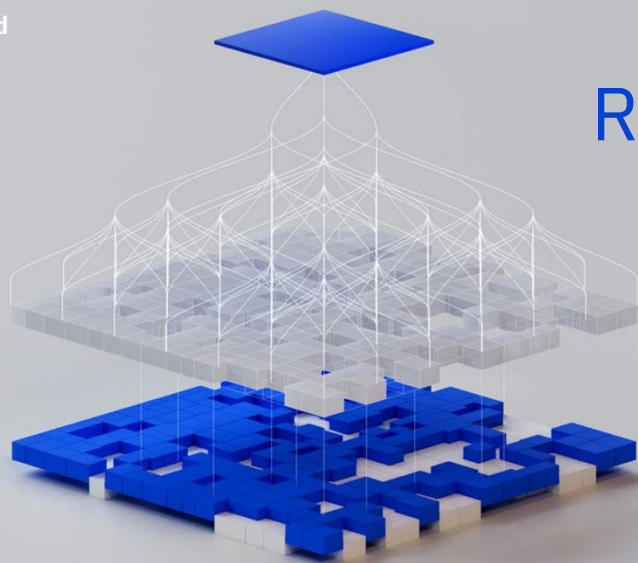


Red Hat AI on IBM Cloud

Accelerate AI Adoption and Innovation
for IT Leaders



Highlights

Red Hat AI on IBM Cloud is a portfolio of products and services to build and run customized AI solutions across hybrid cloud environments

[Red Hat AI Inference](#) on IBM Cloud providing fast, consistent and cost-effective inference for real-time and agentic workloads

[OpenShift AI](#) Kubernetes based MLOps platform for managing the AI/ML to expand and scale across distributed cluster environments

[RHEL AI](#) a foundation model platform for LLMs delivered as a bootable image.

Enterprises across industries are constantly exploring new ways that artificial intelligence (AI) can support business innovation and growth. According to research from IBM, 79% of executives say AI will significantly contribute to their revenue by 2030, but only 24% can clearly see where that revenue will come from. Not only that, 69% of CEOs say AI is already changing the aspects of their business they consider core. Still, many organizations lack the infrastructure and skills to successfully manage the data and deployment challenges that come with these compute intensive workloads. Many organizations lack the infrastructure and skills to manage not only the data and training pipelines, but also the production-grade inference capabilities required to operate GenAI reliably. As adoption grows, enterprises face challenges such as inconsistent latency under high concurrency, unpredictable cost growth tied to token usage, and the operational complexity of running inference across heterogeneous hardware and environments.

IBM Cloud is helping our clients boost competitiveness with the combined power of our [hybrid cloud](#) and [AI](#) stack. Our enterprise cloud platform is built for the most regulated industries and coupled with our heritage in high performance computing (HPC). IBM Cloud is uniquely positioned to support the compute-intensive nature of GenAI—particularly the scalable, low-latency inference workloads that drive real-world business value.

In collaboration with Red Hat and responding to our client's needs, IBM brings Red Hat AI on IBM Cloud, a portfolio of enterprise-built products and services that accelerates time to market and reduces the operational cost of delivering AI solutions across hybrid cloud environments. It enables secure, efficient creation of small, fit-for-purpose models using the clients' own enterprise-relevant data and provides the flexibility to deploy them wherever the business need resides. With integrated inference capabilities, organizations can overcome deployment constraints, eliminate idle GPU waste, manage demand spikes, and deliver consistent performance for both real-time and agentic AI applications.

Red Hat AI on IBM Cloud aids organizations to manage and monitor the lifecycle of both predictive and gen AI models at scale, from single-server deployments to highly-scaled out distributed platforms. This includes distributed inference to maximize throughput and cost-efficiency, along with telemetry for visibility into token usage, GPU utilization, and model behavior. The portfolio is powered by open-source technologies and a partner ecosystem that focuses on performance, stability, and GPU support across various infrastructures.

There's no such thing as an all-purpose gen AI model

Generative AI helps organizations move faster with precision and agility—if they use the right model, running in the right environment, for the right purpose.

While technology leaders are best positioned to decide which gen AI model to use where, understanding the pros and cons of different model types—as well as where the competition is headed—helps CEOs make more informed investment decisions.²

IBM IBV

Red Hat AI on IBM Cloud portfolio includes **Red Hat AI Inference on IBM Cloud** as a service (SaaS) for fast, consistent and cost-effective inference for real-time and agentic workloads, **Red Hat OpenShift AI on IBM Cloud** for building, deploying, and managing AI-enabled applications at scale, and **Red Hat Enterprise Linux (RHEL) AI** for individual Linux server environments.

AI Infrastructure Use Cases

Enterprises are exploring a wide range of generative AI applications to improve customer engagement, automate complex processes and unlock insights from large volumes of unstructured data. To support this diversity, organizations require an AI platform that can handle everything from model development and tuning to high-performance inference and governance across distributed environments. Red Hat AI on IBM Cloud provides this foundation, enabling secure, scalable and efficient deployment of AI across real-time, high-throughput and regulated workloads. The following use cases demonstrate how the platform helps teams operationalize AI at scale—regardless of whether models run on a single server or across a distributed hybrid cloud.

- **Real-time copilots and chatbots**

Build and serve intelligent assistants for customer and employee interactions. Red Hat AI enables secure, responsive conversational experiences across hybrid cloud environments, supporting enterprise access controls and model governance.

- **Agentic AI and workflow automation**

Run or build models that support planning, reasoning, and tool-orchestration. The Red Hat AI portfolio provides the foundation for building agent-driven applications, from data pipelines and model tuning to scalable serving and integration.

- **Knowledge tasks and enterprise search**

Support summarization, retrieval-augmented generation (RAG), content processing, and search across large document sets. Through optimized model execution and lifecycle management, teams can deliver high-throughput, low-latency knowledge applications.

- **Regulated industry AI**

Meet the strict security, governance, privacy, and audit requirements of financial services, healthcare, public sector, and other regulated industries. Red Hat AI enables policy-controlled model creation, deployment, and serving anywhere the data resides.

- **Model optimization and cost-efficient scaling**

Use model compression, right-sizing, and hardware optimization techniques to reduce compute needs and improve cost-efficiency across training, tuning, and inference.

- **Hybrid and heterogeneous deployment**

Run models consistently across cloud, on-premise, and edge environments with support for diverse accelerators and standardized APIs. Red Hat AI provides flexibility to deploy AI workloads wherever the business requires.

- **Multi-team, multi-application AI platforms**

Standardize AI services across teams or business units using shared infrastructure, governance, and model management. Platform and SRE teams can build scalable internal AI platforms that support multiple applications and lifecycle stages.

- **Vertical-specific AI applications**

Enable tailored GenAI solutions for industries such as retail, telecom, healthcare, and manufacturing with reliable performance, integrated governance, and flexible deployment options.



Red Hat AI Inference on IBM Cloud

Red Hat AI Inference on IBM Cloud delivers fast, consistent and cost-effective generative AI inference across hybrid cloud environments. Red Hat AI Inference on IBM Cloud is an enterprise-ready, fully managed inference service. It extends the core Red Hat AI Inference product—part of the Red Hat AI portfolio—to a managed delivery model on IBM Cloud, preserving the same hardware-agnostic, hybrid-cloud operational characteristics. Lets clients run production-grade AI models without managing GPUs, infrastructure, or AI platforms. It delivers inference at enterprise scale that just works, providing consistent performance, reliability, and availability, and is optimized for real-time and agent workloads. Built on Red Hat OpenShift AI and open APIs, it offers fast, cost-efficient AI without idle GPUs through automatic scale up/down to meet real workloads and consumption-based pricing aligned to usage.

Operated and maintained by IBM Cloud, the service removes the complexity of runtime optimization, GPU management, and platform upgrades, offering a fully governed, production-ready environment where enterprises can securely deploy, operate, and monitor inference workloads. It integrates vLLM for high throughput, low-latency serving, along with llm-d, a Kubernetes-native distributed inference engine designed for real-time performance during traffic surges and multi-cluster workloads.

Model deployment is streamlined through an optimized model repository with compression to reduce compute and costs while preserving accuracy. The catalog includes Granite 4.0 H Small (IBM), Mistral-Small-3.2-24B-Instruct, Llama 3.3 70B Instruct, GPT-OSS-120B, expanded with Nemotron-3-Nano-30B-FP8 for broader workload coverage.

Red Hat AI Inference service powered by Red Hat AI software and delivered on IBM Cloud's enterprise-grade infrastructure gives users the foundational, efficient, AI-driven, and secure capabilities to their infrastructure:

FOUNDATIONAL

- Powered by Red Hat AI software and delivered on IBM Cloud's enterprise-grade infrastructure, the service is purpose-built to integrate real-time inference directly into production workflows with predictable performance, security, and cost.
- Built on llm-d, a Kubernetes-native distributed inference engine, the platform delivers sustained low latency and high throughput even during traffic surges.
- Production-ready inference that just works with consistent performance, reliability, and availability, optimized for real-time and agent workloads.
- Open, industry-standard integration using an OpenAI-compatible API protocol and native Hugging Face integration for rapid adoption without client rework.
- Hybrid and future-ready foundation built on Red Hat OpenShift AI and open APIs, running consistently across cloud, on-prem, and hybrid while keeping your data under your control.
- Low lock-in middleware approach—do it your way, where you want it to run—avoiding proprietary workflows and enabling multi-team reuse.

EFFICIENT

- Automatic scale up/down to meet real workloads with consumption-based pricing so you don't pay for idle GPUs, eliminating upfront infrastructure investment.
- vLLM-powered high performance with advanced algorithms for high QPS serving, from single GPU to distributed multi-GPU, already comparable to NVIDIA TRT-LLM.

- The model catalog now includes enterprise-ready optimized models such as Nemotron-3-Nano-30B-A3B-FP8, expanding cost-efficient options for real-time and agentic workloads.
- Model compression and optimization to reduce compute and costs while preserving accuracy, leveraging a validated, optimized model repository ready for inference.
- SLA-backed reliability, full observability, and out-of-the-box integration via widely adopted APIs to accelerate time to value.

AI-DRIVEN

- Distributed inference runtime (vLLM/vLLM-d) delivering faster next-token prediction, continuous batching, and memory-efficient methods for consistent low latency at scale.
- Llm-d enables intelligent GPU utilization and efficient distributed scheduling across multi-cluster deployments, supporting high-throughput and mission-critical workloads.
- Built-in governance and trust with IBM Cloud IAM, enterprise-grade access controls, auditing, and privacy controls to support compliance (e.g., HIPAA, GDPR).
- Curated, right-sized model catalog (e.g., Granite 4.0 H Small, Mistral-Small-3.2-24B-Instruct, Llama 3.3 70B Instruct, GPT-OSS-120B, and Nemotron-3-Nano-30B-A3B-FP8) to match workload needs.
- Composable AI middleware services—AI Gateway/API management, model control plane, and orchestration—integrating easily with enterprise systems and agentic workflows.
- The service is explicitly designed to help enterprises scale real-time inference with secure performance and predictable cost, as emphasized in the press release.

SECURE

- Enterprise-grade governance with built-in IAM, role-based access controls, audit logging, and privacy controls to support regulated workloads and enforce consistent policies across hybrid environments.
- SLA-backed reliability on IBM Cloud's enterprise-grade infrastructure, with IBM-operated upgrades, patching, automated recovery, and runtime hardening to ensure secure and predictable production operations.
- Designed to securely integrate real-time inference into mission-critical workflows, with platform-level safeguards that protect data in motion and at rest, prevent unauthorized model access, and maintain compliance for industries such as financial services, healthcare, and the public sector.

Red Hat OpenShift AI on IBM Cloud

Large Scale AI Cluster

Red Hat OpenShift AI provides a complete AI platform for managing predictive and generative AI (gen AI) lifecycles across the hybrid cloud, including machine learning operations (MLOps) and LLMops capabilities. The platform provides the functionality to build predictive models and tune gen AI models, along with tools to simplify AI model management, from data science and model pipelines and model monitoring to governance and more.

Red Hat OpenShift AI on IBM Cloud is a managed cloud service that builds on the proven capabilities of Red Hat OpenShift to provide a trusted hybrid AI/ML platform for building, training, tuning, deploying, and monitoring AI-enabled applications and ML models in a secure manner, consistently and at scale across public cloud, on-premise, and edge environments.

RHEL AI offers enterprises the following benefits for Large Scale AI clusters:

- Distributed serving

Delivered through the vLLM inference server, distributed serving enables IT teams to split model serving across multiple graphical processing units (GPUs). This helps lessen the burden on any single server, speeds up training and fine-tuning and makes more efficient use of computing resources, all while helping distribute services across nodes for AI models.

- An end-to-end model tuning experience

Using InstructLab and Red Hat OpenShift AI data science pipelines, this new feature helps simplify the fine-tuning of LLMs, making them more scalable, efficient and auditable in large production environments while also delivering manageability through the Red Hat OpenShift AI dashboard.

- AI Guardrails

Red Hat OpenShift AI 2.18 helps improve LLM accuracy, performance, latency and transparency through a technology preview of AI Guardrails to monitor and better safeguard both user input interactions and model outputs. AI Guardrails offers additional detection points in helping IT teams identify and mitigate potentially hateful, abusive or profane speech, personally identifiable information, competitive information or other data limited by corporate policies.

- Model evaluation

Using the language model evaluation (lm-eval) component to provide important information on the model's overall quality, model evaluation enables data scientists to benchmark the performance of their LLMs across a variety of tasks, from logical and mathematical reasoning to adversarial natural language and more, ultimately helping to create more effective, responsive and tailored AI models.



Red Hat Enterprise Linux (RHEL) AI

Small Scale AI

RHEL AI is a foundation model platform to more consistently develop, test and run LLMs to power enterprise applications. RHEL AI provides customers with Granite LLMs and InstructLab model alignment tools that are packaged as a bootable Red Hat Enterprise Linux server image and can be deployed across the hybrid cloud.

RHEL AI offers enterprises the following benefits for small AI scaling:

- Empower innovation by efficient, open-source gen AI models

Red Hat Enterprise Linux AI offers organizations access to enterprise-grade, open-source Granite language and code models that are fully indemnified by Red Hat. The open-source Granite models provide organizations cost- and performance-optimized models that align with a wide variety of gen AI use cases. Granite models were released under Apache 2.0 License with transparency on datasets used for training and include the Granite 7b English language model and Granite 3b, 8b, 20b, and 34b code models.

- Streamline alignment of gen AI models to business requirements

Red Hat Enterprise Linux AI includes InstructLab model alignment tooling to help organizations more efficiently contribute skills and knowledge to their gen AI models to address the needs of their AI-enabled applications and business. InstructLab is more accessible to developers and domain experts that understand business requirements but lack the data science expertise normally required to tune models, allowing them to collaborate on this process and help realize business results faster.

- Train and deploy anywhere

Red Hat Enterprise Linux AI helps organizations accelerate the process of going from proof of concept to production server-based deployments by providing all the tools needed and the ability to train, tune, and deploy these models where the data lives, anywhere across the hybrid cloud. When organizations are ready, it also provides an on-ramp to Red Hat OpenShift AI, for training, tuning, and serving these models at scale across a distributed cluster environment using the same Granite models and InstructLab approach used in the Red Hat Enterprise Linux AI deployment.

Conclusion

In all, **Red Hat AI on IBM Cloud** is a comprehensive suite of products and services designed to accelerate the development and deployment of AI solutions across hybrid cloud environments. It helps organizations reduce operational costs and time to market using highly efficient custom AI models that are aligned with your own enterprise-relevant data, privately and securely.

Red Hat AI on IBM Cloud empowers organizations to manage and monitor the entire lifecycle of both predictive and generative AI models, whether deployed on single servers or large-scale distributed systems. It is built on open-source technologies, and leverages a partner ecosystem focused on optimizing performance, stability, and support for GPUs and AI accelerators.

For more information

To learn more about Red Hat AI on IBM Cloud, contact your IBM representative or IBM Business Partner, or visit ibm.com/products/redhat-ai.

Why IBM?

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs, and gain a competitive edge in their industries. Thousands of governments and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently, and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity, and service.

© Copyright IBM Corporation 2026
IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
May 2026

¹ IBM, Institute of Business Value (IBV) Jan 2026
The enterprise in 2030, <https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/enterprise-2030/>

² IBM, Institute of Business Value (IBV) Jan 2026
The CEO's guide to generative AI The CEO's guide to generative AI.
<https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ceo-generative-ai/ceo-ai-open-innovation>

IBM and the IBM logo are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on [ibm.com/trademark](https://www.ibm.com/trademark).

Red Hat®, OpenShift®, InstructLab™, and Linux®.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

