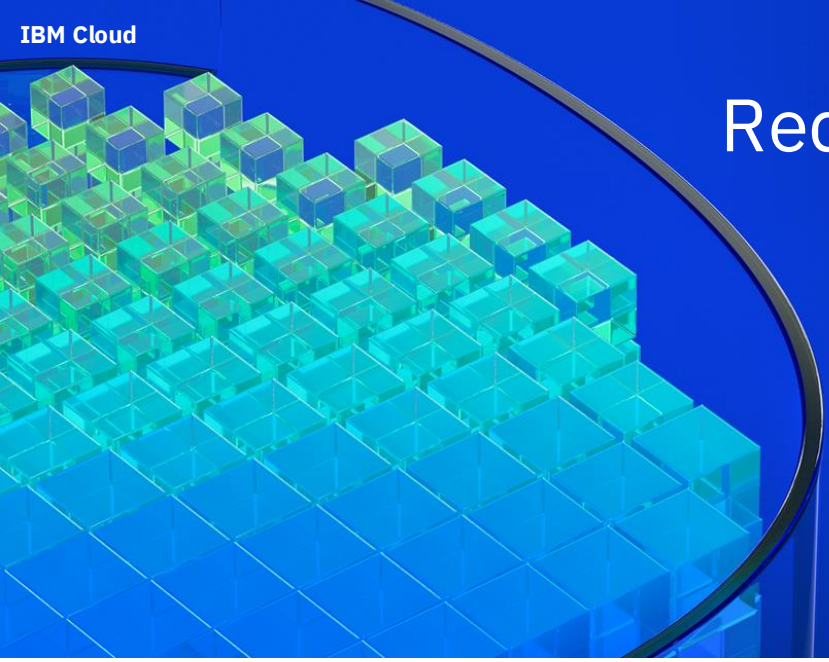


Red Hat AI on IBM Cloud

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 customed AI solutions across hybrid cloud environments

Red Hat AI InstructLab as a Service to democratize LLMs development to enterprises

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, executives say their organization needs to quickly adopt generative AI (Gen AI) to accelerate innovation. However, only 39% of organizations are currently implementing or operating generative AI for innovation and research[1]. Many organizations lack the infrastructure and skills to successfully manage the data and deployment challenges that come with these compute intensive workloads.

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 AI infrastructure needed for performance intensive workloads.

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.

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. The portfolio is powered by open-source technologies and a partner ecosystem that focuses on performance, stability, and GPU support across various infrastructures.

Red Hat AI on IBM Cloud portfolio includes **Red Hat AI InstructLab on IBM Cloud** as a service (SaaS) to build fit-for-purpose Large Language Models (LLMs) models, **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.

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 InstructLab on IBM Cloud

InstructLab fit-for-purpose AI modeling

Increasingly, enterprises are looking for AI solutions that prioritize accuracy and data security, while also keeping costs and complexity as low as possible. Red Hat AI InstructLab deployed as a service on IBM Cloud is designed to simplify, scale and improve the security footprint for the training and deployment of AI models. By simplifying InstructLab model tuning, organizations can build more efficient models tailored to the organizations' unique needs while retaining control of their data.

InstructLab on IBM Cloud allows enterprises to retain data and model ownership while minimizing risk of catastrophic forgetting.

Red Hat AI InstructLab on IBM Cloud offers enterprises the following open-source benefits for AI:

- Retained data and model ownership.
- Minimized risk of catastrophic forgetting.
- Simplified AI development and deployment
- Improved collaboration between data scientists and IT teams
- Enhanced security and compliance

Table 1. InstructLab market summary

	Existing commercial fine-tuning services	InstructLab. ai Open Source	Red Hat AI	Red Hat AI InstructLab on IBM Cloud
Modelling Accuracy and Scale	Lower accuracy, may not be production ready	Higher Accuracy , but requires customization	Higher Accuracy for production usage	Higher Accuracy for production usage
Minimize Risk of Catastrophic Forgetting	-	✓	✓	✓
Portability - Retain Ownership of Your Data and the Model	-	✓	✓	✓
Service model	SaaS	On-prem	On-prem & SaaS	SaaS

InstructLab as a Service (SaaS) only at IBM Cloud

InstructLab as-a-Service (SaaS)

Red Hat AI InstructLab on IBM Cloud delivers an AI model to the client that is customized to their business use case. Among the reasons to use an AI model as a service model, the following ones offer the most impact to enterprises:

- Reduced LLM fine-tuning complexity

Fine-tuning LLMs is a complex workflow process, including monitoring and optimizations. As a SaaS service, IBM Cloud provides a user-friendly interface, managed pipelines and expert support allowing enterprises to focus on business outcomes rather than AI infrastructure.

- Cost Efficiency and predictable Pricing

Running InstructLab Model alignment in-house requires substantial investment in infrastructure, specialized hardware like GPU's, and skilled personnel. A SaaS eliminates capital expenses and offers predictable, subscription-based pricing, democratizing advanced AI to enterprises by removing large upfront cost, freeing infrastructure and staff.

- Faster time-to-value

Enterprises can fine-tune and deploy models for real-world applications reducing procurement, installation, or configuration processes. It allows to foresee value faster and flexibility to changing business needs. IBM Cloud handles the complexity of set up, scaling and maintenance.

- Seamless Scalability and Reliability

Organization benefit from enterprise-grade uptime, performance, and security while IBM Cloud infrastructure is designed to handle increases in data volumes, user, or workloads.

- Continuous Operation

IBM Cloud value add is to continuously update system versioning, security patches and compliance standards.

Table 2. InstructLab as-a-Service (SaaS) market summary

	Existing commercial fine-tuning services	InstructLab.ai Open Source	Red Hat AI	Red Hat AI InstructLab on IBM Cloud
Dedicated hardware & software	-	-	Via IBM Cloud	✓
Always available infrastructure	-	-	Via IBM Cloud	✓
Secure by Default	-	-	Via IBM Cloud	✓
Up-to-date fully managed solution	-	-	Via IBM Cloud	✓

Red Hat AI InstructLab on IBM Cloud is the enterprise ready fit-for-purpose solution on the market.

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

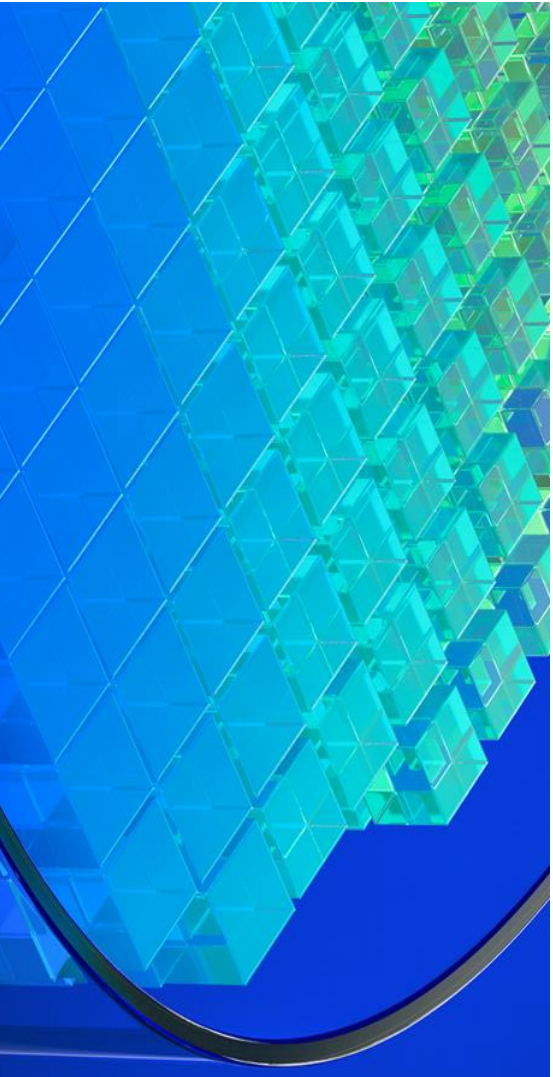
Red Hat Enterprise Linux AI helps organizations accelerate the process of going from proof of concept to production server-based deployments by providing all the tools needed and the ability to train, tune, and deploy these models where the data lives, anywhere across the hybrid cloud. When organizations 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/cloud/redhat.



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.

[1,2] IBM, Institute of Business Value (IBV) Nov 2023, 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 Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
May 2025