



Building Enterprise AI for Choice

Build an AI architecture that can evolve as models, technologies, and business priorities change.

AMD 

together we advance_



TABLE OF CONTENTS

3	Enterprise AI Is an Architectural Journey
4	The Decisions That Define Enterprise AI
6	Futureproofing Your AI Architecture
7	How Open Architecture Preserves Flexibility
8	How AMD Brings the Ecosystem Together
9	Operationalizing Enterprise AI
11	The Open Path Forward





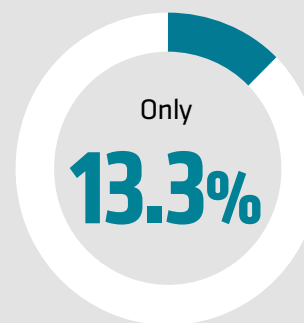
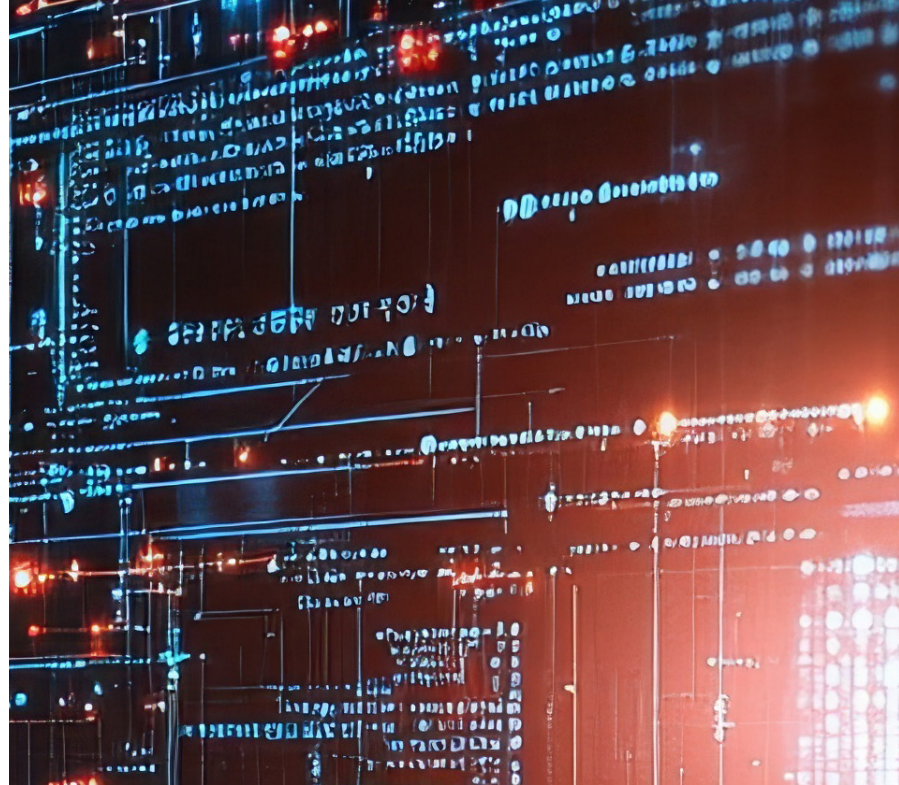
Enterprise IT leaders are being asked to make long-term technology decisions in one of the fastest-moving technology markets we've ever seen.

Enterprise AI Is an Architectural Journey

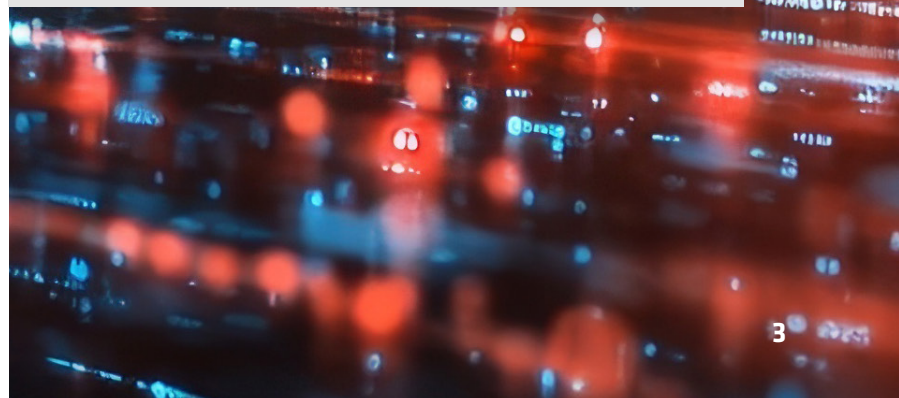
AI models continue to evolve, regulations are still taking shape, and new workloads are emerging as quickly as organizations can deploy them. Yet business leaders are under growing pressure to move beyond experimentation and make AI a lasting source of competitive advantage.

Most organizations aren't there yet.

According to The Futurum Group's 1H2026 AI Platforms Decision Maker Survey, only 13.3% of enterprises have reached the highest level of AI maturity, where AI is embedded in the business and drives how the organization creates value. Most remain in the middle stages of experimentation and optimization, working to bridge the gap between promising pilots and production deployments.¹



of enterprises have reached the highest level of AI maturity



That gap exists because production AI is fundamentally different from experimentation.

Early AI initiatives were often limited to chat-based interactions or narrowly defined use cases. Today's enterprise AI deployments increasingly rely on agentic systems that reason, plan, invoke tools, and operate continuously under governance. A single business process may involve dozens of model interactions, making scalability, reliability, security, and operational management just as important as model performance. Infrastructure designed for a chatbot pilot may struggle to support production-scale AI.

Technology is just one of the big challenges.

Forty-three percent of IT leaders cite unclear business value as one of the biggest barriers to moving AI pilots into production. Reliability, data privacy, team skills, and governance create additional friction.² For both IT and line-of-business leaders, the takeaway is straightforward: successful AI initiatives require clear business ownership, measurable outcomes, and an architecture designed to scale.

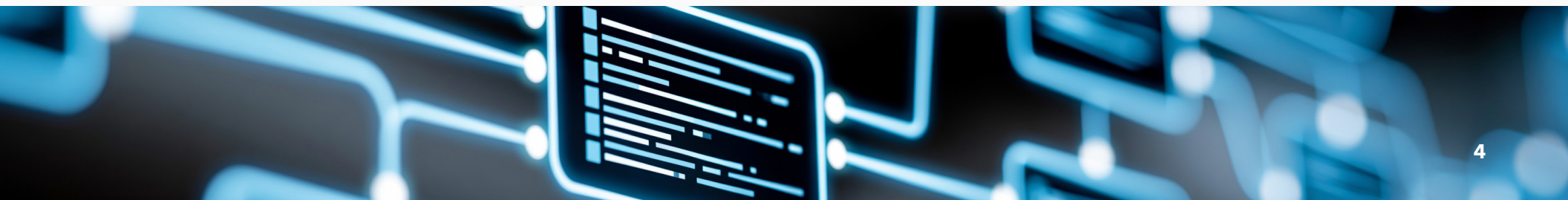
As organizations move from experimentation to production, the conversation changes. The early question was, "Which model should we use?" Today, the more important question is, "How do we build an AI architecture that can evolve?" With 78% of enterprises expecting AI budgets to increase over the next year, today's architectural decisions will influence tomorrow's flexibility, costs, and ability to adapt and scale for years to come.³

The Decisions That Define Enterprise AI

Every enterprise AI strategy is unique, but the architectural decisions behind it are remarkably consistent. Organizations may arrive at different answers based on their business priorities, existing investments, and regulatory requirements. The objective is to make deliberate choices that preserve flexibility as technologies, workloads, and business needs evolve.

Compute Foundation

Every AI workload has different performance, latency, and cost requirements. Modern enterprise AI environments rarely rely on a single type of compute. Retrieval, orchestration, routing, and many inference workloads run efficiently on today's server CPUs, while training and large-model inference benefit from GPUs. The most effective architectures match each workload to the appropriate compute tier, balancing performance with cost and operational efficiency instead of assuming every workload requires the same infrastructure.



Infrastructure Software

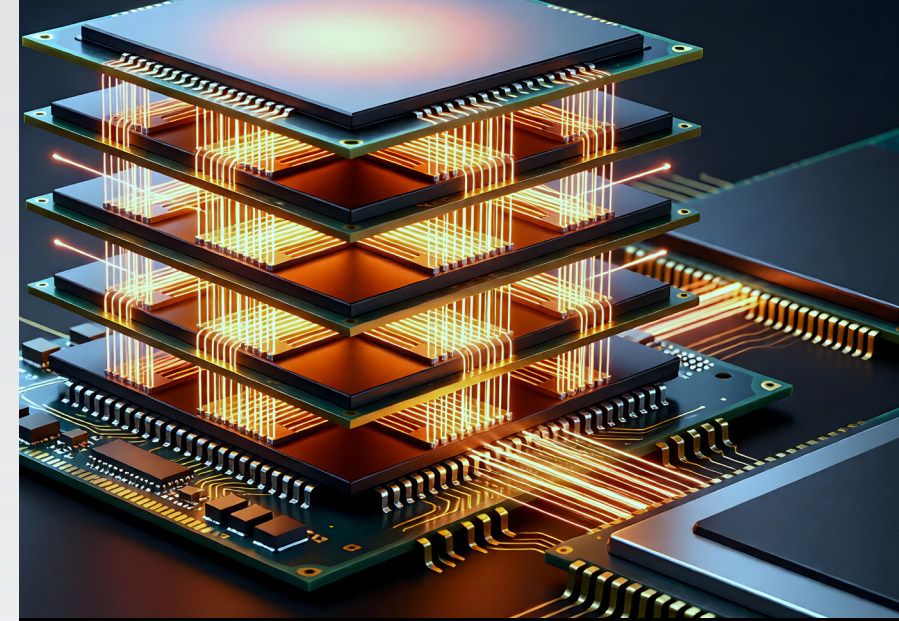
The infrastructure software layer, which includes virtualization, Kubernetes, AI platforms, operating systems, and development frameworks, determines how portable and adaptable AI environments remain over time. Flexibility matters because enterprises already operate in a multi-model world. Futurum research found OpenAI (57%), Azure OpenAI (56%), and Google Gemini (48%) are all widely deployed across organizations.⁴ Rather than standardizing on a single model, enterprises increasingly need infrastructure that supports multiple models and evolving AI frameworks.

Data, Governance, Security, and Sovereignty

Data strategy, governance, security, and regulatory requirements should shape architectural decisions from the beginning. Data residency, access controls, auditability, privacy, and compliance requirements all influence where AI workloads run and how they are managed. Addressing these considerations early helps organizations avoid costly redesigns as AI initiatives scale.

Deployment Model

Enterprise AI rarely lives in a single environment. Organizations deploy AI across a combination of on-premises infrastructure, private cloud, public cloud, edge environments, and, in some cases, sovereign or fully disconnected environments. Futurum research reflects this reality: 64% of enterprises run GenAI on provider-managed cloud platforms, 40% use private cloud or virtual private cloud (VPC) environments, and 43% embed GenAI within SaaS applications.⁵ The challenge is building an operating model that allows these environments to work together.



Questions IT Leaders Should Ask

- Which architectural decisions are we making intentionally and which are simply the result of existing technology investments?
- If our AI strategy changes over the next two years, where would we need to rebuild instead of adapting?
- Which parts of our architecture give us flexibility, and which create unnecessary lock-in?
- Are we designing for today's AI workloads or for the workloads we expect to run three years from now?

Futureproofing Your AI Architecture

Enterprise AI isn't standing still, and neither are the technologies that support it. Models evolve rapidly. Pricing changes. Regulations emerge. New workloads appear almost overnight. In this environment, one of the most valuable architectural characteristics is flexibility: the ability to adapt without rebuilding.

That's what optionality provides. Rather than locking organizations into a single technology path, optionality preserves the freedom to adopt new models, infrastructure, and deployment approaches as business requirements evolve.

Lock-in has always carried costs, but those costs increase as the pace of innovation accelerates. Decisions that optimize for today's requirements can make tomorrow's choices more expensive. A closed technology stack may limit the models an organization can adopt, the infrastructure it can deploy, the partners it can integrate, or the speed with which it can respond to changing business needs.

Enterprise buyers are already responding to this reality. In The Futurum Group 1H 2025 Chipsets and Accelerators Decision-Maker Survey, 69% of organizations said they planned to change or add AI server vendors within the next year.⁶ Six months later, budget and capital expenditure constraints had become the leading barrier to scaling AI infrastructure.⁷ Together, these findings reinforce an important point: organizations need architectures that make change manageable rather than prohibitively expensive.

Emerging Best Practice: Keep Context, Tools, and Skills Under IT Control

As organizations adopt agentic AI, a new form of lock-in is beginning to emerge. Models can be replaced relatively easily. The harder challenge is preserving the context, tool integrations, workflows, and agent skills that define how AI operates inside the business.

When those assets become tightly coupled to a specific model or provider, organizations risk recreating the same lock-in they were trying to avoid. Defining these assets in open, portable formats gives IT the flexibility to change models without rebuilding the operational knowledge that makes AI valuable. The key is to make decisions that leave room for better decisions later.

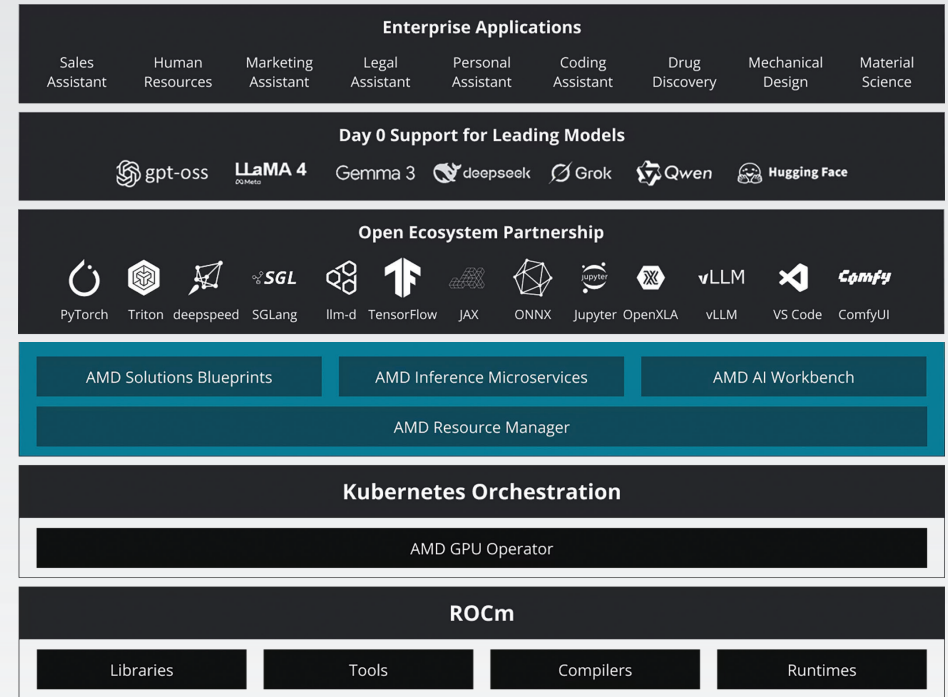
Questions IT Leaders Should Ask

- If we wanted to change models next year, what would actually need to change?
- Which parts of our AI architecture are portable, and which depend on a single vendor?
- Are we building reusable AI capabilities, or are we tying business knowledge to a single platform?
- How would a major change in regulation or pricing affect our current architecture?

How Open Architecture Preserves Flexibility

Flexibility must be designed into the architecture from the beginning. That's where open infrastructure and an open ecosystem come together. Open standards, interoperable software, and broad partner support allow organizations to adopt new technologies without rebuilding the layers around them. When every layer communicates through open interfaces, enterprises can evolve one part of the architecture without disrupting the rest.

Enterprise AI has become a team sport. According to Futurum research, 37% of enterprise AI procurement flows through channel and ecosystem partners, and buyers rank AI expertise as their most important vendor selection criterion.^{8, 3} No single technology provider can deliver every capability an organization needs. An open ecosystem allows enterprises to combine the expertise of multiple partners while preserving the flexibility to evolve as technologies and business priorities change.



EXPERTISE WITH AI IS THE
#1 CRITERION
ORGANIZATIONS USE WHEN
SELECTING AI VENDORS.³

Partner Spotlight: Bringing Open AI to Enterprise Software Development

AI-assisted coding can improve developer productivity, but agentic coding also drives rapid growth in token consumption, cloud costs, and governance demands. Many organizations also need proprietary code, repository context, and sensitive data to remain inside their own environments.

The AMD Instinct™ Coder Solution brings together AMD, Spectro Cloud, and Supermicro in a validated enterprise solution, optimized and pre-configured for rapid time-to-value and fully private for tight control of cost, data, and operations. It ships as a single server that can be stood up in days rather than the months a do-it-yourself build takes, and it does not require a specialist AI infrastructure team to run: the platform handles the GPU enablement, drivers, and day-2 operations that would otherwise need scarce in-house expertise.

AMD provides the compute foundation, Spectro Cloud manages the AI platform and intelligent model routing, and Supermicro provides the validated server platform. Everyday coding requests can run locally, while requests that require a hosted frontier model are routed through a cloud API.

The solution is designed to deliver up to 95% accuracy parity with a leading hosted proprietary frontier model on coding tasks. A single server can support up to 50 users, 30 concurrently. From a CFO's perspective, the illustrative model is compelling. For a team of 50 developers whose cloud inference costs roughly \$1.2M a year, moving everyday work

on-premises with a 10% hosted-frontier fallback can cut the annual cost by around 70%, on the order of \$830K a year, providing payback in approximately 6 months.*

How AMD Brings the Ecosystem Together

AMD provides the open infrastructure foundation that allows enterprises to build AI architectures that can evolve. Rather than prescribing a single technology stack, AMD works with infrastructure software providers, AI platforms, and application partners to give organizations validated paths to deployment while preserving architectural flexibility.

The AMD portfolio spans the compute requirements of modern enterprise AI. AMD EPYC™ Server CPUs efficiently support orchestration, retrieval, routing, and many inference workloads, while AMD Instinct™ GPUs accelerate training and large-model inference. Together with the open-source AMD ROCm™ software stack, they provide an open foundation that supports today's AI frameworks and evolving workloads.

To shorten the path from planning to production, AMD provides two solutions that address different parts of the problem: AMD Solution Blueprints and AMD Inference Microservices (AIMs).



Operationalizing Enterprise AI

Building the right AI architecture is only part of the journey. Long-term success depends on operational discipline: matching infrastructure to workloads, embedding governance from the beginning, and creating processes that allow AI to scale without sacrificing cost control, security, or flexibility.

Match the workload to the infrastructure

Production AI introduces a new operational challenge: every workload has different performance, latency, security, and compliance requirements. Retrieval, orchestration, tool calls, inference, and reasoning don't all require the same compute resources.

Successful organizations match each workload to the lowest-cost compute tier that still meets business, security, and service-level requirements. Intelligent routing automates those decisions, directing each request to the most appropriate infrastructure. Over time, that discipline improves both scalability and economics.

Operational discipline also extends beyond infrastructure. Governance, security, access controls, model provenance, and auditability become increasingly important as AI moves into production. Organizations that establish these capabilities early avoid costly redesigns later as AI initiatives expand across the business.

AMD Solution Blueprints and AMD Inference Microservices (AIMs)

- AMD Solution Blueprints: ready-to-deploy templates that pair validated AI models with reference code, so a team starts from a working pattern instead of an empty environment. The value is time. The integration work that normally sits between a decision and a running deployment is already done.
- AMD Inference Microservices (AIMs): optimized AI models behind industry-standard APIs. The value is optionality made concrete. Because the interface stays constant when the model behind it changes, the applications built on top do not have to be rewritten each time a better model arrives.

AMD Inference Microservices (AIMs) illustrate how AMD approaches optionality in practice. By exposing optimized AI models through OpenAI-compatible APIs, AIMs allow organizations to change or upgrade models without rewriting the applications built around them. That separation helps preserve flexibility as AI models continue to evolve.

Practical Guidance for Scaling Enterprise AI



Start with measurable business outcomes. Enterprise AI investments are increasingly tied to business results rather than experimentation alone.



Keep critical AI assets portable. Maintaining control over context, tools, and agent skills preserves flexibility as models evolve.



Design for continuous change. AI architectures should accommodate future workloads without requiring wholesale redesign.

Customer Spotlight: AT&T Builds AI at Enterprise Scale

AT&T demonstrates what enterprise AI looks like after organizations move beyond pilots and into production. The company processes more than one trillion AI tokens per month and operates more than 100 generative AI models in production across customer care, fraud detection, network planning, and other business-critical workloads. Approximately 40% of its specialized AI models run on AMD, illustrating a heterogeneous infrastructure strategy built for flexibility rather than dependence on a single technology stack.

Speaking at AMD Advancing AI in July 2026, AT&T Chief Technology Officer Jeremy Legg emphasized that enterprises should avoid locking themselves into a single chipset, model, or toolset. Instead, organizations should build on open, flexible infrastructure that allows AI platforms, models, and workloads to evolve as business needs change.

AT&T offers OTel 2.0, an advanced open-source telecom AI model co-developed with AMD and Microsoft. The company is also applying agentic AI to transform internal operations, using AI agents to automate complex business processes such as employee onboarding. Together, these initiatives demonstrate that enterprise AI success depends on more than selecting the right model. It requires an open architecture that can scale, integrate new technologies, and adapt as AI continues to evolve.

Common Scaling Mistakes

- Treating every workload as though it requires the same compute resources.
- Waiting until after deployment to establish governance and security.
- Allowing an early proof of concept to become the long-term architecture.

Questions IT Leaders Should Ask

- Can our existing IT team operate this environment?
- Which workloads truly require accelerated compute?
- If we changed models next year, what parts of our architecture would remain unchanged?
- Where are we creating unnecessary complexity?
- Are we measuring business outcomes as rigorously as technical performance?

The Open Path Forward

Enterprise AI is both an architectural and ecosystem journey. Organizations that preserve flexibility from infrastructure to applications are better positioned to adopt new models, technologies, and deployment approaches as the market evolves. That's why AMD invests in an open ecosystem spanning infrastructure, operating systems, virtualization, Kubernetes, AI platforms, and application partners: to give every organization the freedom to choose the path that best fits its business.

Start Where You Are

Still experimenting? Choose an architecture that can support production without requiring a major redesign.

Scaling successful pilots? Align compute resources with workload requirements and establish governance before expanding deployment.

Optimizing production? Preserve optionality so you can adopt emerging AI capabilities such as agentic AI without rebuilding your environment.

Enterprise AI will continue to evolve, and no organization can predict every change ahead. The organizations that succeed will build architectures capable of adapting as technologies, business priorities, and opportunities evolve. Open architectures, open ecosystems, and thoughtful operational practices provide that foundation.

Continue the Conversation

Explore the AMD and partner enterprise AI solution briefs to see how this approach is being implemented across diverse use cases or connect with an AMD enterprise AI architect to discuss the path that best fits your organization's goals.

About AMD

More than 50 years driving innovation in high-performance computing. AMD Instinct accelerators power AI workloads for leading cloud providers, enterprises, and sovereign AI initiatives worldwide. The AMD commitment to open standards and an open ecosystem ensures enterprises can deploy AI with control, flexibility, and long-term vendor independence.

*All performance and/or cost savings claims are provided by Spectro Cloud and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein may not be typical. GD-181a.

CLAIMS:

- 1 The Futurum Group, March 18, 2026, 1H 2026 AI Platforms Decision Maker Survey (n=820), AI maturity press release, <https://futurumgroup.com/press-release/organizations-with-a-chief-ai-officer-are-more-likely-to-reach-ai-maturity/>
- 2 The Futurum Group, July 4, 2026, "Is AI Ready for Real Work, or Are Enterprises Still Stuck in Experimentation?", <https://futurumgroup.com/insights/is-ai-ready-for-real-work-or-are-enterprises-still-stuck-in-experimentation/>
- 3 The Futurum Group, June 17, 2026, "Databricks' 2026 Partner Awards Reveal the Power Struggle for AI Services Dominance," <https://futurumgroup.com/insights/databricks-2026-partner-awards-reveal-the-power-struggle-for-ai-services-dominance/>
- 4 The Futurum Group, April 25, 2026, "GPT-5.5 Raises the Stakes: Can OpenAI Maintain Its Lead as Enterprise AI Matures?", <https://futurumgroup.com/insights/gpt-55-raises-the-stakes-can-openai-maintain-its-lead-as-enterprise-ai-matures/>
- 5 The Futurum Group, July 3, 2026, "Can DataRobot's Unified AI Governance Break the Silo Trap for Enterprise AI?", <https://futurumgroup.com/insights/can-datarobots-unified-ai-governance-break-the-silo-trap-for-enterprise-ai/>
- 6 The Futurum Group, April 5, 2025, 1H 2025 AI Chipsets and Accelerators for Data Centers Decision Maker Survey, <https://futurumgroup.com/press-release/decision-maker-survey-sharp-increase-in-ai-compute-services-spending-for-2025/>
- 7 The Futurum Group, May 18, 2026, 1H 2026 Data Center Semiconductor Decision Maker Survey (n=824), <https://futurumgroup.com/press-release/52-of-compute-decision-makers-use-reasoning-models-as-primary-ai-model-type/>
- 8 The Futurum Group press release, April 17, 2026, "The Orchestration Era: Why Your GSI Program Is Already Behind", <https://futurumgroup.com/press-release/the-orchestration-era-why-your-gsi-program-is-already-behind/>

©2026 Advanced Micro Devices, Inc. all rights reserved. AMD, the AMD arrow, and combinations thereof, are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

September 2026

PID: 265253815