Red Hat AI 2025

Introduction to Red Hat AI

Red Hat AI is a portfolio of products and services that accelerates the development and deployment of AI solutions across hybrid cloud environments

Last Updated: 2025-05-28

Red Hat AI 2025 Introduction to Red Hat AI

Red Hat AI is a portfolio of products and services that accelerates the development and deployment of AI solutions across hybrid cloud environments

Legal Notice

Abstract

With Red Hat AI, organizations have the flexibility and consistency to deploy and manage both predictive and generative AI models wherever it makes the most sense for their AI workload strategy.

Table of Contents

CHAPTER 1. INTRODUCING RED HAT AI

Red Hat AI is a product and services portfolio that accelerates time to market and reduces the operational cost of delivering artificial intelligence (AI) solutions across hybrid cloud environments. With Red Hat AI, you can efficiently tune small, fit-for-purpose models by using enterprise-relevant data and to flexibly deploy models wherever your data is stored.

Red Hat AI helps you manage and monitor the lifecycle of both predictive and generative AI (gen AI) models at scale, from single-server deployments to highly 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.

With Red Hat AI, you can deploy and manage both predictive and gen AI models for your AI workload strategy. The portfolio supports each stage of the AI adoption journey, from initial single-server deployments to highly scaled-out distributed platform architectures. It also provides support for multiple hardware accelerators, original equipment manufacturers (OEMs), and cloud providers, to deliver a stable, optimized, and high performance platform across various infrastructures.

Access to the latest innovations is complemented by Red Hat's AI partner ecosystem, which offers tested, supported, and certified partner products and services that work with Red Hat technologies and help you solve your business and technical challenges.

Red Hat AI includes:

Red Hat Enterprise Linux AI

A foundation model platform for large language model (LLM) development, testing, and deployment with optimized inference capabilities.
Red Hat Enterprise Linux AI can support you at the beginning of your AI journey if you haven't defined your business use cases yet. The AI platform is built to develop, test, and run generative AI (gen AI) foundation models.

Red Hat OpenShift AI

An integrated MLOps platform that helps you manage your artificial intelligence and machine learning (AI/ML) lifecycle across hybrid cloud and edge environments, which helps you bring models from experimentation to production faster.
Red Hat OpenShift AI can support you if you are ready to scale your AI applications. This AI platform can help manage the lifecycle of both predictive and generative AI models across hybrid cloud environments.

Red Hat AI Inference Server

A container image that optimizes serving and inferencing with LLMs. Using AI Inference Server, you can serve and inference models in a way that boosts performance while reducing costs.

1.1. UNDERSTANDING RED HAT ENTERPRISE LINUX AI

Red Hat Enterprise Linux AI (RHEL AI) empowers you to customize and contribute directly to large language models (LLMs). RHEL AI is built from the InstructLab project, which uses a fine-tuning approach called LAB (Large-Scale Alignment for Chatbots). The LAB method uses synthetic data generation (SDG) with a multi-phase training framework to produce high-quality, fine-tuned LLMs.

You can install RHEL AI as a bootable Red Hat Enterprise Linux (RHEL) container image. Each image is configured for specific hardware accelerators, including NVIDIA, AMD, and Intel, and contains various inference-serving and fine-tuning tools.

You can use your own data to create seed files, generate synthetic data, and train a Granite starter model that you can interact with and deploy.

1.1.1. Key benefits of RHEL AI

1.1.1.1. Installation and deployment

- RHEL AI is installed using the RHEL bootable containerized operating system. The RHEL AI image contains various open source fine-tuning tools so you can customize the Granite starter models provided by Red Hat.

- RHEL AI provides images for you to deploy on bare metal, Amazon Web Services (AWS), Azure, IBM Cloud, and Google Cloud Platform (GCP).

- You can purchase RHEL AI from the AWS and Azure marketplaces and deploy it on any of their GPU-enabled instances.

- You can locally download, deploy, and chat with various models provided by Red Hat and IBM.

1.1.1.2. Model customization

- You can use the Synthetic Data Generation (SDG) process, where teacher LLMs use human-generated data to generate a large quantity of artificial data that you can use to train other LLMs.

- You can use multi-phase training, a fine-tuning framework where a model is trained on a dataset and evaluated in separate phases, called checkpoints. The final phase of training provides the most efficient, fully fine-tuned model.

- You can use various model evaluation benchmarks, including **MMLU**, **MT_BENCH**, and **DK_BENCH**.

1.2. UNDERSTANDING RED HAT OPENSHIFT AI

Red Hat OpenShift AI is a comprehensive MLOps platform designed to streamline artificial intelligence and machine learning (AI/ML) development and operations across hybrid cloud environments and the edge. It fosters collaboration between data scientists and developers while ensuring IT oversight, which empowers organizations to efficiently build, train, fine-tune, and deploy predictive and generative AI models.

Offered as a self-managed or cloud service, OpenShift AI builds on the robust foundation of Red Hat OpenShift, providing a trusted platform for securely deploying AI-enabled applications and ML models at scale—across public clouds, on-premises, and edge environments.

By leveraging a broad technology ecosystem, Red Hat OpenShift AI accelerates AI/ML innovation, ensures operational consistency, enhances hybrid cloud flexibility, and upholds transparency, choice, and responsible AI practices.

1.2.1. Key benefits of OpenShift AI

- **Simplified AI adoption**: Reduces the complexities of building and delivering AI models and applications that are accurate, reliable, and secure.

- **Enterprise-ready open source tools**: Provides a fully supported, secure enterprise version of open-source AI tools, ensuring seamless integration and interoperability.

- **Accelerated innovation**: Gives organizations access to the latest AI technologies, helping them stay competitive in a rapidly evolving market.

- **Extensive partner ecosystem**: Enables organizations to select best-of-breed technologies from a certified AI ecosystem, increasing flexibility and choice.

1.2.2. Features for data scientists, developers, and MLOps engineers

- **Integrated development environments (IDEs)**: Provides access to IDEs like JupyterLab, with pre-configured libraries like TensorFlow, PyTorch, and Scikit-learn.

- **Data science pipelines**: Supports end-to-end ML workflows by using containerized pipeline orchestration.

- **Accelerated computing**: Integrated support for GPUs and Intel Gaudi AI accelerators to speed up model training and inference.

- **Model deployment and serving**: Deploy models in a variety of environments and integrate them into applications by using APIs.

1.2.3. Features for IT operations administrators

- **Seamless OpenShift integration**: Leverages OpenShift identity providers and resource allocation tools for secure and efficient user management.

- **Accelerator management**: Enables efficient resource scheduling for GPU and AI accelerator usage.

- **Flexible deployment**: Available as a self-managed solution or as a managed service in Red Hat OpenShift Dedicated and Red Hat OpenShift Service on AWS (ROSA).

- **Scalability and security**: Provides enterprise-grade security features and governance controls for AI workloads.

1.3. UNDERSTANDING RED HAT AI INFERENCE SERVER

Red Hat AI Inference Server provides advanced inferencing features with enterprise-grade stability and security building on the open source vLLM project.

AI Inference Server uses continuous batching and tensor parallelism to provide reduced latency and higher throughput. Continuous batching processes model requests as they arrive instead of waiting for a full batch to be accumulated. Tensor parallelism distributes LLM workloads across multiple GPUs.

To reduce the cost of inferencing models, AI Inference Server uses paged attention. LLMs use a mechanism called attention to understand conversations with users. Normally, attention uses a significant amount of memory, much of which is wasted. Paged attention addresses this memory wastage by provisioning memory for LLMs similar to the way that virtual memory works for operating systems. This approach consumes less memory, which lowers costs.

Red Hat AI Inference Server has the following features:

- **Inference runtime for the hybrid cloud** Run your choice of models across accelerators, Kubernetes, and Linux environments.

- **LLM Compressor**: Compress models to optimize accelerator and compute usage. Reduce costs while maintaining high model accuracy.

- **Optimized model repository**: Gain access to a collection of optimized models ready for inference deployment, with support for both NVIDIA and AMD accelerators.
- **Certified for use with Red Hat products** Integrate with RHEL AI and OpenShift AI.