Five shifts reshaping the agentic AI economy in 2026

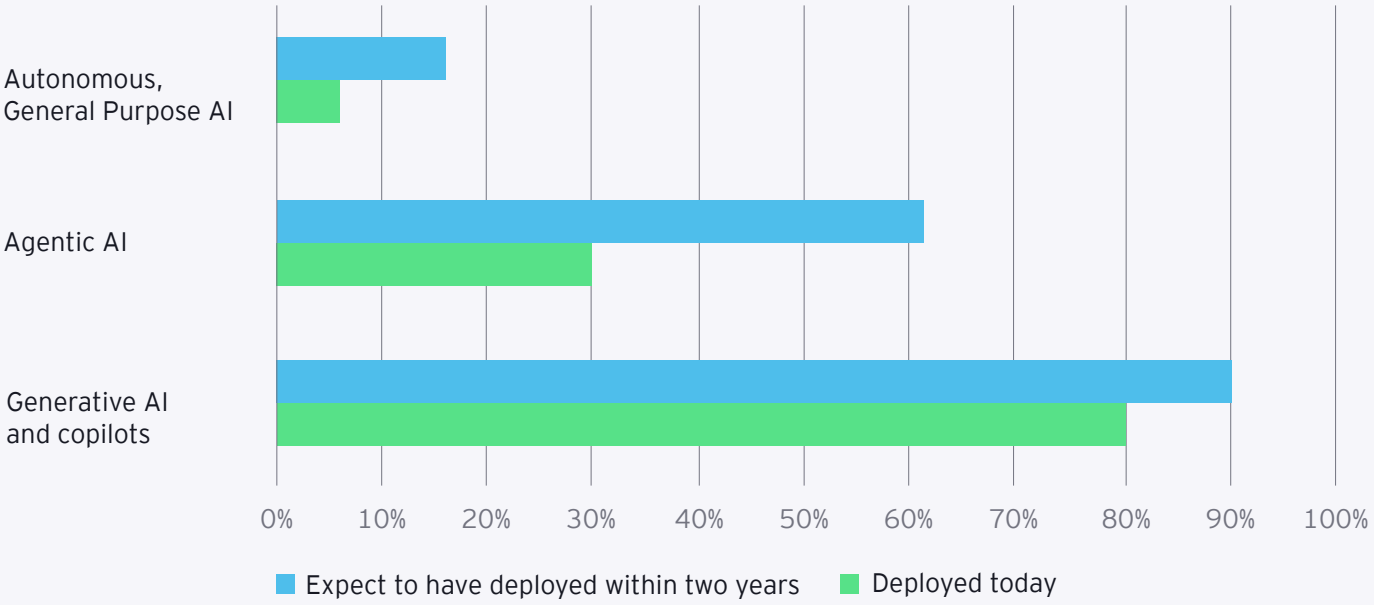
The better the prompt.
The better the answer.
The better the world works.

Shape the future with confidence

At the start of 2026, the question facing technology leaders was whether agentic AI would actually arrive on the timelines people were predicting. Today, that question is largely settled. The more important question is whether companies can industrialize it.

The clearest signal is that artificial intelligence (AI) has moved from experimentation to scaling. The EY May 2026 global technology executive survey found that 69% of respondents identify the shift from experimentation to scaling and production as the most significant change in AI strategy, while 76% expect AI investment to rise by at least 20% over the next 12 months.

Level of AI deployment in Global Tech Enterprise



What follows is a layer-up look at five structural shifts now defining the landscape. The story is less about which companies are winning today and more about what the playing field itself is becoming.

AI pricing is shifting from subsidies to value capture

The first half of 2026 brought what some have called the "SaaSpocalypse" — a reset across legacy software as enterprise buyers confronted a basic mismatch: Per-seat licensing breaks down when autonomous agents, not human employees, do the work. Flat subscriptions assumed relatively bounded human usage; agentic workflows can plan, retrieve, call tools, retry, summarize and run continuously.

Frontier model companies and AI platforms are responding by removing subsidies and capturing more of the value they create — the early contours of a new tokenomics in which price is tied directly to consumption and outcomes rather than seats. Free tiers are tightening, enterprise pricing is shifting toward base access plus metered consumption. The future is not simply seats giving way to usage; it is layered pricing — subscriptions for access, tokens or credits for activity, outcome fees for completed work, and take-rates where agents execute transactions.

Tokens are the visible meter for this transition, but they are not the full cost of an agent. They are where inference intensity, model choice, context strategy, tool retrieval, caching quality and upstream compute scarcity first show up, often after the work is done. That creates "token anxiety": Fixed software budgets become variable consumption pools, and spend can rise nonlinearly as agents reload context, summon subagents, run evaluations and retry failed actions.

It also creates behavioral distortions. "Tokenmaxxing" rewards visible AI usage even when token burn is not linked to output. "Localmaxxing" is the opposite response: routing routine work to local, edge or open-weight models to avoid cloud API costs and latency. Both patterns show why raw usage is a poor proxy for productivity.

Open-source and open-weight models are intensifying this pressure. They have become a primary competitive lever, giving developers and enterprises a way to build specialized applications at a fraction of the cost of proprietary systems. They are also changing global competition: Chinese models recently accounted for 17.1% of downloads on a leading global AI model repository, surpassing the US share for the first time. That is not just an adoption statistic. It shows how open-weight releases can commoditize parts of the foundation model layer, attract developer talent and create ecosystem lock-in even when access to advanced hardware is constrained.

That does not make the model layer irrelevant. It changes where advantage accumulates. When capable models become cheaper, more open and more interchangeable, differentiation shifts toward proprietary data, workflow context, model routing, prompt caching, quantization, observability and cost per resolved task.

The operating discipline comes down to token strategy — a set of practical rules for keeping AI spend in check. Route premium models to high-value work, and lean on cheaper, open or local models for commodity inference. Compress context, cache prompts and cap runaway loops. And measure cost per resolved task, not cost per query. That pushes enterprises toward agentic FinOps — spend ceilings, call-volume caps, chargebacks, kill switches and a named owner for agent economics. It also pushes vendors away from pure token billing toward value meters such as resolved conversations, qualified leads or completed workflows.

Why the AI stack is becoming more open and more complex

At the start of 2026, debates about agentic interoperability still leaned heavily on the assumption that incumbents would protect their walled gardens. That assumption has not survived 2026. Large enterprise platforms are exposing core systems through APIs and open agent protocols, treating products less as destination applications and more as programmable substrates that third-party agents can compose.

Agentic commerce is the clearest example. In our global survey, 15% of technology executives said agentic commerce is already live in production, 27% have it in pilot and 44% are actively evaluating it. The next customer journey may begin with an agent discovering, evaluating, negotiating and transacting through machine-readable interfaces rather than with a human search query or website visit.

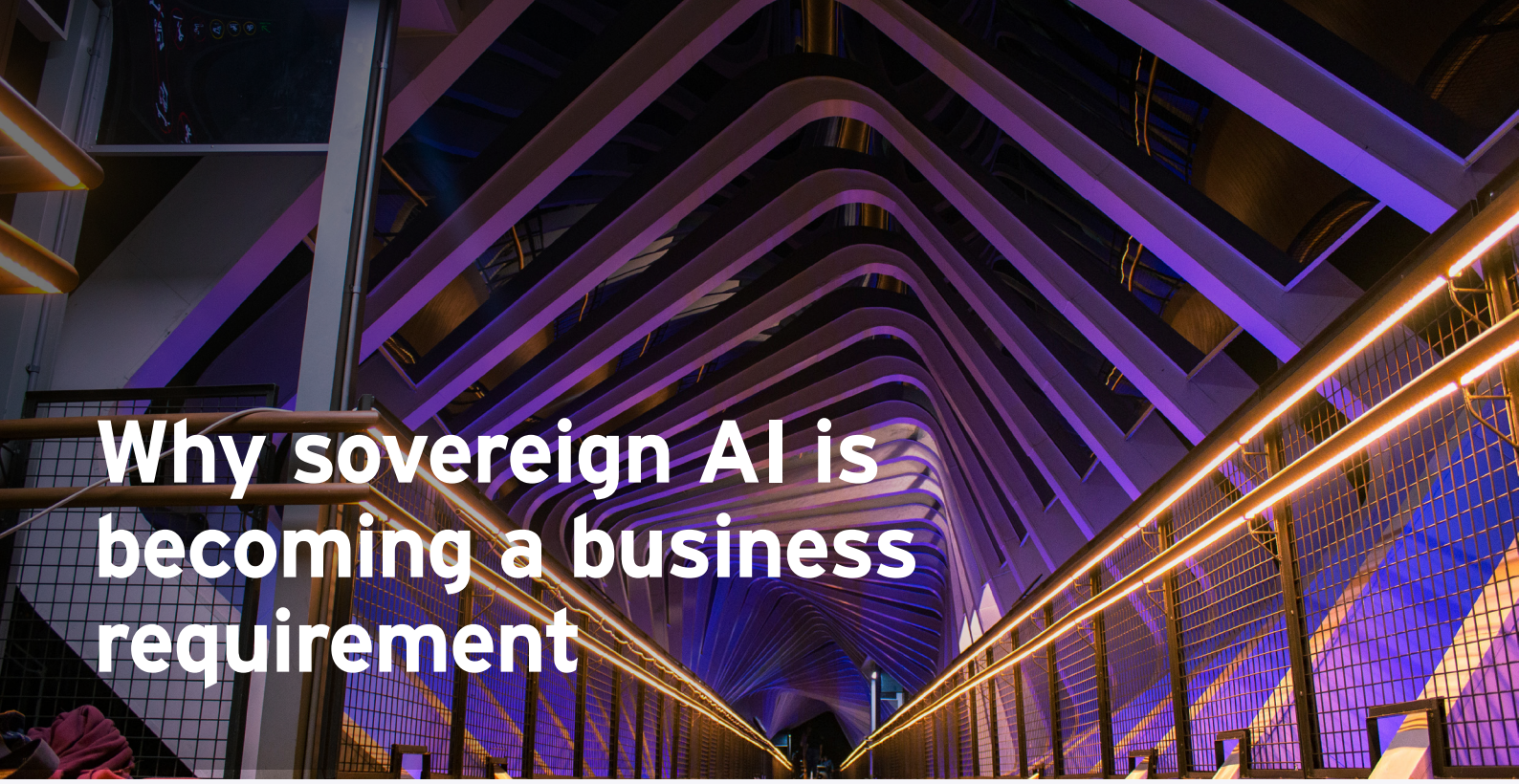
The same dynamic is reshaping the compute layer. AI infrastructure is no longer a single-stack GPU procurement story. Training, inference, agentic reasoning, retrieval, video, robotics and edge deployment stress different combinations of GPUs, custom ASICs, CPUs, memory, networking, cooling and software compilers. Inference economics are forcing buyers to optimize for latency, utilization, power and cost per token, while hyperscalers push vertically optimized stacks around custom silicon.

Open models are an important part of that heterogeneity. Enterprises are no longer choosing only between one proprietary model and another. They are building portfolios of closed, open, local, edge and specialized models, then routing work across them based on cost, sensitivity, latency, accuracy and control. That creates more flexibility, but it also makes architecture more complex.

The strategic question becomes less "which model is best?" and more "which model is best for this task, under these constraints, at this cost?"

The unresolved control point is the agent orchestration layer — the harnesses that coordinate models, tools, memory, permissions, evaluations and actions. Open protocols point toward interoperability, but economic value may still pool inside optimized stacks where a vendor controls the model, data, workflow graph, identity layer, observability and billing meter. Token strategy reinforces this: The winning harness may be the one that reduces wasted inference through prompt caching, context compression, model routing and permission-aware tool use, not the one with the cleanest developer abstraction.

The same systems-design pressure is now extending into physical AI and space-based compute. Physical AI is moving from research toward deployment as foundation models are paired with sensors, motor control and embodied reasoning. But the hardest near-term constraint is energy: power availability, not chips or capital, is emerging as the binding limit on AI buildout, as grid interconnection queues, generation capacity and transmission struggle to keep pace with data center demand. Space-based compute remains earlier, but it is no longer a purely speculative metaphor. The appeal is easy to understand: Power, water, permitting and land constraints are becoming harder terrestrial bottlenecks. The barriers are just as real: thermal management in vacuum, radiation hardening, launch economics, satellite servicing and orbital debris risk. The strategic signal is not that the cloud is moving off-planet; it is that AI scale is forcing companies to rethink not only what compute they buy, but where compute physically lives and how it is powered.



Why sovereign AI is becoming a business requirement

"AI sovereignty" started 2026 as a political buzzword and is now a concrete procurement and infrastructure reality. In our survey, 65% of respondents report strict sovereign AI requirements across most workloads or requirements for select geographies or functions, and more than 92% either have sovereign AI requirements or are actively evaluating them.

Countries and regions are building native AI companies around a wide set of unique enablers – local languages, cultural context, regulated data sets, procurement regimes and industrial priorities. The key opportunity may be less one national champion model per country and more a dense layer of smaller labs, applied AI vendors and integrators that make the local ecosystem self-reinforcing. Full domestic self-sufficiency in frontier AI is unrealistic for most countries. The concentration of compute, capital, chips and engineering talent means many governments are moving toward strategic interdependence: domestic control over critical systems, data and procurement combined with managed reliance on foreign providers.

Geopolitical risk has caught up with the infrastructure reality. For technology CEOs, geopolitical tensions, instability and conflicts ranked as the top business risk over the next 12 months, cited by 51% of survey respondents. The pattern emerging is what some are calling "sovereign vaults" – localized model, data and cloud environments housed within national or regional digital infrastructure. For companies operating across borders, jurisdictional flexibility is no longer a hedge; it is a baseline requirement for staying eligible to do business in fragmented markets.

AI security is no longer just a cyber problem

The security threat surface in 2026 has evolved in two directions at once.

On the physical layer, attacks on data center infrastructure in the Gulf made explicit that AI compute is now treated as kinetic critical infrastructure by nation-states. The implications go beyond resilience planning: They reshape siting, redundancy, insurance, energy security and the political risk profile of physical infrastructure. The CEO data reinforces the point: 39% of technology CEOs agreed or strongly agreed that their business would struggle to absorb prolonged and significant disruptions across critical third-party providers, infrastructure or global operating networks, while 48% said sustained energy price shocks would create significant headwinds.

On the digital layer, autonomous agents with write access to firewalls, endpoint quarantine systems, customer workflows, financial systems and other privileged operations introduce a new class of identity and privilege-abuse risk. Machine identities already outnumber human ones, and many carry sensitive permissions. Yet only 25% of technology executives surveyed say they have enterprise-wide AI governance and ethical oversight in place for agentic AI, 34% have monitoring and control mechanisms, and one-third have ensured high-quality, accessible and governed data.

Agents do not merely generate recommendations; they take actions across systems, APIs and workflows. Anthropic's Project Glasswing and restricted Claude Mythos Preview make the issue tangible: The same model capabilities that improve coding and defensive cybersecurity can also accelerate vulnerability discovery, exploit simulation and autonomous security operations if controls are weak.

Quantum computing adds another complicating factor. The near-term issue is not that quantum systems will immediately displace classical infrastructure; it is that post-quantum cryptographic migration is a multiyear systems program that must begin before fault-tolerant quantum machines arrive.

The required posture is no longer perimeter-and-permission. It is behavior-based identity management, infrastructure resilience and cryptographic agility operating on machine-time scales.

Why AI partnerships and infrastructure deals are accelerating

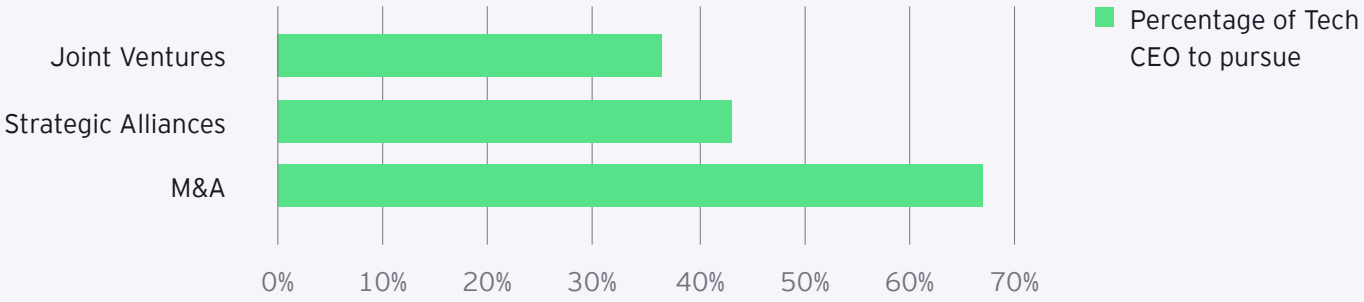
The first half of 2026 has been defined by historically large and structurally complex AI transactions. Hyperscalers, foundation model builders, chipmakers, neoclouds and sovereign capital providers are locking in compute access, distribution, model rights and infrastructure capacity through deals that look less like classic M&A and more like industrial policy. The most important transaction may be a long-term GPU lease, cloud commitment, sovereign joint venture, model-access agreement or data partnership rather than the acquisition of a software company.
Among companies planning M&A, 90% expect appetite to increase versus 2025 and 41% say enhancing technology or AI capabilities is one of the most important factors in acquisition or divestment decision-making. At the same time, 65% of technology executives say they have reassessed or are reviewing AI investment strategy because of debt-funded AI infrastructure or long-dated lease

commitments, while more than a quarter are tightening gating or ROI criteria.

This is where the AI capital-intensity gap becomes visible. AI-native companies are scaling investment faster than AI-enabled incumbents and the capability stack is becoming too broad to build organically.

Model access, inference optimization, proprietary data, agent orchestration, security, edge deployment, physical AI, sovereign compliance and specialized infrastructure all require different partners and assets. The build-buy-partner question is no longer episodic; it is part of operating model design.

Tech CEOs appetite for transactions in the next 12 months



For companies below hyperscaler scale, the strategic lesson is that the window for purely organic AI capability buildout has narrowed. Capability is being assembled through infrastructure commitments, model access agreements, cloud partnerships, targeted acquisitions, government-backed funding structures and specialized ecosystems. Structured alliances, joint ventures and integration-capability acquisitions are becoming the default operating mode.

What this means for technology leaders

The throughline across all five shifts is the same:
The agentic era is not a feature release. It's an enterprise-level refresh.

Pricing models, technology architectures, geopolitical assumptions, security postures and corporate development strategies are all being rewritten simultaneously and the rewrite now reaches beneath software into power availability, cooling, chip manufacturing, fiber networks, cryptography and the physical geography of compute.

The first six months of 2026 have sharpened the pattern. Subsidies are giving way to value capture. Tokens have become the operating meter — and increasingly the allocation unit — for agentic AI, but the strategic issue is not token minimization; it is disciplined allocation of agent capacity against measurable value. Compute is becoming more heterogeneous, coopetition is becoming structural, the agent harness layer remains unresolved, open-source and open-weight models are changing the economics of developer adoption and sovereign AI is moving from policy rhetoric to funded local ecosystems.

Those that treat this moment as a full-system redesign will be better positioned to convert AI scale into durable advantage — with faster execution, greater resilience and new room to create value.

Authors:

Ken Englund
EY Americas Technology Sector Growth Leader

James Brundage
EY Global and Americas Technology Sector Leader

Andrew Young
EY Global Technology Industry Lead Analyst

BMC Agency

GA 17347660

SCORE number: SCORE 113153-26-GBL
ED None