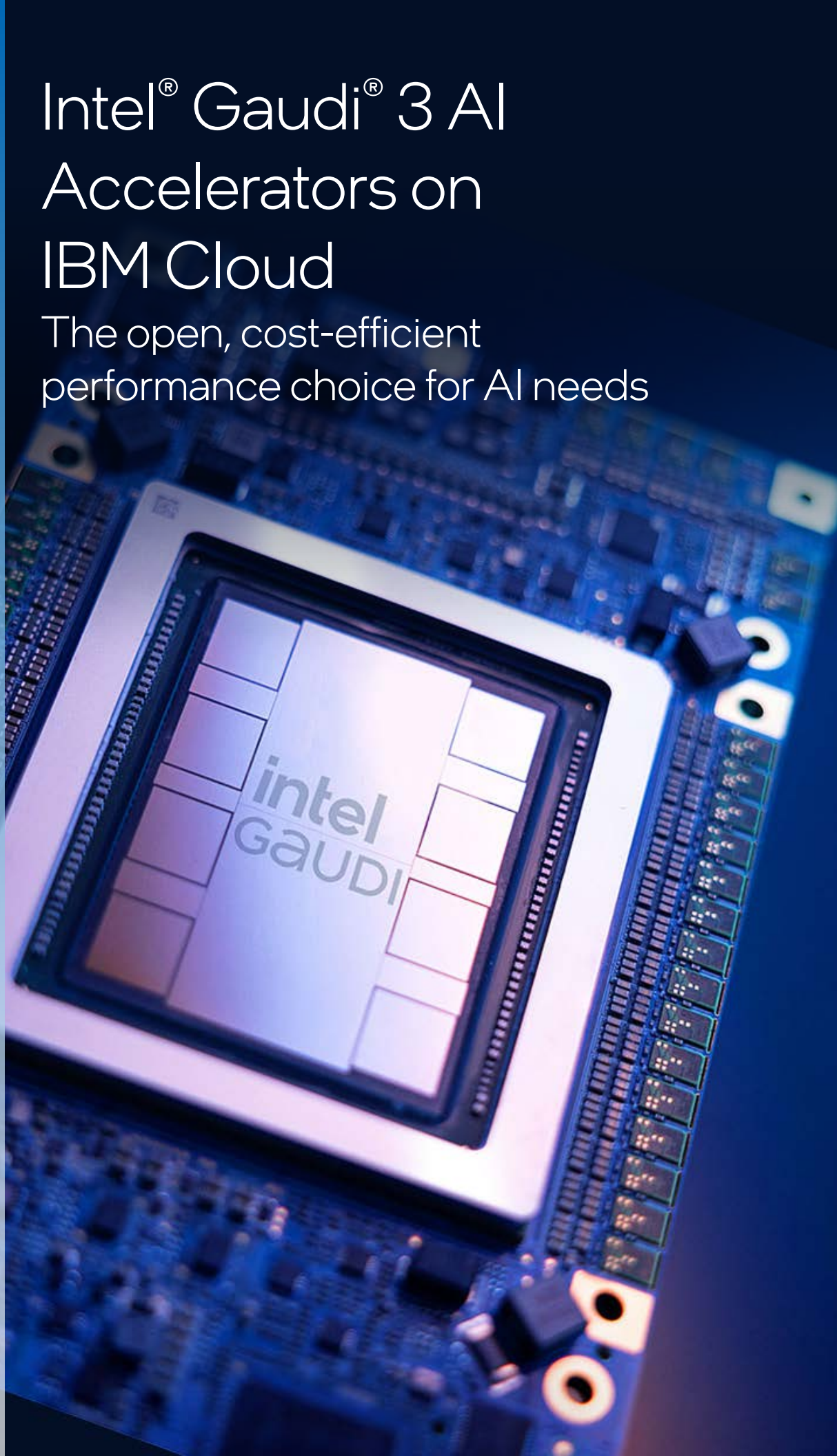
Intel® Gaudi® 3 AI Accelerators on IBM Cloud

The open, cost-efficient performance choice for AI needs



Solution Guide

- General AI benefits
- Healthcare and life sciences
- Financial services

Intel® Gaudi® 3 AI accelerators on IBM Cloud

Table of contents

Intel® Gaudi® 3 AI accelerators on IBM Cloud

All AI needs are not alike: choose a scalable AI infrastructure designed for cost-effective, open, secure performance.

Myth:
AI model inferencing should be run on the same expensive infrastructure used to train the model.

Reality:
Scalable, high-performance inferencing can be performed more cost-effectively with Intel Gaudi 3 AI accelerators.

Next-generation AI performance and cost efficiency

Intel Gaudi 3 AI accelerators on 5th Gen Intel® Xeon® processor-based instances on IBM Cloud are designed to deliver open, cost-effective scalability for your enterprise AI inferencing and fine-tuning workloads. Support for Intel Gaudi 3 with IBM's watsonx AI and data platform provides additional infrastructure resources for scaling gen AI workloads, while aiming to optimize price-performance for model inferencing.

IBM
Infrastructure designed for security

\+

intel
Intel Gaudi 3 AI accelerator

\+

Enterprise AI stack and models with IBM watsonx

Intel Gaudi 3 AI accelerator advantages

Cost efficient
Intel Gaudi 3 provides cost-effective AI acceleration.

Choice
Helps reduce hardware and software licensing costs.

Open
Access support for hundreds of models, frameworks, and libraries with an open-source software stack.

Scalable
Flexibility to scale enterprise AI workloads with efficiency and freedom from closed system lock-in.

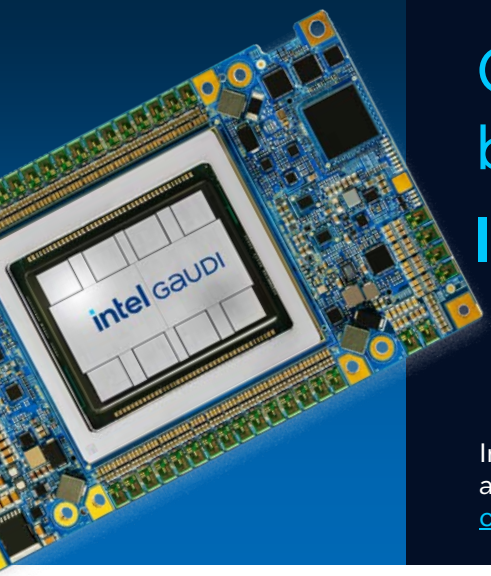
IBM Cloud + Intel: A history of thought and technological leadership

Intel and IBM have collaborated to deliver industry firsts that help drive innovation and results, starting with the inception of the IBM PC in 1981.[1]

Now, IBM Cloud is the first cloud services provider to make Intel Gaudi 3 AI accelerators available to enterprise customers to help provide:[2]

- Cost-effective AI, security, and resiliency.
- Flexibility with an open ecosystem.
- Large model hosting using fewer accelerators due to large memory.[3]

1. The Intel® 8086 and the IBM PC, Behind the Black Curtain, Intel.com.
2. Intel and IBM Collaborate to Provide Better Cost Performance for AI Innovation, IBM Newsroom, August 29, 2024. Intel and IBM Deliver Enterprise AI in the Cloud, Intel Newsroom, August 29, 2024.
3. Intel Gaudi 3 AI Accelerator White Paper, October 23, 2024.

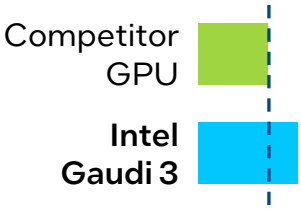


Can exceptional AI performance be cost-effective?

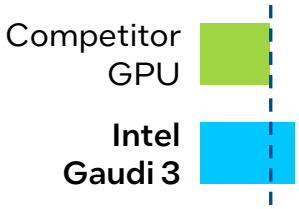
It can! Choosing Intel® Gaudi® 3 AI accelerators on IBM Cloud helps enable your organization to experience all the advantages and outcomes of AI with breakthrough cost efficiency.

Independent AI inferencing performance testing of Intel Gaudi 3 AI accelerators on IBM Cloud demonstrates exceptional performance and cost efficiency benefits vs. competitor GPUs:

Performance:

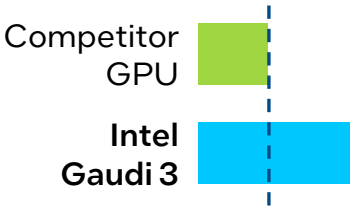


Up to **43%** more tokens per second than competitor GPU when running IBM Granite-3.1-8B-Instruct for small AI workloads.[4]

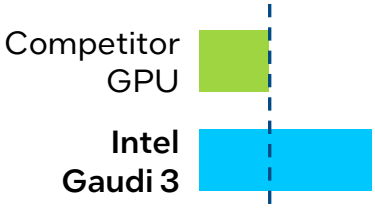


Up to **36%** more tokens per second than competitor GPU when running Llama-3.1-405B-Instruct-FP8 with large context sizes.[4]

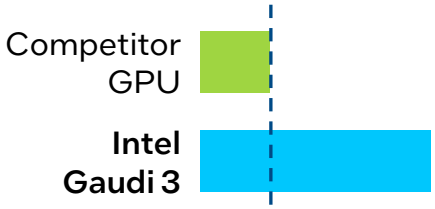
Performance per Dollar:



Up to **120%** increase in tokens per dollar than competitor GPU when running Mixtral-8x7B-Instruct-v0.1 with long input and short output sizes.[4]



Produce up to **2.5x** more work per dollar than competitor GPU, across the same set of inferencing workloads.[5]



Deliver up to a **335%** increase in tokens per dollar than competitor GPU when running Llama-3.1-405B-Instruct-FP8.[4]

4. Source: Signal65 Lab Insight Report, "Intel Gaudi 3 AI Accelerator at Scale on IBM Cloud," Intel-commissioned study by Signal65, published April 2025. Reported numbers are inferencing results on Intel Gaudi 3 vs. NVIDIA H200. See the source for workloads and configurations. Results may vary.
5. Source: Signal65 Lab Insight Report, "The New AI Accelerator Economic Landscape," Intel-commissioned study by Signal65, published February 2025. Reported numbers are inferencing results on Intel Gaudi 3 vs. NVIDIA H100. See the source for workloads and configurations. Results may vary.

Can AI development be based on open standards?

It can! Speed AI development and ensure future scalability with open Intel® Gaudi® 3 AI accelerators.

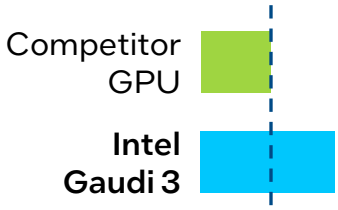
- Intel Gaudi 3 AI accelerators support popular AI models and AI frameworks, such as PyTorch, to simplify integration with your existing AI workloads and pre-trained models.
- Open software support helps speed development while providing future-ready AI flexibility, scalability, and choice for today and tomorrow.
- Intel provides a free, comprehensive, full-stack set of software tools and supports a vast array of AI models to simplify and speed time to value.
- Over 500K models are available on Hugging Face, enabled for ease of integration with the Optimum Habana software library.

Can AI flexibly scale for the future?

It can! Future-proof your AI scalability with the flexible, cost-effective, open standards-based Intel Gaudi 3 accelerators.

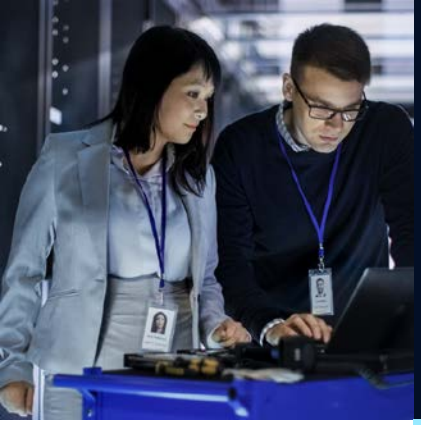
Intel Gaudi 3 accelerators deliver outstanding energy efficiency to help reduce your AI operating costs and meet your sustainability objectives.

Intel Gaudi 3 accelerators deliver up to **92%** more tokens per dollar vs. competitor GPU when running Llama-3.1-405B-Instruct-FP8.[6]



Intel Gaudi 3 accelerators provide flexible, cost-effective scalability from a single accelerator to mega-clusters to handle your most demanding inference, training, and fine-tuning needs today—and tomorrow.

6. Source: Signal65 Lab Insight Report, "Intel Gaudi 3 AI Accelerator at Scale on IBM Cloud," Intel-commissioned study by Signal65, published April 2025. Reported numbers are inferencing results on Intel Gaudi 3 vs. NVIDIA H200. See the source for workloads and configurations. Results may vary.



Deployment options: Intel® Gaudi® 3 accelerators on IBM Cloud

Choose from an array of flexible consumption options designed for inferencing and fine-tuning.

Deployable architectures

Designed for simplified user deployment, scalability, and modularity.

Production-ready, pre-configured applications from Intel:

- Intel AI for Enterprise Inference
- Intel AI for Enterprise RAG

Faster time to deployment

IBM watsonx	▪ Bring-your-own-license or deploy on premises. ▪ Support for watsonx software-as-a-service is planned.
Red Hat on IBM Cloud	▪ Leverage fully managed, containerized infrastructure with Red Hat OpenShift and Red Hat OpenShift AI. ▪ Support for IBM Kubernetes Service is planned.
IBM Cloud Virtual Servers for VPC	▪ Deploy isolated cloud instances inside a software-defined network with support for Red Hat Enterprise Linux AI images.

Discover available foundation models for IBM watsonx.ai.

Healthcare and Life Sciences

Healthcare and life sciences AI challenges

Integrating AI in healthcare and life sciences promises transformative impact—but it comes with unique and complex hurdles. Protecting patient safety and privacy, managing costs, and navigating regulatory landscapes are critical for successful adoption.

Chronic resource shortages

Healthcare systems face persistent shortages—from funding and primary care providers to nurses and specialists. Limited resources create bottlenecks that make it difficult to deliver timely, high-quality care.

Patient outcome pressures

Chronic conditions like diabetes, heart disease, and hypertension are on the rise, demanding continuous monitoring and management. These pressures strain already stretched resources and challenge traditional care models.

Coordination of care

Fragmented communication between providers can lead to errors, duplicate testing, or conflicting treatment plans. Lack of integrated care negatively impacts patient outcomes and provider efficiency.

Privacy and security risks

Patient data is sensitive, and cyber threats are growing. Strong cybersecurity safeguards are essential to protect confidentiality, prevent misuse, and maintain trust.

Data accessibility challenges

Healthcare generates massive volumes of data—up to 97% remains unused. Siloed, unstructured information in legacy systems prevents organizations from scaling AI solutions effectively.

Explainability and transparency

“Black box” AI models can be hard to interpret, undermining trust in decisions and limiting error detection. Clear, explainable AI is essential for clinicians and patients to adopt insights confidently.

Regulatory complexity

AI in healthcare is governed by evolving regulations that differ across regions. Navigating these standards while deploying AI at scale adds operational and compliance complexity.

AI opportunities and use cases for healthcare and life sciences

These are just a few of the AI-powered use cases that can transform healthcare and life sciences.

	Use case examples	Data source
	▪ Clinical decision support ▪ Disease prediction ▪ Disease management ▪ Chatbot/telehealth ▪ Clinician/patient experience ▪ Reduce administrative workloads	▪ EHR clinical notes ▪ Discharge summaries ▪ Lab reports ▪ Online forum posts ▪ Live audio/vision (phone/video)
	▪ Adverse event detection ▪ Drug screening/discovery ▪ Clinical trial optimization	▪ Adverse event reports ▪ Biomedical literature ▪ Clinical trial documents
	▪ Protein sequence modeling/structure prediction	▪ Molecules/protein sequence data
	▪ Patient positioning ▪ Diagnosis screening ▪ Prognostication of outcomes and treatment response ▪ Pathology segmentation ▪ Disease monitoring ▪ Converting low-dosage, low-resolution CT images into high-resolution images	▪ Medical images: X-rays, CT, PET, MRI, FMRI, etc.

Spotlight AI solution: Generate from Iterate.ai

The Generate platform on IBM Cloud is designed to help healthcare organizations reclaim revenue and boost efficiency by leveraging generative AI to:

- Ingest and parse complex claims data.
- Work with claims across disparate EMR systems.
- Process raw data without requiring structure.
- And more.

Better together

IBM Cloud + Intel® Gaudi® 3 AI accelerators for healthcare and life sciences

Intel Gaudi 3 AI accelerators on IBM Cloud are designed to deliver cost-efficient, scalable, and secure AI performance tailored for healthcare and life sciences. By optimizing models, Gaudi 3 helps accelerate results at a lower total cost of ownership. Together, Intel and IBM Cloud design solutions to empower organizations to innovate confidently and compete intelligently in an AI-driven world.

A firm foundation on IBM Cloud

Designed for security and compliance

In healthcare and life sciences, data protection is everything. IBM Cloud is purpose-built for regulated industries—designed for advanced encryption, identity management, and continuous threat detection to help safeguard sensitive data. With Intel Gaudi 3 AI accelerators, AI workloads can run efficiently within IBM Cloud's infrastructure—helping institutions meet strict requirements while accelerating innovation safely.

Hybrid cloud for sensitive data

Health and life sciences organizations often need the flexibility of the cloud with the control of on-premises infrastructure. IBM Cloud's hybrid-by-design architecture can help organizations keep mission-critical data on premises while seamlessly extending AI workloads to the cloud. Paired with Intel Gaudi 3 accelerators' scalable AI performance, firms can aim to analyze, predict, and automate securely—without disrupting existing operations or compromising data governance.

Resilience, recovery, and continuity

Downtime isn't an option in healthcare and life sciences. IBM Cloud's enterprise-grade disaster recovery and business continuity capabilities are engineered to minimize disruption and protect confidence. Together with Intel Gaudi 3 AI accelerators' optimized performance and efficiency, institutions can help ensure critical AI-driven operations remain available—even in the face of unexpected events.

Accelerate insight with watsonx + Intel Gaudi 3 AI accelerators

IBM watsonx is designed to unify enterprise AI and analytics. When paired with Intel Gaudi 3, organizations can create integrated AI workflows across training, tuning, and deployment within a consistent on-premises or off-premises environment.

Global reach and reliability

IBM Cloud's global network of data centers and partners is designed to help health and life sciences institutions deploy AI solutions closer to their customers with lower latency and consistent compliance onboard. Intel Gaudi 3 accelerators add scalable, high-performance AI—helping organizations to innovate securely and reliably.

Healthcare and life sciences deployment options

Flexible, efficient, and cost-saving AI

Intel® Gaudi® 3 AI Accelerators on IBM Cloud are designed for secure, high-performance AI solutions that support healthcare and life sciences. The mission of this collaboration is to empower organizations to place workloads where they make the most impact—optimizing costs, improving operational efficiency, and accelerating innovation.

Stand-alone server on IBM Cloud Virtual Private Cloud (VPC)

Build an isolated private cloud with Intel Gaudi 3 accelerators while retaining key public cloud benefits. Ideal for organizations with specialized software stacks or strict compliance requirements, this option provides dedicated infrastructure, high resiliency, and support for Red Hat Enterprise Linux AI images.

Container worker node

Intel Gaudi 3 accelerators can serve as worker nodes for Red Hat OpenShift AI clusters or Red Hat OpenShift on IBM Cloud, helping teams create a flexible, containerized environment. Ideal for scaling workloads efficiently, simplifying orchestration, and accelerating the deployment of AI applications.

Bring-your-own watsonx software license

Deploy IBM watsonx.ai directly on Intel Gaudi 3 accelerators to help manage your AI stack. An integrated developer toolkit and end-to-end lifecycle management can help organizations build, train, and deploy AI services faster while maintaining control and compliance.

Accelerate adoption with Deployable Architectures (DAs)

Help speed AI deployment without friction. Intel Deployable Architectures on IBM Cloud help enable fast adoption of Intel Gaudi 3 accelerator capabilities with pre-built, validated design modules. Options include Intel AI for Enterprise Inference and Intel AI for Enterprise RAG.

Customer success stories

Healthcare and life sciences

- Drugs.com deployed its services on IBM Cloud with Intel® technology to provide reliable information about more than 24,000 different medications—their indications, side effects, potential drug interactions, and alternative drug treatment options.
  Learn more
- University of Fukui Hospital in Japan trusts IBM Cloud with Intel technology to protect its sensitive data and reduce its on-premises data center spend.
  Learn more
- The University of Maryland, Baltimore County, takes advantage of IBM Cloud with Intel technology to store and analyze confidential healthcare data and meet strict compliance regulations, as well as to store and process other education data.
  Learn more

Financial Services

Financial services AI challenges

Regulatory compliance has always been central to the financial services industry. Many day-to-day operations have traditionally required significant human oversight—making them slow, costly, and prone to error. The prospect of using AI to automate and streamline these processes is therefore highly attractive from an operational standpoint. However, the adoption of AI also introduces a new set of risks and challenges, including:

Data privacy and security

Cyberattacks, data breaches, and unauthorized access can severely damage customer trust, incur significant financial losses, and lead to non-compliance with critical regulations such as GDPR and CCPA.

Money laundering and fraud prevention

Fraudulent activities have skyrocketed, with the FTC reporting a 25% increase in fraud cases, totaling over $12.5 billion in losses in 2024. AI systems need to stay ahead of evolving fraud tactics to protect both customers and institutions.

Bias and ethical concerns

AI models can unintentionally reinforce existing biases, resulting in unfair or discriminatory decisions. This raises serious ethical concerns about the implications of automated decision-making, especially when it impacts individuals' financial futures.

Lack of transparency

Many AI systems, particularly those using deep learning (DL), operate as "black boxes," making it difficult to trace how decisions are made. This lack of explainability can undermine trust and complicate audits and compliance checks.

Regulatory compliance

Navigating the shifting regulatory landscape in financial services is an ongoing challenge. AI systems must be continuously updated to ensure they comply with complex and ever-changing laws and regulations.

Model overfitting

Overfitting occurs when an AI model becomes too specialized in its training data, reducing its ability to generalize to new, unseen data. In the dynamic world of financial services, this can result in inaccurate predictions and poor decision-making.

AI opportunities and use cases for financial services

These are just a few of the AI-powered use cases that can transform financial services.

Customer onboarding and Know Your Customer (KYC)

- Automate verification: AI-driven KYC rapidly processes and verifies customer information, eliminating manual bottlenecks and accelerating account activation.
- Predictive insights: Customer profiles, transaction histories, and behaviors power predictive models to recommend relevant products and cross-sell opportunities.
- Smart support: Virtual assistants and chatbots handle queries, provide instant support, and escalate complex issues to staff when needed.

Anti-Money Laundering (AML) and fraud detection

- Detect complex patterns: AI identifies suspicious activity across large datasets, uncovering cross-border transfers, shell company usage, pass-through funds, and other fraud schemes in real time.
- Proactive alerts: Systems flag potential risks early, helping institutions act faster and minimize financial and regulatory exposure.

Augmented financial risk management and compliance

- Simplify compliance: AI manages massive datasets and adapts to evolving regulations, keeping operations aligned with the latest rules.
- Free up staff: Routine compliance tasks are automated, allowing teams to focus on strategic risk planning and decision-making.

Improved algorithmic trading and quantitative analysis

- Faster market response: Machine learning and deep learning algorithms analyze financial data in real time, predicting trends and informing trading decisions.
- Next-gen AI: Emerging GenAI applications enhance speed and accuracy beyond traditional methods, giving institutions an edge in fast-moving markets.

European Instant Payment Regulation (IPR) fraud detection

- Real-time monitoring: AI-powered systems detect anomalies instantly, analyzing multiple data points to minimize false positives.
- Regulatory compliance: Supports PSPs in meeting EU requirements for 24/7, instant euro payments while maintaining robust fraud prevention.

Better together

IBM Cloud + Intel® Gaudi® 3 AI accelerators for financial services

Intel Gaudi 3 AI accelerators on IBM Cloud are designed to deliver cost-efficient, scalable, and secure AI performance tailored for financial services. By optimizing large language, multimodal, and RAG models, Intel Gaudi 3 accelerators can help institutions detect fraud faster, assess risk more accurately, and uncover deeper insights—all at a lower total cost of ownership. Together, Intel and IBM Cloud help empower financial firms to innovate confidently and compete intelligently in an AI-driven world.

A firm foundation on IBM Cloud

Designed for security and compliance

In financial services, data protection is everything. IBM Cloud is purpose-built for regulated industries—designed for advanced encryption, identity management, and continuous threat detection to help safeguard sensitive financial data. With Intel Gaudi 3 AI accelerators, AI workloads can run efficiently within IBM Cloud's infrastructure, helping institutions meet strict requirements such as GDPR and PCI DSS while accelerating innovation safely.

Hybrid cloud for sensitive data

Financial institutions often need the flexibility of the cloud with the control of on-premises infrastructure. IBM Cloud's hybrid-by-design architecture helps organizations keep mission-critical data on premises while seamlessly extending AI workloads to the cloud. Paired with Intel Gaudi 3 accelerators' scalable AI performance, financial firms can help analyze, predict, and automate securely—without disrupting existing operations or compromising data governance.

Resilience, recovery, and continuity

Downtime isn't an option in finance. IBM Cloud's enterprise-grade disaster recovery and business continuity capabilities are engineered to minimize disruption and protect customer confidence. Together with Intel Gaudi 3 AI accelerators' optimized performance and efficiency, institutions can help ensure critical AI-driven operations remain available—even in the face of unexpected events.

Accelerate insight with watsonx + Intel Gaudi 3 AI accelerators

IBM watsonx is designed to unify enterprise AI and analytics—turning vast amounts of financial data into actionable insights. When paired with Intel Gaudi 3, organizations can create integrated AI workflows across training, tuning, and deployment within a consistent on-premises or off-premises environment.

Global reach and reliability

IBM Cloud's global network of data centers and partners is designed to help financial institutions deploy AI solutions closer to their customers with lower latency and consistent compliance onboard. Intel Gaudi 3 accelerators add scalable, high-performance AI—helping organizations to innovate securely and reliably.

Financial services deployment options

Flexible, cost-effective, and secure

Intel® Gaudi® 3 AI accelerators on IBM Cloud help empower financial institutions to deploy AI wherever it delivers the greatest impact—securely, efficiently, and on their terms. These options can help provide scalable performance, cost savings, and operational control for a variety of AI workloads.

Stand-alone server on IBM Cloud Virtual Private Cloud (VPC)

Build an isolated private cloud with Intel Gaudi 3 accelerators while retaining key public cloud benefits. Ideal for organizations with specialized software stacks or strict compliance requirements, this option provides dedicated infrastructure, high resiliency, and support for Red Hat Enterprise Linux AI images.

Container worker node

Intel Gaudi 3 accelerators can serve as worker nodes for Red Hat OpenShift AI clusters or Red Hat OpenShift on IBM Cloud, helping teams create a flexible, containerized environment. Ideal for scaling workloads efficiently, simplifying orchestration, and accelerating the deployment of AI applications.

Bring-your-own watsonx software license

Deploy IBM watsonx.ai directly on Intel Gaudi 3 accelerators to help manage your AI stack. An integrated developer toolkit and end-to-end lifecycle management can help organizations build, train, and deploy AI services faster while maintaining control and compliance.

Accelerate adoption with Deployable Architectures (DAs)

Help speed AI deployment without friction. Intel Deployable Architectures on IBM Cloud help enable fast adoption of Intel Gaudi 3 accelerator capabilities with pre-built, validated design modules. Options include Intel AI for Enterprise Inference and Intel AI for Enterprise RAG.

Customer success stories

Financial services

- GEVA Group deploys critical workloads on IBM Cloud with Intel® technology to deliver its innovative services at scale, while protecting highly sensitive customer data.
  Learn more
- Circeo relies on IBM Cloud Bare Metal Servers running Red Hat OpenShift on Intel technology to power their innovative online banking platform, providing valuable services—such as automated, online loans—to customers, while keeping all their sensitive data protected and secure.
  Learn more

Legal

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software, or service activation. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. All product plans and roadmaps are subject to change without notice.