Red Hat AI

Introduction to Red Hat AI

Accelerate development and delivery of AI solutions

Martin Isaksson
AI BU

Red Hat

Gen AI adoption rates across the enterprise are rapidly increasing

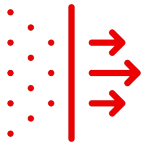
Gen AI is a core part of the products we use, at home and at work

Frontier AI services grow new capabilities at a relentless speed

Generative AI customer adoption challenges

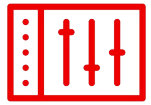
Cost

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 AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical

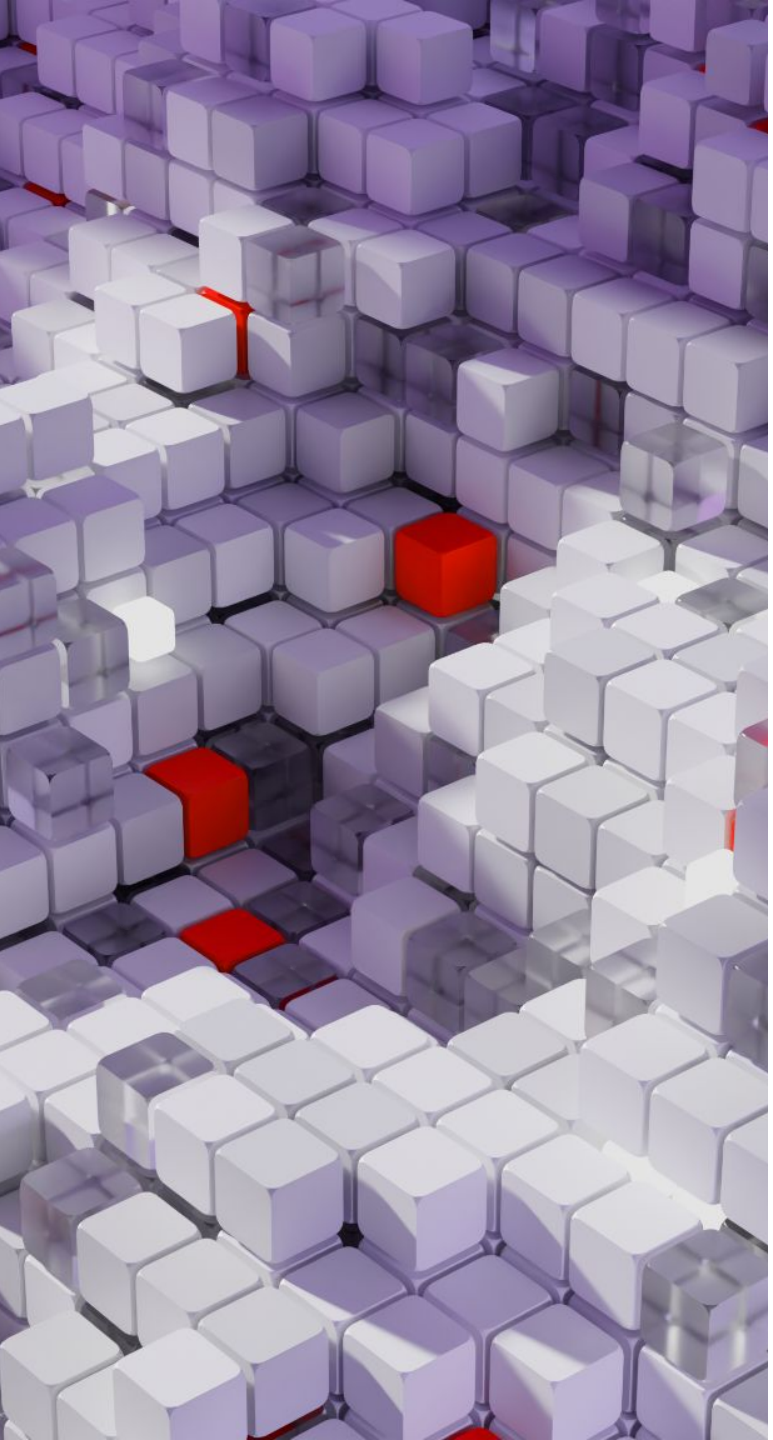
Virtual

Private Cloud

Public Cloud

Edge

Red Hat



Fast, flexible and scalable inference

Red Hat

Expanding choice of models

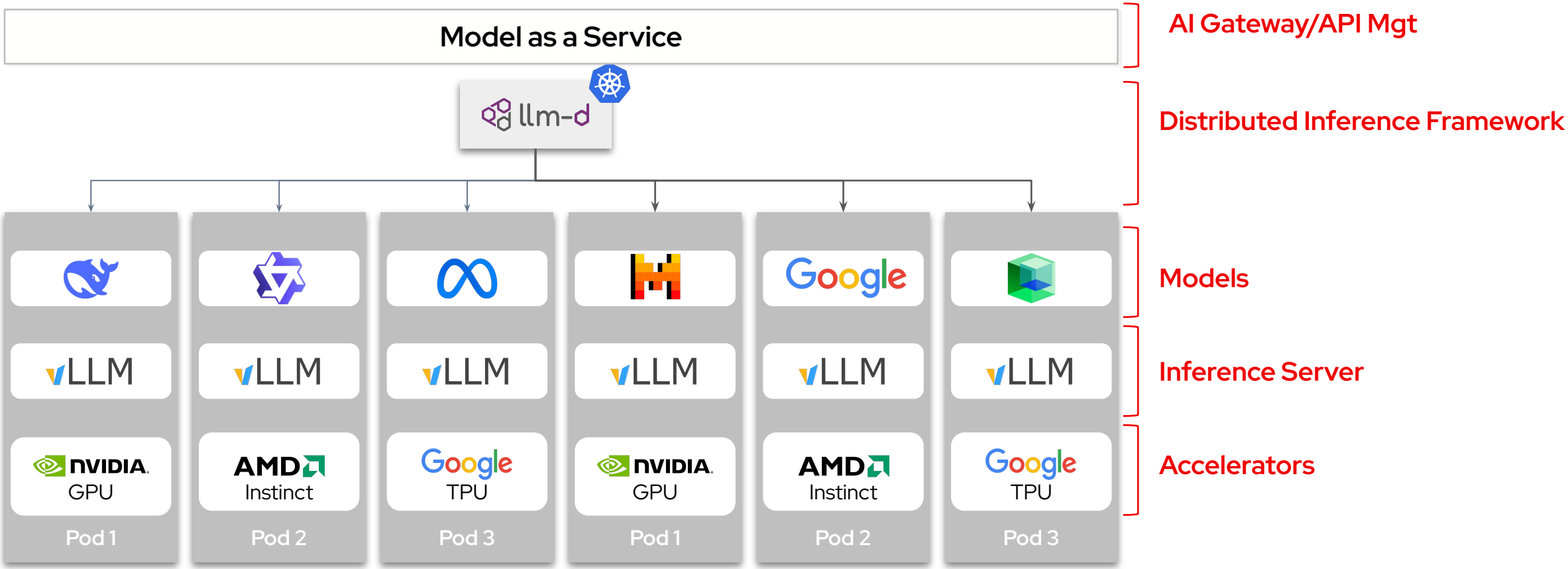
There has been an explosion of capability from open-source over the last 2 years

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 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

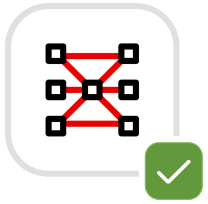
- Tested using realistic scenarios
- Assessed for performance across a range of hardware
- Done using GuideLLM benchmarking and LM Eval Harness

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 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

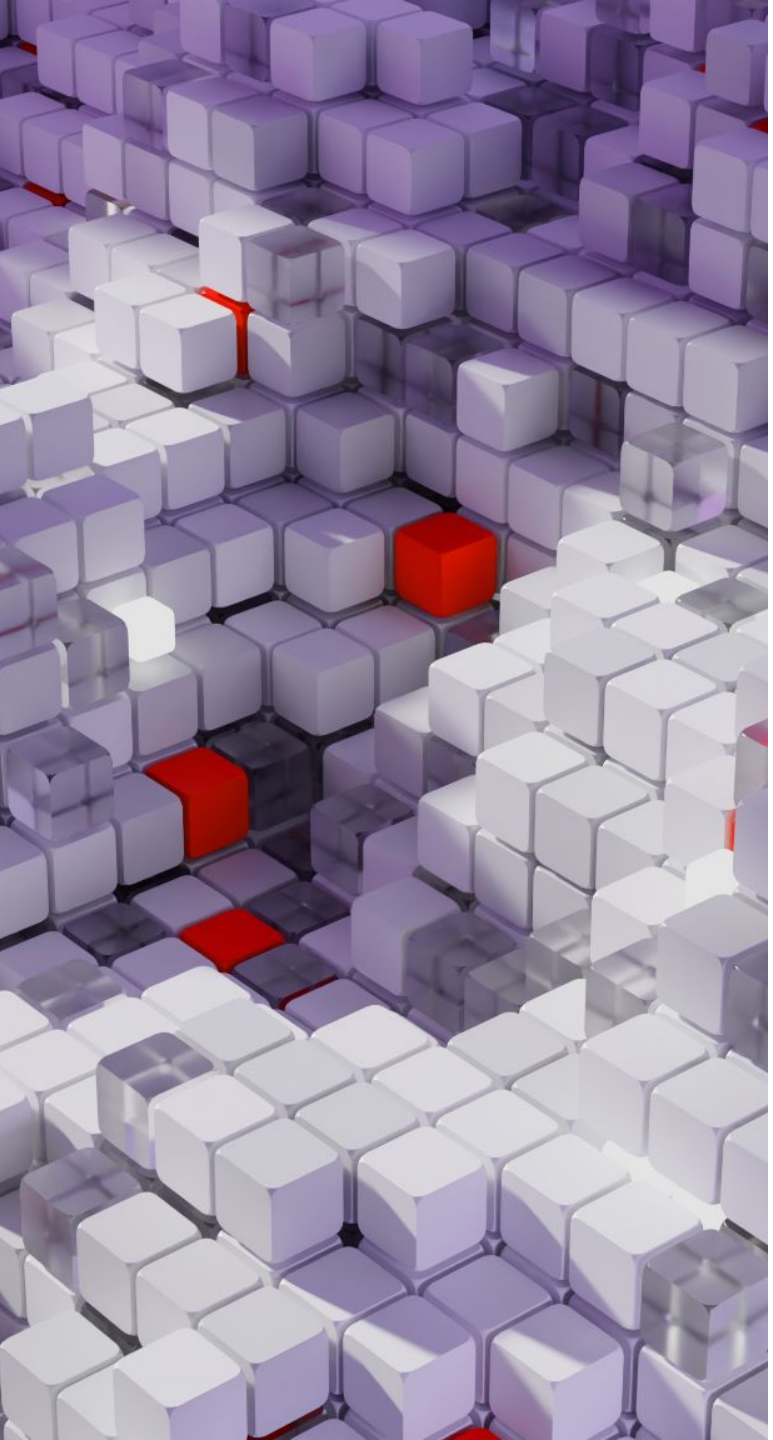
vLLM connects model creators to accelerated hardware providers

Scale AI inference when ready

llm-d

Llm-d provides consistent, distributed, inference at scale

Red Hat

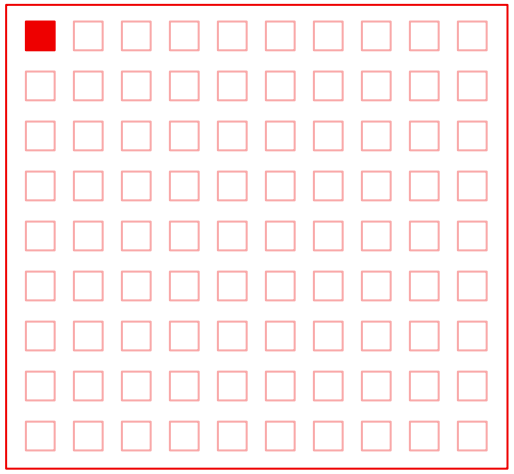


Connecting models to data

Red Hat

Enterprises need models aligned to their private data

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

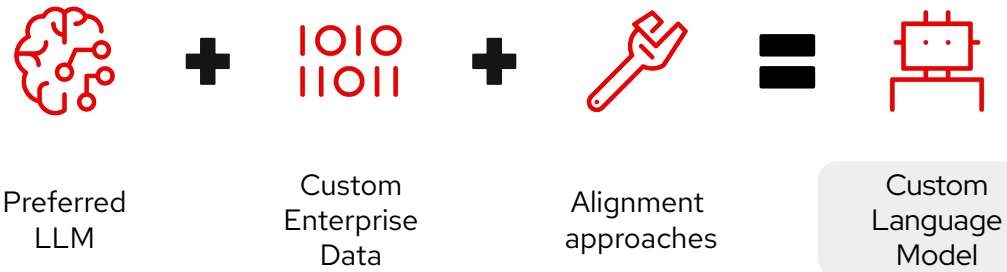


Less than 1% of all enterprise data is represented in foundation models

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

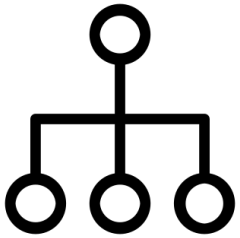
What's next: multiple approaches?

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

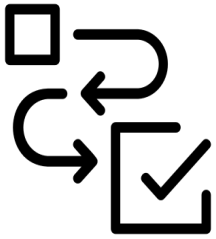
Retrieval Augmented Generation

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

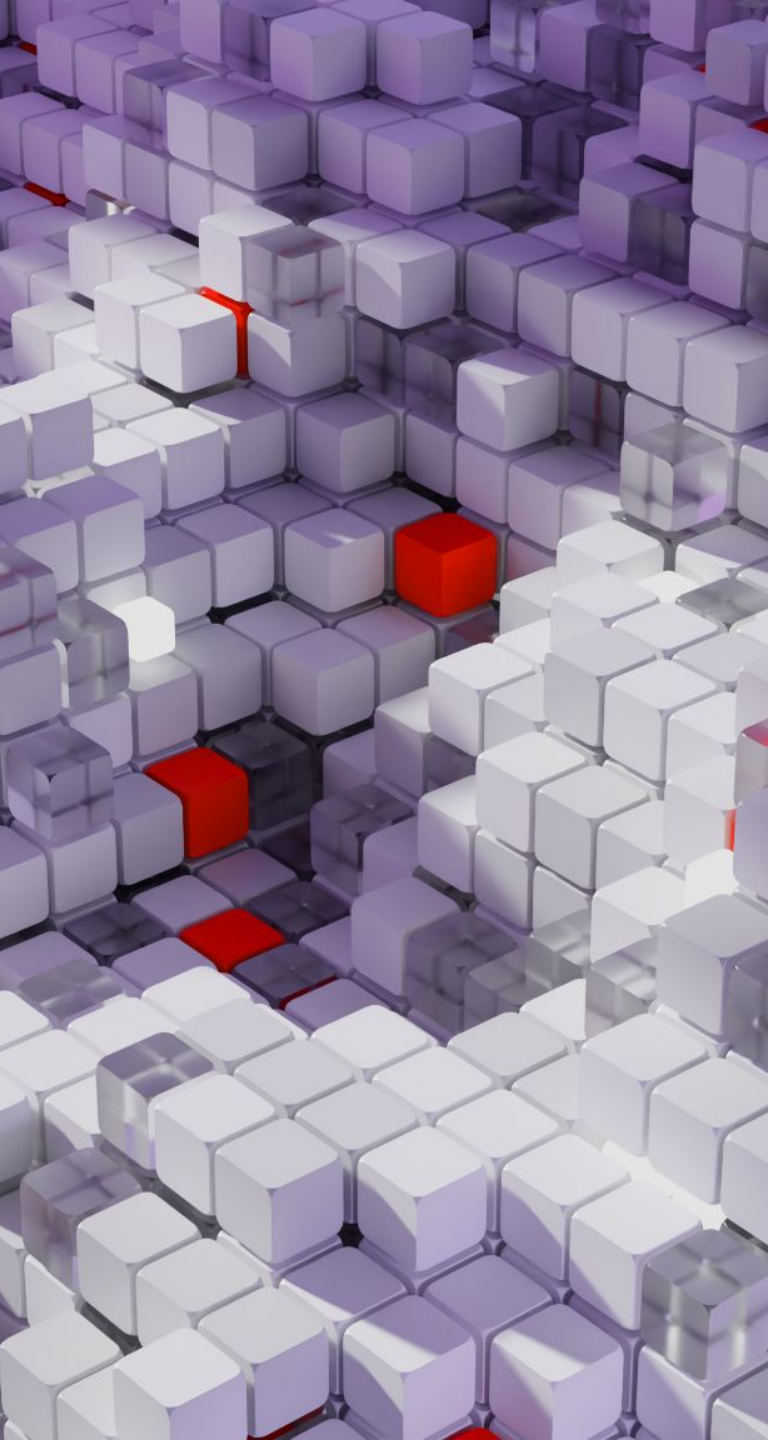
Enhanced

Fine tuning

InstructLab, OSFT, LoRA and QLoRA



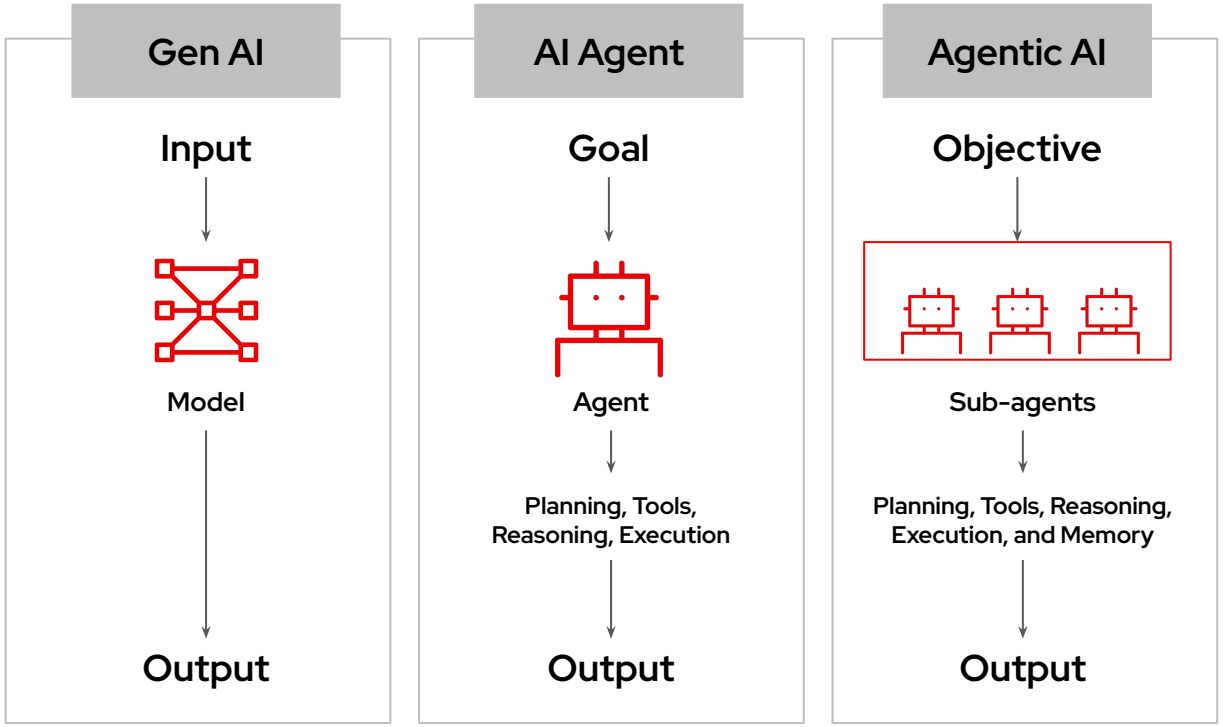
Customize a base model for specific tasks or private data, using a range of approaches—from full fine-tuning to parameter-efficient methods—to balance performance and efficiency.



Accelerate Agentic AI innovation

Red Hat

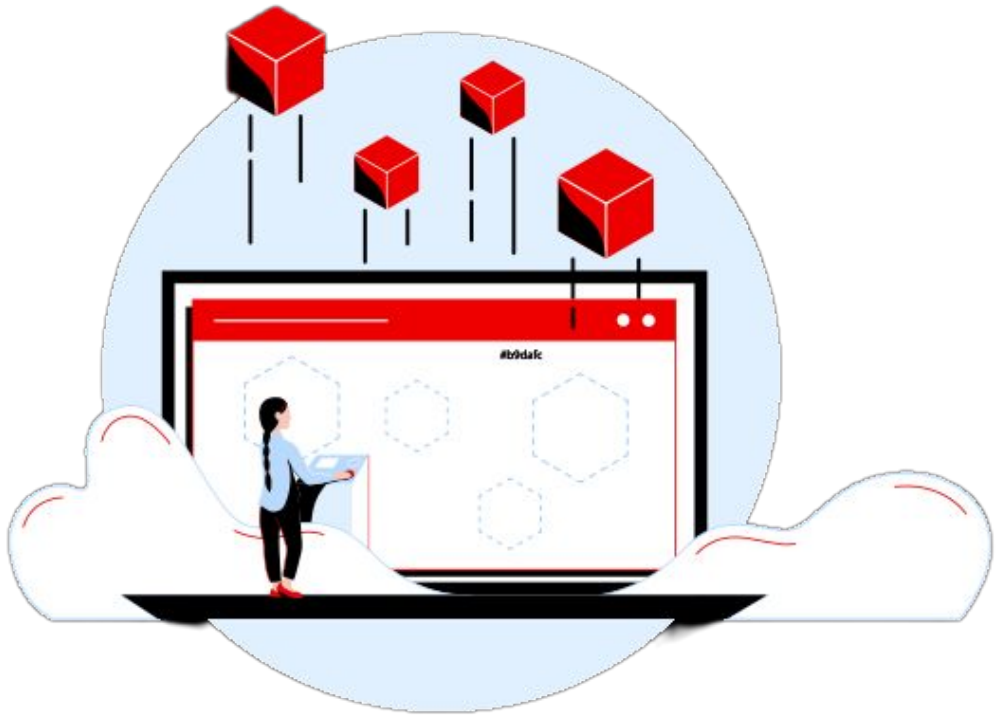
How does the next evolution in AI look?



- **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 development and deployment 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.

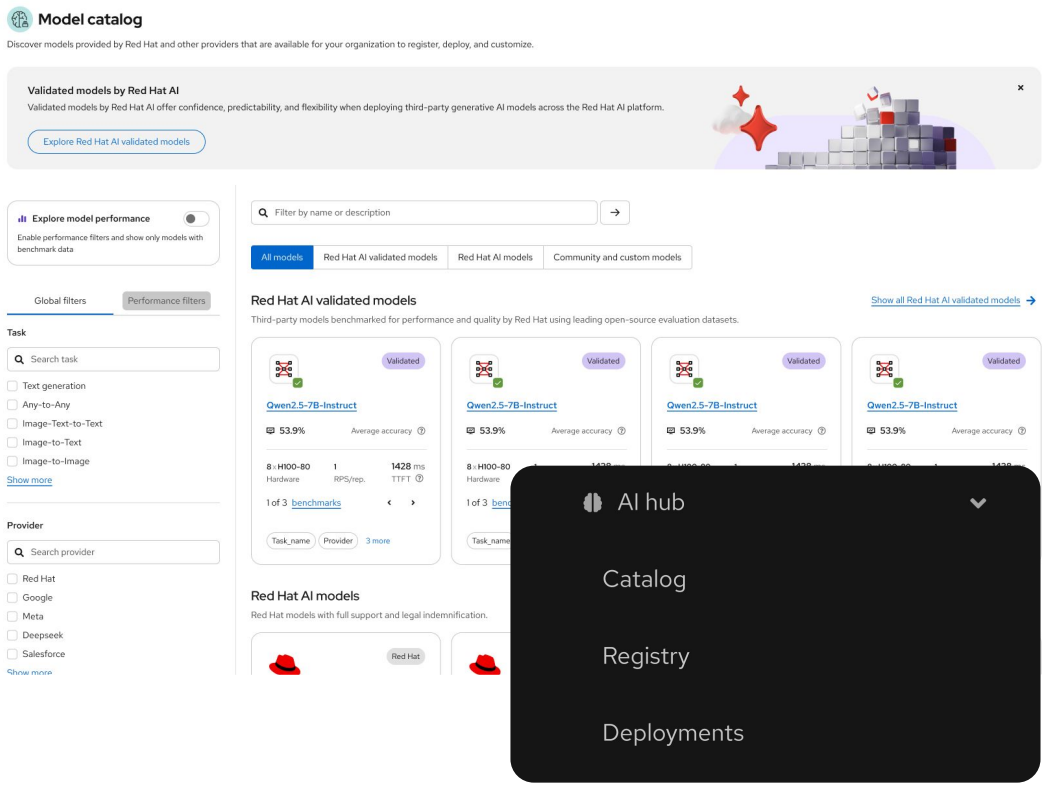


Red Hat

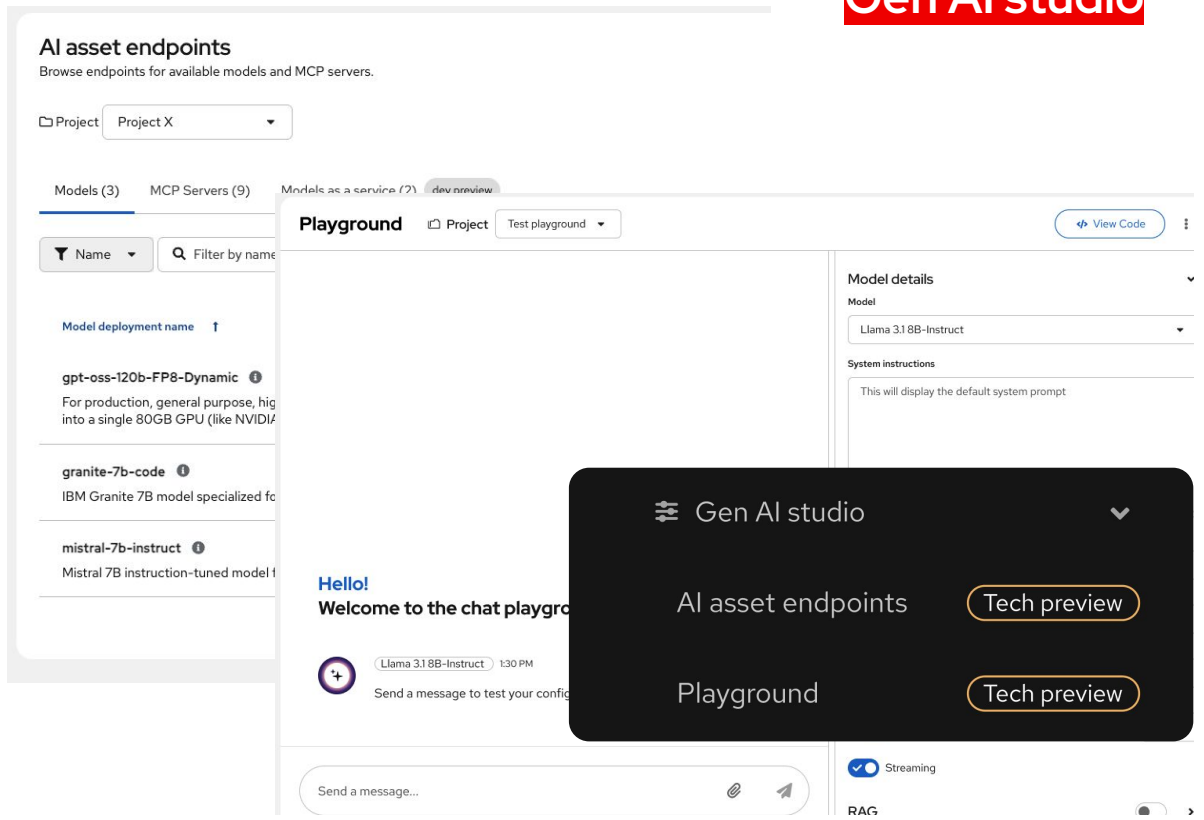
AI dedicated experiences

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

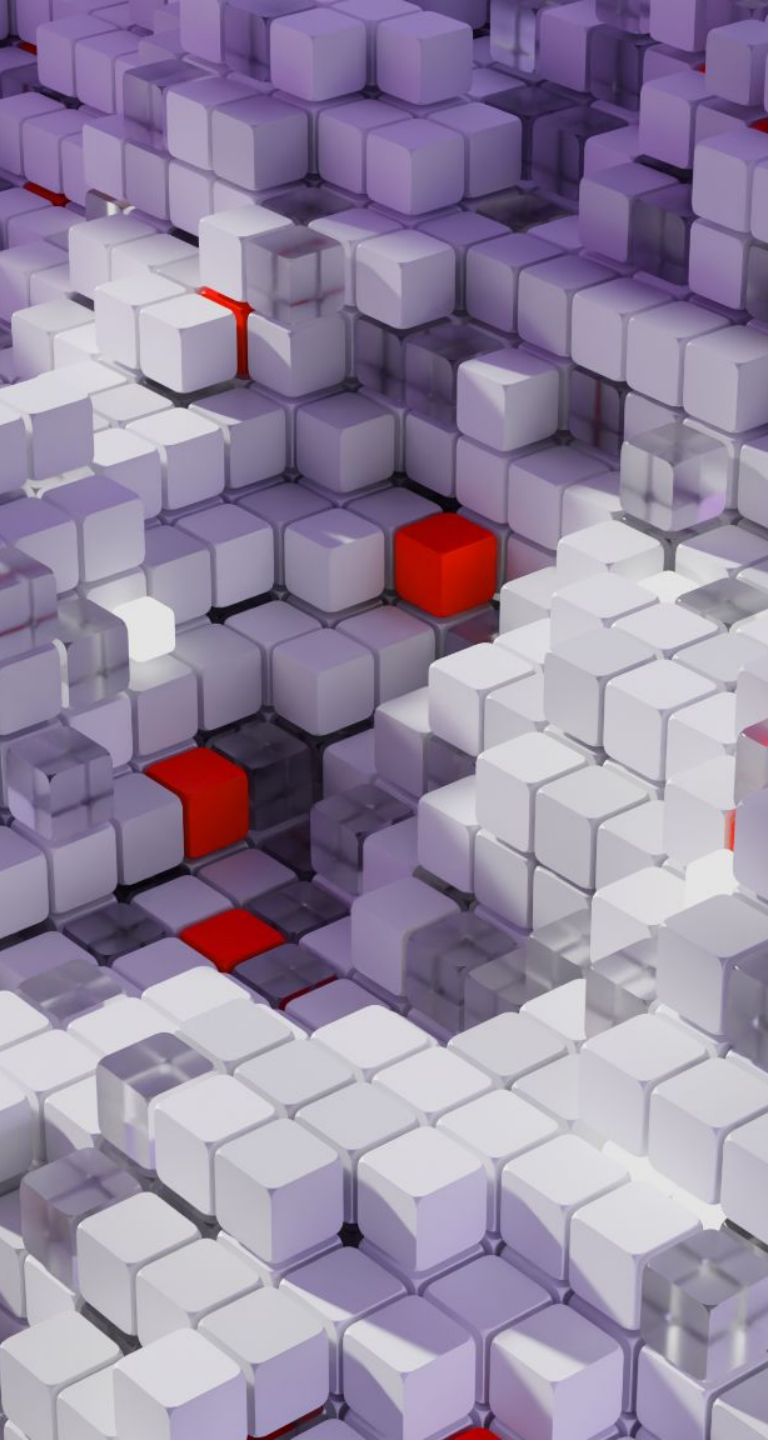
AI hub



Gen AI studio



Red Hat



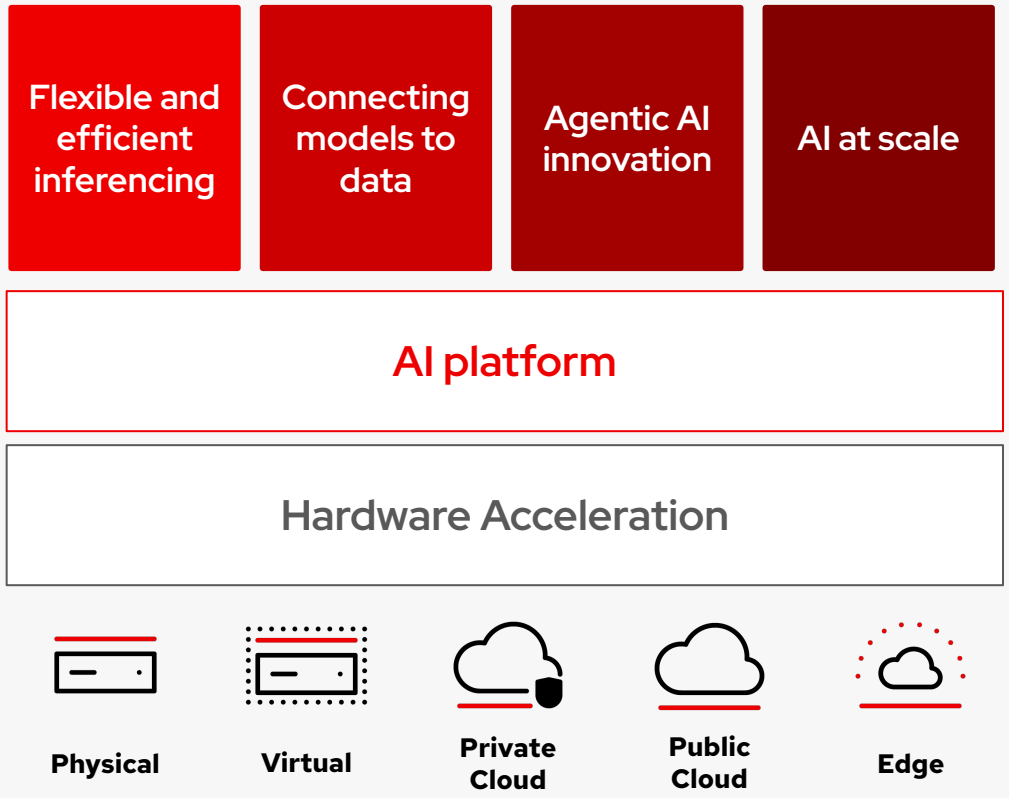
Scaling AI across the hybrid cloud

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

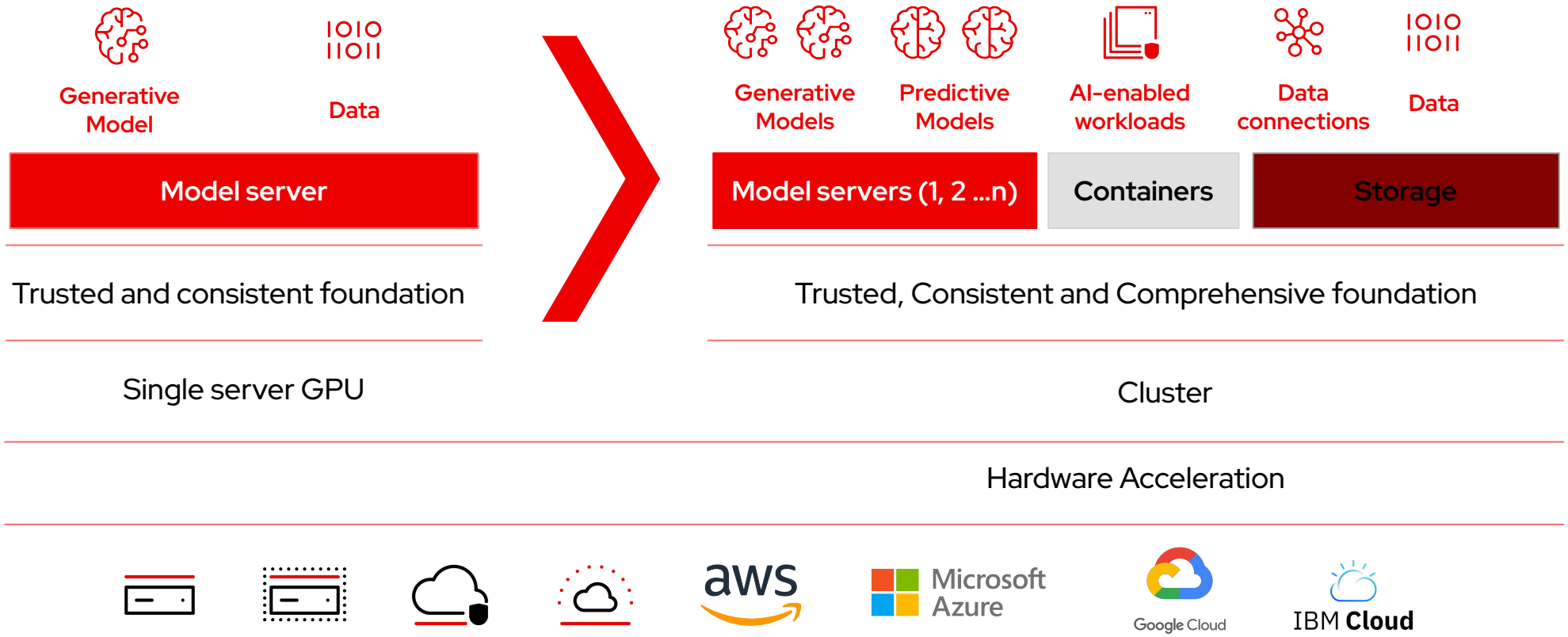
It includes:

- 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.



Red Hat AI supports each stage of the AI adoption journey

From single server deployments to highly scaled-out platform architectures



Red Hat

Hybrid cloud deployment for AI

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

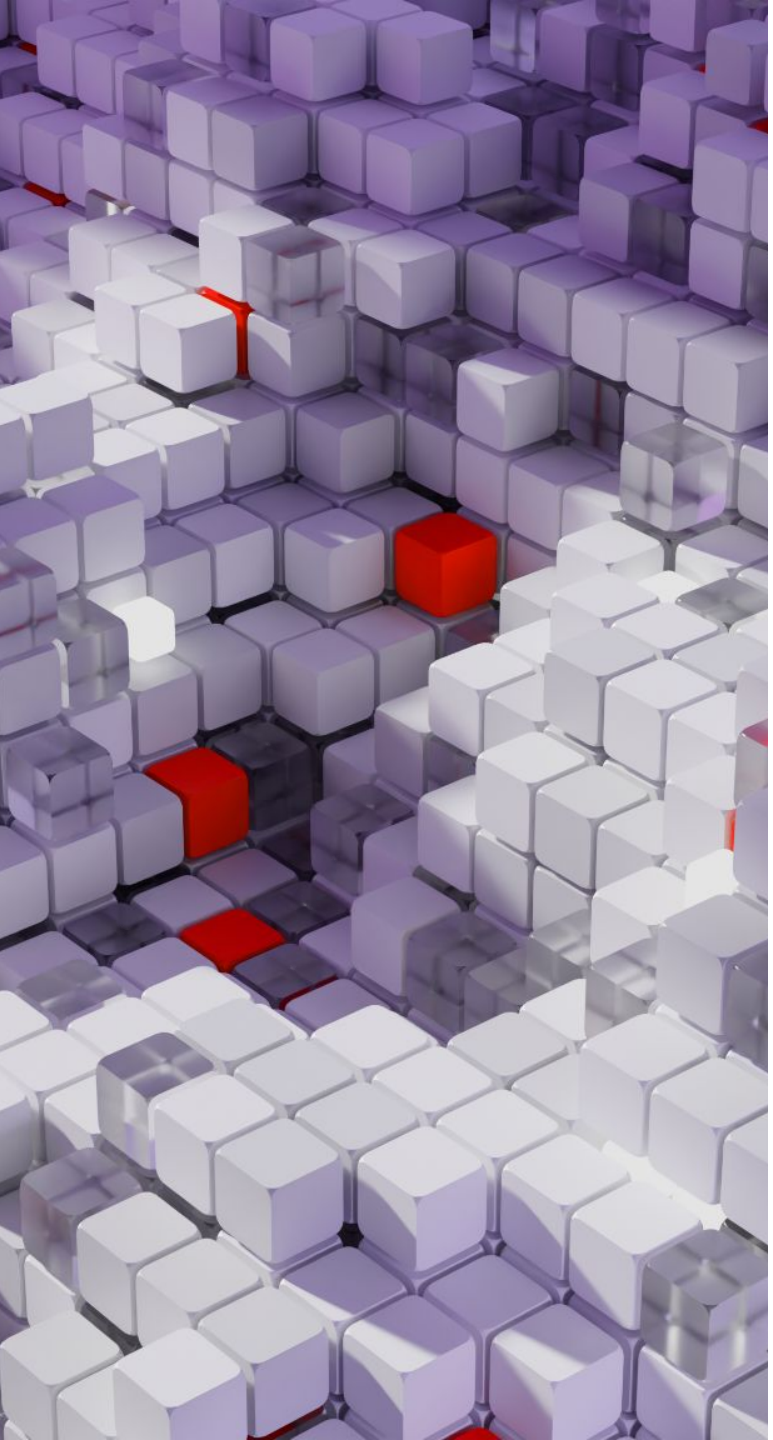
Hardware Accelerators

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

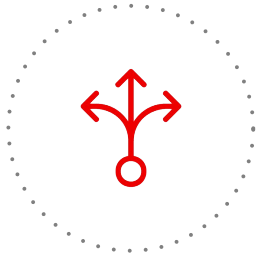


Why Red Hat AI?

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

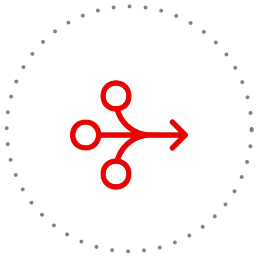


Flexibility

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

Choice

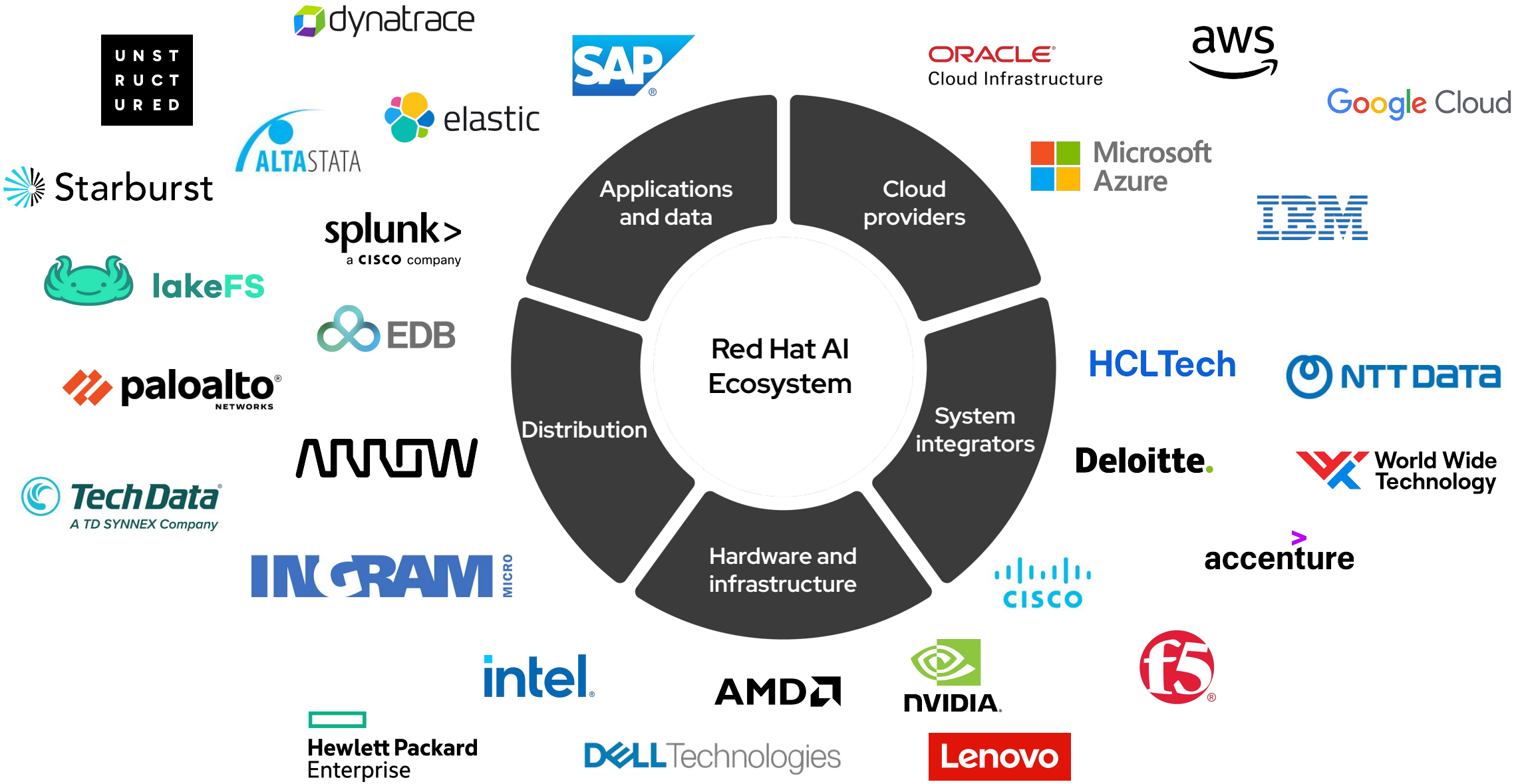
Access to an open ecosystem of communities, technology providers, ISVs and customers.



Abstract the complexity

Reduce the complexity of switching and adapting to new technologies.

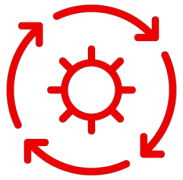
Red Hat



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

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

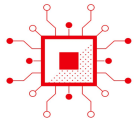
- Rapidly prototype AI applications and services
- Develop RAG+RAFT based patterns for model tuning and training
- Prototype AI Assistants & chatbots
- Develop evaluations for model accuracy and speed
- Prototype data ingestion pipelines

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

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