Latest updates: Feb, 2026

Red Hat AI

Introduction to Red Hat AI

Accelerate development and delivery of AI solutions

Presenter's Name
Title

Presenter's Name
Title

Red Hat

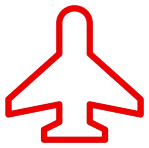
Gen AI adoption rates across the enterprise are rapidly increasing

Gen AI is a core part of the products we use, at home and at work

Frontier AI services grow new capabilities at a relentless speed

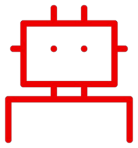
Artificial Intelligence innovation in action

Red Hat customers benefit from both predictive to generative AI



Enabling AI-innovation in the airline industry

Enabled teams to use AI in their everyday operations by reducing the time it takes to develop and deploy models to production environments.



Virtual assistants and RAG for environmental assessments

Streamlined document review and resolution, resulting in greater efficiency, minimized errors and reduced response times.

Comprehensive, mature and governed internal AI platform

Enabled teams to build AI models resulting in improved call center efficiency and accelerated development of paramount systems.

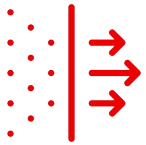
Sources:
Turkish Airlines Pioneers AI-led Innovation for Aviation with Red Hat OpenShift AI | Government of Castilla-La Mancha and Red Hat Power Public Services with AI | Red Hat Empowers Hitachi, Ltd. to Evolve into an AI-Driven Enterprise with Red Hat OpenShift AI

Red Hat

Generative AI customer adoption challenges

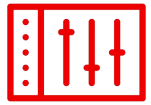
Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise customer use cases.



Complexity

Tuning models with private enterprise data for customer use cases is too complex for non-data scientists.



Control

Increasing concerns with data privacy, security, and latency are compelling organizations to adopt hybrid strategies.

Red Hat AI

Accelerate the development and delivery of AI solutions across hybrid-cloud environments

Increase efficiency with **fast, flexible and efficient inferencing**

Simplified and consistent experience for **connecting models to data**

Accelerate Agentic AI deployments

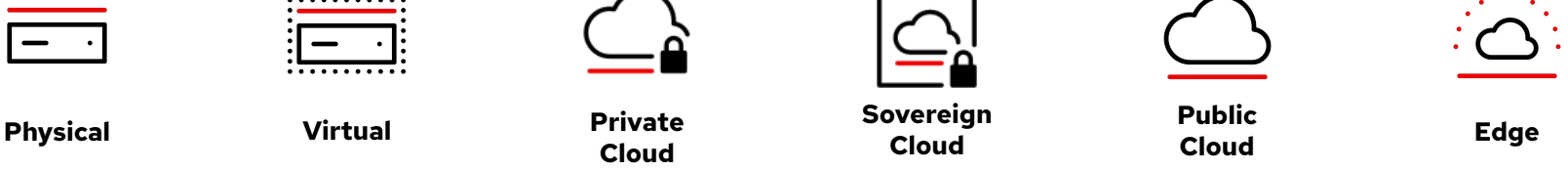
Flexibility and consistency when **scaling AI across the hybrid cloud**

Red Hat

Red Hat AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM



Red Hat

The products that power Red Hat AI

Red Hat AI meets customers where they are in their adoption journey providing choice of models, hardware, and location—whether on premise, in the cloud, or at the edge.

Red Hat AI Enterprise

For enterprise AI, deploy and scale efficiently, anywhere

Red Hat AI Inference Server

For optimized inference of large language models

Red Hat OpenShift AI

For the distributed deployments in OpenShift

Red Hat Enterprise Linux AI

For inference in individual Linux server environments

Red Hat AI: Customer Success

Accelerate Time-to-Market

Time to prepare a model development environment for **DenizBank** data scientists was slashed from **one week to just 10 minutes**

AI applications for **Clalit Health Services** went from request to production **in just two weeks**

The City of Vienna can reduce the time taken for certain processes from **15 minutes to 5 seconds**

Reduce Costs

ARSAT achieved a **30% reduction in operational costs** and expenses

Red Hat internally realized over **$5 million** in cost avoidance through AI-powered initiatives

Turkish Airlines improved fuel efficiency by 0.2% leading to **significant fuel consumption savings**

Increase Productivity

The Government of Castilla-La Mancha's AI assistant streamlines document review minimizing errors and **reducing response times by two months**

Hitachi operationalized AI across business units, resulting in **improved call center efficiency**

AGESIC reduced the time required to resolve service tickets to **less than 1% of the time** previously needed

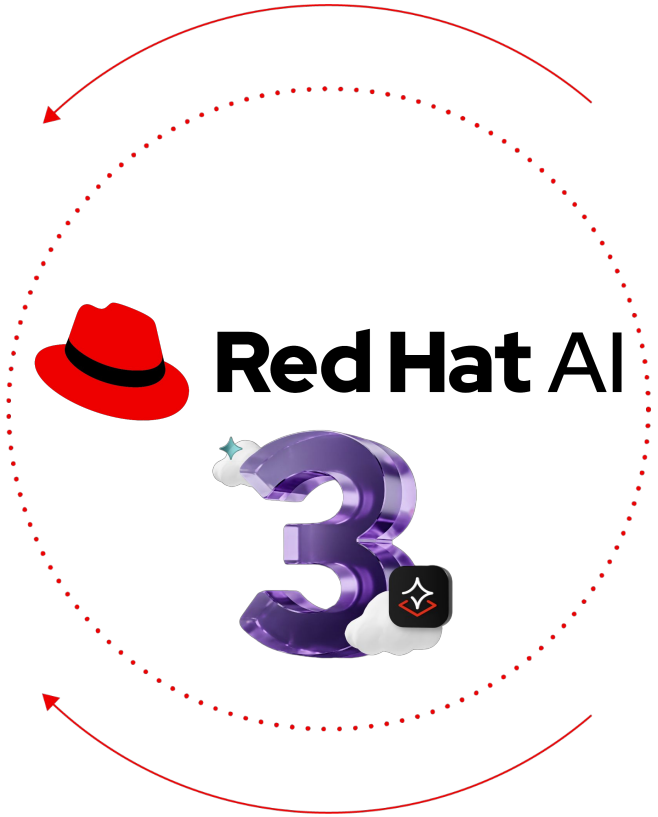
Red Hat

Flexible and Efficient Inference

- GA distributed inference (llm-d)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Models to Data

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Feature Store GA



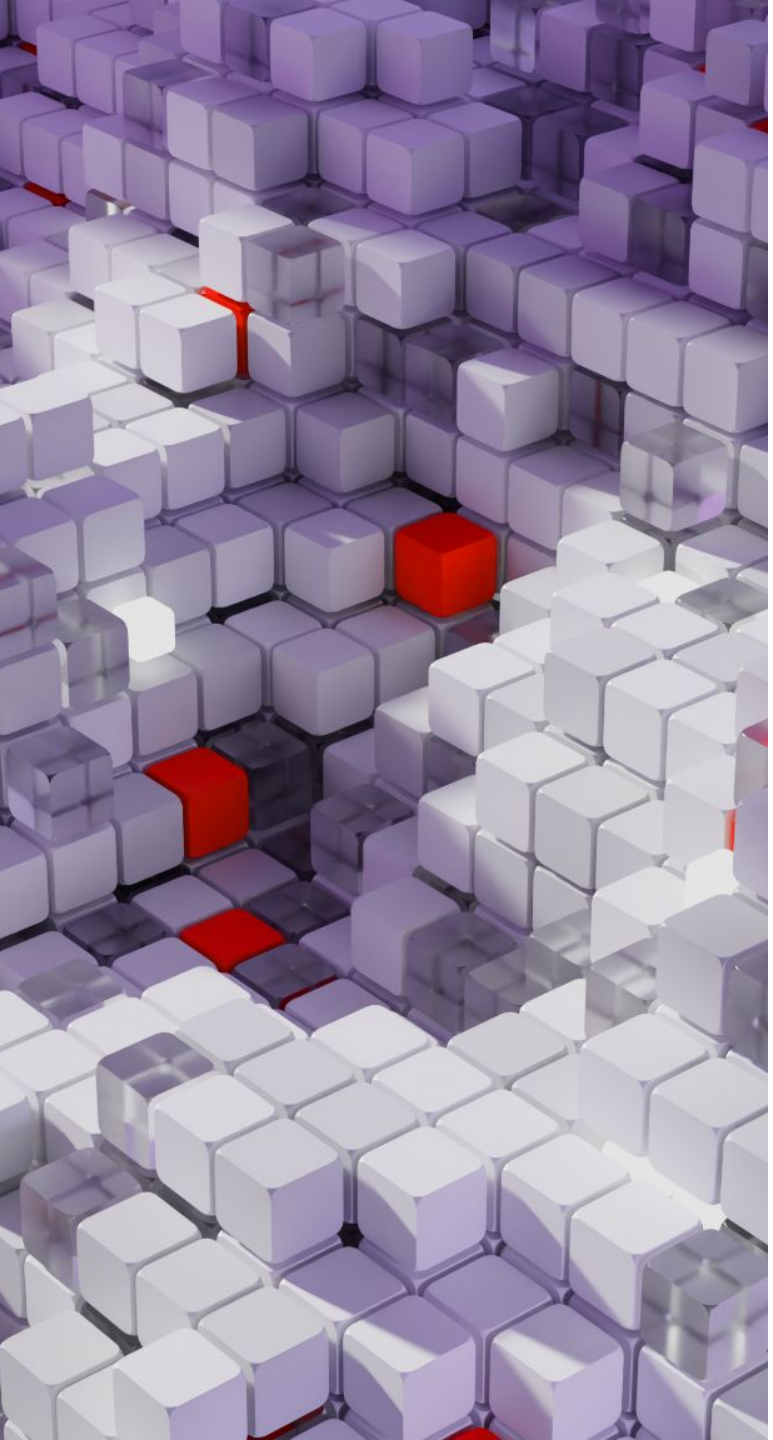
Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integration
- GPU as a Service enhancements

Single platform to run any model, on any accelerator, on any cloud



Fast, flexible and scalable inference

Red Hat

Expanding choice of models

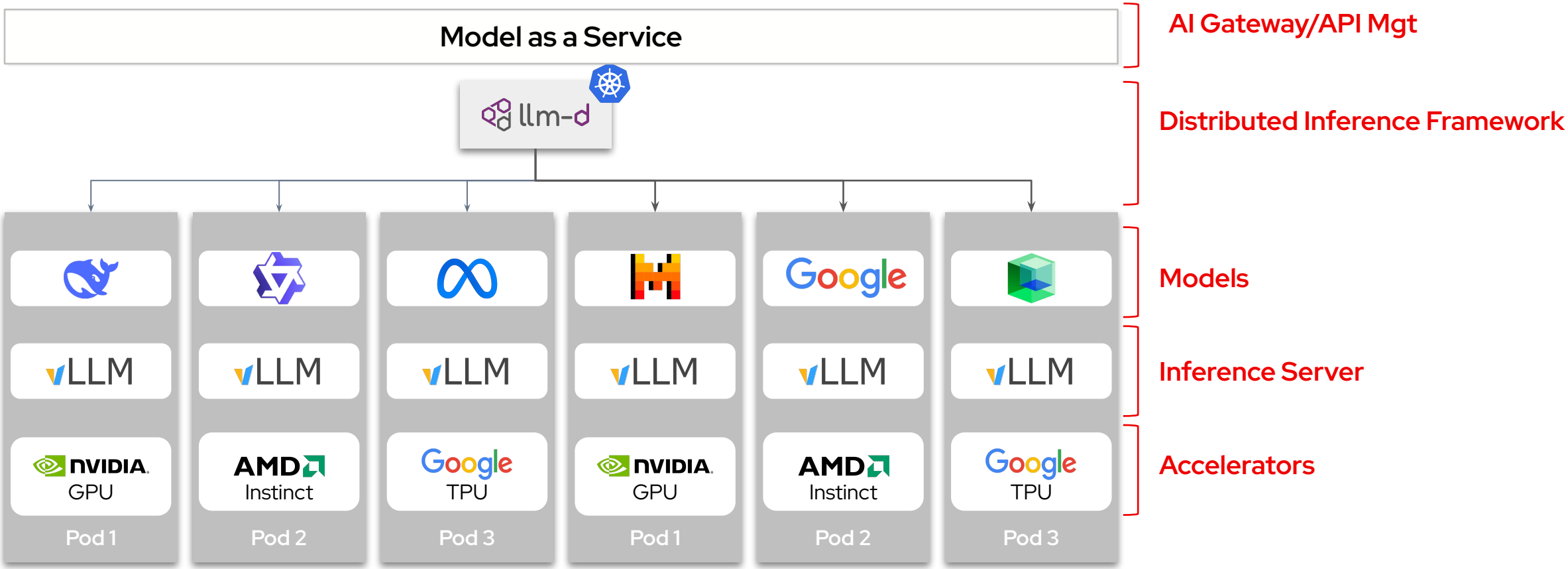
There has been an explosion of capability from open-source over the last 2 years

Meta · Ai2 · NVIDIA · MISTRAL AI_ · Hugging Face · databricks · snowflake · IBM · Google · Microsoft · deepseek

Red Hat

Enterprise GenAI inference platform

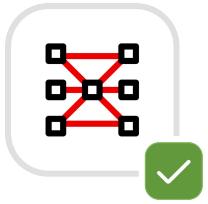
Holistic approach to optimize and operationalize deployment and scaling of open-source LLMs



Red Hat

Red Hat AI delivers consistent, fast and cost-effective inference

Select a large language model



A catalog of ready-to-use, third party validated and optimized models

Choose an inference runtime

vLLM

An optimized engine to deliver fast, cost-effective, and consistent inference

Choose the hardware that works best for you

AMD NVIDIA Google intel aws IBM

vLLM connects model creators to accelerated hardware providers

Scale AI inference when ready

llm-d

Llm-d provides consistent, distributed, inference at scale

Red Hat

Red Hat AI repository on Hugging Face

Collection of third-party models

Llama

Qwen

Gemma

Mistral, Voxtral

DeepSeek

Microsoft Phi

Molmo

Granite

NVIDIA Nemotron

OpenAI GPT-oss

KIMI K2

SMOLI M3 3B

Choice of Models

- Transformers (Dense, MOE), Multi-modal LLMs, Embeddings Models, Hybrid / Novel Attention, Vision
- Hugging Face compatible (safe tensors), OCI-compatible containers

Validated models

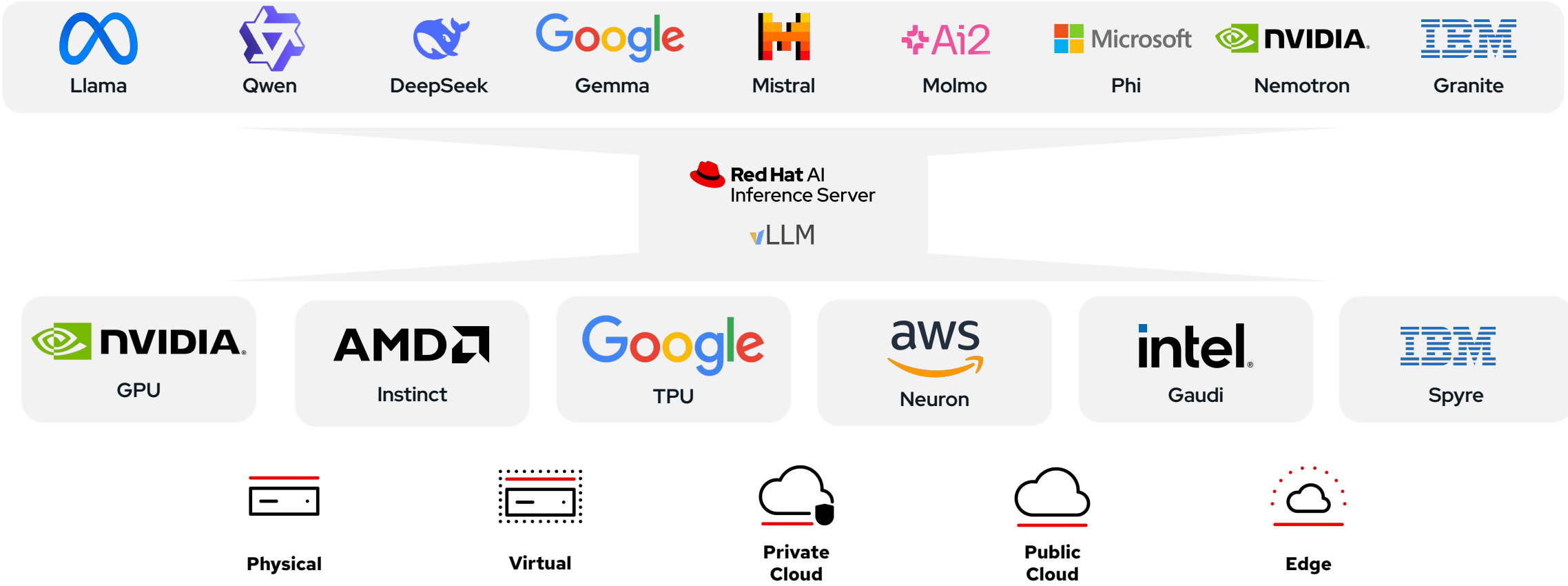
- Tested using realistic scenarios
- Assessed for performance across a range of hardware
- Done using GuideLLM benchmarking and LM Eval Harness

Optimized models

- Compressed for speed and efficiency
- Designed to run faster, use fewer resources, maintain accuracy
- Done using LLM Compressor with latest algorithms

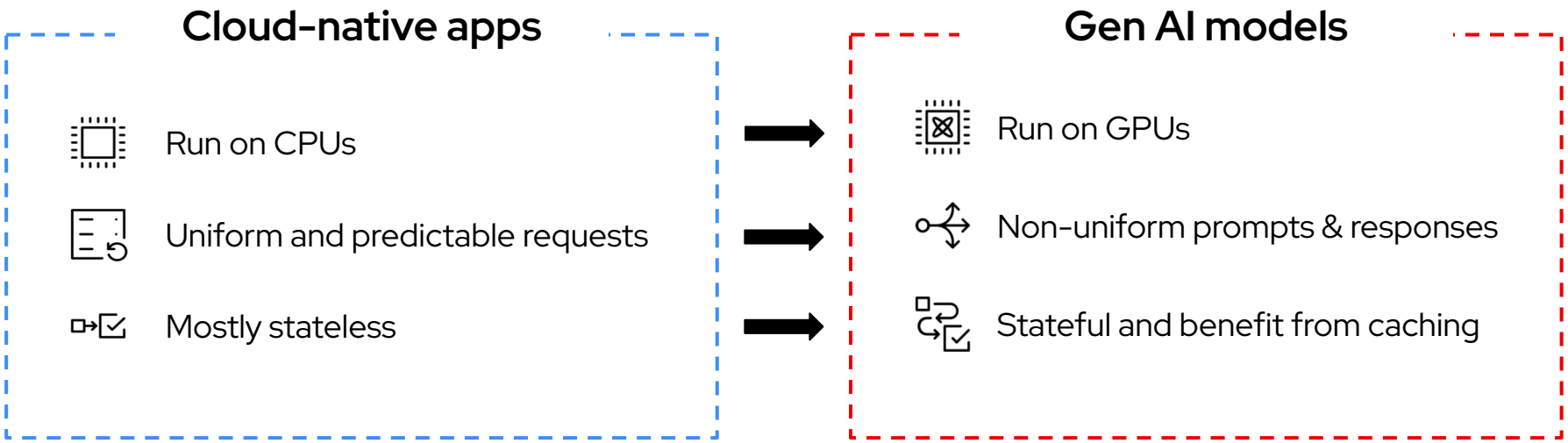
Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



What makes model workloads unique

Managing LLMs vs cloud-native apps in Kubernetes



Red Hat

Inference at scale everywhere

Distributed, scalable gen AI inference for Enterprise AI

Red Hat AI

Now includes llm-d

llm-d reimagines how LLMs run on Kubernetes

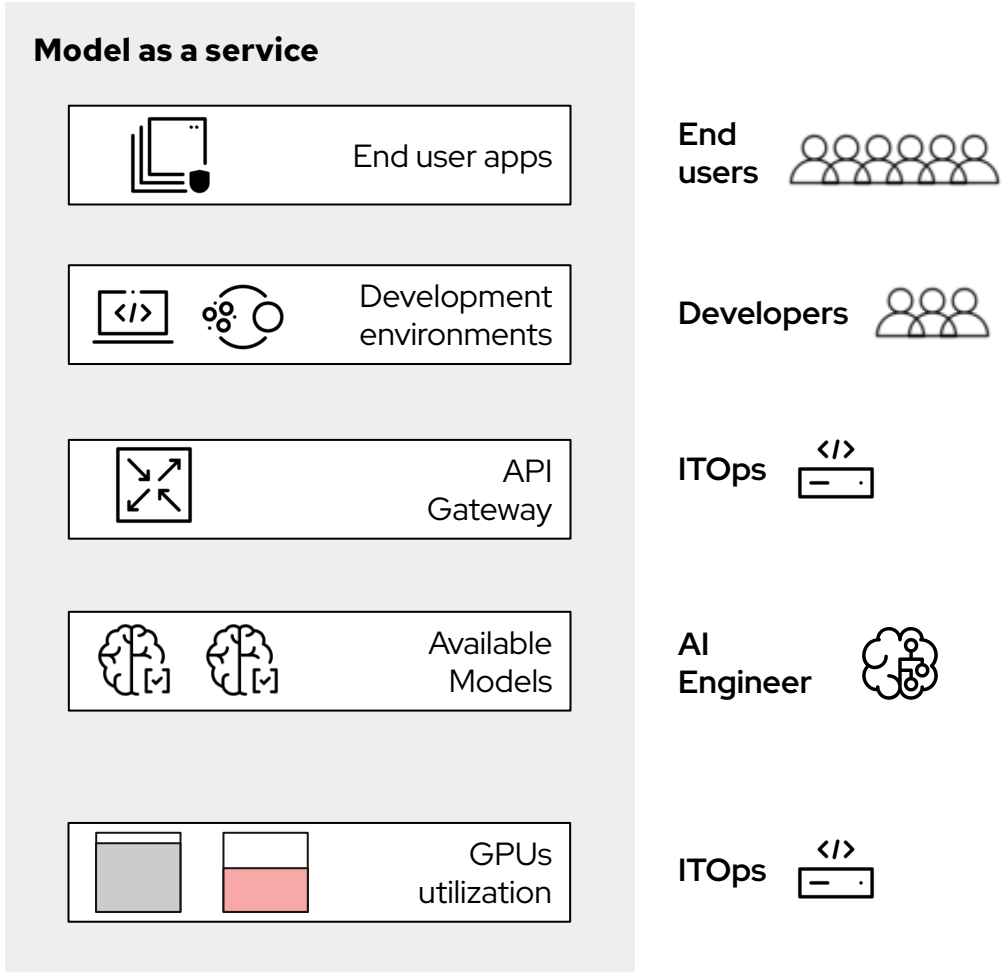
- Lower infrastructure cost & increased efficiency
- Faster response times for multi-turn & agent workloads
- Simplified management for platform administrators

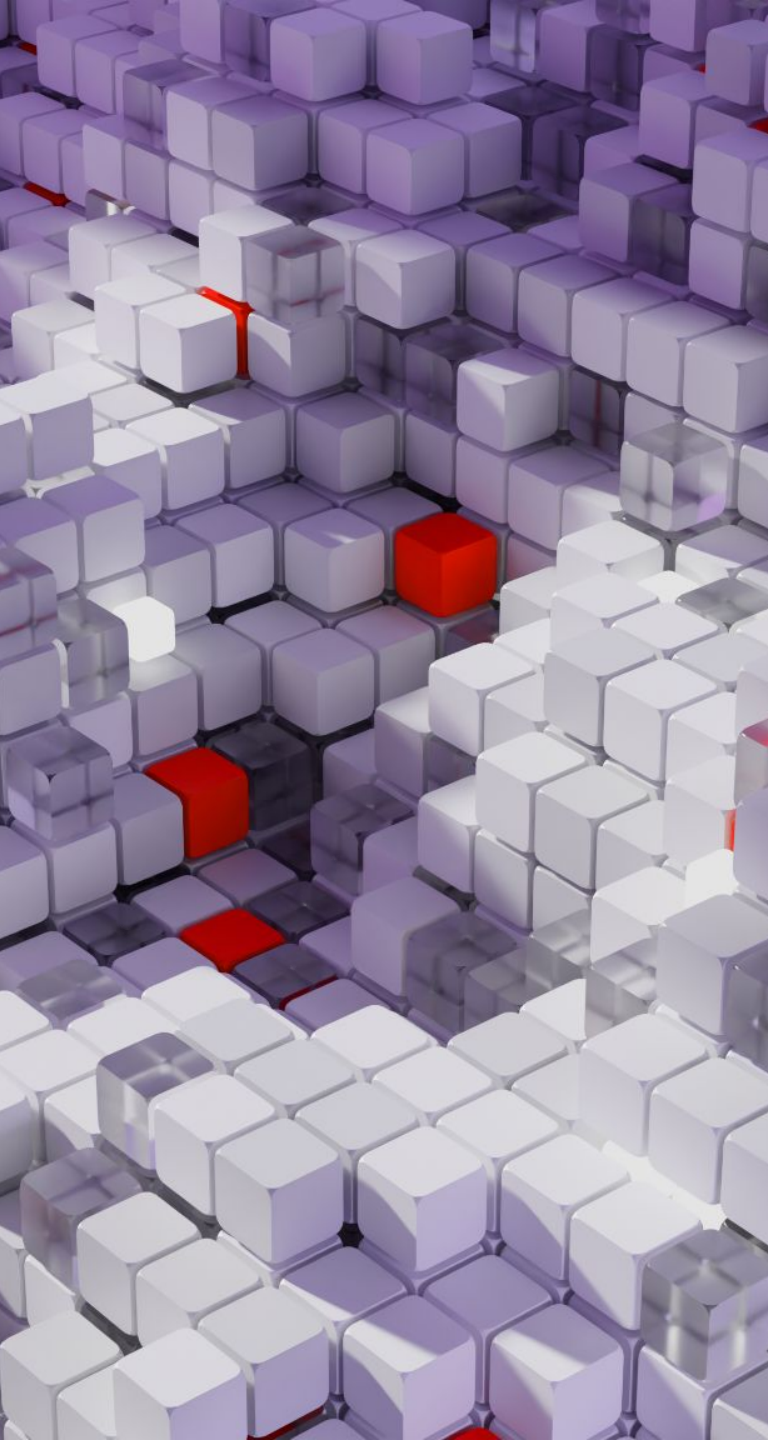
Deliver faster, cheaper, and more manageable AI systems for enterprise production

Model-as-a-Service (MaaS)

Offering AI **models as *the* service** to a larger audience

- **IT serves common models centrally**
  - Curated, open models available through your console
  - Centralized pool of hardware including GPUs
  - Dedicated UI to support Platform Engineering for AI
- **Developers consume models via APIs, build AI apps**
  - Models exposed through an API Gateway
  - For end users (AI assistants, etc)
  - Embed AI into existing products or services
- **Shared resources business model keeps costs down**
  - Enforce access policies
  - Chargeback and set quotas



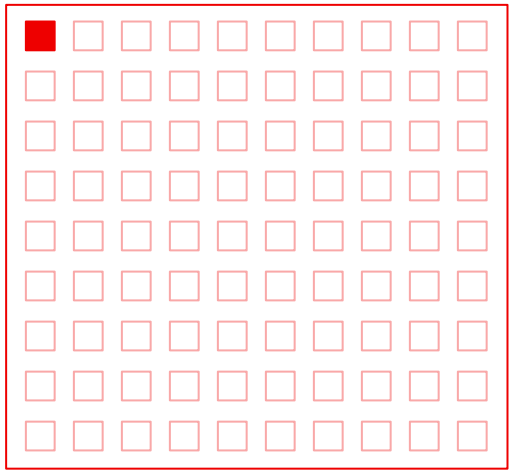


Connecting models to data

Red Hat

Enterprises need models aligned to their private data

LLMs are trained with a range of public data, not enterprise-relevant data

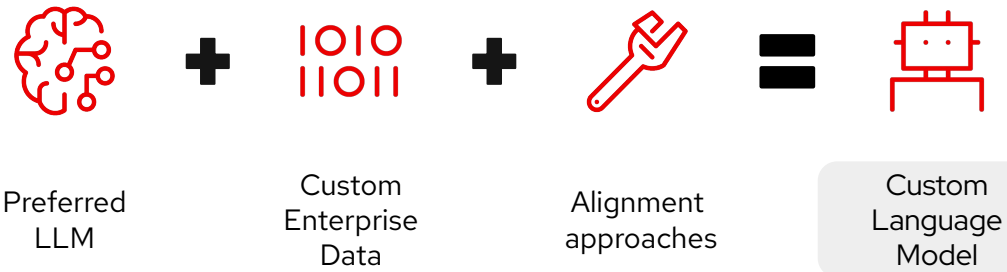


Less than 1% of all enterprise data is represented in foundation models

Enterprise organizations need to

1. Start from a trusted base model
2. Create a new representation of their data
3. Deploy, scale, and create value with their AI

Customize your preferred model using enterprise data to build an efficient, cost-effective solution.



Red Hat AI provides:

- ✓ Validated and optimized models ready-to-use
- ✓ Data ingestion capabilities
- ✓ Synthetic data generation pipelines
- ✓ Multiple alignment techniques

Red Hat

New model customization approach offers a modular flexible architecture

Data processing

Simplifies document processing and parsing into AI-readable data for model customization and RAG applications

Synthetic data generation

Generate high-quality data, with dynamic parameters, run-time visibility, and multilingual support

Training hub

An algorithm-focused interface for common llm training, continual learning, and reinforcement learning techniques

Evaluate

Simplifies the distributed execution of Evaluation jobs from popular eval frameworks or tasks

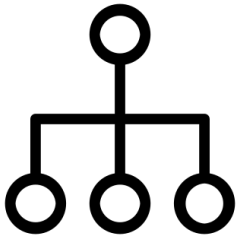
Multiple customization approaches

Build customized AI solutions that address domain specific business cases

Coming soon

Prompt design

Prompt tuning and engineering



Design and engineer the prompts to enhance GenAI model responses and achieve more specific and accurate outcomes.

Enhanced

RAG

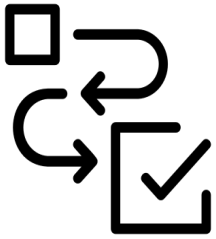
Retrieval Augmented Generation

Enhance Gen AI model generated text by retrieving relevant information from external sources, improving accuracy and depth of model's responses.

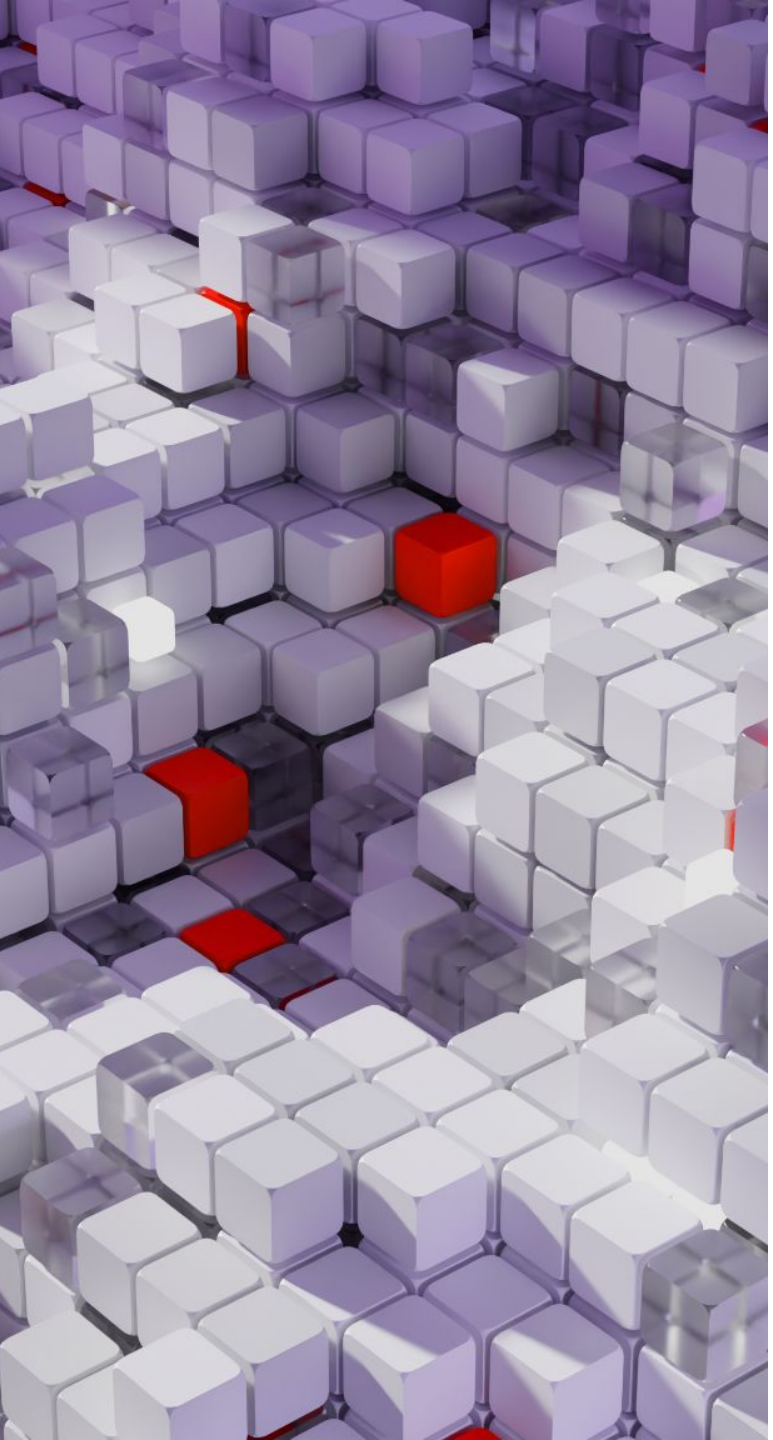
Enhanced

Fine tuning

InstructLab, OSFT, LoRA and QLoRA



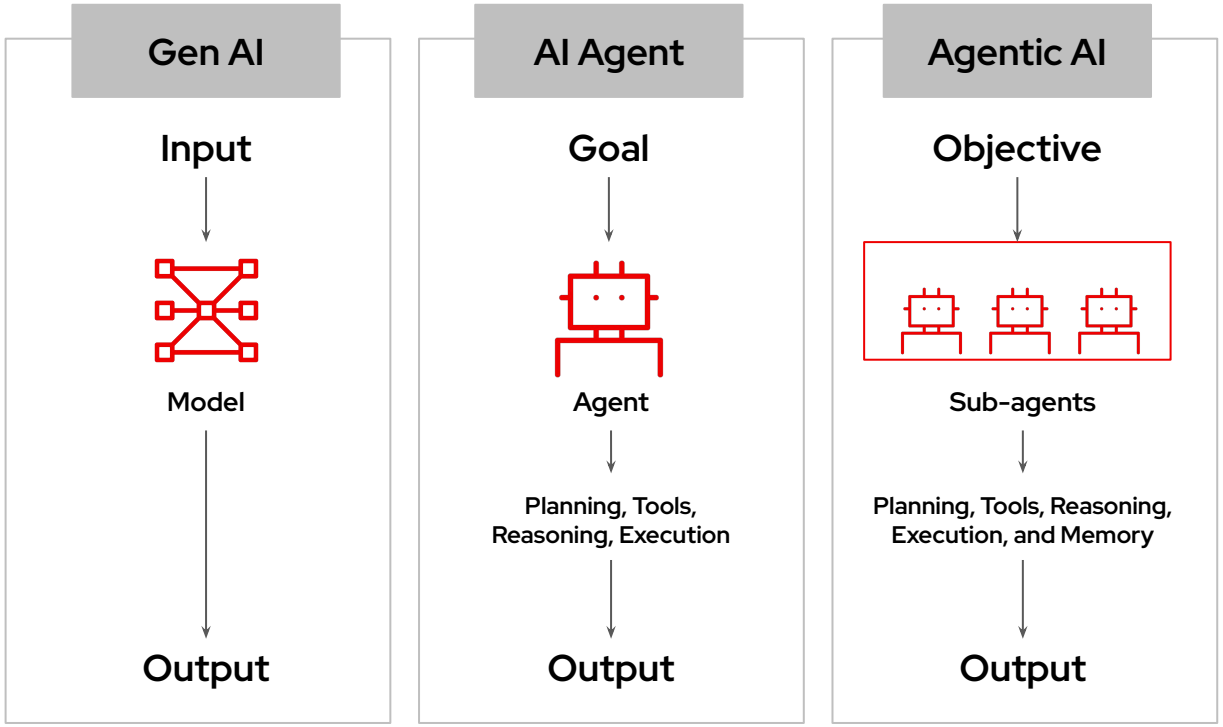
Customize a base model for specific tasks or private data, using a range of approaches—from full fine-tuning to parameter-efficient methods—to balance performance and efficiency.



Accelerate Agentic AI innovation

Red Hat

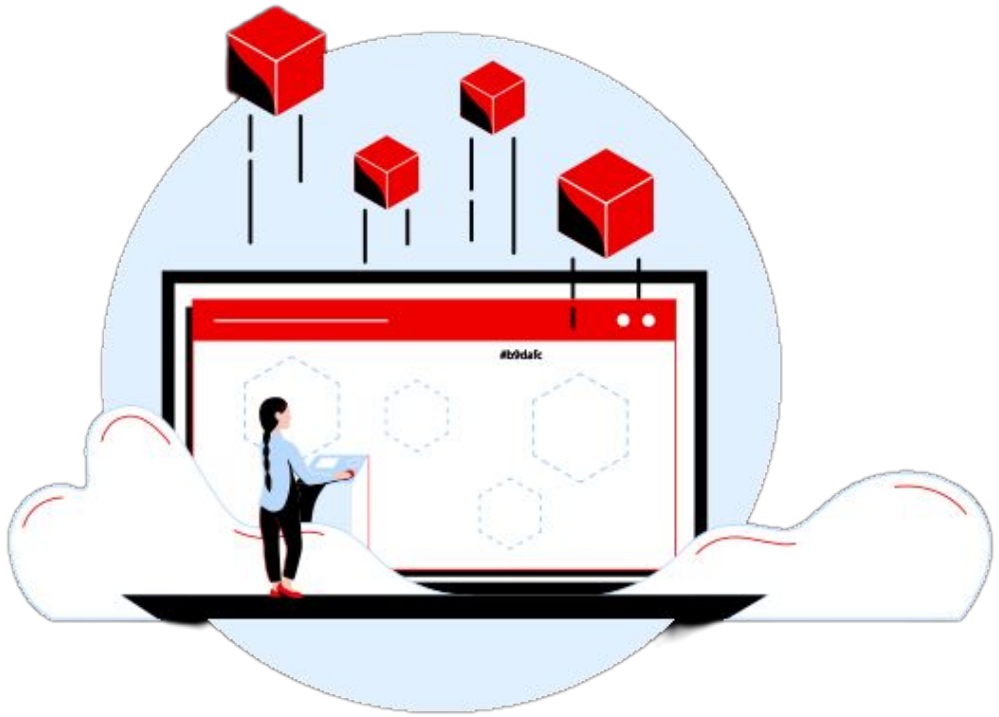
How does the next evolution in AI look?



- **GenAI is the creator.** Generates an output by following simple, predefined rules.
- **AI Agent is the doer.** Executes tasks and makes decisions.
- **Agentic AI is the manager.** It's the framework that enables multiple agents to collaborate and adapt to solve complex problems on their own.

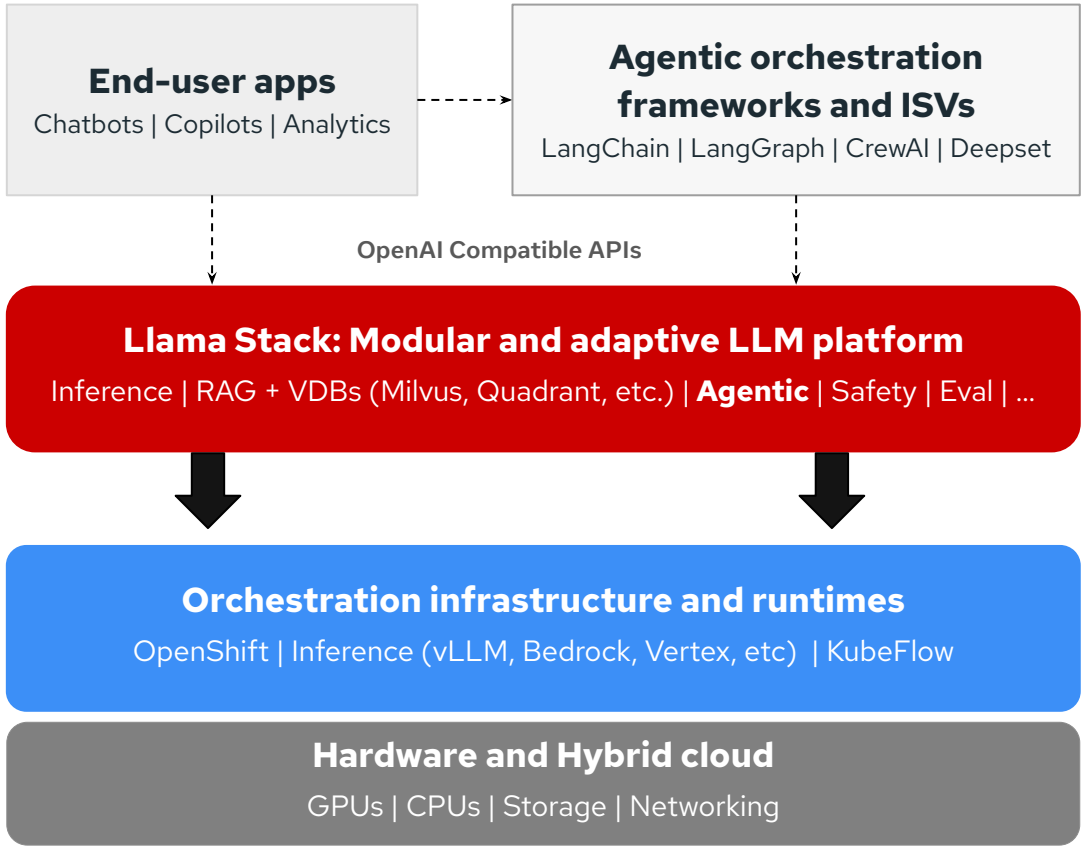
Red Hat AI provides an agile, stable foundation to accelerate the deployment and management of AI agentic workflows.

- Offers built-in agent frameworks with Llama Stack, and standardized communication protocols (MCP).
- Provides the flexibility to integrate preferred tools like LangChain and Crew AI.
- Allows running and managing agents as microservices.
- Simplifies production deployment by managing LLM serving and scaling.



Red Hat

Llama Stack: The Foundation Layer for Agentic AI



- **Enhanced flexibility.** Provides standardized APIs and a plug-in architecture for seamless integration with tools, hardware and external providers.
- **Simplified operations.** Reduces friction for developers and IT teams, making it easier to deploy, manage and scale agentic AI workflows.
- **Open and extensible architecture.** Enables users to mix and match best-in-class technologies, ensuring adaptability and control over their stack.

Model Context Protocol (MCP): standardize tool calling

Open standard for connecting AI apps and models to external systems

01

Creation Guide

Clear steps to design & configure MCP Servers

02

Try before you buy

Evaluate MCP Servers with loaded models in the Playground

03

AI assets

Quickly discover and launch what's ready to use

04

Llama Stack tool calling

Connect models to external tools (via **MCP servers**) or built-in tools.

Why it Matters?

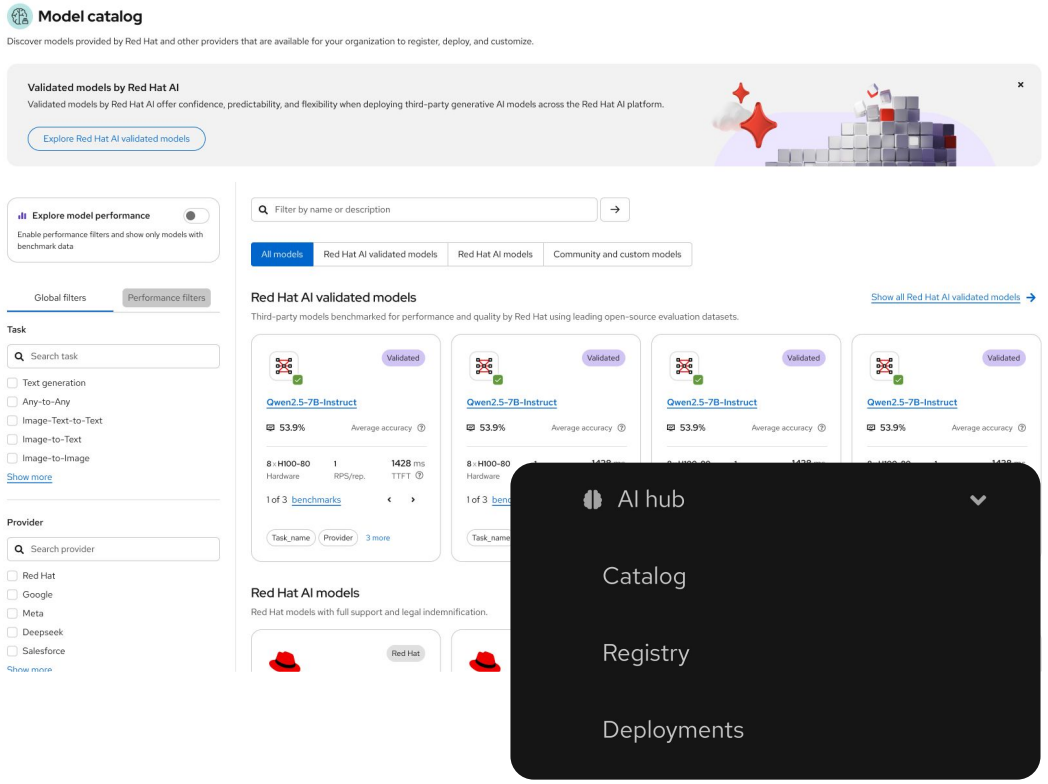
- Instant Experimentation: **Start building today!**
- Model and tool synergy: **Bring real-world capabilities to your AI workflows**
- Foundation for Red Hat AI 3: **Governed today, certified assets tomorrow**

Red Hat

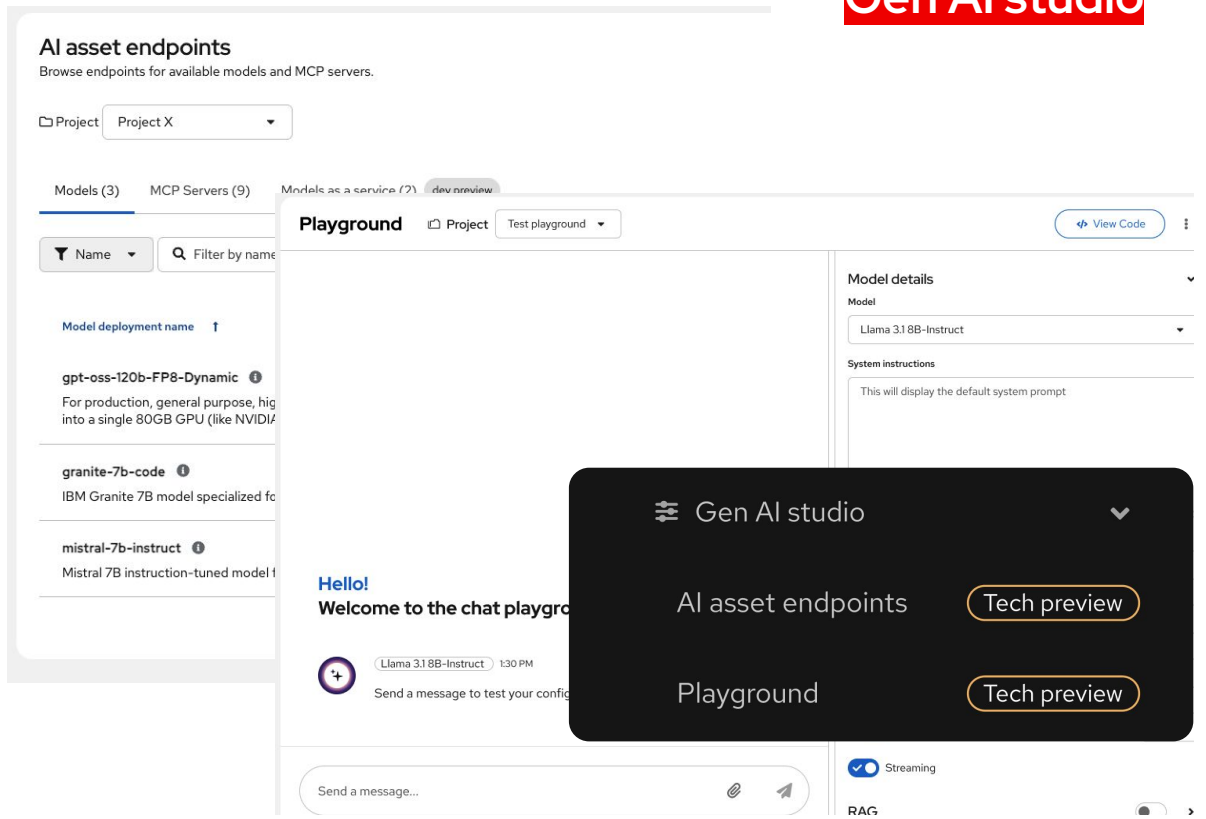
AI dedicated experiences

Dedicated dashboard experiences provide a seamless experience to platform and AI engineers

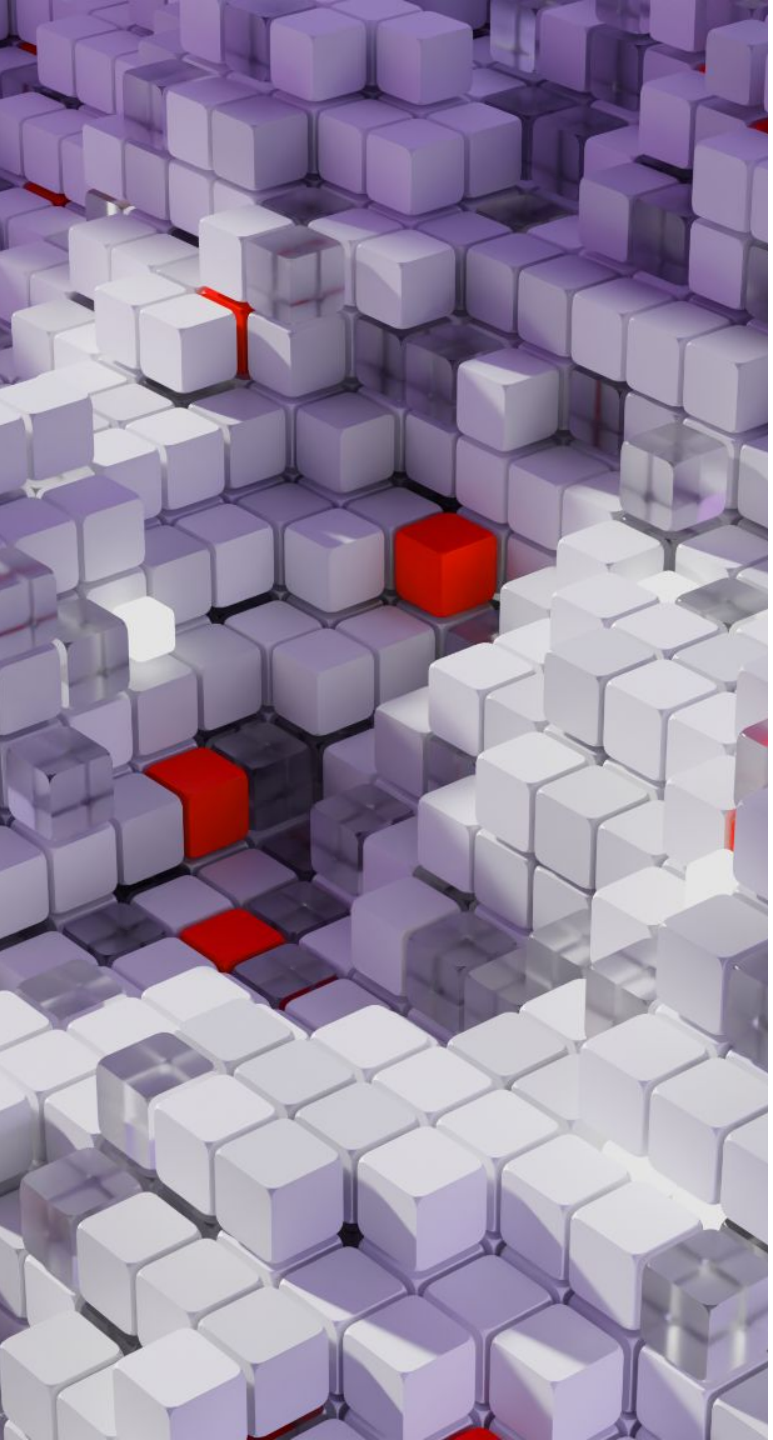
AI hub



Gen AI studio



Red Hat



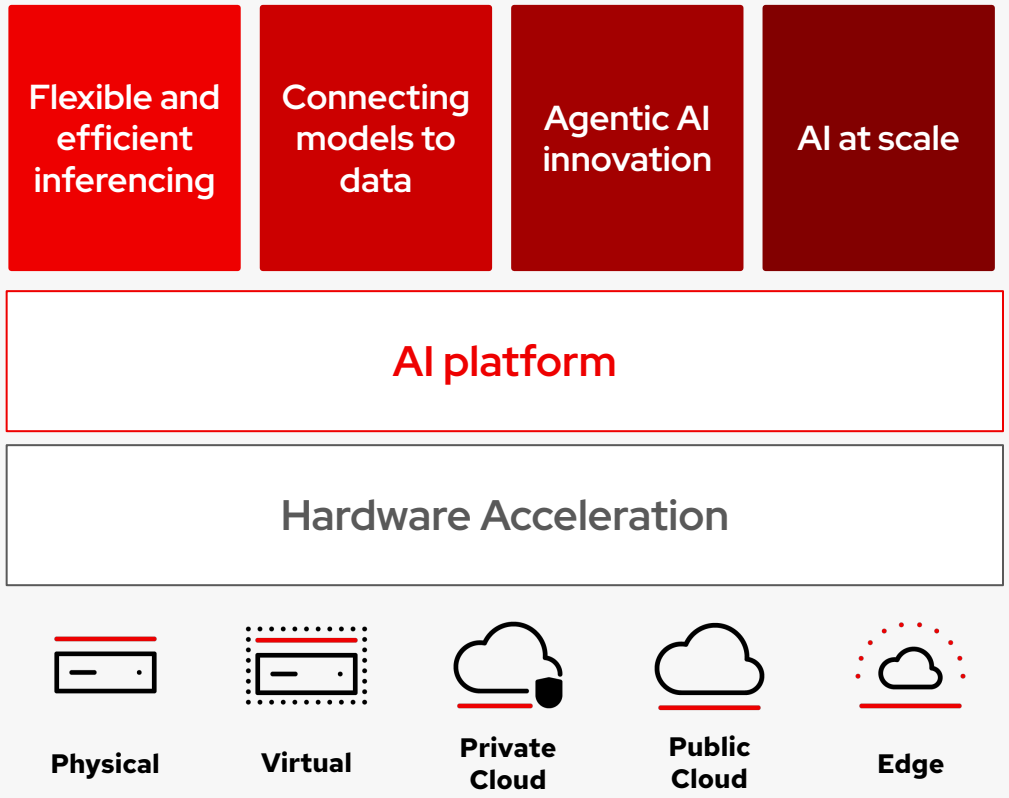
Scaling AI across the hybrid cloud

Red Hat

Red Hat AI provides a platform to consistently build, deploy and manage AI models and agentic apps at scale across the hybrid cloud

It includes:

- Enterprise-grade, flexible and secure AI platform
- Private, sovereign AI capabilities and practices
- Enhanced observability: platform metrics, zero-configuration GPU, and AI performance metrics
- Full MLOps and Gen AI Ops lifecycle support
- Model catalog and registry for increased reusability and governance
- Enhanced capabilities for GPU utilization: slicing, partitioning, scheduling, prioritization.



Hybrid cloud deployment for AI

Across different hardware accelerators, on-prem OEM servers, and cloud environments

Hardware Accelerators

OEM Servers

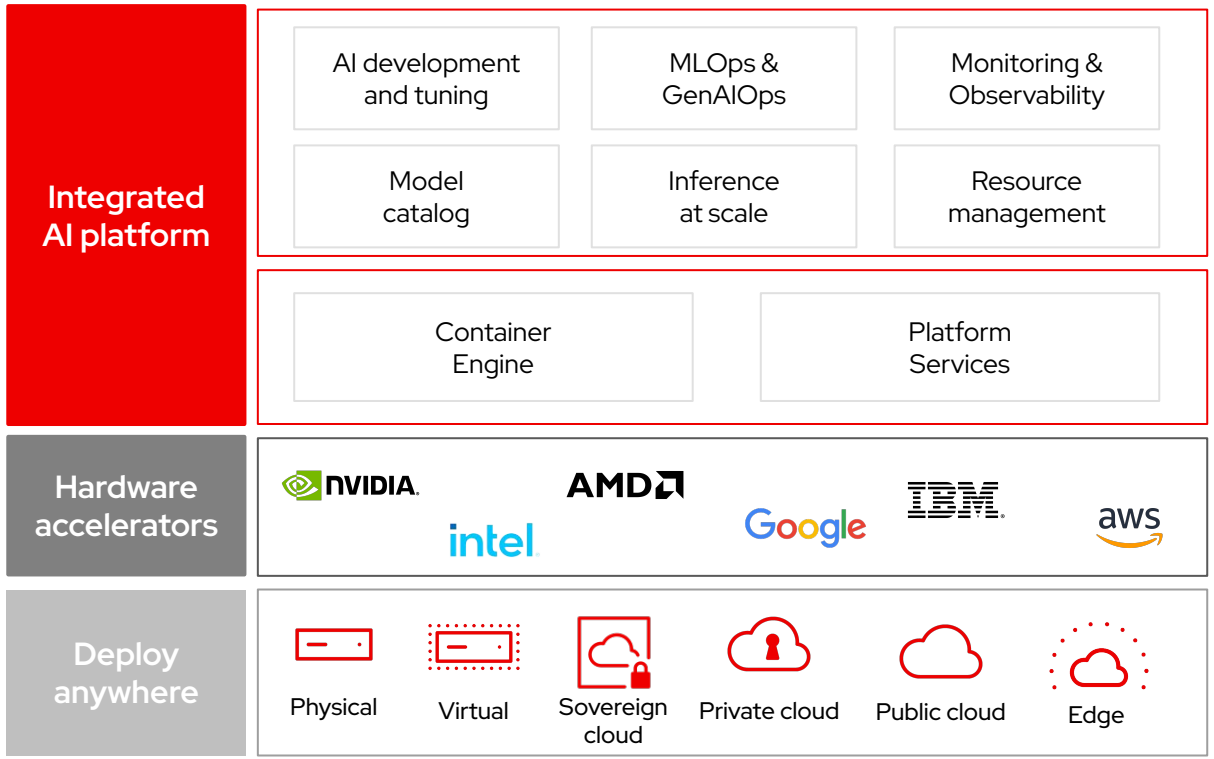
Cloud Environments

*NVIDIA, AMD and Intel are supported in Red Hat AI. Google TPU and IBM AIU supported in Red Hat AI Inference Server only and support for IBM AIU is coming in RHOAI 3.0. AWS Neuron is on on our roadmap.

Red Hat

Scale and optimize your AI and application deployments

Existing investments must work in support of AI

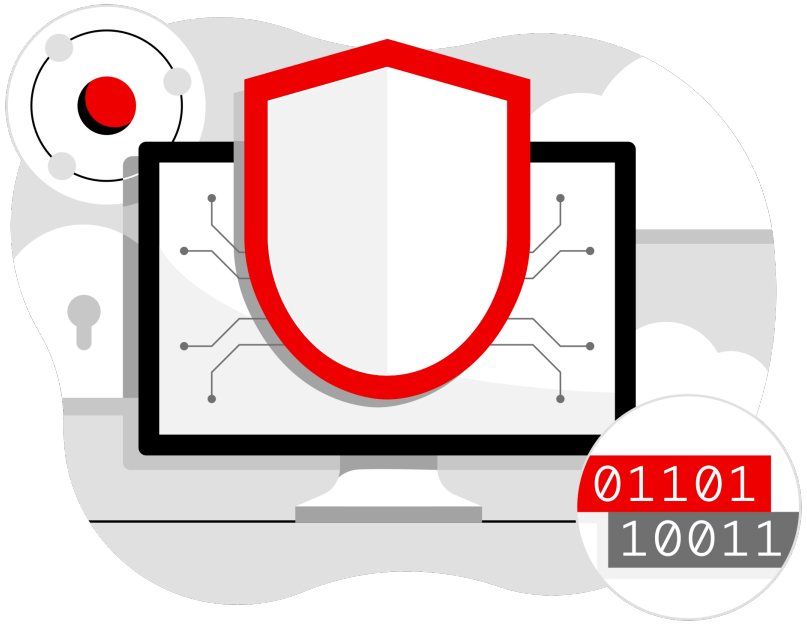


- **Integrate to real workflows** with access to data sources, workloads and applications.
- **Think of day 2 operations** for governance, management and automation.
- **Scale AI workloads dynamically** across hybrid cloud using Kubernetes, including horizontal and GPU scaling with automated resource management to meet fluctuating demands.

Red Hat

AI safety, monitoring and observability

Track accuracy, biases, performance, and more



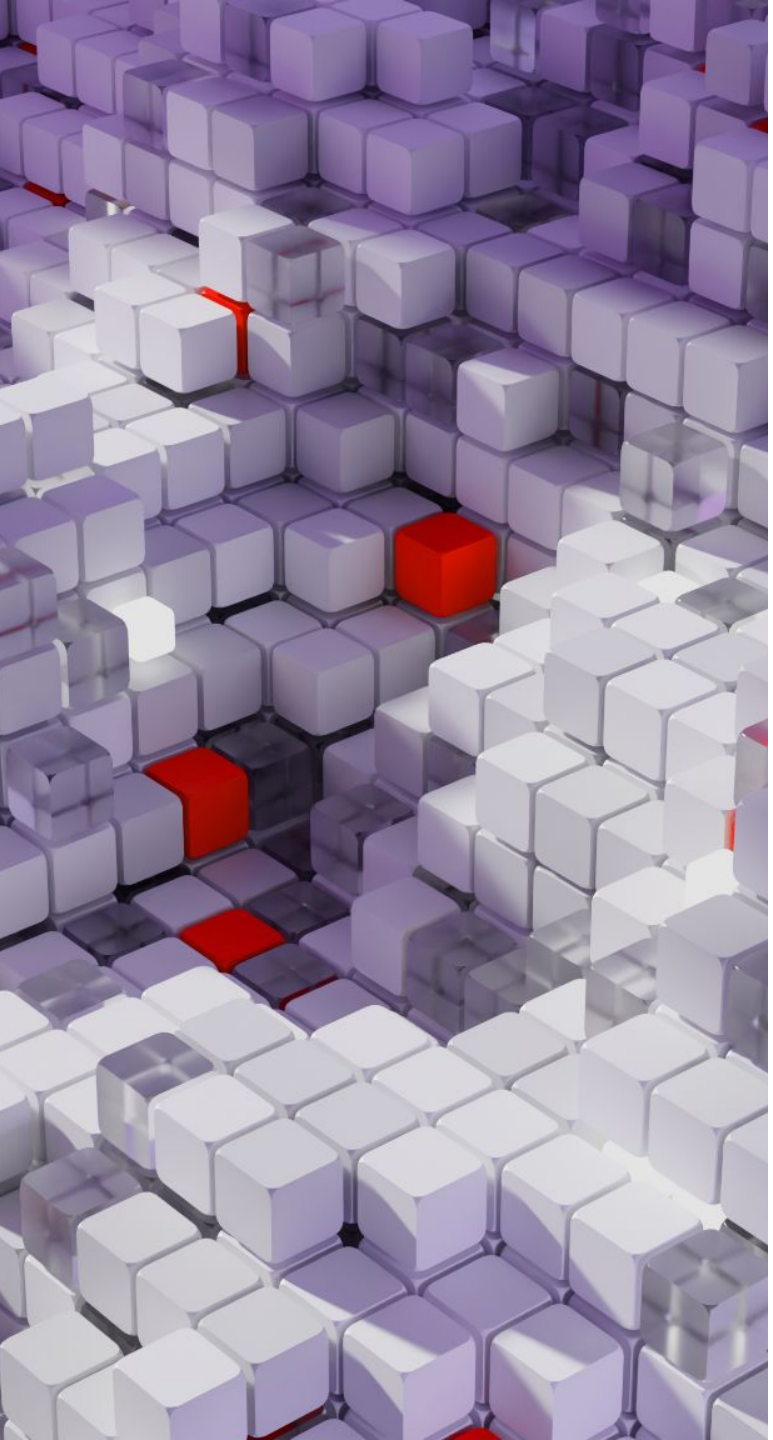
Detect bias and drift: Monitor for outcome disparities and training data differences.

Guardrails: A customizable framework to moderate interactions between users and generative AI.

Model monitoring: Track operations and performance metrics.

Accuracy evaluation: Measure model knowledge across topics.

Experiment tracking: Gain visibility and confidence when experimenting with models (visuals, metrics, UI/UX)

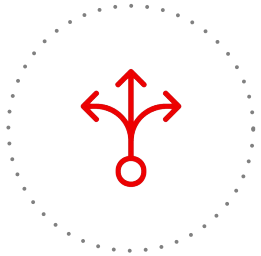


Why Red Hat AI?

Red Hat

Increasing flexibility and choice with an open source approach

Red Hat prioritizes investments on open source AI and building a certified AI partner ecosystem

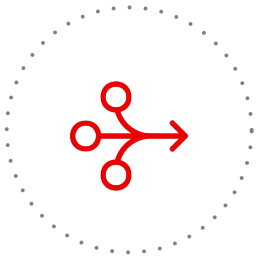


Flexibility

Access to cutting-edge open source innovations to keep up with a fast moving market.

Choice

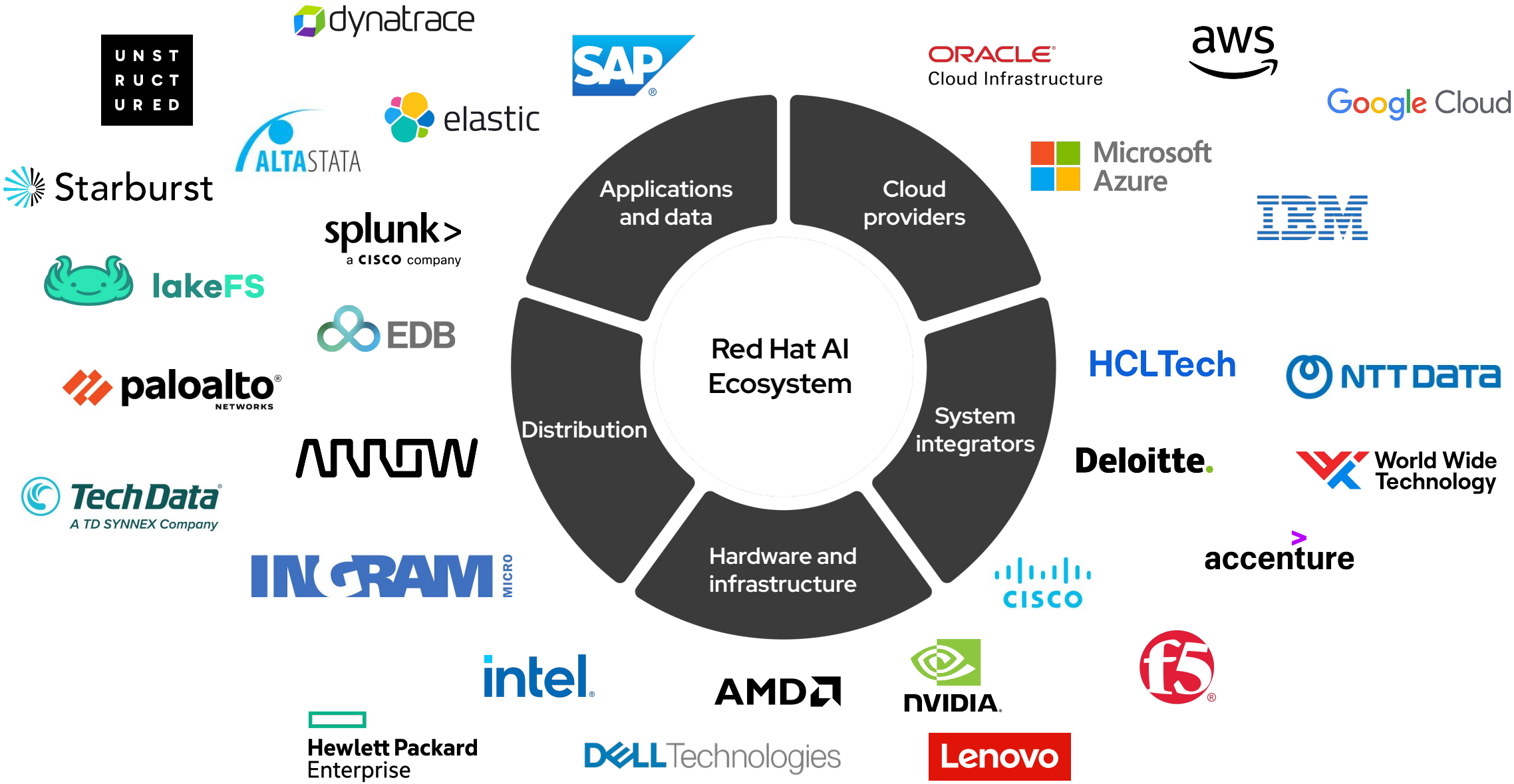
Access to an open ecosystem of communities, technology providers, ISVs and customers.



Abstract the complexity

Reduce the complexity of switching and adapting to new technologies.

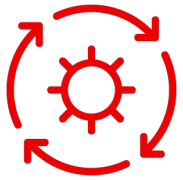
Red Hat



Note: This is not an exhaustive list of Red Hat AI partners, visit catalog.redhat.com for the full list.

Red Hat

The value of Red Hat AI



Increased efficiency

Reduce development and deployment costs with access to optimized open source models

Simplified experience

Enable developers, data scientists and domain experts to tailor models more efficiently

Flexibility to deploy anywhere

Mitigate risks, reduce costs, and scale your AI deployments across the hybrid cloud

Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce veteran suicides

U.S. Department of Veterans Affairs

Suicide has no single cause, and no single strategy can end this complex problem. That's why Mission Daybreak is fostering solutions across a broad spectrum of focus areas.

A diversity of solutions will only be possible if a diversity of solvers answer the call to collaborate and share their expertise.

Challenge

Develop new data-driven means of identifying veterans at risk for suicide

Solution

Red Hat teamed with global consulting services provider Guidehouse and Philip Held, Ph.D. of Rush University Medical Center, to develop a new data-driven means of identifying veterans at risk for suicide running on Red Hat technologies.

Results

- Allows providers to more **easily identify and help specific veterans in need,** using artificial intelligence and machine learning to sift through vast volumes of data.
- Offers an API-first approach that streamlines integration into existing systems, **providing ready access** to medical histories that are key to identify veterans at risk in support of timely interventions.
- Uses a managed cloud service for data scientists and developers to **rapidly develop, train and test machine learning models** in the public cloud before deploying to production

Source: Red Hat blog. Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce Veteran suicides, March 2023.

Red Hat Summit

"Red Hat's work with AGESIC exemplifies our dedication to improving the user experience for both our and their customers."

Steven Huels
Vice President and General Manager – AI Business Unit, Red Hat

Presentation abstract

AGESIC, Uruguay's Agency for Electronic Government and Information and Knowledge Society, is responsible for e-government strategy and implementation. With Red Hat®, it led Uruguay's AI strategy and provided a more consistent, hybrid AI/ML platform to build and host models while delivering innovative applications.

Presentation summary

- With the proliferation of AI, AGESIC knew that infusing it into its operations would be key to meeting Uruguay's evolving needs.
- AGESIC optimized its AI infrastructure with Red Hat OpenShift®, which brought a containerized approach to workload management and automation of key processes while also bringing development, operations, and systems security functions together on a centralized platform.
- AGESIC evolved its offerings to include Platform as a Service (PaaS), enabling other government agencies to develop, run, and manage applications without the build and maintenance of complex infrastructure.
- AGESIC has begun automating the creation and development process of its AI models and managing model lifecycles, which has enabled standardization of AI usage across all Uruguayan governmental agencies

Products and services

Red Hat OpenShift

Red Hat OpenShift AI

Red Hat

DenizBank transforms AI operations and empowers innovation

"As an invaluable AI-driven solution, Red Hat OpenShift AI provides a streamlined environment that enables our data scientists to build and deploy more robust and secure models."

Okan Çetinkaya
CDO – CAO
DenizBank

Challenge

Intertech - IT subsidiary of DenizBank - wanted to build a comprehensive, standardized, holistic solution for data scientists that would improve time to market while delivering AI/ML process cost efficiencies across multiple business lines, including risk management, marketing and customer relations.

Solution

Red Hat Consulting helped the team design and architect the Red Hat OpenShift AI solution - on premise - providing self-service capabilities and capacity to scale model serving while improving operational efficiency.

Results

- Provided more than 120 data scientists, from different lines of business, with greater autonomy and more consistent standards
- Accelerated time-to-market while ensuring more robust and secure models
- Optimized GPU usage with slicing

Source: Red Hat case study DenizBank empowers its data scientists. Jan, 2025.

Services offerings for Red Hat AI

Learn how to maximize your technology investments

Red Hat Skills Assessment

Training and Certification: Developing and Deploying AI/ML Applications on Red Hat OpenShift AI with Exam (AI268)

Prototype

AI Incubator

Rapid prototyping of use cases in a controlled environment

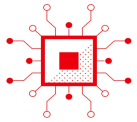
- Rapidly prototype AI applications and services
- Develop RAG+RAFT based patterns for model tuning and training
- Prototype AI Assistants & chatbots
- Develop evaluations for model accuracy and speed
- Prototype data ingestion pipelines

Deployment

AI Platform Foundation

Automated deploy of Red Hat AI Platform while advancing your AI practices

- Upskill customer's ML Platform team and data scientists
- Help customers adopt new AI capabilities
- Layout future roadmap of skills and capabilities
- Increase teams core MLOps competency



Scale

MLOps Foundation

Roll out automated MLOps pipelines and practices throughout your organization

- Establish self-service of MLOps platforms
- Automate and template ML pipelines
- Establish patterns and best practices for managing production ready solutions

Operational guidance & advisory services from TAM Services for Red Hat AI Platform (yearly subscription)

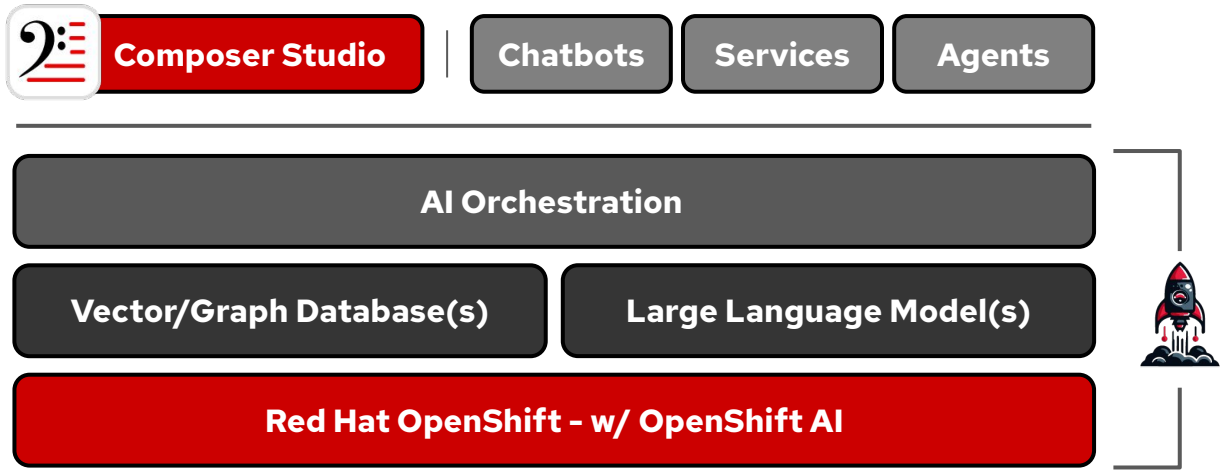
Red Hat

AI Incubator powered by Red Hat Composer AI

Generative AI Use Case Development Made Easy

The AI Incubator is a consulting approach designed to accelerate innovation by providing a unified, scalable platform designed to rapidly transform new ideas into impactful AI solutions. Streamline the entire lifecycle of AI development, from concept to deployment, while benefiting from a robust architecture that ensures seamless integration and top-tier security. Our platform empowers your enterprise to harness cutting-edge AI technologies quickly and efficiently, all while maintaining complete control over your data and infrastructure.

- **Centralized AI Platform:** Accelerate AI use case development with an intuitive, secure, and scalable environment
- **Tailored for All Roles:** Empower executives, developers, and infrastructure managers with tools designed for innovation, efficiency, and control
- **Seamless Integration:** Works with open source models and manages proprietary data securely across various databases
- **Automate & Innovate:** Streamline workflows, automate routine tasks, and focus on what matters most—driving your organization forward



Red Hat

Are you ready to get ahead of the curve with **Red Hat AI Foundations**

No-cost training certificates offers two learning paths:

Red Hat AI Foundations offers no-cost learning paths with certificates for executives and technologists to help enhance their understanding of AI.

Whether you are just **starting to understand AI** or you are ready to get **more technical**, Red Hat AI Foundations has a **certificate for you**.

Key Benefits:

- Gain understanding of AI terminology and ethics to enhance skills quickly
- Learn how Red Hat accelerates AI solution delivery.
- Empower strategic business outcomes with valuable AI insights

Red Hat AI

Next best steps you can take

Learn more and get hands-on experience

TRY RED HAT ENTERPRISE LINUX AI

A single, 60-day, self-supported subscription to Red Hat® Enterprise Linux® AI

red.ht/RHELAI-trial

TRY RED HAT OPENSHIFT AI

A single, 60-day, self-supported subscription to Red Hat® OpenShift® AI (Self-Managed)

red.ht/RHOAI-trial

TRY RED HAT AI INFERENCE SERVER

A single, 60-day, self-supported subscription to Red Hat® AI Inference Server

red.ht/ai-inference-server-trial

Red Hat

Red Hat AI

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat

Product slides

Red Hat

Red Hat AI Inference Server

Red Hat AI Inference Server

Optimizes model inference across the hybrid cloud, creating faster and more cost-effective model deployments.

- **Optimized model inference with vLLM** - a runtime for maximizing throughput and minimizing latency.
- **Accelerate model serving** - access to a model repository with validated and optimized models.
- **Reduce model size and compute requirements** - Optimized model while preserving accuracy using the LLM compressor.
- **Runs anywhere** - Run in any Linux or Kubernetes distribution, any hardware and any cloud.

Red Hat AI Inference Server

Gain consistent, fast and cost-effective inference at scale

Inference runtime for the hybrid cloud

Run your models of choice across any accelerator and any environment

Compress Models

Reduce compute and costs while preserving accuracy

Red Hat AI Hugging Face repository

Access a collection of third-party validated and optimized models ready for inference.

Certified for all Red Hat products

Deployable across non-Red Hat Linux and Kubernetes platforms

Red Hat AI Inference Server is included in OpenShift AI and RHEL AI

Red Hat AI Inference Server

vLLM is emerging as the Linux of GenAI Inference

HIGH PERFORMANCE

- Advanced algorithms for high QPS serving
- Single server/GPU to distributed/multi GPU
- Already comparable to Nvidia (TRT-LLM)

EASY TO USE CAPABILITIES DRIVING DEVELOPER AND IT PRODUCTIVITY

- Native Hugging Face integration
- Simple APIs for online and offline inference
- OpenAI-compatible API protocol

Scalable inference across the hybrid cloud

Red Hat

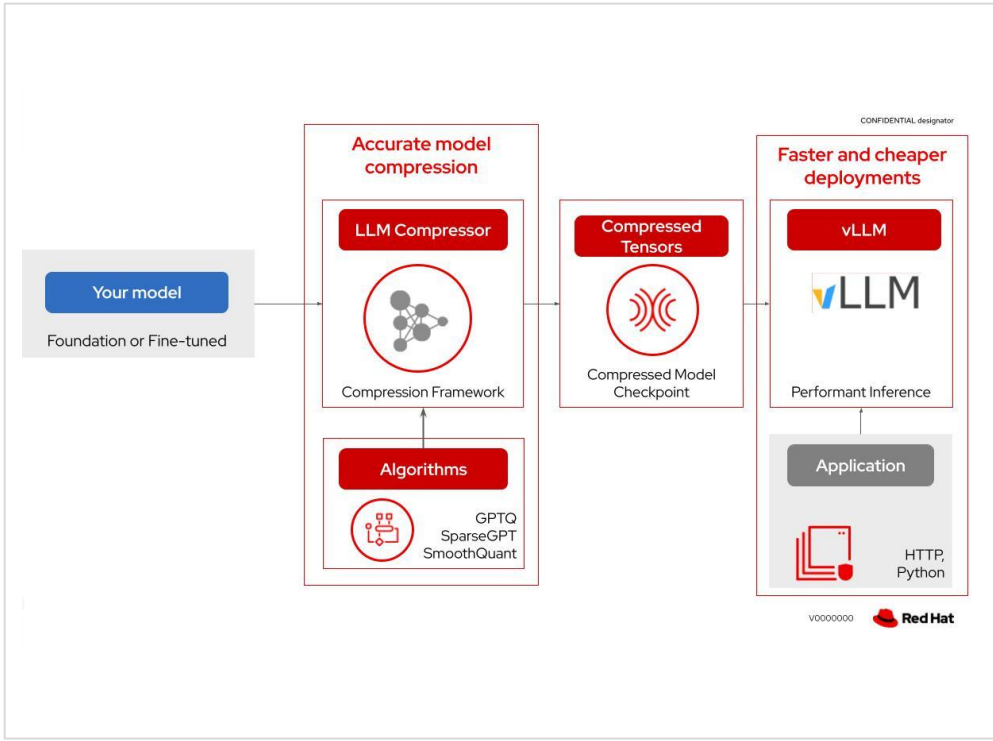
Red Hat: Leaders in OSS GenAI Inference

Red Hat has built a comprehensive set of model optimization capabilities to drive operational efficiencies

Third-party validated and optimized models

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

LLM Compression Tools



Hosted on the Red Hat AI repository on Hugging Face

Red Hat AI tooling for model optimization

Optimize and validate your choice of model

Inference benchmarks with GuideLLM

Tool for evaluating LLM performance to guarantee efficient, scalable, and affordable inference serving.

Accuracy evaluation with LM-eval-harness

A unified framework for evaluating the accuracy of LLMs across a variety of tasks and benchmarks.

LLM Compression tools

Framework for reducing the size and computational requirements of a LLMs while preserving accuracy

Receive tailored capacity planning guidance from our experts

Red Hat AI Enterprise

Red Hat

Red Hat AI Enterprise

Integrated AI platform for deploying and running efficient and cost-effective AI models, agents and applications in the hybrid cloud

- **Unified AI lifecycle:** Manages e2e process (develop, tune, infer) for predictive, generative and agentic AI on a single, centralized platform.
- **Intelligent scale & performance:** Optimizes AI inference at scale ensuring efficient GPU utilization and intelligent resource allocation.
- **Enterprise governance & trust:** Guarantees comprehensive, layered security and safety across the entire AI lifecycle.
- **Hybrid cloud agility:** Enables flexible deployment of AI use cases across the entire hybrid cloud, diverse hardware, and the edge.

Deliver AI faster, efficiently and lower the risk

Using a trusted, comprehensive and consistent platform

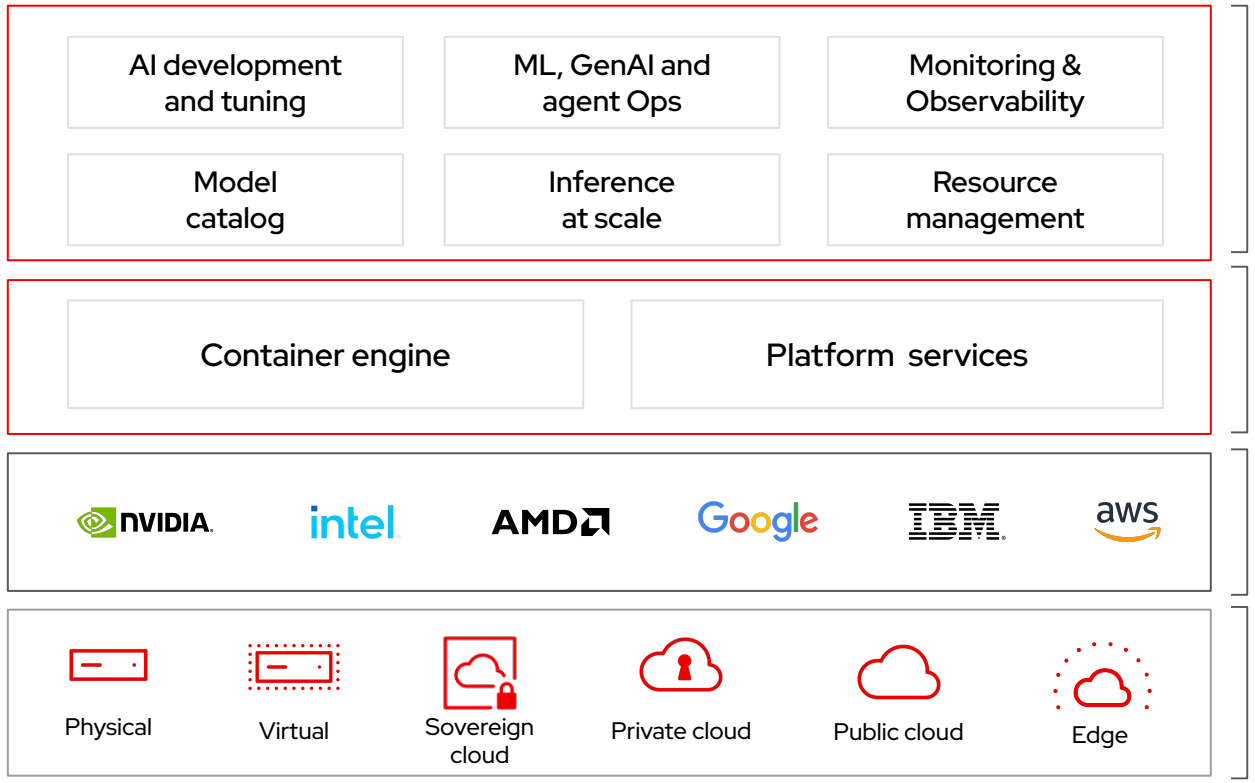
With Red Hat AI Enterprise, you can:

- **Accelerate time-to-value**
  Integrated, enterprise-ready AI platform for quickly building and delivering cloud-native AI-powered apps.
- **Increase operational efficiency**
  Streamline workflows and intelligently optimize resource allocation to maximize value.
- **Mitigate risk**
  Offer a tested, supported stack that ensures business continuity and helps meet regulatory needs.

Red Hat

Red Hat AI Enterprise

An integrated AI platform for enterprise deployments



AI services layer: Build, tune, and manage all AI models and agent workflows.

Platform services layer: Core foundation for consistency, security, and scalable operations.

Any hardware: support for various hardware accelerators for optimized performance and cost.

Deploy anywhere: choose where to train, tune, deploy and run AI models, agents and applications.

Red Hat

Needs update

Build without compromise. Scale without complexity.

Build and tune

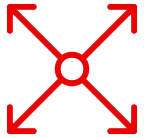
Build predictive AI models, tune LLMs and assemble AI-powered apps

ML, GenAI and Agents Ops

Manage the lifecycle of predictive, generative and agentic AI

Safety, monitoring and observability

Control models, monitor performance, and guarantee compliance across all AI.

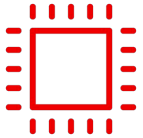


Inference at scale

Deploy models using efficient inference runtimes (vLLM), manage the serving process and scale with intelligent scaling

Catalog and registry

Access and experiment with validated and optimized models, MCP servers and AI agents



Optimize resource allocation

Run models across different GPU platforms and the hybrid cloud while optimizing resource allocation

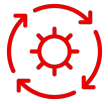
Red Hat

Real-world benefits from choosing the right AI platform

Comprehensive: Provides the AI models, tooling, and capabilities on top of Kubernetes infrastructure to deploy and scale AI

Hybrid: Control where you build, train, and run AI workloads in alignment with regulatory, cost, and strategic goals

Flexibility: Provides support for various models, hardware accelerators, and clouds enabling choice to align to your architecture



Consistency: Helps standardize the process of building, deploying and managing AI models, agents and applications

An AI platform that supports enterprise-grade, production AI agentic workflows and AI-enabled applications efficiently and cost-effectively

Red Hat

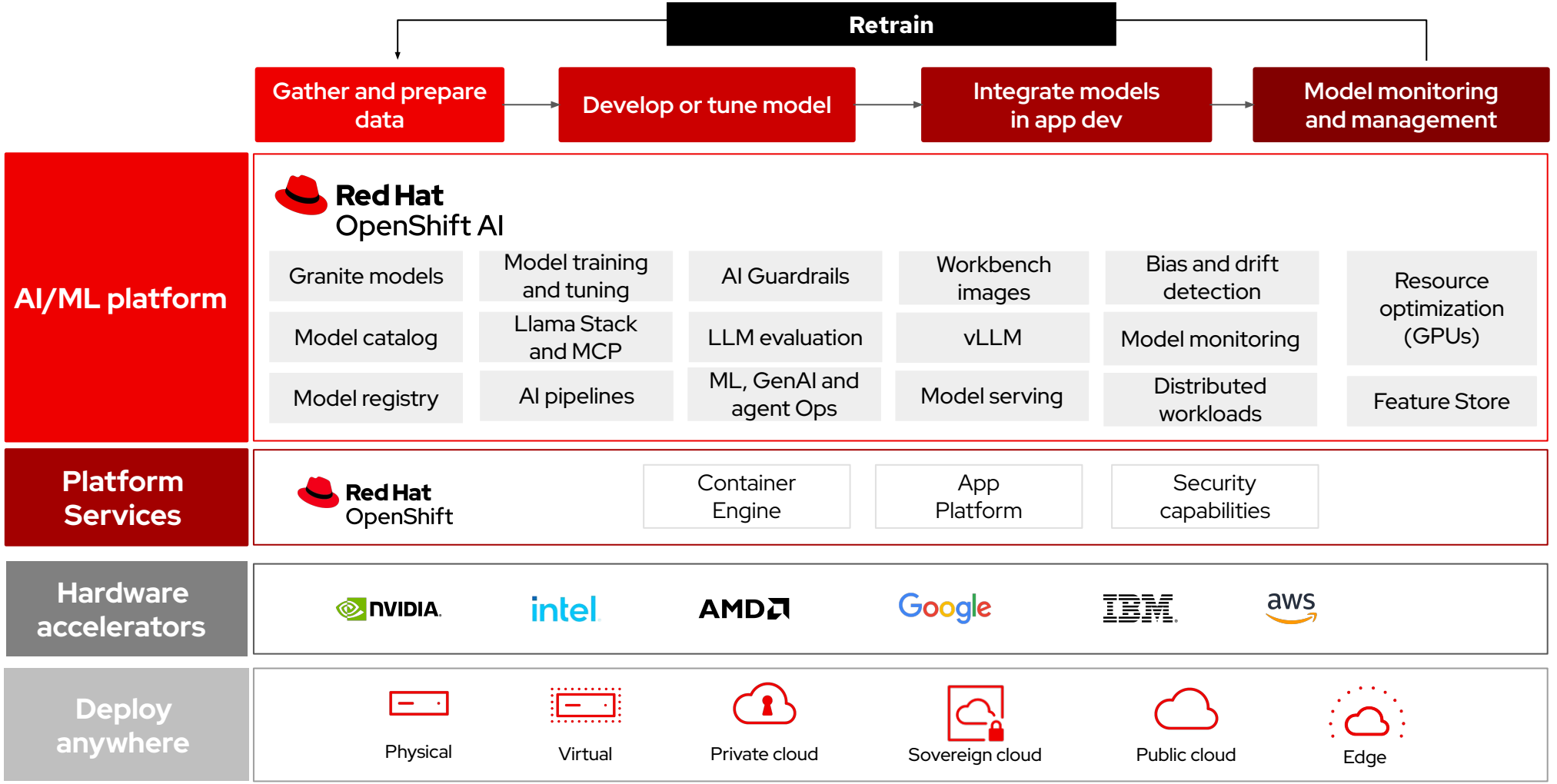
Red Hat OpenShift AI

Red Hat OpenShift AI

AI platform for managing the lifecycle of predictive and generative AI models at scale across hybrid-cloud environments.

- **Unified MLOps/LLMOps** - Manages the entire AI lifecycle and expands DevOps practices.
- **Model development** - Access to AI models and tooling for Gen AI tuning and predictive AI development.
- **Intelligent resource optimization** - Efficient GPU utilization via intelligent scheduling and workload scaling.
- **Trusted deployment & safety** - Supports air-gapped deployments and monitors models for transparency, fairness, and reliability.
- **Hybrid cloud and sovereign agility** - Flexible deployment across hybrid cloud, diverse hardware, and the edge.

Red Hat OpenShift AI - One platform for gen and predictive AI



Red Hat

Red Hat OpenShift AI - Key features

Model development

Interactive, collaborative UI for **seamless access** AI/ML tooling, libraries, frameworks, etc.

Model customization

Access to **RAG** and **tuning capabilities** (LoRa, QLora, fine tune), **SDG and evaluation** tools

Model monitoring

Centralized monitoring for **tracking models performance and accuracy**

Model serving

Model serving routing for **deploying models to production** environments

AI pipelines

Automate AI tasks into repeatable pipelines

vLLM & llm-d

Inference runtimes and **distributed, scalable AI inference** for performance and cost-efficiency

Trust & AI guardrails

Improve LLM accuracy, performance, latency and **transparency**

Agentic AI

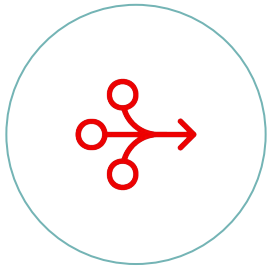
Assemble, manage and run agentic AI workflows using Llama Stack API and MCP

The value of Red Hat OpenShift AI

What differentiate us?

Simplify AI adoption

Promotes freedom of choice and access to latest innovation on AI



Drive AI/ML operational consistency

Streamline the process of moving models from experiments to production

Gain hybrid cloud flexibility

Deploy models in containerized format across on-prem, clouds and edge, including disconnected environments

Red Hat Enterprise Linux AI

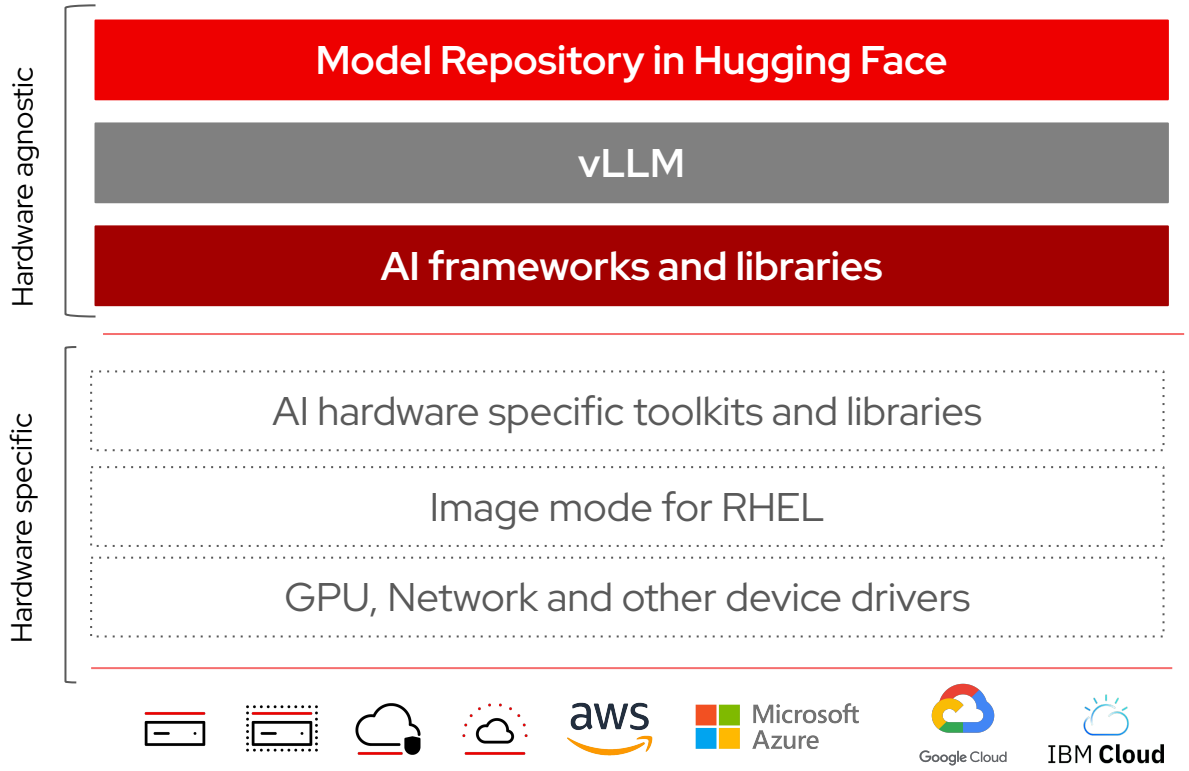
Red Hat Enterprise Linux AI

Red Hat Enterprise Linux AI (RHEL AI) is a foundation model platform to consistently run LLMs in individual server environments

- **Purpose-built AI appliance** – Delivers an immutable RHEL image for stable, optimized single-server LLM inference.
- **Performance & cost control** – Optimizes AI performance and reduces compute costs using vLLM and the LLM compressor.
- **Trusted models & IP assurance** – Provides indemnified Granite models with full transparency under the Apache-2.0 license.
- **Hybrid cloud agility** – Enables flexible deployment across the entire hybrid cloud, diverse hardware, and the edge.

Red Hat Enterprise Linux AI

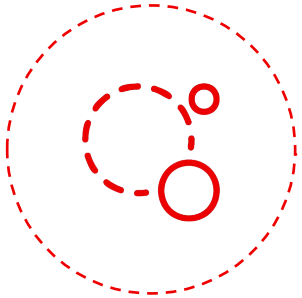
Seamlessly develop, test, and run large language models (LLMs) for enterprise applications



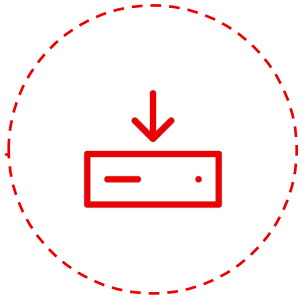
Red Hat

RHEL AI benefits

Streamline adoption of generative AI



Unlock the power of efficient, AI Inference



Deploy AI quickly in a single server environment

Manage and monitor AI models anywhere

Customer stories

Red Hat

AI is a strategic enabler across industries

Predictive AI runs businesses today, Generative AI brings innovation to the enterprise

Revenue Generation

- Chatbots
- Campaign and Content Marketing
- Developer assistants
- Guided selling
- Drive product innovation

Cost Optimization

- Automated AI support
- Knowledgebase Search & Summarization
- Doc summarization
- AI-optimized logistics
- Augmented Product R&D

Risk Management

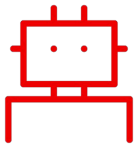
- Sentiment analysis
- Predict employee attrition
- Contract risk assessment
- Fraud detection
- AI-assisted Security Operations

Artificial Intelligence innovation in action

Red Hat customers benefit from both predictive to generative AI

Enabling AI-innovation in the airline industry

Enabled teams to use AI in their everyday operations by reducing the time it takes to develop and deploy models to production environments.



Virtual assistants and RAG for environmental assessments

Streamlined document review and resolution, resulting in greater efficiency, minimized errors and reduced response times.

Comprehensive, mature and governed internal AI platform

Enabled teams to build AI models resulting in improved call center efficiency and accelerated development of paramount systems.

Sources:
Turkish Airlines Pioneers AI-led Innovation for Aviation with Red Hat OpenShift AI | Government of Castilla-La Mancha and Red Hat Power Public Services with AI | Red Hat Empowers Hitachi, Ltd. to Evolve into an AI-Driven Enterprise with Red Hat OpenShift AI

Red Hat OpenShift AI leads AI-innovation for aviation

Accelerate development and deployment of AI solutions

- Minutes vs. hours → Development Environments
- 2x Deployment Speed

"Red Hat OpenShift AI allow us to keep data on-premises while accelerating model development and enabling business units to leverage AI capabilities in a flexible way. This approach is helping us move AI beyond isolated use cases and into a scalable, organization-wide capability. As Turkish Airlines, we believe that the use of AI will play a vital role in enhancing our key achievements in the aviation industry"

Serdar Gürbüz
General Manager, Turkish Technology, a Turkish Airlines subsidiary

Source:
Turkish Airlines Pioneers AI-led Innovation for Aviation with Red Hat OpenShift AI

Revolutionizing government efficiency with AI

AI can improve the efficiency of processes and deliver better outcomes for citizens

"This project demonstrates how AI can be used to improve the efficiency of government processes and deliver better outcomes for citizens"

Juan Pedro de Ruz Ortega
General director of digitalization and artificial intelligence, Junta de Comunidades de Castilla-La Mancha

- Improved quality of service for citizens
- Increased operational efficiency at scale
- Reduced operational complexity
- Optimized productivity for officials

Source:
Government of Castilla-La Mancha and Red Hat Power Public Services with AI

Evolve into an AI-Driven Enterprise

Operationalize AI across business units, and set a new standard for governance and innovation

"Hitachi isn't simply experimenting with AI–we're industrializing it. The success of our internal platform and integrating AI tools like Red Hat OpenShift AI into our daily operations exemplifies how we're making AI the heart of our business, from call centers to large-scale system development. With Red Hat OpenShift AI, we were able to achieve this level of growth"

Masahiro Kikuchi,
Director, platform service department, Managed & Platform Services Business Division, Hitachi, Ltd.

Sources:
Red Hat Empowers Hitachi, Ltd. to Evolve into an AI-Driven Enterprise with Red Hat OpenShift AI