MAKE YOUR COMPANY AN AI COMPANY

Breaking Through AI Roadblocks

Overcoming 6 of the Most Common AI Business Challenges in 2025

Complimentary research from

Gartner®

The potential for AI to transform how we work is undeniable.

From smarter lead scoring to content generation to expense forecasting, businesses are using AI to save time and money, expand their offerings, and reach new markets. According to Gartner, AI is the number one technology that CEOs believe will most significantly impact their industry in the next three years.[1]

Yet, as excitement grows, businesses are facing serious roadblocks when it comes to implementing AI. AI is complex, and navigating this landscape isn't easy. Companies must overcome challenges related to cost, data privacy, and integration. Leaders must be able to prove their AI projects have real business value and can scale beyond a pilot. The following six challenges are among the most common to businesses implementing AI today.

> "Successful value creation with AI requires more than technology. A lack of the right investments in data, change management, AI literacy, risk mitigation, trust and governance represents a significant obstacle to AI success and value realization."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

6 of the Most Common AI Business Challenges in 2025

1

Businesses aren't sure which AI projects are worth the investment.

According to Gartner, despite the outlandish AI hype, turning the promise of AI into reality is not a given: 49% of leaders highly involved in AI report that their organizations struggle to estimate and demonstrate the value of AI.[1] So, how can they know which projects will pay off?

It's critical to start with the desired outcome—for example, are you looking to boost productivity or reduce costs? You'll also want to research which AI use cases are popular in your industry and identify the stakeholders who will help determine what success looks like.

2

Decision makers must justify the cost of AI investments and avoid overspending.

Too many businesses are surprised by the unexpected costs that often come with AI projects. According to Gartner, through 2028, at least 50% of GenAI projects will overrun their budgeted costs due to poor architectural choices and lack of operational know-how.[2]

Your leadership and IT decision makers will need to work together to decide which AI infrastructure—whether on prem or in the cloud—will bring you the highest return. Notably, you may be able to get more from your investments by leveraging existing hardware and cloud platforms, software, and data.

3

Data security and privacy remain a major concern.

Among leaders whose organizations have implemented AI, 3 in 10 report their organizations had an AI privacy breach or security incident, of which over one third were malicious attacks, reports Gartner.[3]

When launching new AI projects, security is paramount. You'll need to ensure robust risk mitigation and governance policies to help keep data protected. Silicon-enabled security also plays a role. Your team will want to consider security features built into PCs, servers, and cloud services.

> "There is a risk that an organization creates a roadmap for its AI operating model that is completely disconnected from the AI portfolio, creating a high probability that business value won't be achieved. A balance is required between tangible value that business stakeholders can envision (via a portfolio of use cases) and the foundational capabilities required to deliver this portfolio."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

4

Even for businesses that are committed to AI use cases, getting enough compute performance is not always a clear path.

It almost goes without saying—if you want to use AI, you need performance. But exactly how much performance is enough? Not everyone will need a rack full of GPUs.

Gartner emphasizes that many AI workloads—including enhancements for communications and personal assistance—will move to the client from the cloud due to cost, efficiency, privacy concerns, and functionality.[4] As a result, IT operations teams must decide which PCs are best for running AI locally while also delivering a comprehensive suite of business-caliber manageability and security features.

In the data center, architects must determine how to add enough compute for AI where power and space is already stretched to the limit. If running AI in the cloud, you'll want to determine the best instances to maximize performance per VM so that you don't overpay.

5

Hardware and off-the-shelf models or software alone are not enough. Businesses must be able to combine their data with new AI software.

Having powerful compute is one thing. But it's software that makes AI work. And it's software plus your data that makes AI work for your business.

As you explore how to use the latest AI models, software, and agentic AI tools, you'll want to consider how your computing platform can better support the integration of your data and give you the most choice in software selection.

6

Finally, getting a prototype into production has its own challenges.

Gartner says that organizations that have successfully created value with AI have managed to go beyond the phase of experimentation and piloting. These companies have built an AI strategy that is well aligned with their business strategy.[1]

Early in planning, your team will need to consider how your AI applications will scale across your organization and accommodate different kinds of environments—cloud-based, on the PC, and possibly in power-constrained edge locations. Migrating out of CUDA can offer some flexibility, while building on x86 architecture specifically can help streamline your integration with existing workflows.

> "The enormous potential business value of AI is not going to materialize spontaneously. AI leaders should guide their organization toward an era in which AI is not only creating tangible business value, but goes beyond to become a critical competitive differentiator and industry disruptor."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

CHALLENGE 1

Determining the business value of AI

intel ai

When large language models (LLMs) broke onto the scene, business gurus were quick to position them as the ultimate way to do more with less.

But determining the right models, integrating the right data, and getting the right employees to adopt AI tools isn’t easy. After all, there’s no one-size-fits-all way to use AI.

To avoid getting distracted with shiny new offerings, start with the business outcome: What do you actually want to do with AI? Help your employees save time? Increase revenue from existing customers? Reach new markets? Respond more quickly to security threats? Here are a few realistic ways you can use AI today (hint: not all of them involve LLMs):

- **Use machine learning to score and prioritize sales leads**
- **Use computer vision to monitor product shelves in the store to ensure timely restocking**
- **Use LLMs to generate narratives for financial reports**
- **Use natural language processing to analyze social media conversations about your brand**
- **Use generative AI models to create marketing text and images**
- **Use LLM-powered chatbots to enhance automated customer service**

Some of the best use cases will come from others in your industry. Use real-life examples of what your peers (and competitors) are doing as a short-cut to realizing your own value with AI. Here are some examples:

- **Healthcare organizations** are using machine learning to predict disease progression and treatment plans, computer vision to analyze medical images, and natural language processing to automate information from medical records.
- **Financial companies** are using machine learning to detect and prevent fraud, deep learning to perform algorithmic trading, and generative models to simulate scenarios for risk management.
- **Retailers** are using machine learning to analyze customer behavior and optimize inventory management, computer vision to learn about product placement, and LLMs to deploy customer service chatbots.
- **Manufacturers** are using machine learning to predict equipment failures, deep learning to identify defects, and generative models to optimize processes.
- **Insurance companies** are using machine learning to assess risk profiles and natural language processing to extract data from claims.
- **Educators** are using machine learning to create personalized learning plans, natural language processing to automate grading, and generative models to provide virtual tutoring.
- **Legal firms** are using natural language processing to assist in research, generative models to simulate scenarios and predict case outcomes, and LLMs to generate and review legal documents.
- **Telecommunications providers** are using machine learning to optimize network resource allocation, predictive analytics to identify factors leading to customer churn, and generative AI to optimize network topologies.
- **Energy companies** are using machine learning to optimize energy distribution and predict consumption patterns and deep learning to predict equipment failures.
- **Media and entertainment companies** are using machine learning to recommend content and generative AI to create artwork, music, and other content.

Of course, before you can identify your goals and explore use cases, you need to bring the right stakeholders together. Gartner recommends including senior, C-level managers to weigh in on business priorities and opportunities. It's also wise to include architects, IT operations leaders, and the employees who will be using your AI applications.

> "Through 2025, 30% of generative AI projects will be abandoned after proof of concept (POC) due to poor data quality, inadequate risk controls, escalating costs or unclear business value."
>
> **Gartner**, "How to Calculate Business Value and Cost for Generative AI Use Cases"

Gartner: Identifying your strategic value priorities for AI

What are the levels of ambition with respect to applying AI?
Will AI be used mostly to improve existing business, or to extend or even disrupt business? How much funding should go to each of these ambition levels and in which business areas?

How do these priorities relate to business objectives?
For example, if the business objective is to cut costs, then AI use cases that impact costs take priority. If it is to improve customer engagement, then AI use cases that support customer engagement take priority. In other words, in which business areas are there the most important opportunities for AI to create real value?

In each of these areas, what business goals is the use of AI related to?
In which areas can AI be a catalyst for new or currently unaddressed business opportunities or challenges?

Which KPIs will be impacted for which stakeholders?
What are the key metrics for measuring the value that AI creates?

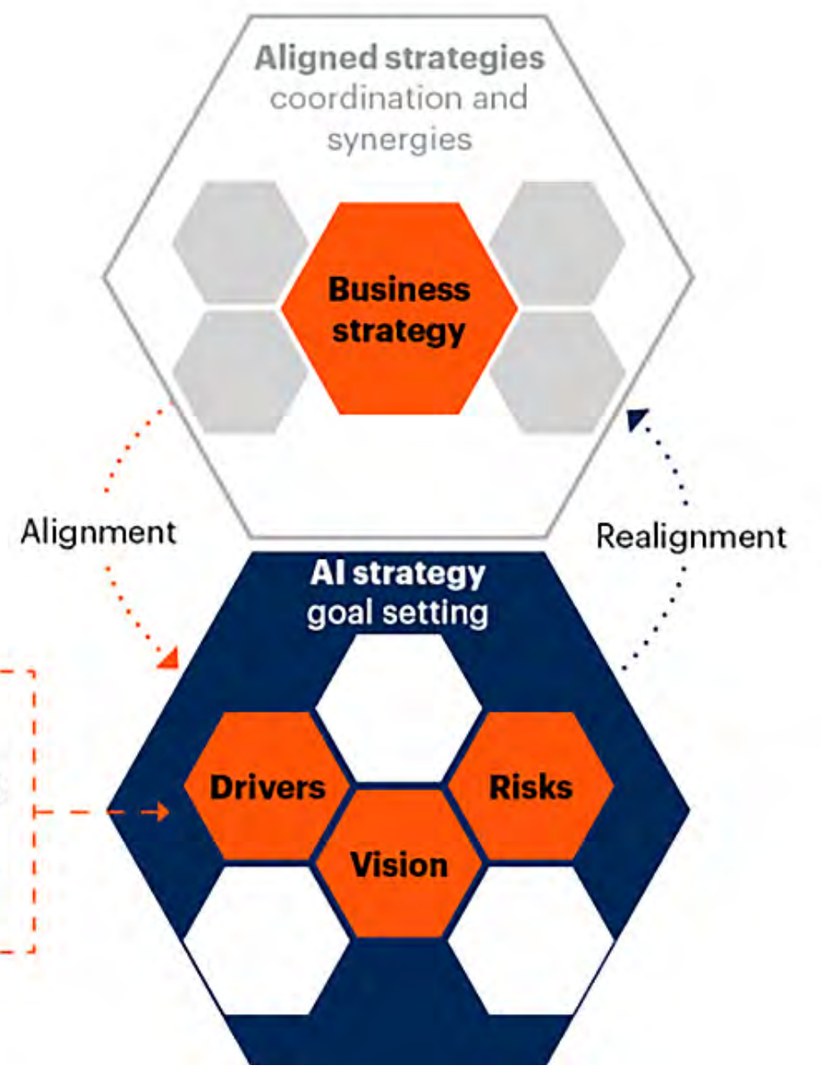
What are some practical examples of AI use cases?
Which business objectives and metrics are these examples related to? What is the art of the possible from within and outside the organization's sector or industry?

Gartner, "The Pillars of a Successful Artificial Intelligence Strategy"

Gartner

Core Elements of an AI Strategy

Source: Gartner
805859_C



"The foundation of an AI strategy is about vision, drivers and risks, focused on the essence of what an organization wants to achieve with AI, fully aligned with its business strategy.

Preferably, this emerges from (repeated) discussions that the AI leader has with relevant stakeholders. In the case of the AI strategy, this involves senior, C-level managers. Their business priorities, opportunities, challenges and concerns related to AI should find common ground in the AI strategy."

Gartner, "The Pillars of a Successful Artificial Intelligence Strategy"

The emerging value of generative AI

There's a reason businesses are excited about GenAI, which includes popular LLM services like ChatGPT, image generation tools like Midjourney and Stable Diffusion, and even audio and 3D model generators. GenAI has the potential to take hours' worth of human work to just minutes, and in many cases works simply by creating the right prompt.

"Many organizations have initiated or expanded their AI (including generative AI) activities, and even more have announced very significant new or further investments.

As for generative AI, a recent survey revealed that 18% of business leaders are piloting, implementing or have implemented it for their functions, while 47% will do so in the coming 12 months."

Gartner, "The Pillars of a Successful Artificial Intelligence Strategy"

Gartner: New business value with GenAI

Earlier adopters across industries and business processes are reporting a range of business improvements that vary by use case, job type and skill level of the worker. A large majority of business executives who are implementing or actively planning to implement GenAI have anticipated or realized benefits from their implementations, according to the Gartner Generative AI 2024 Planning survey of 822 business leaders.

On average, survey respondents report:

15.8% REVENUE INCREASE

15.2% COST SAVINGS, **4.6%** THROUGH REDUCTION IN HEADCOUNT

22.6% PRODUCTIVITY IMPROVEMENT

Worker productivity has broadly been found to improve when GenAI is employed:

- **ChatGPT has been shown to improve worker productivity by 37%.**
- **GenAI coding assistants can result in 7% to 55% worker productivity improvements.**

GenAI conversational assistants can improve customer service and support agents' productivity (studies show a range of 14% to 35% improvement).

Gartner, "How to Calculate Business Value and Cost for Generative AI Use Cases"

Gartner

Top Types of Business Value From Applying Generative AI

Multiple responses allowed

n = 78 CIOs, excluding "not sure"

Q: What are the top three types of business value your enterprise seeks from applying generative AI?

Source: 2024 Gartner CIO Generative AI Survey

818765_C

Gartner, "Top Strategic Technology Trends for 2025: Agentic AI"

Discover real-world success stories, explore cutting-edge AI tools, and find powerful solutions to enhance your AI capabilities:

- **Intel Customer Spotlight**
- **AI use cases and applications**
- **Intel® AI Software Catalogue**
- **Intel® AI Inference Software & Solutions Catalogue**
- **YouTube playlist: What Can You Do with an AI PC?**

CHALLENGE 2

Overcoming cost barriers

intel ai

> "Business shapes AI. AI shapes business. Playing it safe has never been riskier, as competitors or new entrants may overtake incumbents, leveraging the growing power of AI."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

Realizing business value with AI starts with smart investments and avoiding wasteful overspending. Initial costs can be steep, but ongoing expenses add up too. According to Gartner, through 2028, at least 50% of GenAI projects will overrun their budgeted costs due to poor architectural choices and lack of operational know-how.[2] To truly benefit from AI, businesses need strategic investments that balance expenses while allowing you to push the envelope of innovation.

A significant portion of AI-related costs stems from training, which involves teaching a model. However, most businesses will not need to develop models from the ground up. Instead, their primary AI workload will involve inference—the process of running a pretrained model. Fortunately, while training generally requires specialized AI accelerators, inference can often run on general-purpose processors. Gartner projects that through 2028, the aggregated costs of model inference will be at least 70% of the total model lifetime costs, eclipsing the training costs by a considerable margin.[2]

Adding inference to your data center or cloud

Even if you're not sure yet which AI applications your business will run, you can gain an advantage by being prepared for the next wave of AI use cases. While GPUs are often considered the standard for AI tasks, there are more cost-effective alternatives available, particularly for inference.

For on-prem data centers, build inference into your plan for regular infrastructure refreshes. The last few generations of general-purpose servers have come a long way in inference, with CPUs like the latest Intel® Xeon® processors including built-in acceleration for AI. If you expect to support some inference but won't have heavy usage, it's likely that your Xeon-based servers will be enough to easily handle inference alongside all your other enterprise workloads, saving you from having to purchase specialized hardware.

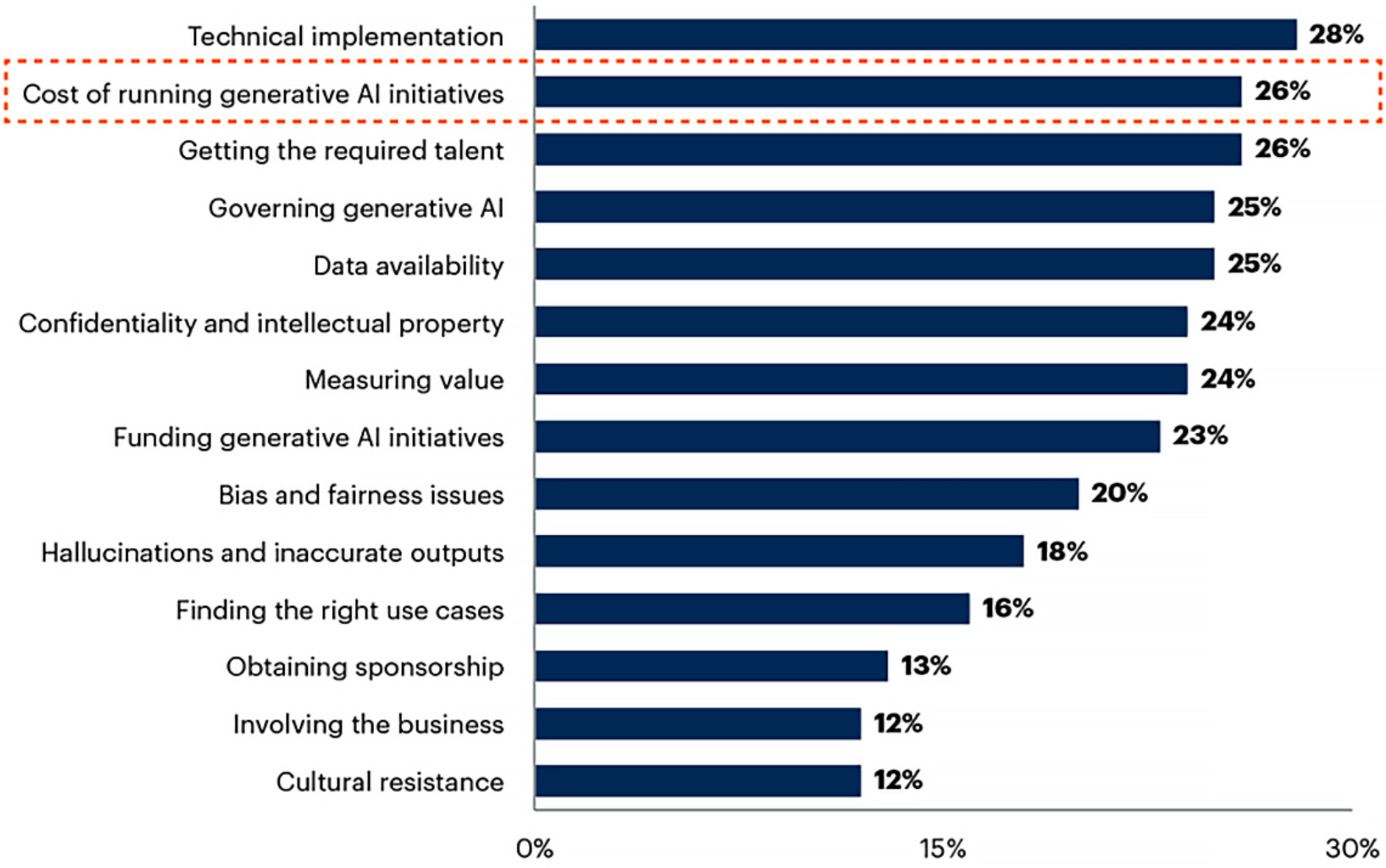
If you need more performance in the data center, start with Xeon as the host CPUs for your AI accelerated system. Beyond GPUs, you might consider Intel® Gaudi® AI accelerators. These are a cost-effective alternative for inference and training that are also designed to be power efficient.

In the cloud, VMs running on the latest Xeon CPUs give you enough performance for many inference workloads without the need for a specialized instance, which can help keep costs in check. Cloud services running on Gaudi accelerators offer extra performance at a competitive cost. You can select these instances through public cloud providers or the Intel® Tiber™ AI Cloud.

Gartner

Top Barriers to Implementing Generative AI Initiatives

Sums of top three ranks



n = 114, leaders highly involved in AI, whose orgs are advanced in generative AI adoption; excluding "unsure"

Q. What are the top 3 challenges that your organization has come across when implementing Generative AI initiatives?

Source: 2023 Gartner AI in the Enterprise Survey

810685_C

Gartner, "10 Best Practices for Optimizing Generative AI Costs"

Gartner: Understanding the trade-offs of self-hosting

Self-hosting GenAI models (often on-premises) can seem attractive for businesses seeking increased control and data privacy. It is also true that model inferencing will be more hybrid in the future, driven by costs, performance and privacy needs.

However, it's crucial to be aware of the potential trade-offs, as the list of cost drivers for self-hosting is extensive.

Gartner, "10 Best Practices for Optimizing Generative AI Costs."

IT leaders should:

- **Consider the complexity and cost implications before opting to self-host generative AI models. If you decide to self-host, ensure your supplier can deliver Opex-based pricing models and/or managed services for it.**
- **Evaluate your organization's capacity for upfront investment, ongoing maintenance and expertise before opting for self-hosting, considering that the costs and complexities can escalate—especially with larger models and high usage volumes.**

Getting the most from your space, power, and budget

Performance is about balance. Too little, and you don't have enough to support new workloads. Too much, and you're overpaying for hardware and facing expensive ongoing power costs. Furthermore, in today's data centers, rack space and wattage are at a premium. How do you add enough performance for inference within these already constrained data centers?

The latest generation of CPUs offer single-threaded "efficient core" options designed to help lower power consumption for scale-out workloads. For example, Intel Xeon processors with Efficient-cores (E-cores) feature a high number of cores per socket and high performance per watt. By replacing 5-year-old servers with Intel Xeon 6 processors with E-cores, you can consolidate racks up to 3:1.[5] Over four years, this can save 80k MWh of fleet energy and reduce CO2 emissions by 34k metric tons.[5] After consolidating racks, you can repurpose space and power to support new AI use cases.

Adding new servers with "performance cores" gives you the performance for AI inference alongside support for your other enterprise workloads, including database and analytics. Xeon 6 processors with Performance-cores (P-cores) deliver up to 18% to 69% lower total cost of ownership on a range of workloads, including AI image classification, generative AI (LLM), encrypted web serving, data services, and high-performance computing, compared to AMD EPYC 9654 servers.[6]

Starting with x86 infrastructure for AI not only helps you reduce the complexity of your systems, but it also helps you maximize your hardware utilization, leaving no silicon behind and unused. This gives you a greater return on your infrastructure investments.

The fast path to deployment

Gartner projects that by 2028, more than 50% of enterprises that have built their own models from scratch will abandon their efforts due to costs, complexity, and technical debt in their deployments.[2]

The truth is, most businesses won't need to build a model from scratch. AI has one of the strongest open source ecosystems the developer community has ever seen, with plenty of sample code to save time and development costs.

Intel has aggregated many of these resources into a library of pre-built AI models and customizable code recipes to jumpstart development for popular enterprise AI use cases. Integrations with the popular Hugging Face AI platform are developer-friendly and help ensure your applications will get excellent performance on Intel platforms. Intel® AI for Enterprise RAG includes a software catalog with models and tools for chat Q&A, code generation, content summarization, and more.

For PC development, Intel sponsors several programs to give developers resources and toolkits to get up and running faster, without the steep learning curve.

For enterprises looking for fully integrated solutions, Intel works with the world's leading global system integrators to optimize AI software and solutions. These offerings eliminate the guesswork and put businesses on the fast track with AI.

CHALLENGE 3

Addressing security and privacy

intel ai

> "Organizations that invest in AI risk and security programs and tools report more positive business outcomes in terms of revenue growth, cost optimization and regulatory compliance than those that don't invest."
>
> **Gartner**, "AI Survey Reveals AI Security and Privacy Leads to Improved ROI"

Cyberattacks are evolving, perhaps as quickly as AI. Businesses looking to leverage AI to create value must be careful to not wipe out those gains by exposing data. According to Gartner, among leaders whose organizations have implemented AI, 3 in 10 report their organizations had an AI privacy breach or security incident, of which over one third were malicious attacks.[3]

Businesses need a carefully considered approach to risk mitigation, trust frameworks, and strong governance policies in order to leverage AI securely. Software alone and traditional perimeter-based security are no longer enough to protect against the latest security threats. Silicon-enabled security is the foundation of an advanced security approach that protects infrastructure, people, and data.

Gartner: Identifying strategic risks

As a starting point for setting trust and governance policies, Gartner recommends exploring the following questions:

- **What is the organization's vision for the responsible, ethical and secure use of AI?**
- **What are the main risks related to the use of AI in the organization?**
- **In addition to compliance, ethics, security and reputation, what other areas will the use of AI pose risks to?**
- **What are the main action plans to mitigate those risks?**
- **Who or which council is mandated to make decisions regarding the use of AI?**

Gartner, "The Pillars of a Successful Artificial Intelligence Strategy"

Gartner: Managing AI trust and risk

Here are some actions organizations can take to manage AI trust, risk and security:

- **Form AI teams across the organization to manage AI privacy, security and risk to achieve higher AI maturity and more positive AI outcomes**
- **Assign budget for AI privacy, security and/or risk when an AI project is approved**
- **Invest in AI trust, risk and security management (TRiSM) tools for more positive business results such as increased revenue and greater efficiencies**

Gartner, "AI Survey Reveals AI Security and Privacy Leads to Improved ROI"

Silicon-enabled security technologies

Software can be spoofed by security breaches at lower layers—if the firmware, BIOS, OS, or hypervisor are compromised, hackers can gain privileged access to systems. It takes a combination of software and silicon-enabled security features to help keep IT infrastructure secure, starting with platform root of trust at the silicon level.

In the data center or cloud, silicon-enabled security can help you put AI to work while securing data and models. Intel® Xeon® processors have built-in security engines that improve your ability to respond to threats if they occur. These engines enable high levels of cryptographic security with enhanced control—so only those who are authorized can access and work with your data. You can isolate your data into secure enclaves to help protect against attacks.

On the PC, Intel helps businesses defend against threats and accelerate remediation with multilayer security features and enabled partner software. Intel is working with some of the world's leading cybersecurity ISVs, like CrowdStrike, Proofpoint, and Trend Micro, to leverage its AI capabilities and help enhance security. By design, Intel® Core™ Ultra processors and Intel vPro® provide a more secure foundation for modern computing and AI through a collection of technologies that start at the hardware level and continue through the entire compute stack.

Intel vPro® has been at the forefront of AI-powered security for years with Intel® Threat Detection Technology, which can help detect attacks before they happen or allocate monitoring tasks to the GPU to free up CPU performance. You can also update and patch your devices, no matter where they reside, with Intel® Active Management Technology, available exclusively on Intel vPro® Enterprise.

> "Security is threatened by AI-enhanced malicious attacks. To combat this new AI-enhanced threat, detection and mitigation techniques are emerging. Continuous behavioral analysis is already being added and will become much less taxing on systems when it can leverage an NPU. Additionally, continuous biometric identification using various techniques can better secure PC access (e.g., SessionGuardian or education proctoring software)."
>
> **Gartner**, "AI PC Introduces a New Era for Personal Computing"

According to ABI Research,[7] Intel leads the silicon industry in product security assurance and continues to demonstrate superior product security as highlighted in the Intel 2023 Product Security Report.[8]

Learn more about silicon-enabled security features:

- Intel® Confidential Computing Solutions
- Blog article: Protect Your Business with AI-based PC Security
- Blog article: Intel AI PCs Deliver an Industry-Validated Defense vs Real-World Attacks

CHALLENGE 4

Getting performance-ready for AI

intel ai

There's no getting around the fact that in order to run AI, you need high-performance compute. Where this compute comes from depends on your business needs.

> "Not all AI initiatives require the same level of maturity in terms of capabilities. For example, some initiatives may utilize existing tools and require only limited capabilities for implementation whereas other initiatives may be highly complex in terms of technology or change management and require very advanced capabilities."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

In the cloud

The cloud offers a fast and easy way to get the performance for AI applications without large upfront costs. Keep in mind that your ongoing cloud fees may quickly expand as you scale AI across your organization.

In your data center

Hosting AI applications in your data center requires a large initial investment, but can offer cost advantages in the long term, along with plenty of performance to handle AI and your other workloads.

On PCs and edge devices

Running AI locally on PCs or edge servers brings performance closer to where your data is being processed, which can help eliminate latency challenges. Data locality also makes it easier to comply with privacy requirements.

In some cases, CPUs with built-in acceleration may offer enough performance to handle your inferencing needs. The clear advantage is that you can use your existing infrastructure and platforms to add AI capabilities. Where more performance is needed, you'll want to consider your options for dedicated AI accelerators.

Gartner: GPUs or accelerated CPUs?

With most new CPU introductions including AI acceleration logic, more than 70% of new servers and PCs will be able to support local processing of small LLM models without the addition of a dedicated accelerator chip within three years.

Server market companies, such as Numenta, are implementing support to run LLMs on latest-generation Intel Xeon processors. Meanwhile, in the PC market, Microsoft has announced Copilot+, an application that requires the NPU integrated into the CPU to have at least 40 tera operations per second (TOPS) performance.

Gartner, "Emerging Tech Impact Radar: Generative AI Hardware Technologies"

AI in the data center and cloud

Gartner projects that by 2027, 40% of existing AI data centers will be operationally constrained by power availability.[9] With data centers quickly running out of rack space and power, architects must determine how to add enough compute performance for AI within their constraints. If running AI in the cloud, teams must choose the best instances to maximize performance per VM.

The performance you need for AI inference will vary based on the size of your models, the number of users, how many queries per day you must support, and your requirements around response time. While very large models like the ones developed by OpenAI must support billions of queries a day, most enterprises will use much smaller models and have much lighter usage. In many cases, a general-purpose server with AI acceleration built into the CPU may offer enough performance, saving the cost and integration challenges that come with GPUs.

You can make use of CPUs with built-in AI acceleration in both the data center and the cloud, significantly reducing your total cost of ownership by avoiding specialized hardware and instances. In mainstream deployments, the latest Intel® Xeon® processors beat the competition in throughput and efficiency on workloads that matter most to enterprises, including AI, web serving, data services, and high-performance computing.[10] Xeon processors have the most built-in accelerators of any CPU on the market to accelerate the greatest range of workloads and improve performance per watt.

What if you plan to run larger models or support heavy usage? GPUs may be a good option to ensure the performance you need. But many architects don't realize there are other dedicated AI processors that can be more competitive than GPUs in terms of cost and power efficiency.

For example, Intel® Gaudi® AI accelerators have been shown to deliver better performance per dollar than GPUs on many AI workloads. In one test, Intel Gaudi 3 accelerators delivered near-parity inference throughput with 1.7x performance per dollar vs. NVIDIA H100.[11]

> "Utilizing custom AI processors, as opposed to GPUs—internally designed, from a traditional semiconductor vendor or from a startup—can lower the cost of deploying LLM-based workloads at the same or better performance."
>
> **Gartner**, "Emerging Tech Impact Radar: Generative AI Hardware Technologies"

Extra benefits when upgrading software

While data center architects puzzle out how to add performance for AI, they must also plan around regular hardware refreshes and software migrations. Architects can refresh to new Xeon processors in the data center to add performance for AI inference on the same cores that run your other enterprise workloads. But Xeon can also help you get the most value from your software migrations, particularly when upgrading to the latest versions of Microsoft SQL Server and Windows Server. By combining your software migration with new hardware, you can get additional performance, cost, and security advantages that help optimize your data center.

Gartner: Using accelerated CPUs for smaller LLMs

While there is a big focus on deployment of very large general-purpose LLMs, many IT organizations will deploy smaller, domain-specific LLMs targeted for their specific business use case.

The availability of servers with accelerated CPUs will enable many of these smaller, domain-specific LLMs to be easily deployed onto standard servers within a data center or a cloud infrastructure-as-a-service service platform. This will drive total cost of ownership (TCO) benefits for many organizations.

Gartner, "Emerging Tech Impact Radar: Generative AI Hardware Technologies"

AI on the PC

From summarizing meeting notes to writing documents to generating code, if your teams want to use the latest AI tools and agentic AI, they'll need powerful PCs to keep up. Additionally, having the option to run AI on the PC instead of in the cloud helps you keep your data private and sidestep the need to be connected to the internet.

Essentially, an "AI PC" is built with a neural processing unit (NPU) that processes AI workloads locally. In 2024, PC makers released the first mainstream Windows business laptops with NPUs, based on the Intel® Core™ Ultra processor. Gartner estimates that by the end of 2026, 100% of enterprise PC purchases will be an AI PC, up from less than 5% in 2023.[4]

Your IT operations team will need to evaluate and decide which PCs are best for running AI locally while also delivering the business-class manageability and security features that help them do their day-to-day jobs.

Intel Core Ultra processors have three compute engines—CPU, GPU, and NPU—that work in concert to support AI workloads. This platform offers the largest library of supported AI software of all PC processor vendors and the best AI software compatibility, supporting 99% of AI features tested.[12] Intel has engaged more than 100 ISVs to optimize tools for Intel Core Ultra processors, with more than 300 AI-accelerated features available,[12] so you can explore plenty of ways to bring AI to your workforce.

Intel also offers AI Playground, a generative AI app suite for Intel AI PCs. It makes it easy to quickly load AI models, and then build custom features and applications that run on the GPUs built into Intel Core Ultra processor-based PCs. Businesses can use it for product design, storyboarding, presentation or script creation, content summarization, search, and coding assistance, among other uses.

Gartner: The benefits of AI PCs

PCs with local AI acceleration will enable a broad range of AI models and algorithms to run efficiently and locally without turning to the cloud. AI PCs offer:

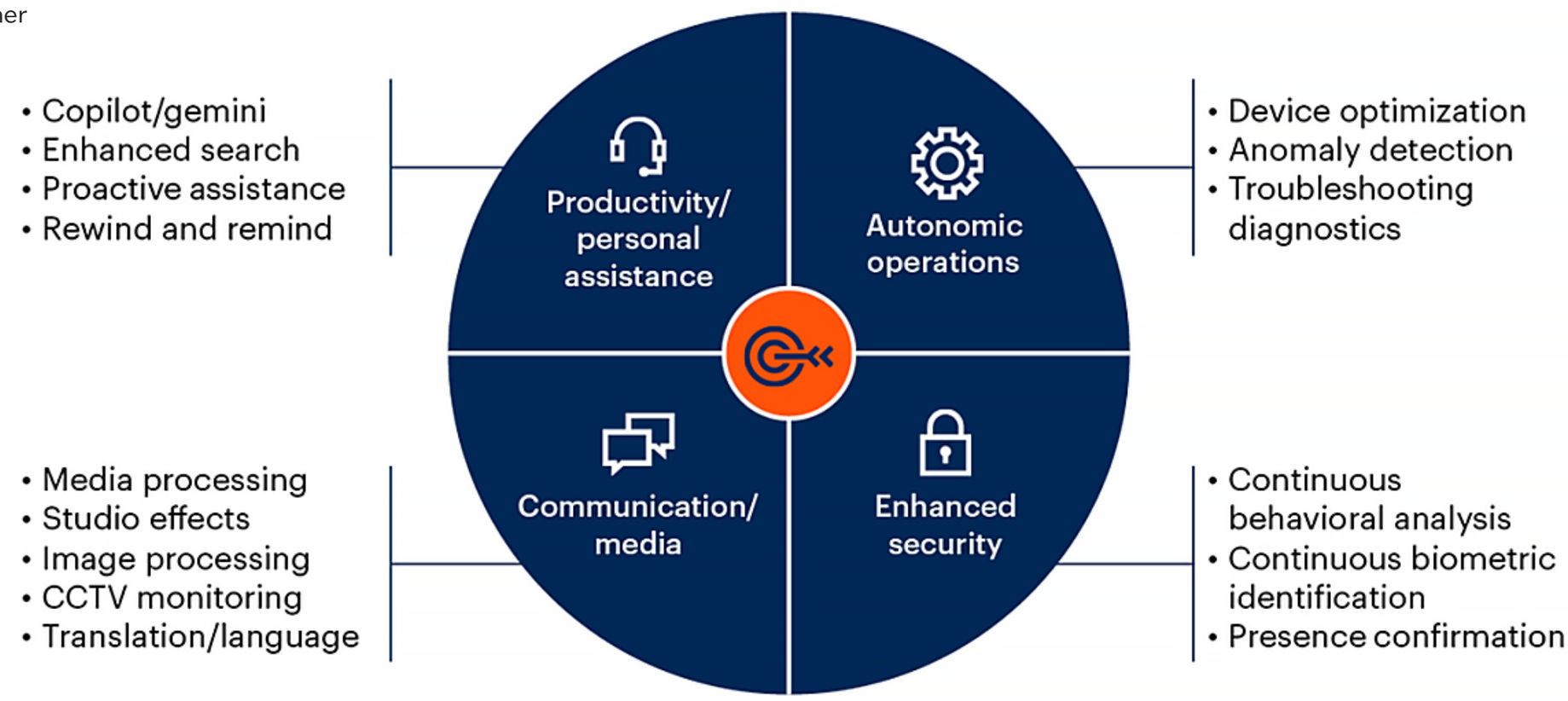
- **Better performance for AI-related tasks—most simple models will run better on an NPU than embedded GPUs**
- **Better power efficiency (battery life, thermal, noise) as NPUs offer better performance per watt**
- **Improved responsiveness, since the CPU and GPU are available for other tasks**
- **The ability to target tasks to the appropriate engine**

Gartner, "AI PC Introduces a New Era for Personal Computing"

Gartner

Local AI Application Targets

Source: Gartner
805646_C



Gartner, "AI PC Introduces a New Era for Personal Computing"

> "While there might be a temptation to hold out for next-generation devices with more performance, the current offerings are all capable of handling basic broad AI applications. Even small targeted GenAI LLMs have been demonstrated running on these first-generation devices with modest performance.
>
> Future applications and OS features will require more NPU performance to execute locally, but for now, there is no advantage in delaying PC refreshes."
>
> **Gartner**, "AI PC Introduces a New Era for Personal Computing"

Gartner: Migrating to Windows 11 with a PC refresh

Enterprises should complete migrations to Windows 11 and only deploy the new AI PC devices with the new OS. This will not only enable users to begin leveraging Copilot and other AI-related features embedded within the OS, but it will also ensure that the OS and applications can exploit the new hardware's performance.

Gartner, "AI PC Introduces a New Era for Personal Computing"

Learn more about accelerated performance for AI:

- **Cloud Performance Benchmarks**
- **Data Center Performance and Efficiency: Intel® Xeon® 6 Processor Family**

CHALLENGE 5

Combining your data with AI

intel ai

Product features. Customer purchase history. Supply chain records. Data is a powerful differentiator that makes it possible for businesses to gain value—and a competitive edge.

> "D&A [data and analytics] capabilities in data management, D&A governance, analytics, organization, roles and data literacy can be leveraged for AI. But only after they have been adapted and extended are they ready for use in AI initiatives. For example, D&A's data management and governance capabilities can be instrumental in providing AI-ready data.
>
> Conversely, for example, D&A can greatly benefit from using AI to generate code or scripts for development and testing, to enable data fabrics or augmented data integration, to generate synthetic data, or to enable user interaction in natural language."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

To make AI work, you'll need to integrate your data into your chosen AI tools and models. Intel offers a scalable computing platform that gives you an extensive choice in software selection and helps you use the latest AI models, software, and agentic AI tools.

Getting ready for agentic AI

Agentic AI is the next wave of enterprise AI. These are software programs designed to use AI to help with specific tasks, acting autonomously and sometimes even proactively. Gartner estimates that by 2028:

- **33% of enterprise software applications will include agentic AI, up from less than 1% in 2024.**[13]
- **At least 15% of day-to-day work decisions will be made autonomously through agentic AI, up from zero percent in 2024.**[13]

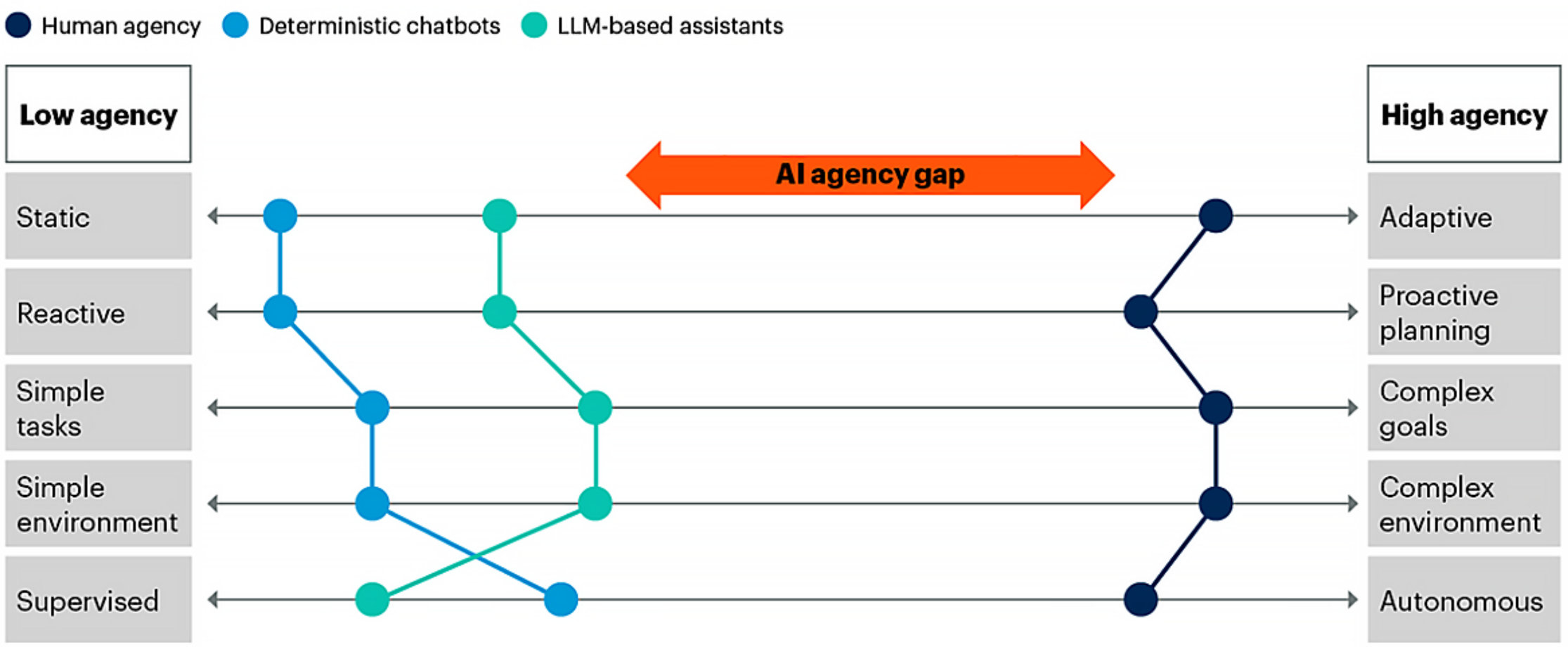
Agentic AI systems can rapidly analyze data sets and act on them, saving time and helping your teams more quickly make decisions. Agents can connect people, applications, and data sources from across your organization, helping break down silos so that more can be shared between teams.

Your success with agentic AI will rely on your ability to connect to data throughout your organization. Intel gives you a flexible, enterprise-ready foundation for agentic AI. On the PC, Intel® Core™ Ultra processors offer the performance to run AI agents locally, so they can access private data without the need to send it to the cloud. Across your data center, cloud, and edge, Intel® Xeon® and Intel® Core™ CPUs are the foundation for data operations, giving agentic AI systems consistent access to the data you need.

> "Agentic AI will dramatically upskill workers and teams, enabling them to manage complicated processes, projects and initiatives through natural language.
>
> However, the orchestration and governance of autonomously acting software entities require advanced tools and strict guardrails."
>
> **Gartner**, "Top Strategic Technology Trends for 2025: Agentic AI"

Gartner

Mind the AI Agency Gap



Source: Gartner
806843_C

Gartner, "Top Strategic Technology Trends for 2025: Agentic AI"

Using AI to improve IT operations

Not only can data help you make better business decisions, but it can also help your IT operations work more efficiently and even automate IT tasks. For example, the Intel vPro® platform on business PCs provides telemetry and device insights that can feed AI Ops. Better data for AI Ops means better business outcomes and makes a smarter, automated PC fleet possible.

Combining data with AI models

In order to get meaningful, custom results with AI, your developers and data scientists must combine in-house data with pre-built models. The ability to use the software or model of their choice is affected by hardware.

With Intel hardware, you can use your favorite AI frameworks and models—with a performance boost. Whether you're working with PyTorch, TensorFlow, ONNX, or other popular frameworks, Intel has worked with the ecosystem on performance optimizations. Intel-optimized libraries provide deep integration with popular AI frameworks, eliminating the need for manual optimization. That means you don't have to modify your code to benefit from hardware-specific performance boosts right out of the box.

Developers can also move from prototype to production faster with pre-built models. For rapid project starts, Intel's extensive library of pre-built AI models, customizable code recipes, and integrations with Hugging Face can jumpstart development. By leveraging Intel-optimized models and tools on Hugging Face, developers can quickly fine-tune models and deploy them across Intel hardware for enhanced performance and efficiency.

> "Mitigate the risks of moving away from GPU-based designs to AI processors by evaluating the associated software ecosystem and the vendor's ability to support migration.
>
> Further, assess the vendor's financial stability, specifically if you are selecting an AI processor developed by a startup, as many are still in early-stage development and rely on venture capital funding rather than stable revenue from product sales."
>
> **Gartner**, "Emerging Tech Impact Radar: Generative AI Hardware Technologies"

CHALLENGE 6

Getting from prototype into production

intel ai

It's one thing to show what's possible with an AI pilot. It's another to scale the same use case across your entire operation.

> "...many current AI initiatives are still experimental and exist as isolated projects. This fragmentation leads to difficulties in scaling, managing risks and realizing business value.
>
> Although experimental siloed approaches may be useful for building skills and learning what AI can and cannot do, they are not enough to create sustainable business value."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

Before you go too far with your AI projects, consider how you'll accommodate different environments and needs, including deployments in the cloud, in the data center, on the PC, and at the edge.

Migrate out of CUDA

CUDA is a favorite development environment for many of today's AI innovators. But too often, the proof of concept that works great on a single GPU or cluster fails to scale out cost-effectively. Migrating out of CUDA and into open tools gives you the ability to scale across more diverse hardware types.

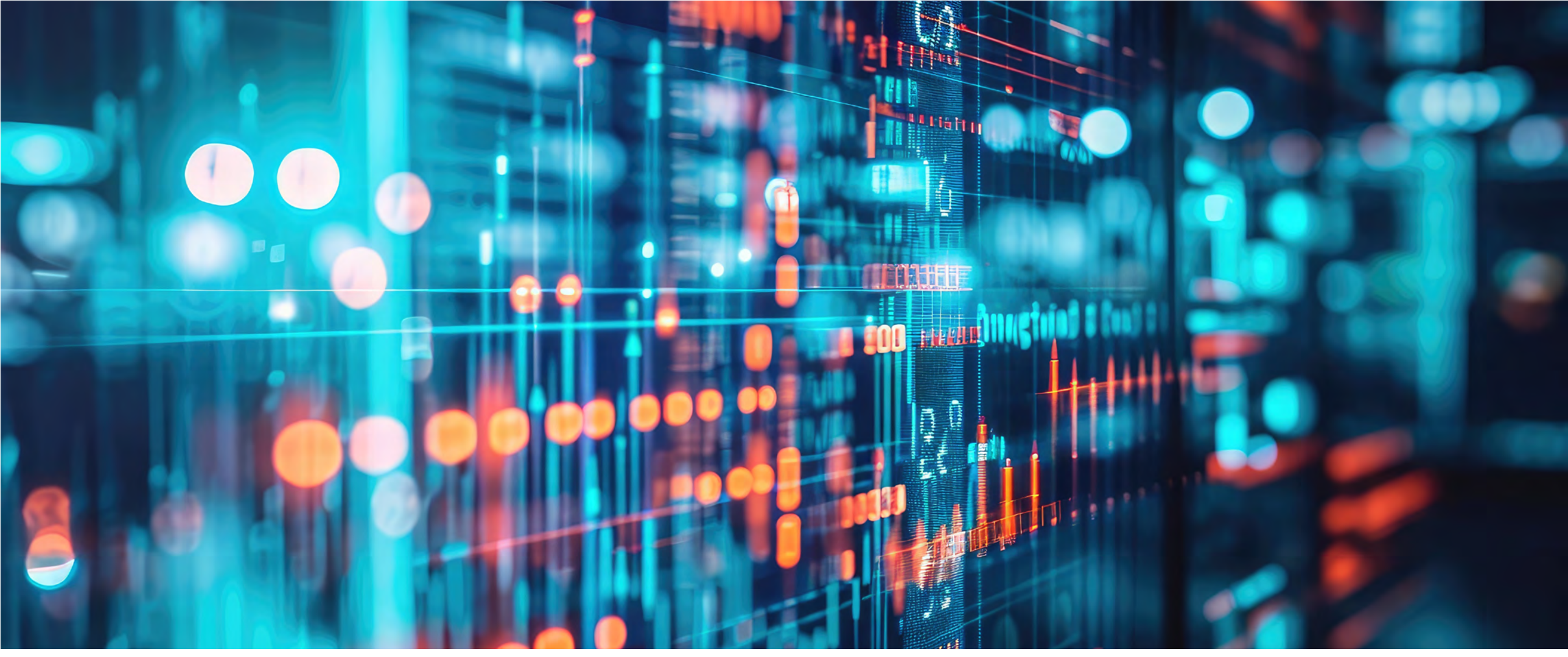
Intel offers tools to help streamline the migration from NVIDIA systems with a CUDA framework to Intel® Gaudi® AI accelerators and Intel® CPUs. Developers can also use an open source, pretrained model from Hugging Face, TensorFlow, Keras, PyTorch, PaddlePaddle, or ONNX, and then convert it for their Intel® hardware of choice using the OpenVINO™ toolkit.

For developers who need the performance for larger models or more users, Intel® Gaudi® AI accelerators provide a cost-effective alternative to competitors, especially in generative AI use cases, ensuring top-tier performance without overspending. You can find an increasing number of optimized models from the Intel Gaudi software catalog and Hugging Face.

Grow on x86 architecture

By developing applications on x86 architecture, you can streamline their deployment across millions of PCs, servers, and edge devices. There are more than 40 million AI PCs running on Intel® Core™ Ultra processors and countless platforms based on Intel® Xeon® and Intel® Core™ CPUs across the enterprise, cloud, and edge.

For AI PC applications, developers can hit the ground running with Intel's toolkits and resources to develop, deploy, and scale. These include more than 500 AI models optimized for Intel® Core™ Ultra processors, wide support for AI frameworks, and compatible AI runtimes. Intel's tools put you at the forefront of AI development, so you can leverage both the



established x86 PC market and the emerging AI PC ecosystem to ensure your applications can reach a global user base.

Building AI applications for x86 architecture can also streamline integration with existing workflows and data. When you develop AI applications on the same Xeon processors that are at the foundation of the enterprise ecosystem, you can maximize compatibility with wide-scale data center and cloud infrastructure and software.

For edge applications, Intel has generations of innovation in edge hardware and software and a broad ecosystem of partners. Intel offers software tools to optimize computer vision, small or large language model (SLM/LLM) inference, and other AI applications to run smoothly across hardware types.

The future is looking bright for x86. Intel is the co-founder of an x86 ecosystem advisory group that is shaping the future of the world's most widely used computing architecture. The advisory group will help enable compatibility across platforms, simplify software development, and provide developers with a platform to identify architectural needs and features to create innovative solutions. Consistent x86 features and programming models will extend across data center, cloud, client, edge, and embedded devices.

In addition, Intel contributes to projects like OPEA (Open Platform for Enterprise AI) to advance AI innovation. OPEA is a collaborative environment where developers can access open source tools, frameworks, and enterprise-level resources to build AI solutions in real-world scenarios.

TAKE THE NEXT STEP

Make your company an AI company with Intel inside

intel ai

As you set the course for your AI strategy, keep in mind that it's not a one-time job.

> "The alignment between the AI strategy and other strategies, in particular the business strategy, should be bidirectional. After all, AI is not just a technology to improve existing business only. In addition, it is increasingly applied to catalyze new business opportunities or disrupt existing business models and even entire markets.
>
> Either way, any changes in the business strategy, perhaps triggered by new competitive activity or changing market conditions, should be reflected in an updated AI strategy. This should in turn result in a reprioritized AI portfolio and updated planning goals for the AI operating model."
>
> **Gartner**, "The Pillars of a Successful Artificial Intelligence Strategy"

Revisit your goals, projects, and outcomes regularly with key stakeholders so you can evaluate results and adjust accordingly. Your industry is changing constantly, and your AI strategy should evolve along with it.

By choosing computing foundations with Intel inside, you can be better prepared to make your company an AI company, no matter how the market evolves. Business-caliber AI PCs can help you put AI tools directly into the hands of your workforce. In the data center and cloud, efficient, powerful compute can help you add the performance you need to tackle the next generation of AI use cases. And Intel's open, flexible platform helps you prepare to handle any AI model, any type of data, and any kind of environment.

Today, Intel is helping you prepare for what happens tomorrow, so you can deploy the use cases that bring your enterprise the most value—on your terms. That's the power of Intel inside®.

To learn more about building your AI strategy, access your complimentary copy of the Gartner report, "The Pillars of a Successful Artificial Intelligence Strategy."

intel ai

intel ai

Footnotes

1. Gartner, The Pillars of a Successful Artificial Intelligence Strategy, Pieter den Hamer, Raghvender Bhati, 23 April 2024.
2. Gartner, 10 Best Practices for Optimizing Generative AI Costs, Arun Chandrasekaran, Leinar Ramos, Alberto Pietrobon, Justin Tung, June 2024.
3. Gartner, AI Survey Reveals AI Security and Privacy Leads to Improved ROI, Avivah Litan, Leinar Ramos, May 2024.
4. Gartner, AI PC Introduces a New Era for Personal Computing, Stephen Kleynhans, Autumn Stanish, Tom Cipolla, May 2024.
5. See 7T2 at intel.com/processorclaims: Intel® Xeon® 6. Your costs and results may vary.
6. See 9T9, 9T8, 9T7 at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
7. As measured by ABI Research: intel.com/content/www/us/en/security/security-as-a-component-of-tech.html.
8. For more information, see the Intel 2023 Product Security Report: intel.com/content/www/us/en/security/intel-2023-product-security-report.html.
9. Gartner, Emerging Tech Impact Radar: Generative AI Hardware Technologies, Alan Priestley, Shrish Pant, Gaurav Gupta, Anushree Verma, Menglin Cao, Uko Tian, Adrian O'Connell, George Brocklehurst, October 2024.
10. See intel.com/processorclaims: Intel® Xeon® 6, 5th Gen Intel® Xeon®, and 4th Gen Intel® Xeon®. Results may vary.
11. NVIDIA H100 comparison based on TensorRT-LLM/docs/source/performance/perf-overview.md at main NVIDIA/TensorRT-LLM-GitHub, October 18, 2024. Reported numbers are per GPU vs. Intel® Gaudi® 3 measurements for LLAMA3-8B, LLAMA2-70B. Results may vary. Based on Intel Gaudi software release 1.18. Refer to this link for the latest published Gaudi 3 performance: https://www.intel.com/content/www/us/en/developer/platform/gaudi/model-performance.html. Pricing estimates based on publicly available information and Intel internal analysis.
12. Based on analysis of publicly available information as of January 2025.
13. Gartner, Top Strategic Technology Trends for 2025: Agentic AI, Tom Coshow, Arnold Gao, Lawrence Pingree, Anushree Verma, Don Scheibenreif, Haritha Khandabattu, Gary Olliffe, October 2024.

Performance varies by use, configuration and other factors. Learn more at intel.com/performanceindex. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation. Availability of accelerators varies depending on SKU. Visit the Intel Product Specifications page for additional product details.

All versions of the Intel vPro® platform require an eligible Intel® processor, a supported operating system, Intel LAN and/or WLAN silicon, firmware enhancements, and other hardware and software necessary to deliver the manageability use cases, security features, system performance, and stability that define the platform. See intel.com/performance-vpro for details.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Data latency, cost, and privacy advantages refer to non-cloud-based AI apps. Learn more at intel.com/AIPC.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

Gartner

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.