the backbone for how work is done and the operating system for the enterprise of the future. **Here, we explain how.**

“As AI systems plan, reason, and act, work can be decomposed, executed, and governed by machines. And when work becomes computational, architecture becomes the operating system of the business.

Value compounds only when every layer is designed to work as a system and each is doing its job.

73%

of KPMG Q1 2026 AI Pulse Survey respondents cite automating workflows across multiple functions as their next priority.

A full-stack AI architecture offers a new approach to value

Leading organizations are moving beyond fragmented AI use cases to orchestrated, enterprise-wide capability. In fact, 73 percent of KPMG Q1 2026 AI Pulse Survey respondents cite automating workflows across multiple functions as their next priority. For more than half, it's mission critical: 53 percent cite routing critical information between teams as essential.[1]

That is not a model problem. It is a work orchestration problem that spans systems, functions, decisions, and accountability. Value compounds only when every layer is designed to work as a system and each is doing its job.

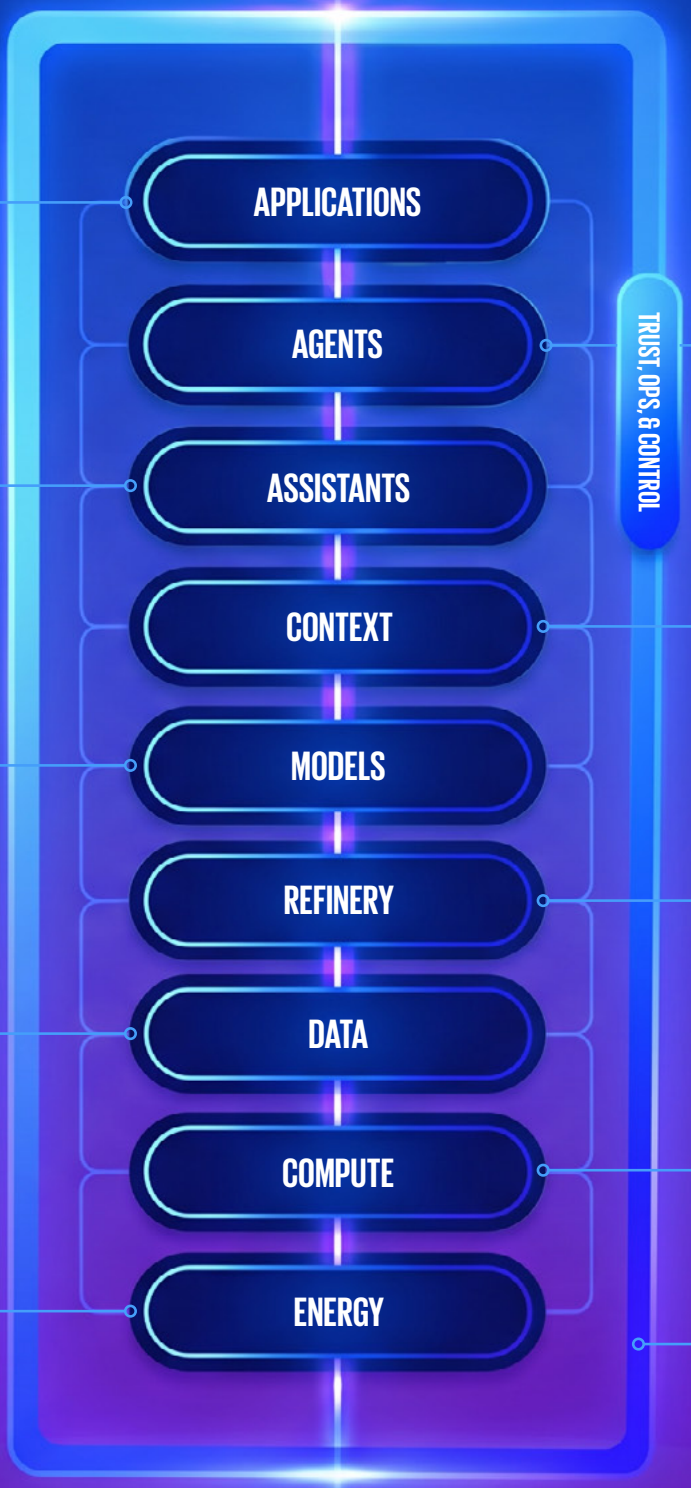
For leading organizations, this shift starts with capability-level alignment, rethinking required outcomes, and reinventing the work itself instead of implementing isolated use cases. Rather than starting with model deployment, they begin by targeting specific business capabilities. From there, they design the technology stack to reimagine the economics and improve performance across the entire capability.

Imagine a full-stack AI architecture that can support reinvented work and achieve the desired outcomes? This approach has 10 layers, each with a defined role, a characteristic failure mode, and decisions that determine whether it becomes an asset or a liability. At the center of the stack is a Work Layer, or the execution plane where intent is translated into governed action by agents operating with shared context. This anchors the organization's knowledge, practices, memory, and rules in a common semantic model.

Underneath, the refinery prepares enterprise data so it's AI-ready. Then, the data layer houses the full enterprise data estate, including systems of record. The governance layer sets the guardrails within which autonomous processes can operate. The result: Success is defined from the outset and is based on value-focused key performance indicators, like error reduction or new revenue generation, instead of technical metrics.

1 KPMG LLP, AI Pulse Survey Q1 2026 (April 2026)

The KPMG full-stack AI architecture encompasses nine layers plus a trust, ops, and control wrapper that delivers transparency and trust throughout.



AI work interfaces and apps that evolve with context, streamline interactions, and transform how work gets done.

Retrieval-driven copilots that answer questions and support tasks by grounding responses in enterprise knowledge and search.

Foundational multi-modal models, including various deterministic AI models and specifically unique algorithms, for advanced inferencing and reasoning.

Data (structured and unstructured) required to train & tune the AI. The data AI models need to access to perform inferencing. The data that AI manages.

Power, cooling, data centers, grid access, and carbon posture at sustainable industrial scale.

Autonomous systems that plan, coordinate multiple steps or tools to achieve user goals with optimal human oversight.

Semantic layer, curated instructions, knowledge, memory, tools, and guardrails provided to models, agents and assistants to ensure accurate, safe, and goal-aligned outcomes.

Acquire, process, enrich, and index data to make it optimized and discoverable for training, inference, and retrieval.

Silicon-optimized hardware including edge devices for training, tuning, and serving of AI models.

End-to-end safeguards, governance, evaluations, system cards, observability, runtime management, and value measurement to ensure AI remains transparent, trustworthy, reliable, scalable, and outcomes-driven across all layers.

Full-stack AI architecture overview

> Without the right orchestration and Agent Skills discipline, organizations wind up with demo agents that look intelligent but cannot reliably complete work.

Applications

These are the work surfaces where humans and systems interact, such as procurement approval or compliance reviews. Ideally, applications should adapt to context and reduce friction across end-to-end work and not simply add a chat panel to an existing screen. In failure mode, the interface looks modern, but if users still need to do heavy lifting outside the tool, then workforce adoption is likely to stall.

Agents

The full-stack AI architecture decomposes work into atomic, reusable units of execution, or Agent Skills, that compose, sequence, and govern with the right level of human oversight. At the surface, this is where people delegate work. Underneath, they execute by composing Agent Skills, sequencing actions, handling exceptions, and driving progress toward outcomes.

As applications evolve from fixed workflows to adaptive, dynamic execution environments, ownership shifts from traditional app teams to agent orchestrators. These orchestrators redesign and reinvent familiar business processes as autonomous systems and step in when automation needs human intervention. In practice, this layer brings in new roles: workflow and autonomy designers who turn business goals into governed agent behaviors and oversee execution at scale.

Without the right orchestration and Agent Skills discipline, organizations wind up with demo agents that look intelligent but cannot reliably complete work. Exception handling reroutes to humans, while autonomy and productivity dwindle. The organization concludes that agents don't work, when the real failure is the architecture.

Assistants

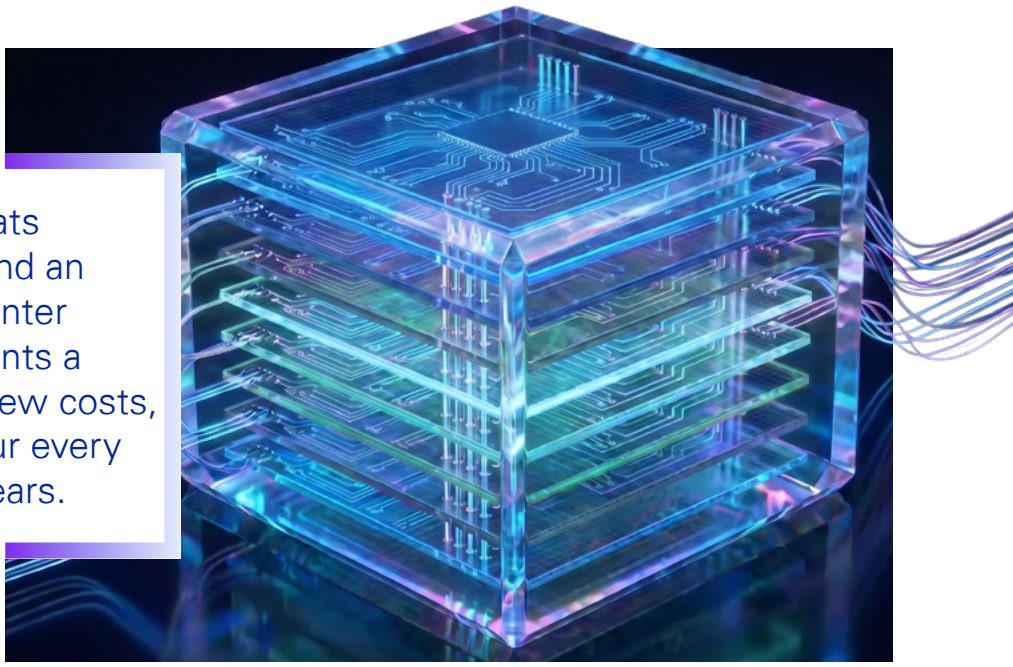
Assistants are retrieval-driven copilots that ground responses in enterprise knowledge by way of policy lookups, document synthesis, or procedure guidance. Increasingly, assistant capability is embedded within agents rather than deployed as a stand-alone tier. The distinction matters: An assistant retrieves and responds. An agent plans and acts. When that line blurs, assistants drift into autonomous action without the controls or execution design to support them. Output becomes unreliable, trust breaks, and user adoption stalls.

Context

The context or semantic layer determines whether agents and assistants behave with accuracy and accountability. Together, context and agents form a significant part of the Work Layer. Context includes curated instructions, approved Agent Skills, institutional knowledge, durable memory, authorized tools, and guardrails provided at runtime. It ensures systems know what matters, what is permitted, what has already happened, and what must be preserved for audit and learning. Without it, the same question yields different answers and the same workflow yields different decisions. Reliability collapses and teams respond with hardcoding, manual checks, and brittle prompt patches.

Models

Models provide inferencing and reasoning capability. The strategic question is not which model to use but how models are routed, versioned, evaluated, and governed as performance and constraints change. A full-stack architecture treats models as capabilities behind an interface rather than the center of the system. It also prevents a model churn cycle where new costs, rework, and disruption occur every time a "better" model appears. A deliberate model strategy spanning edge, local, and cloud deployment is also the primary lever for tokenomics that contain cost sprawl before it compounds across the stack. These decisions determine not only performance, but also unit economics—which decisions can run continuously, which require batching, and which demand human validation to remain financially viable.



"

A full-stack architecture treats models as capabilities behind an interface rather than the center of the system. It also prevents a model churn cycle where new costs, rework, and disruption occur every time a "better" model appears.

Refinery

How enterprise content is indexed, chunked, and surfaced at inference time is a source of real competitive differentiation. Here, the refinery layer makes enterprise data AI-ready. Classical extract, transform, and load pipelines handle acquisition, cleaning, and standardization. Beyond that, the refinery structures knowledge through retrieval-augmented generation, vector indexing, and knowledge graph construction, enabling agents to reason over relationships, not just records. Beware: A refinery that looks clean on paper can still produce brittle pipelines when legacy integration is underestimated, so address it before it becomes a production bottleneck.

Data

The data layer is the factual substrate of the organization and when an automated decision is challenged, the audit trail starts here. It incorporates structured systems of record with unstructured repositories of policies, contracts, communications, and third-party signals. Lineage, cataloging, and provenance are not hygiene exercises but instead determine whether the organization can defend automated decisions and trace outcomes back to sources.

Compute

Compute is the silicon and runtime fabric for training, tuning, and serving. It is a direct driver of unit economics for AI execution. It sets latency, throughput, placement (cloud, on-premises, Edge) and often regulatory posture. Token economics are not an operational concern; they are a design constraint that determines how work is decomposed, which Agent Skills run where, and which execution paths are economically viable at scale. This shifts financial planning from fixed IT budgets to usage-based forecasting, where the marginal cost of each autonomous decision is weighed against the business value it produces. Without deliberate management, tokenomics becomes a monthly surprise rather than a controllable design parameter. Costs spike unpredictably, performance becomes inconsistent, and scale becomes a budget problem instead of an engineering one.

Energy

The energy layer comprises power, cooling, data centers, grid access, and carbon posture. The industry underpriced this layer for a decade, but organizations that treat energy as a facilities issue rather than an architectural input will quickly constrain their AI ambitions. For enterprises with net-zero commitments, the energy layer becomes a gating factor for scale, forcing explicit trade-offs between autonomy, latency, cost, and carbon impact. In other words, sustainability posture starts here, not in the sustainability report.

Trust, operations, and control

Governance bolted on after deployment creates the compliance debt it was meant to prevent. So, while the other layers operate within a boundary, trust, ops, and control are the boundary. This layer also formalizes new accountability structures. Organizations operationalizing autonomy establish AI risk and compliance leads, alongside centralized AI Centers of Excellence, that span IT, data, legal, and the business. These roles do not slow deployment; they create the conditions under which autonomy can scale safely and credibly across the enterprise.

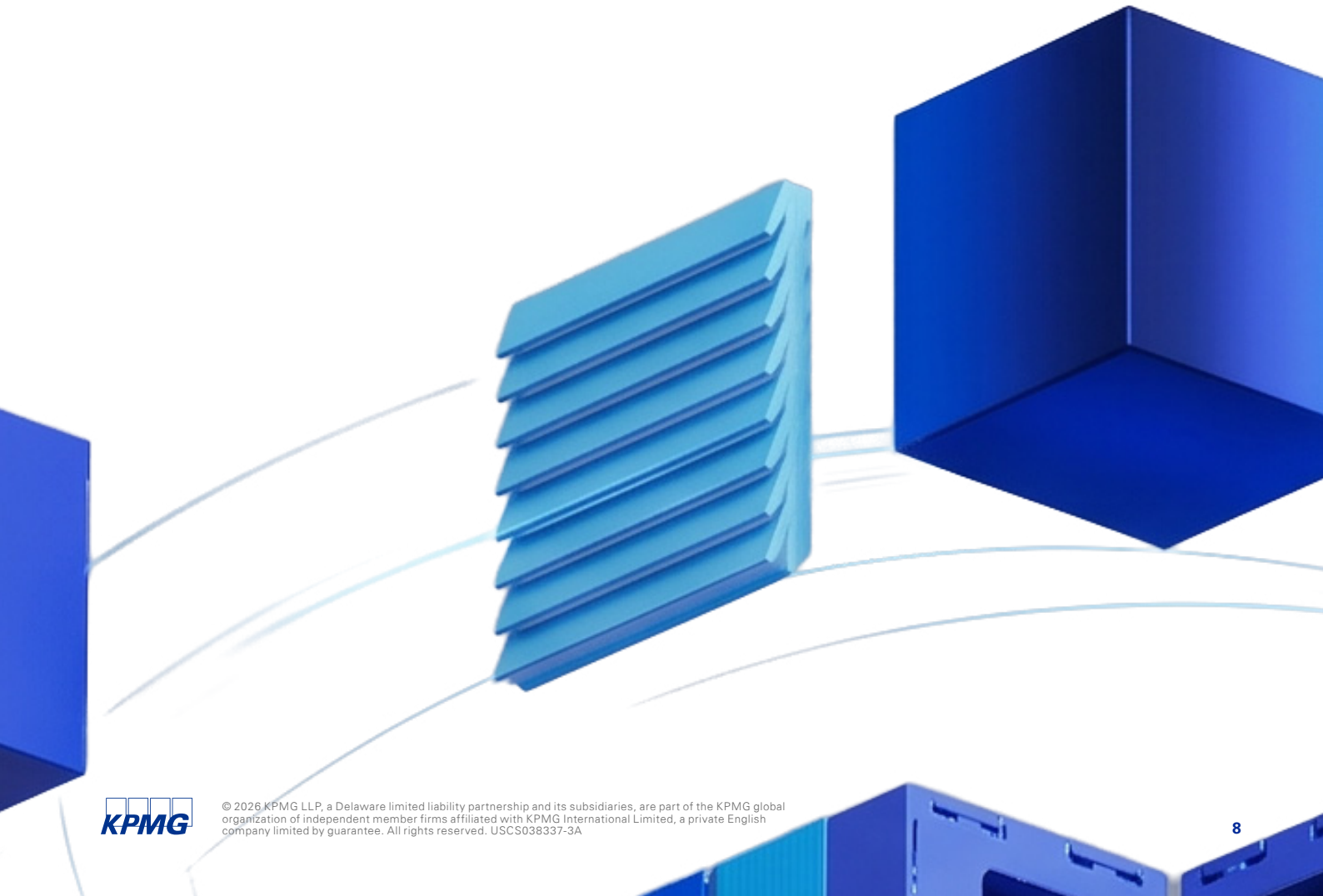
Governance bolted on after deployment creates the compliance debt it was meant to prevent. So, while the other layers operate within a boundary, trust, ops, and control are the boundary.

From architecture to action

The architecture is a reference model, not a sequence. Start with one outcome, rather than a use case or a pilot. Think of a measurable business result with a hard number attached, such as cutting time-to-first-contact, and trace that outcome through the stack to determine which layers are missing or weak. Identify the places where work breaks down because context is absent, Agent Skills don't exist, or data can't be defended. That diagnostic is more valuable than any roadmap because it tells you exactly where to invest and in what order before a dollar is committed.

As AI systems move from assisting humans to planning and executing work, complexity is unavoidable. The common failure is fragmentation. Organizations layer new capabilities onto old structures and hope integration will compensate. It does not. The result is brittle execution, runaway consumption, and outcomes that fail to scale.

Autonomy introduces real trade-offs—speed versus control, flexibility versus cost, scale versus sustainability. Model performance, token economics, infrastructure constraints, context, and governance are no longer secondary concerns. They are design inputs. At this stage, the organizations pulling ahead are not the ones with the most models or the most pilots. They are the ones that decided to rebuild how work gets done—layer by layer, outcome by outcome, with architecture as the discipline, not an afterthought. And the organizations that figure that out first will not play catch-up to anyone.



How KPMG can help

KPMG supports clients in building robust, full-stack AI architectures by providing a thorough, wide-ranging approach that spans strategy, implementation, workforce transformation, and trust. Leveraging established frameworks and advanced tools, KPMG helps organizations shape tailored AI strategies, develop actionable roadmaps, and construct sustainable AI solutions alongside the necessary data infrastructure. An AI factory approach enables rapid piloting and scaling of agentic AI agents, while streamlining integration with emerging technologies and platforms. Throughout the journey, KPMG emphasizes ethical, secure, and compliant AI deployment by embedding trusted AI practices and risk management into every stage. A people-centric methodology helps ensure that both human and AI collaborative efforts are optimized, unlocking value, accelerating adoption, and empowering enterprises to confidently infuse AI capabilities across their operations.

Learn more: https://www.kpmg.us/AIservices

Authors

Swami Chandrasekaran
Global Head of AI & Data Labs at KPMG US
E: swamchan@kpmg.com

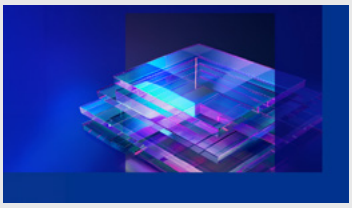
Swami Chandrasekaran is Global Head of AI & Data Labs at KPMG US, where he leads the firm's AI strategy and innovation agenda across Tax, Audit, Advisory, and enterprise functions. He oversees global research and development initiatives spanning full-stack AI architecture, AI agents and Superagents, digital employees, knowledge elicitation, agent control systems, domain-specific small language models, synthetic data, enterprise discovery and search, and embodied AI. A technology executive with more than 25 years of experience in enterprise-scale digital, AI, and cloud transformation, Swami has worked with 500+ clients across 20+ countries on large-scale modernization and AI adoption programs. He chairs the KPMG AI Technology Review Board, helping advance trusted and scalable AI adoption, and serves as a principal investigator with the NIST AI Safety Institute Consortium. Swami holds 30+ patents and was previously an IBM Distinguished Engineer and Master Inventor. He holds a master's degree in electrical engineering.

Matteo Colombo
Global Leader for Cloud, Data, & AI, KPMG US
E: matteocolombo@KPMG.com

Matteo serves as the global leader for Digital Technologies at KPMG. In this pivotal role, he spearheads growth, transformation, and innovation across the KPMG Digital Foundation and the global Centers of Excellence for Cloud, Data, and AI. He plays an instrumental role in shaping the use of technology across the firm's collective strategy and helping ensure its effective implementation in local markets. Additionally, he manages relationships with key strategic technology partners. His experience encompasses both strategy and technology, with a particular emphasis on emerging trends that drive growth and innovation for clients and partners. He is a staunch advocate for the effective and ethical use of AI. With over 20 years of experience, Matteo has adeptly guided large enterprises in harnessing leading technologies to achieve significant, large-scale transformations.

Related insights

Why knowledge engineering is the key to AI agent value



Agentic AI untangled: Navigating the build, buy, or borrow decision

KPMG AI Quarterly Pulse Survey

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Learn about us: **kpmg.com**