Red Hat AI

Introduction to Red Hat AI

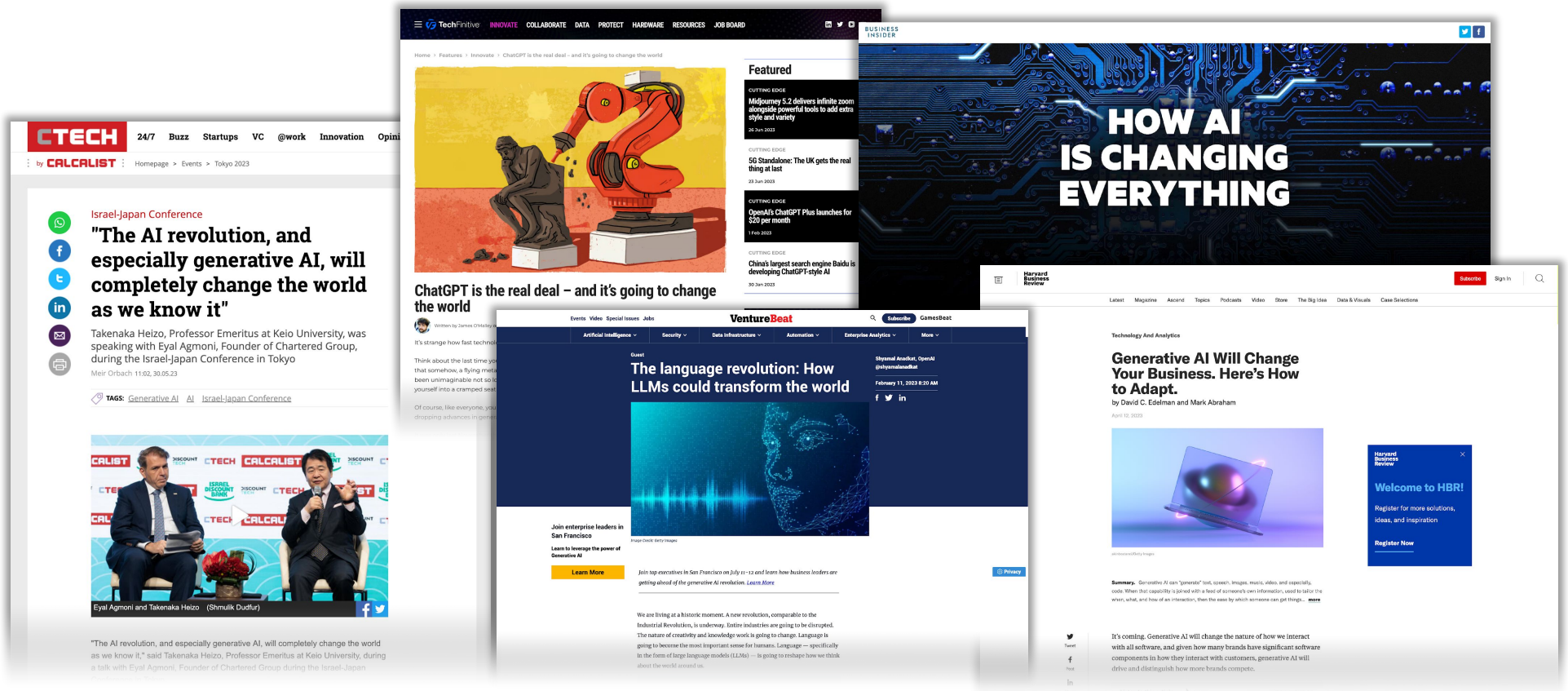
Accelerate development and delivery of AI solutions

Tushar Katarki
Senior Director
Red Hat AI Products

Red Hat

The World Changed in November 2022

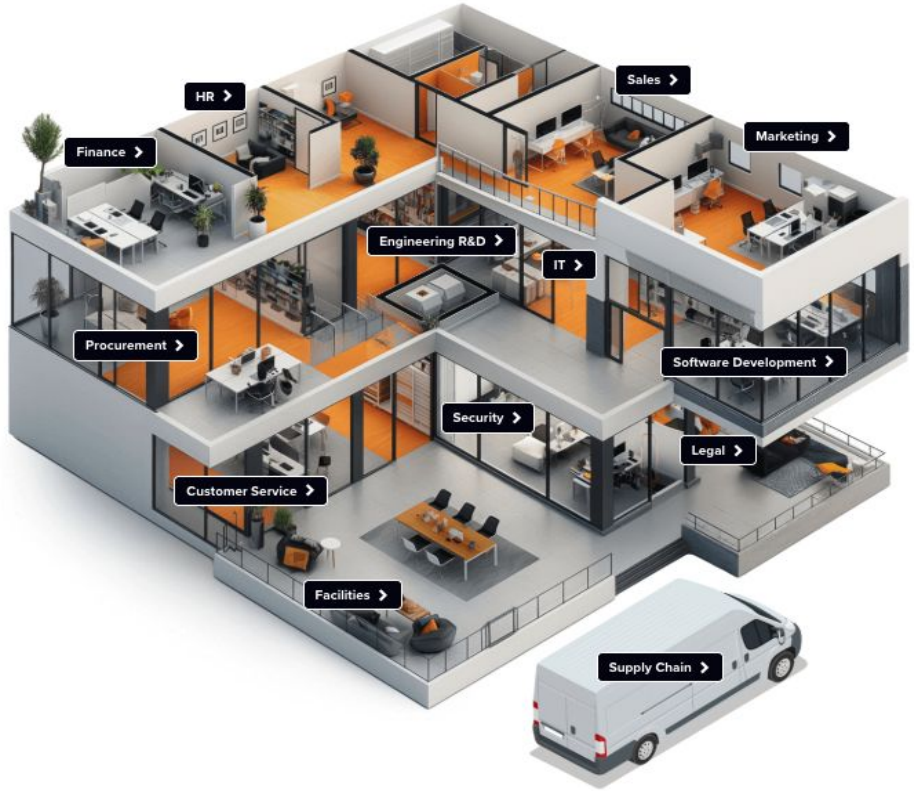
ChatGPT woke the world up to the power of generative AI



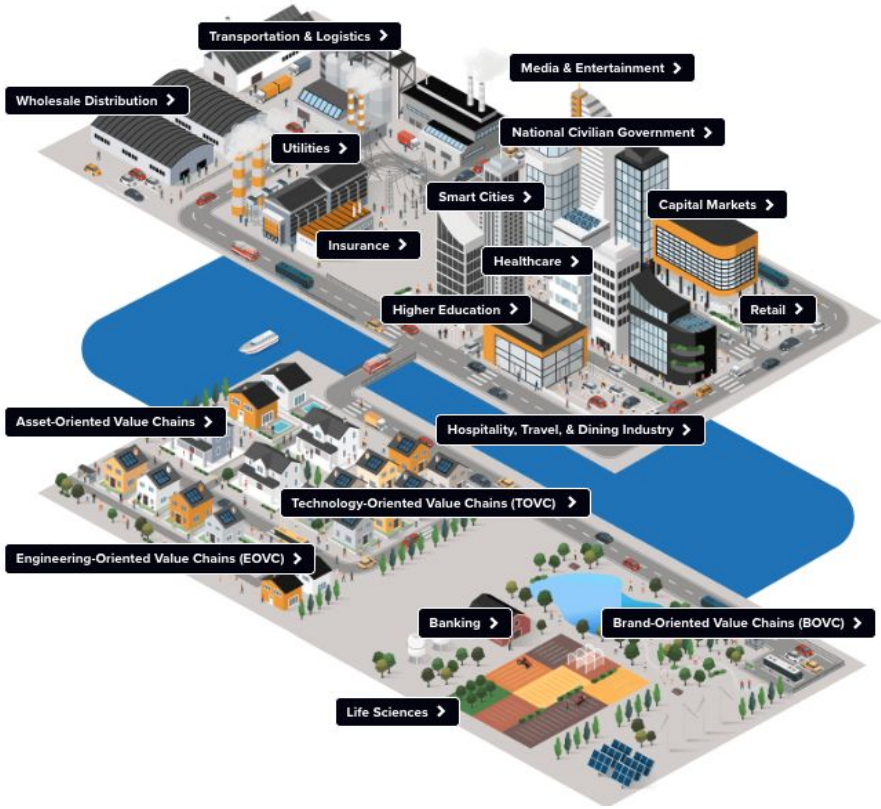
Red Hat

Generative AI is a strategic enabler across industries

AI use cases that drive productivity and efficiency



Every business function



Every vertical industry

IDC: Use cases by industry or function

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

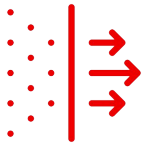
Risk Management

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

Generative AI customer adoption challenges

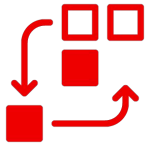
Cost

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.



Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

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**

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

Accelerate Agentic AI delivery and stay at the forefront of innovation

Red Hat

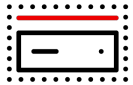
Red Hat AI Inference Server

Red Hat OpenShift AI

Trusted, Consistent and Comprehensive foundation

NVIDIA AMD intel Hardware Acceleration Google aws IBM

Physical



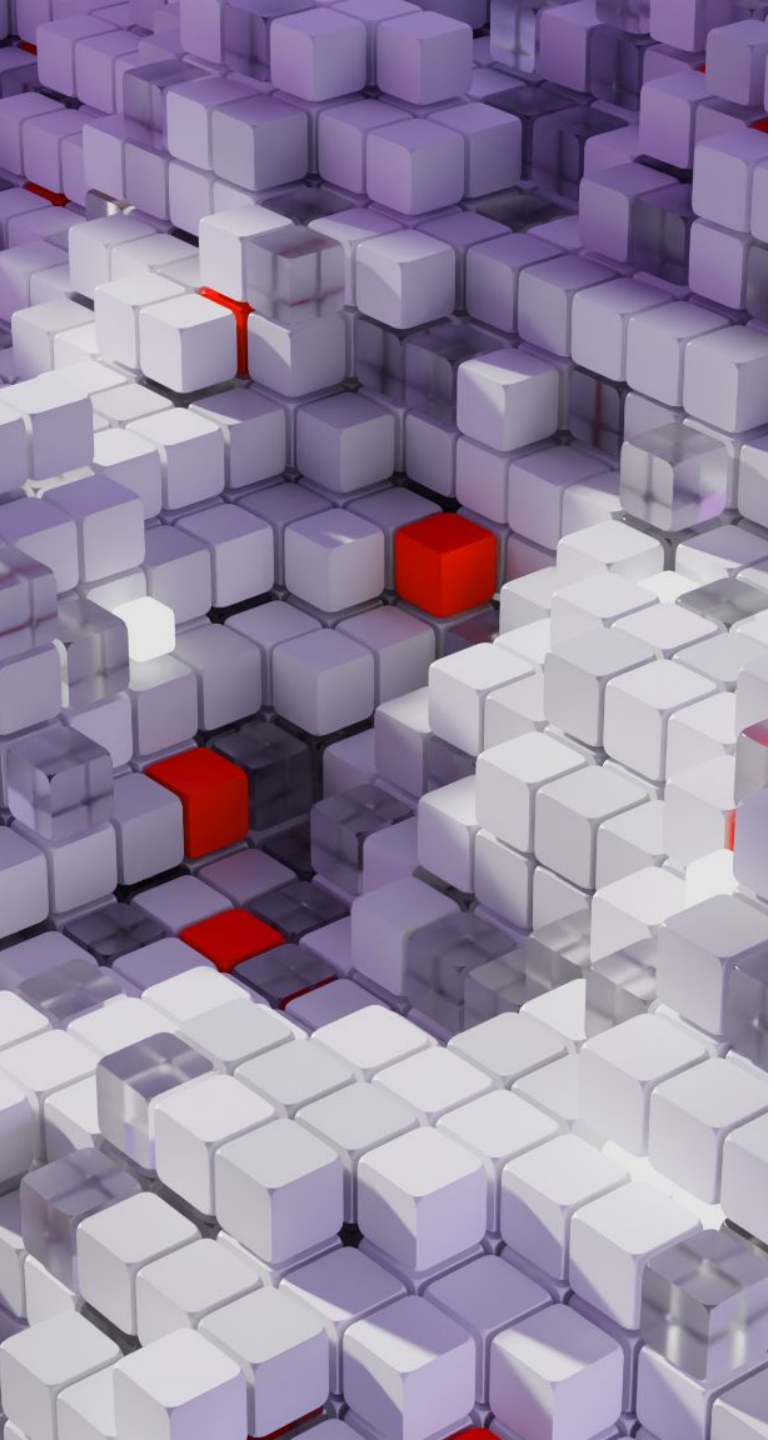
Virtual

Private Cloud

Public Cloud

Edge

* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap



Fast, flexible and scalable inference

Red Hat

The requirements of an enterprise AI production systems

Identifying the tradeoffs of inference

Need to be fast and **accurate** in its responses

Manage processing times and token output to control **cost**

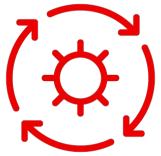
Deliver high throughput and lower latency for best **performance**

Red Hat AI helps address the trade-off challenges of inference

Gain consistent, fast and cost-effective inference at scale with vLLM

Select an optimal LLM

Red Hat AI third party validated and compressed models ready-to-use



An efficient inference runtime

vLLM

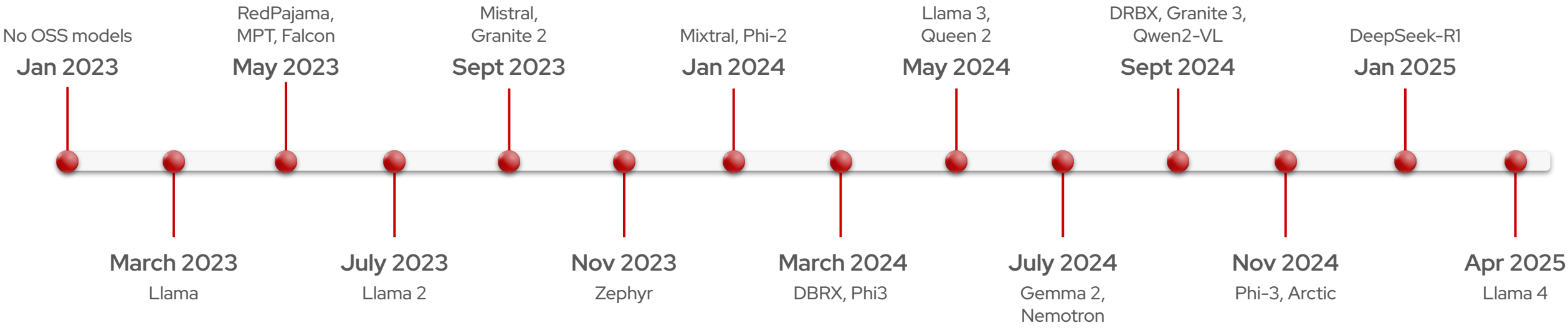
Broad support for hardware

AMD NVIDIA IBM Google intel aws

Red Hat

The power of open

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



Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators

Llama · Qwen · DeepSeek · Gemma · Mistral · Molmo · Phi · Nemotron · Granite

vLLM

GPU · Instinct · TPU · Neuron · Gaudi · Spyre

Physical · Virtual · Private Cloud · Public Cloud · Edge

Red Hat

The value of vLLM

Deliver fast, flexible and scalable inference

Faster response time

vLLM can achieve higher throughput, this translates to processing more tasks or requests within a given amount of time.

Reduce hardware costs

vLLM offers a more efficient use of resources, which is equivalent to fewer GPUs needed to handle the processing of LLMs.

Efficient memory management

vLLM organizes virtual memory, this translates to handling larger models and longer sequences more effectively within a given hardware setup.

Designed for security and scale

Self-hosting an LLM with vLLM provides you with more control over data privacy and usage, as well as an ability to handle growing demand.

Red Hat AI repository on Hugging Face

A collection of third-party validated and optimized large language models

Broad Collection of models

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

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 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

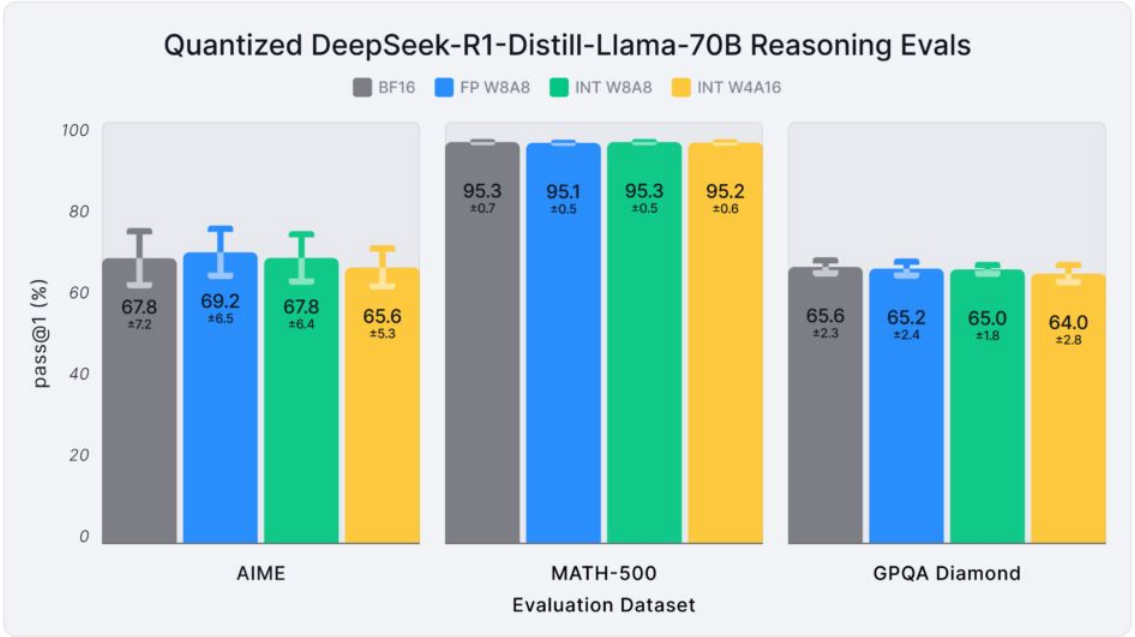
Framework for reducing the size and computational requirements of a LLMs while preserving accuracy

Receive tailored capacity planning guidance from our experts

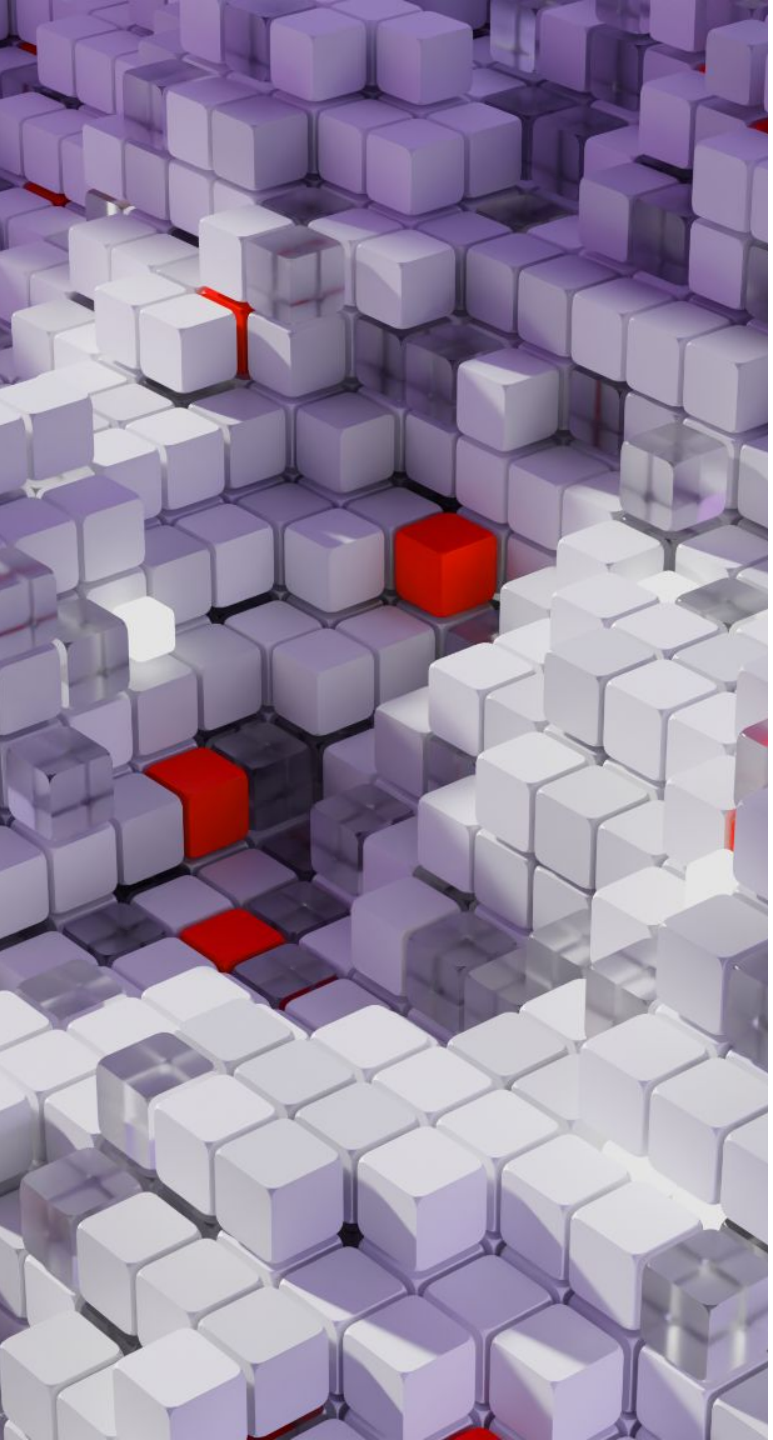
Ex. Compressed DeepSeek-R1 models

State-of-the-art, open-source quantized reasoning models built on the DeepSeek-R1-Distill

- **FP8 and INT8 quantized versions achieve near-perfect accuracy recovery** across all tested reasoning benchmarks and model sizes —except for the smallest INT8 1.5B model, which reaches 97%.
- **INT4 models recover 97%+ accuracy** for 7B and larger models, with the 1.5B model maintaining ~94%.
- With **vLLM 0.7.2,** deliver **4X better inference performance** across many common inference scenarios.
- **Reduce GPU requirements** by (e.g. 50% for INT8)



https://developers.redhat.com/articles/2025/03/03/deployment-ready-reasoning-quantized-deepseek-r1-models#

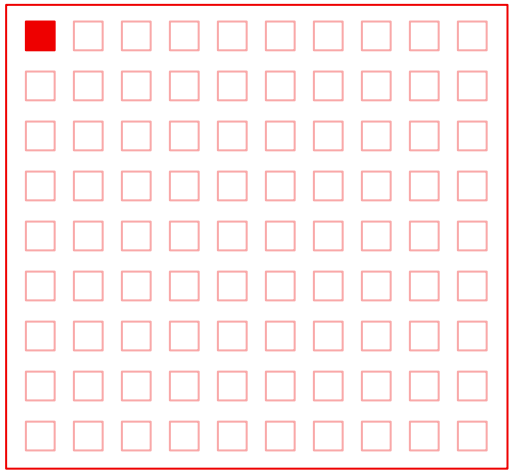


Connecting models to data

Red Hat

Enterprises need models aligned to their private data

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



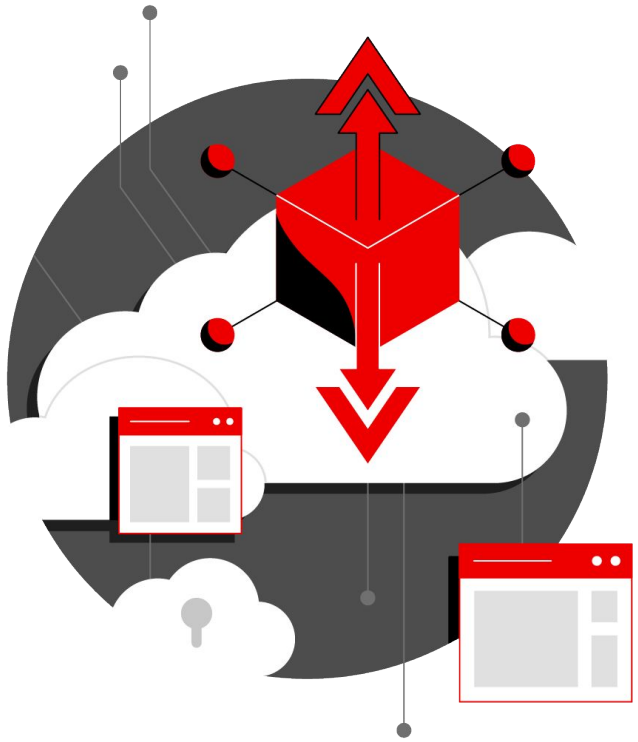
Less than 1% of all enterprise data is represented in foundation models

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

The value of open source and smaller language models

Smaller models are more efficient & customizable



- Open source AI models are catching up to proprietary models.
- Smaller language models, like **IBM Granite**, are orders of magnitude smaller than frontier models.
  - Models with less than 10 billion parameters **are cheaper and faster to run**, and consume less energy.
- These models can be **tuned and customized with private enterprise data** for domain specific tasks.
- **Customers own their own models** and can create multiple instances for different use cases and deployment environments.

Granite models

A family of open, performant and trusted AI models to accelerate enterprise AI adoption

Open

Open source under the apache 2.0 license and available on watsonx.ai, Hugging Face, and other platforms.

Performant

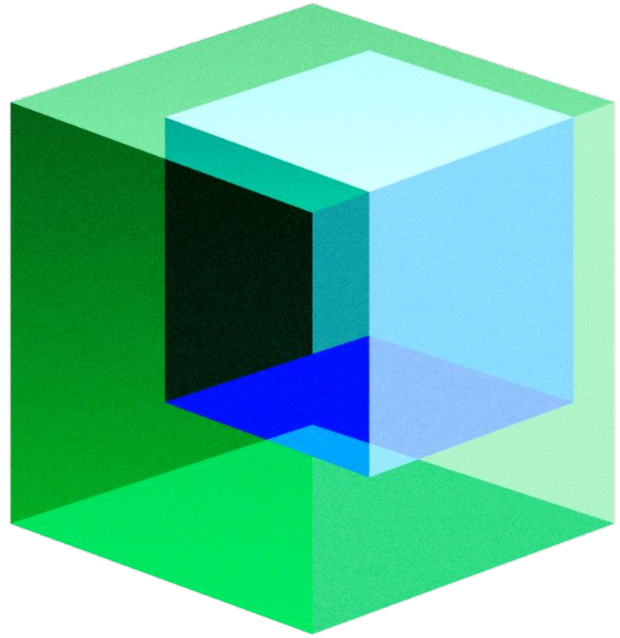
Diverse range of smaller, fit-for-purpose models that deliver performance on par with similar-sized models.

Trusted

Trained on trusted and governed data relevant to enterprise domains. Models offer IP indemnification and support.

Cost-effective

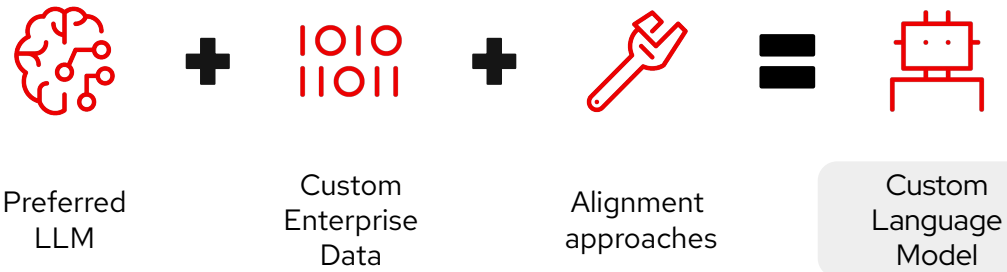
Offer lower cost of inference, and lower infrastructure hosting costs.



Foundation models for: code, language, time series, agents, safety via the Granite Guardian companion model, and even geospatial data

Red Hat

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

Red Hat AI provides multiple model alignment approaches

Build customized AI solutions that address domain specific business cases

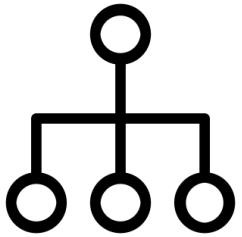
RAG

Retrieval Augmented Generation

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

InstructLab

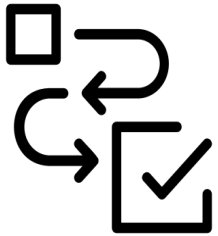
Large-scale Alignment for chatBots



Leverage a taxonomy-guided synthetic data generation process and a multi-phase tuning framework to improve model performance.

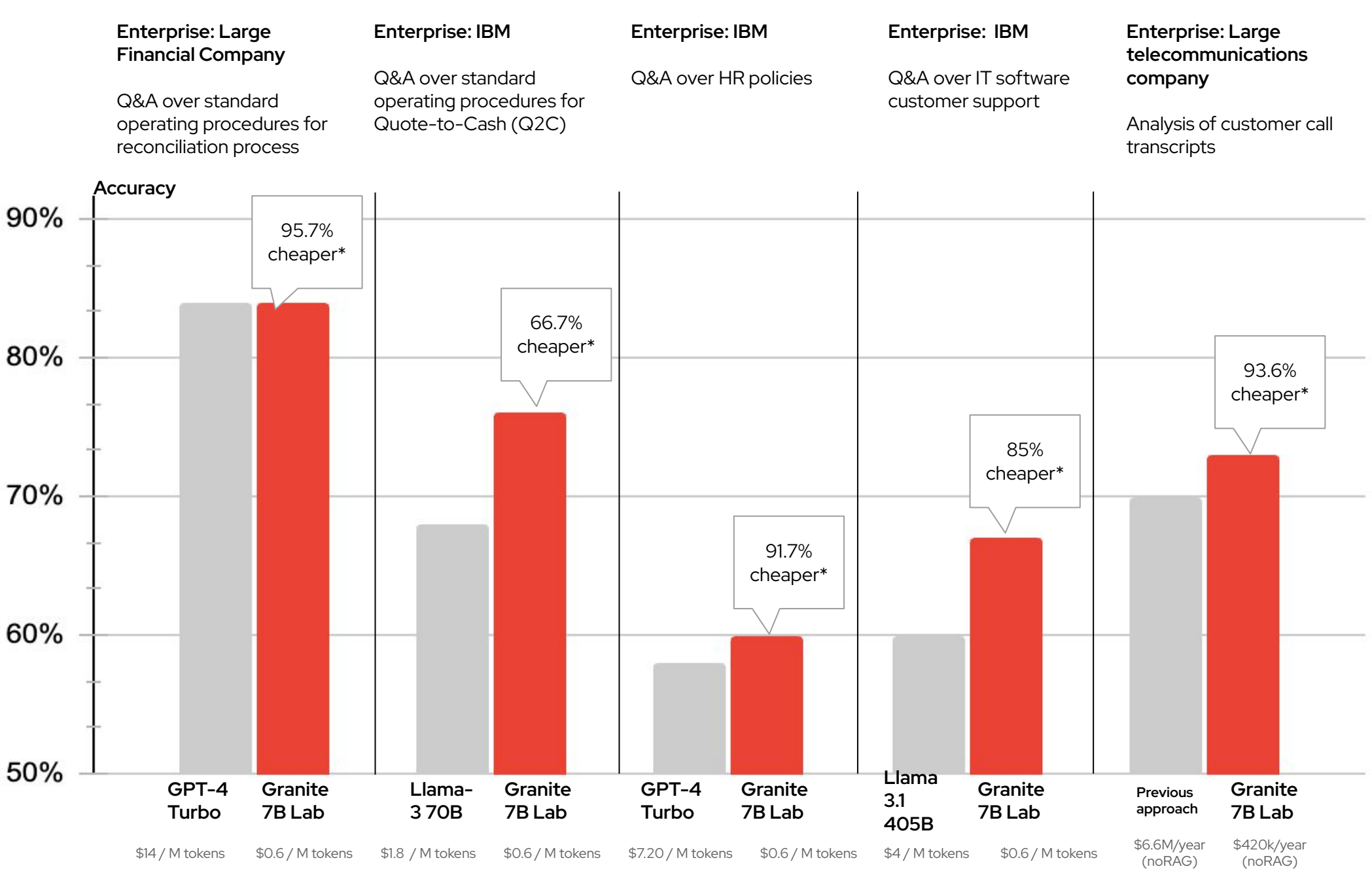
Fine tuning

Fine Tuning, LoRa and QLora

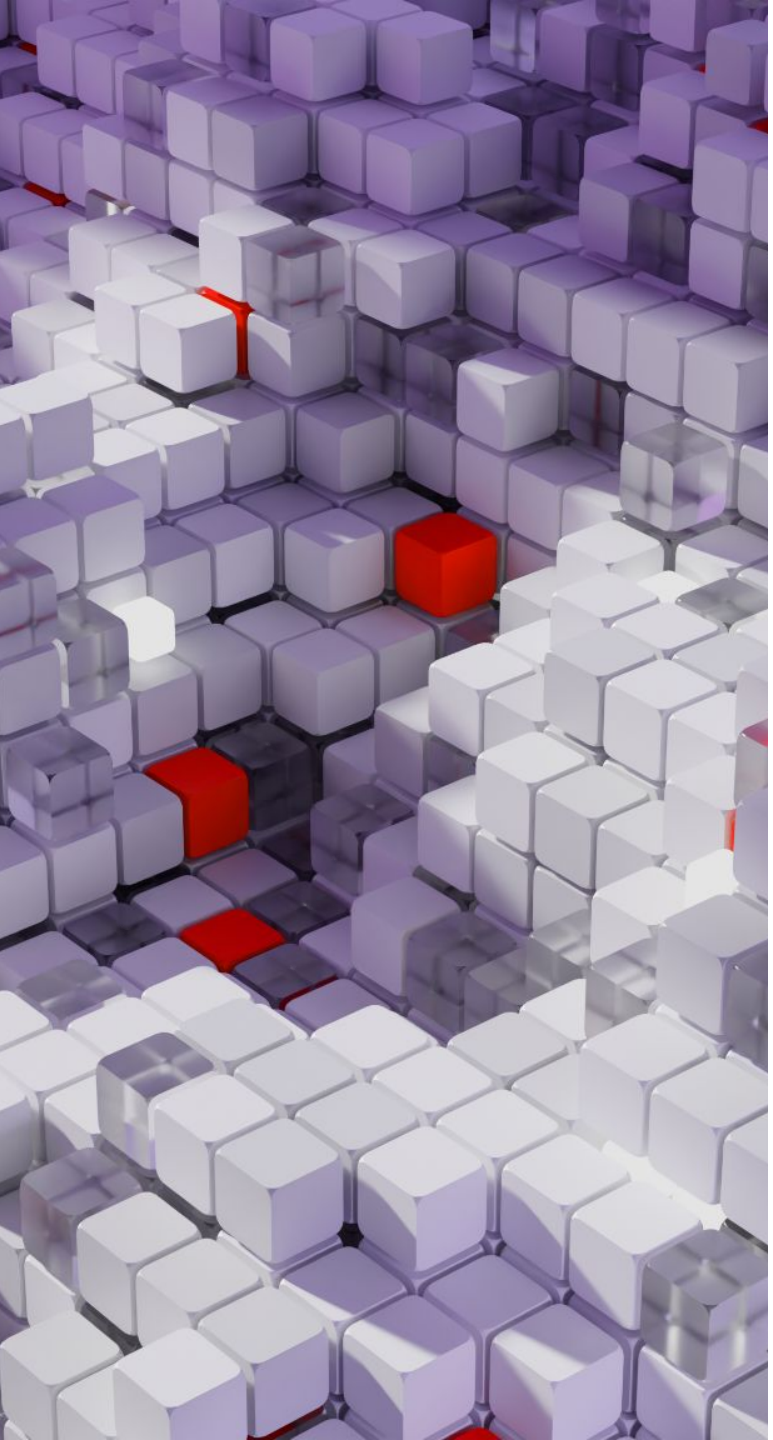


Adjust a pre-trained model on specific tasks or data, improving its performance and accuracy for specialized applications without full retraining.

The value of enterprise data can be seen in tuning smaller, targeted, optimized models to deliver state-of-the-art performance at considerably lower cost.



*SaaS cost per million tokes (assuming blend of 80% input, 20% output), https://artificialanalysis.ai/models/prompt-options/multiple/medium#pricing

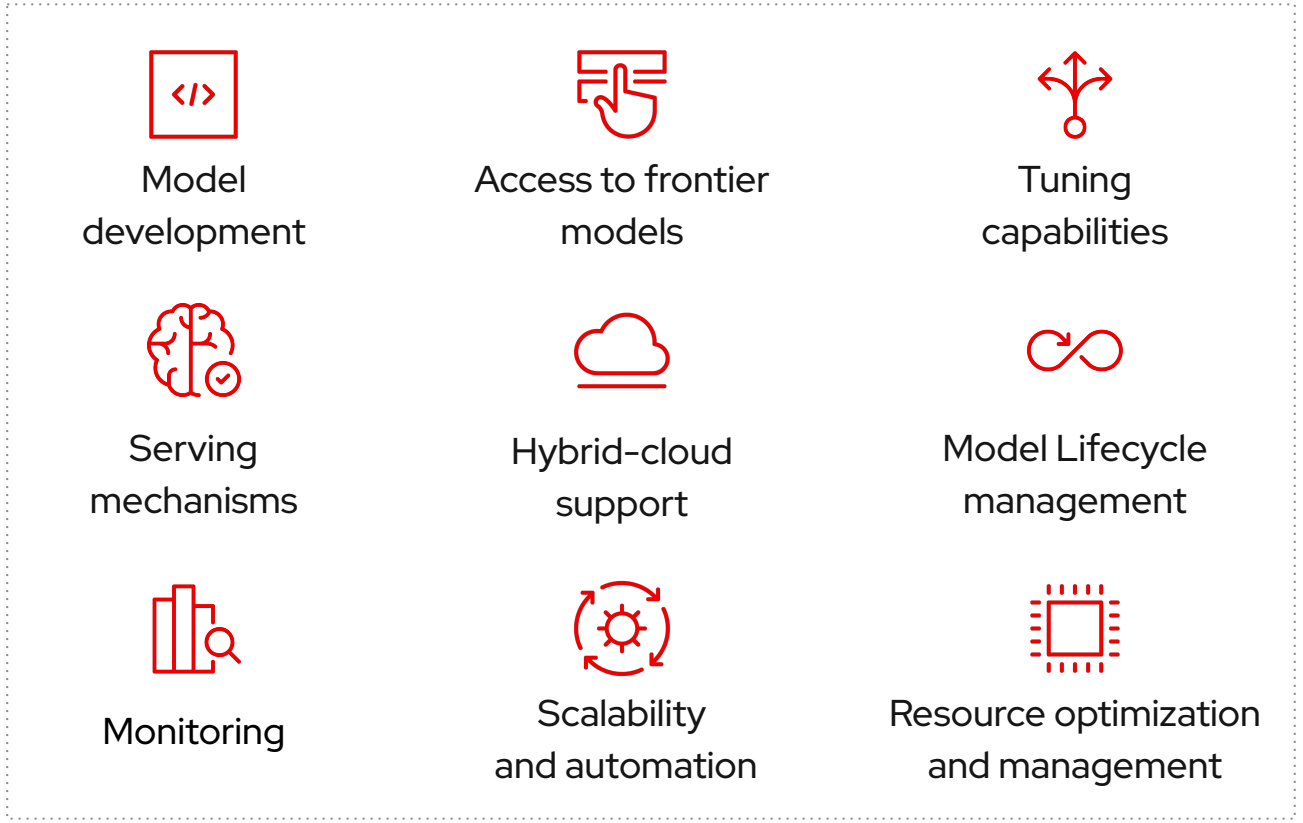


Scaling AI across the hybrid cloud

Red Hat

Components of an AI platform

Successful AI implementations require more than models and GPUs

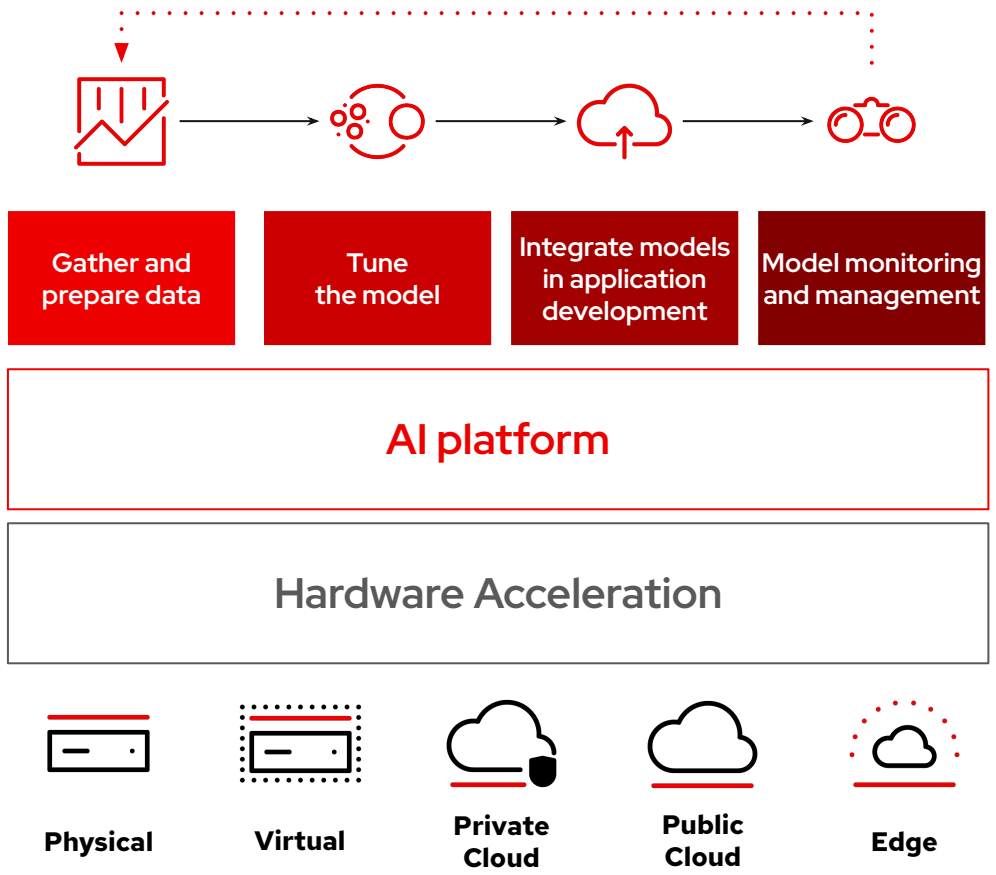


Red Hat

Red Hat AI provides a platform for consistently building, deploying and running AI models, AI-enabled applications, and AI agents across the hybrid cloud at scale.

It provides:

- An efficient inference runtime (vLLM)
- Validated and optimized third-party models
- InstructLab and RAG for customization
- MLOps and LLMOps capabilities
- Monitoring, bias detection and guardrails



Red Hat

Hybrid cloud deployment for AI

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

Hardware Accelerators

NVIDIA GPU

AMD Instinct

intel Gaudi

Roadmap

Google TPU

aws Neuron

IBM AIU

OEM Servers

DELL

Lenovo

CISCO

Hewlett Packard Enterprise

Cloud Environments

IBM Cloud

aws

Google Cloud

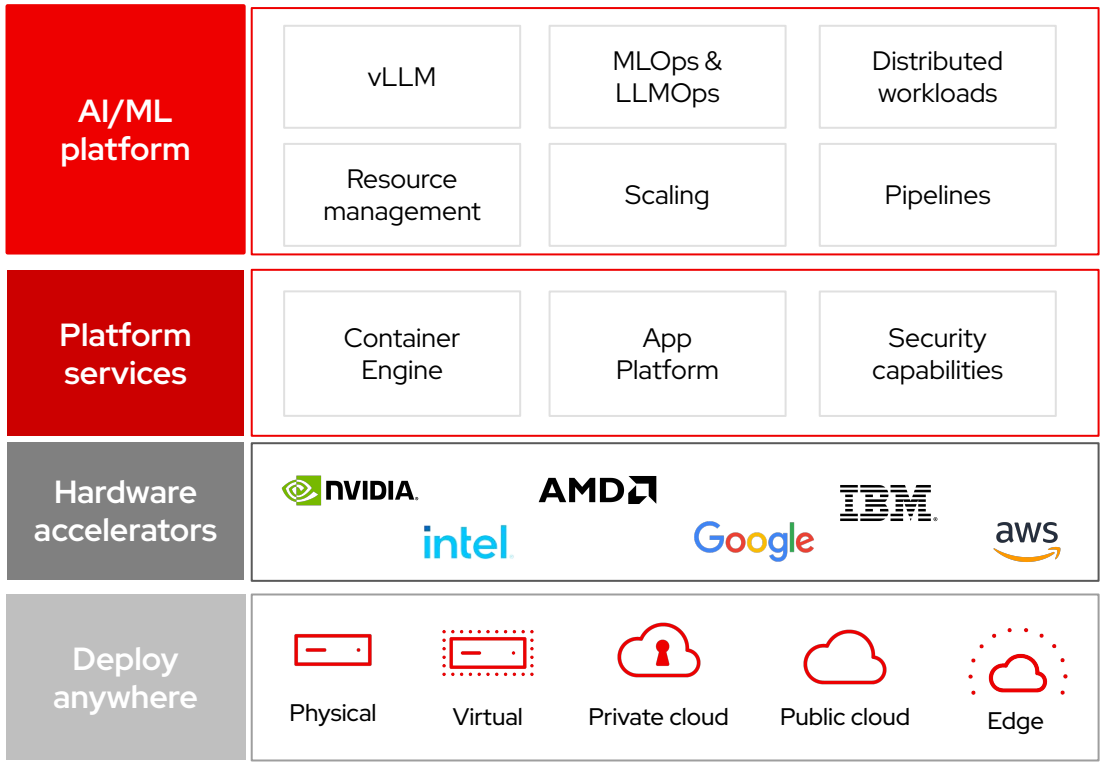
Azure

Red Hat

* NVIDIA GPUs fully supported in Red Hat AI. AMD Instruct $ Intel Gaudi in Preview, AWS Inferentia/Neuron, Google TPU, IBM AIU are on our roadmap

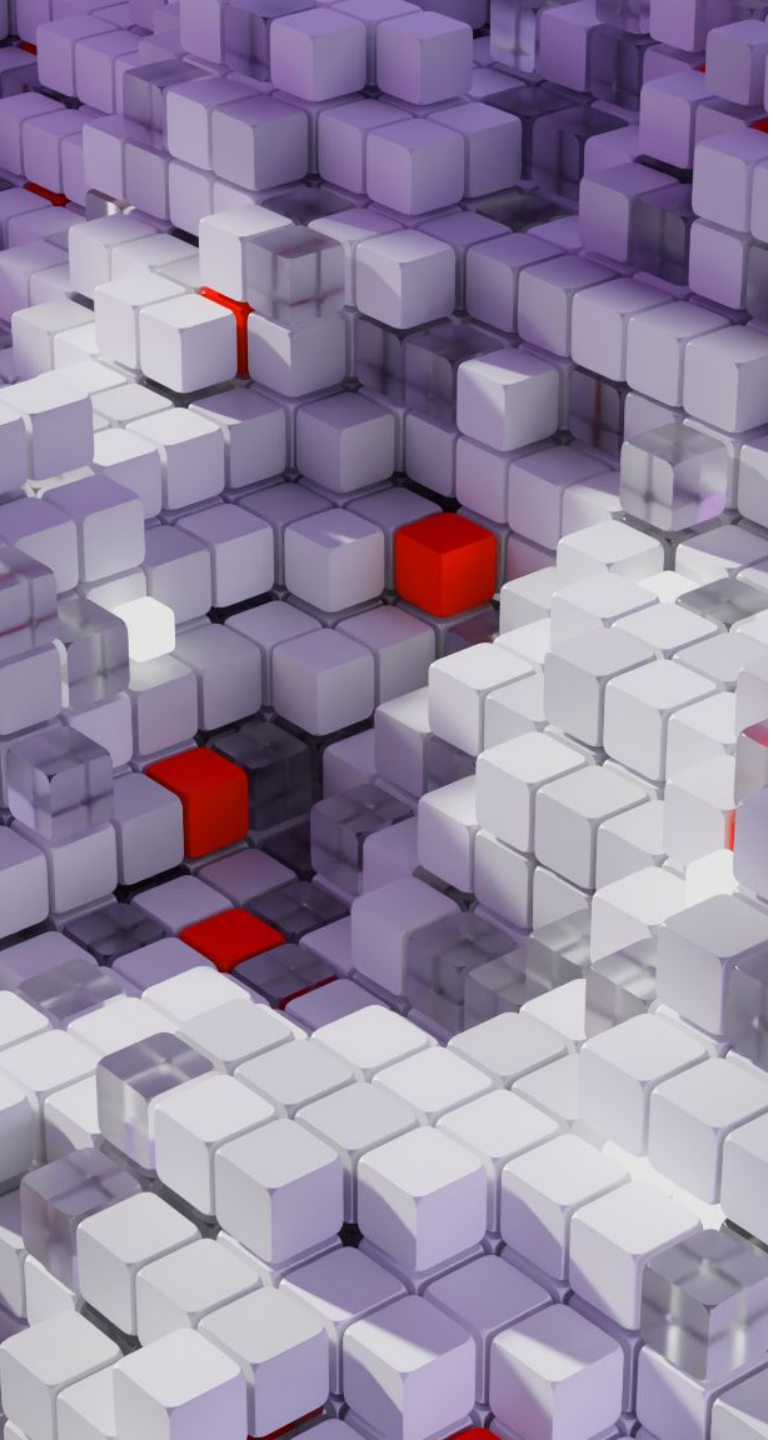
Scale and optimize your AI and application deployments

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

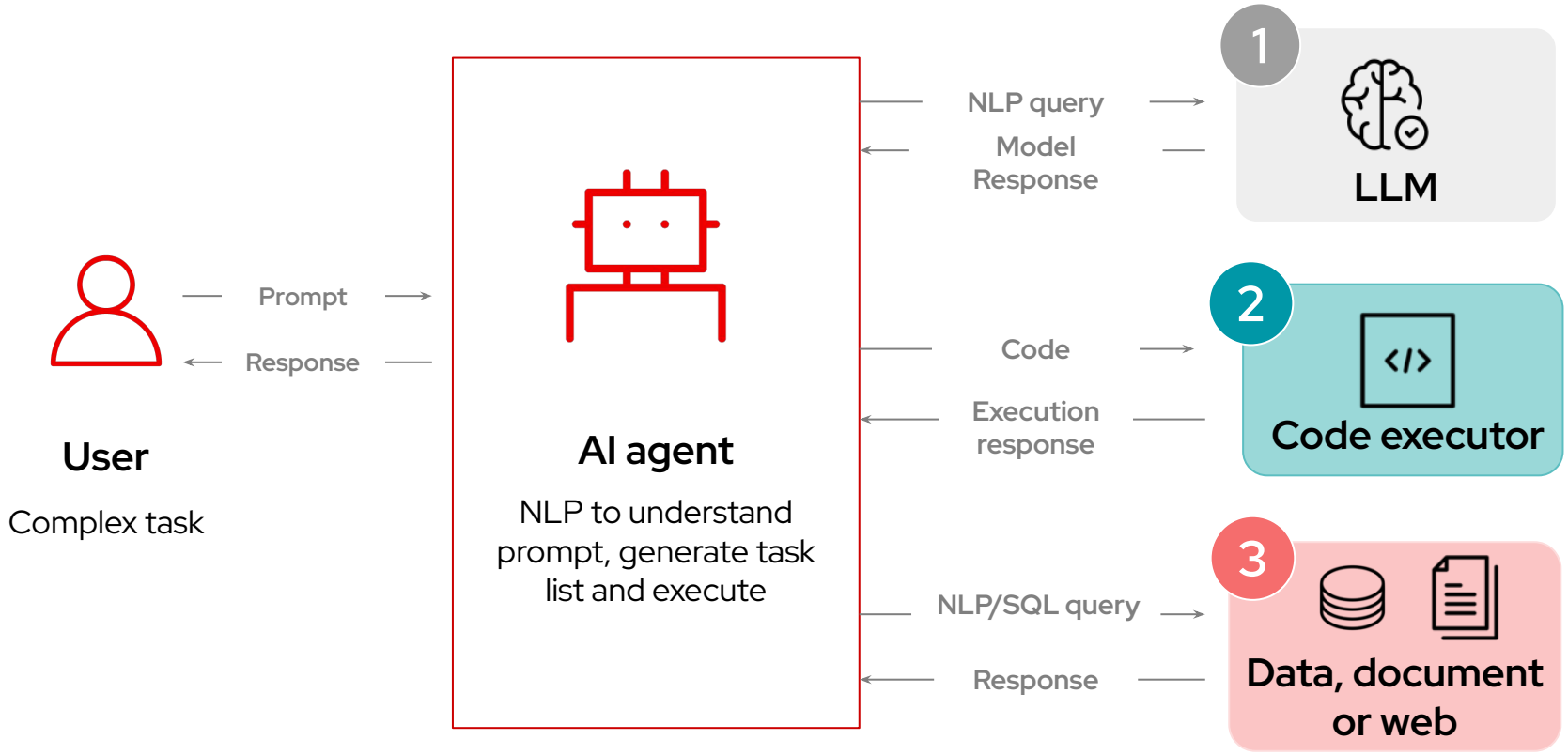


Accelerate Agentic AI innovation

Red Hat

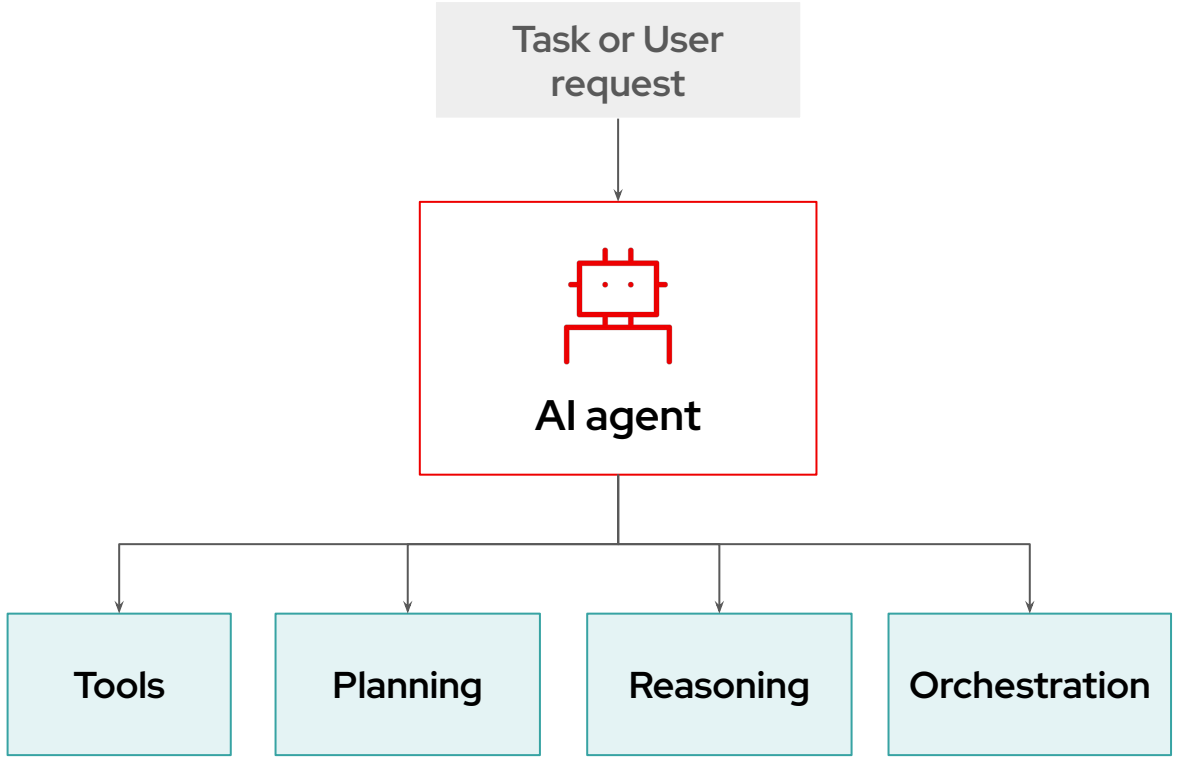
AI agents integrate models, functions & tools

Gen AI Models, Predictive AI Models, Code Functions, Search & more



Red Hat

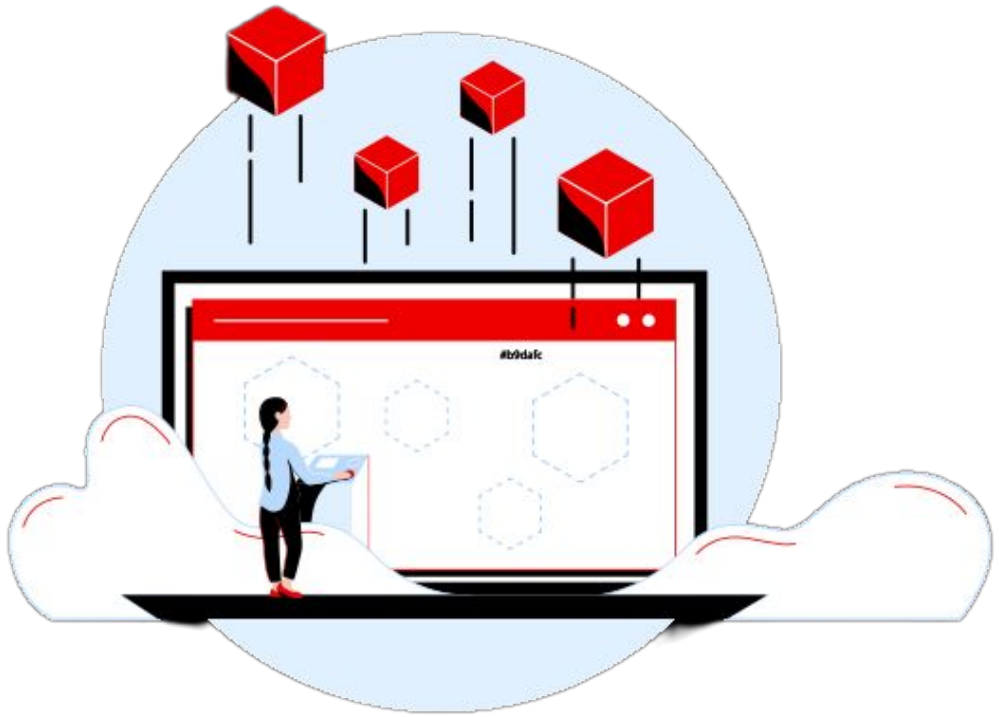
The components of an AI Agent system



- **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- **Orchestration:** Coordinates actions, tools, and agents to dynamically adjust and complete tasks.
- **Communication protocols:** enables the connections between the components.

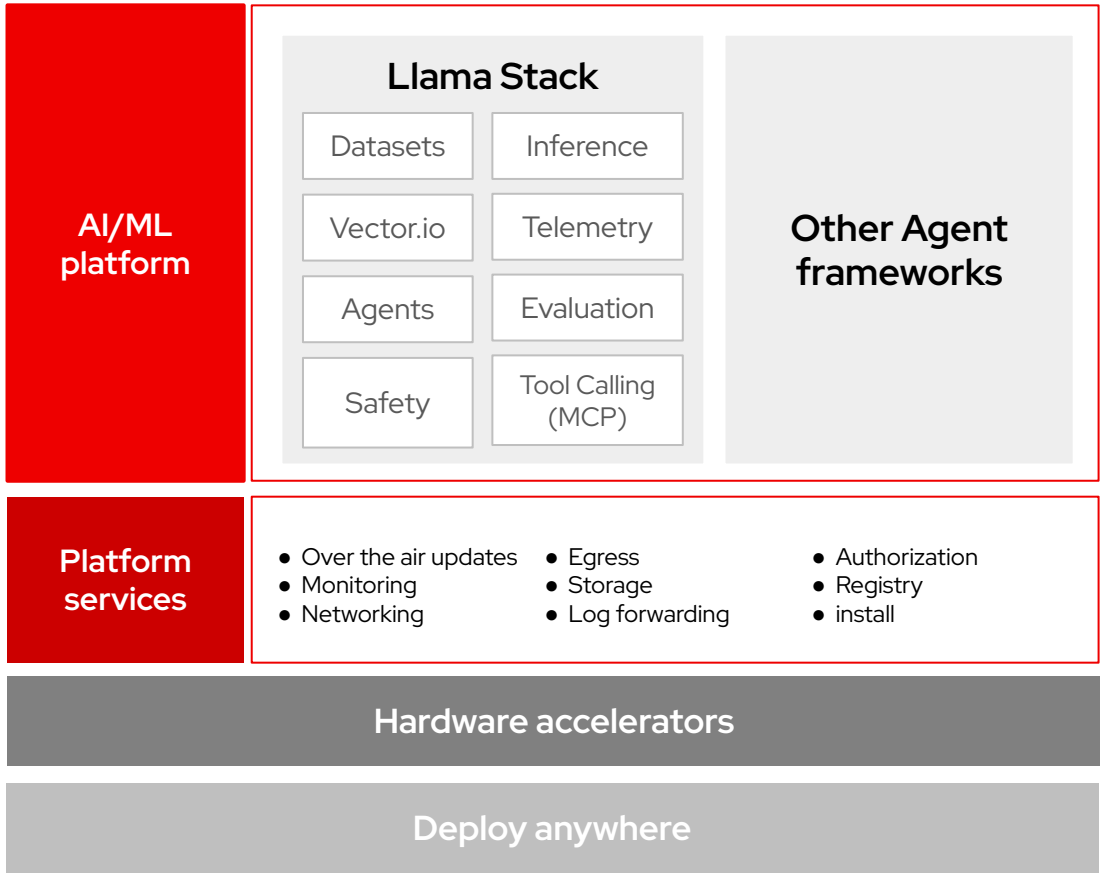
Red Hat AI provides an agile, stable foundation to accelerate the development and deployment of AI agentic workflows.

- 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

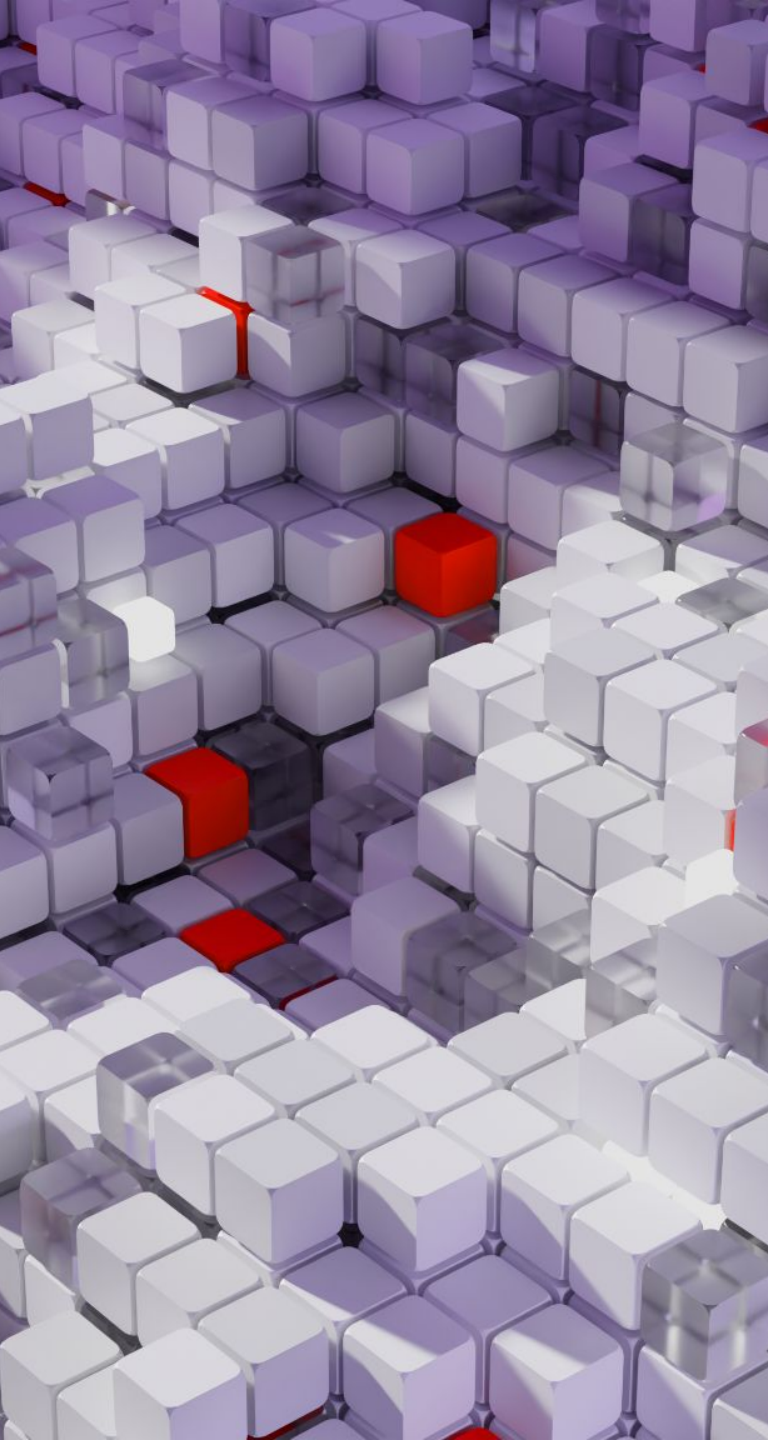
A modular approach to building AI agents



Red Hat AI allows to:

- Build agents using **Llama Stack's native capabilities and implementations**.
- **Bring compatible Llama Stack implementations** to OpenShift AI.
- **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- **Build with Core Primitives** and manage your own agent framework as a standard workloads.

Red Hat

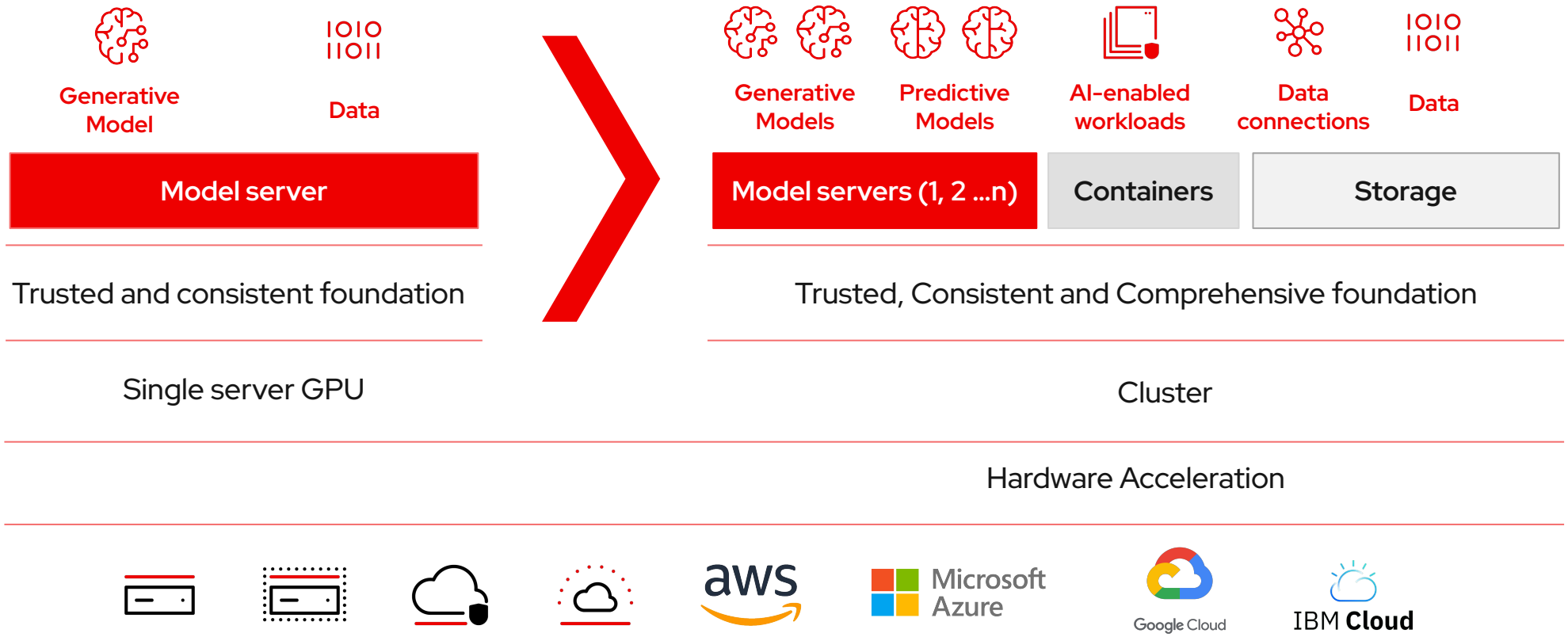


Why Red Hat AI?

Red Hat

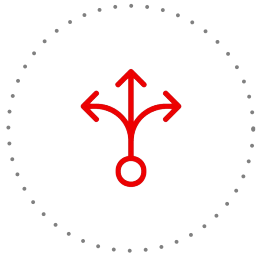
Red Hat AI supports each stage of the AI adoption journey

From single server deployments to highly scaled-out platform architectures



Increasing flexibility and choice with an open source approach

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

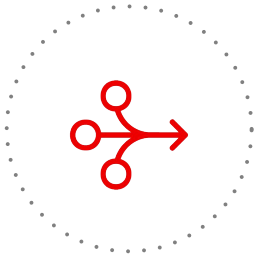


Flexibility

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.

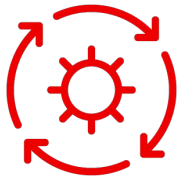


Abstract the complexity

Reduce the complexity of switching and adapting to new technologies.

Red Hat

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

agesic

"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

Source: Red Hat Summit presentation - May 2024

Financial services: EMEA

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

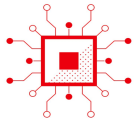
- Rapidly prototype AI applications and services
- Develop RAG+RAFT based patterns for model tuning and training
- Prototype AI Assistants & chatbots
- Develop evaluations for model accuracy and speed
- Prototype data ingestion pipelines

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

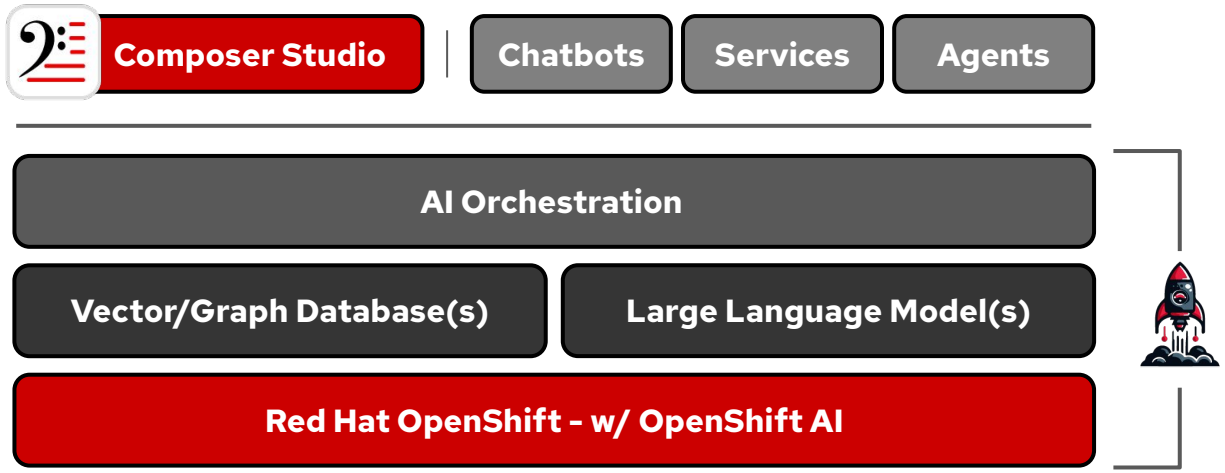
Operational guidance & advisory services from TAM Services for Red Hat AI Platform (yearly subscription)

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

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

TRY RED HAT OPENSHIFT AI

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

PROOF OF CONCEPT (POC) DESIGNED FOR YOU

Let us bring your vision to life: Request your personalized POC today!

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

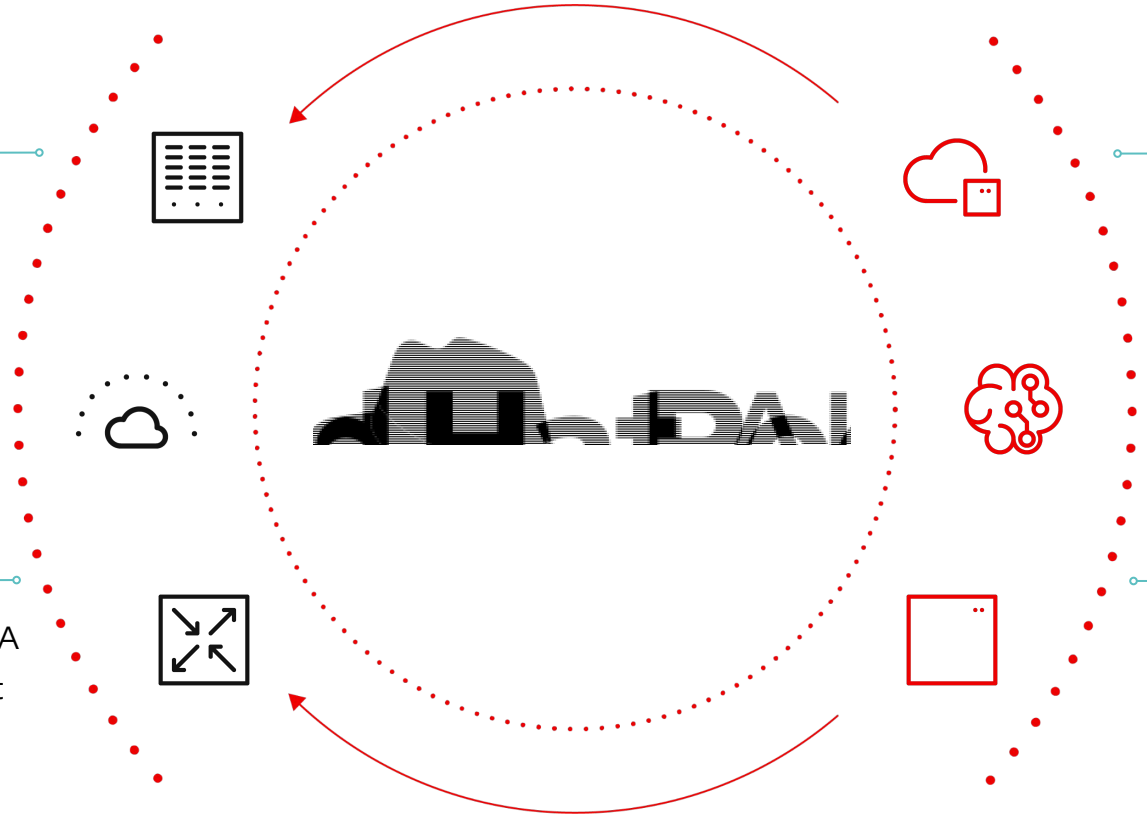
Red Hat AI Announcements

Fast and Efficient Inference

- Introducing Red Hat AI Inference Server
- LLM Compressor tools
- Announcing llm-d for inference at scale

Connecting Data to Models

- Red Hat AI InstructLab on IBM Cloud now GA
- InstructLab distributed training in OpenShift AI with Kubeflow Training Operator
- InstructLab multi-language capabilities



AI Platform

- Red Hat AI Validated models
- Model catalog and feature store in OpenShift AI (Tech Preview)
- RHEL AI now available in Google Cloud Marketplace

Agentic AI

- Llama Stack (Dev Preview)
- Model Context Protocol (Dev Preview)

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

Red Hat

vLLM Inference Server in Red Hat AI

vLLM connects model creators to accelerated hardware providers

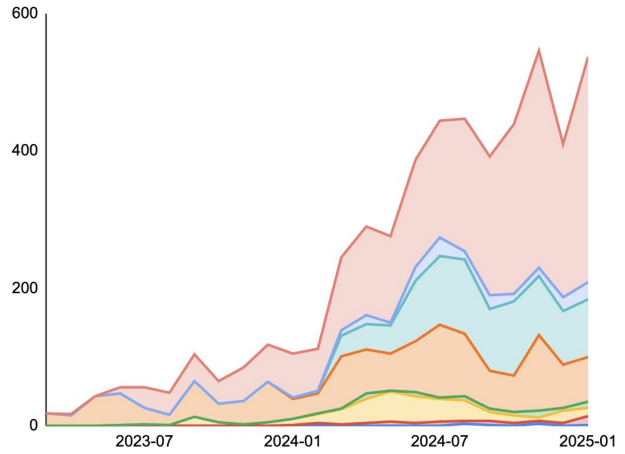
Model creators

Meta Llama
Mistral
Qwen
Microsoft Phi
IBM Granite
Ai2 Molmo
AI21 labs Jamba
snowflake Arctic
databricks DBRX
Hugging Face

Hardware Vendors

NVIDIA — Features for new HW
AMD — Choice for MI300X
intel — Gaudi enablement
Google — TPU enablement
aws — Neuron enablement
IBM — Spyre enablement

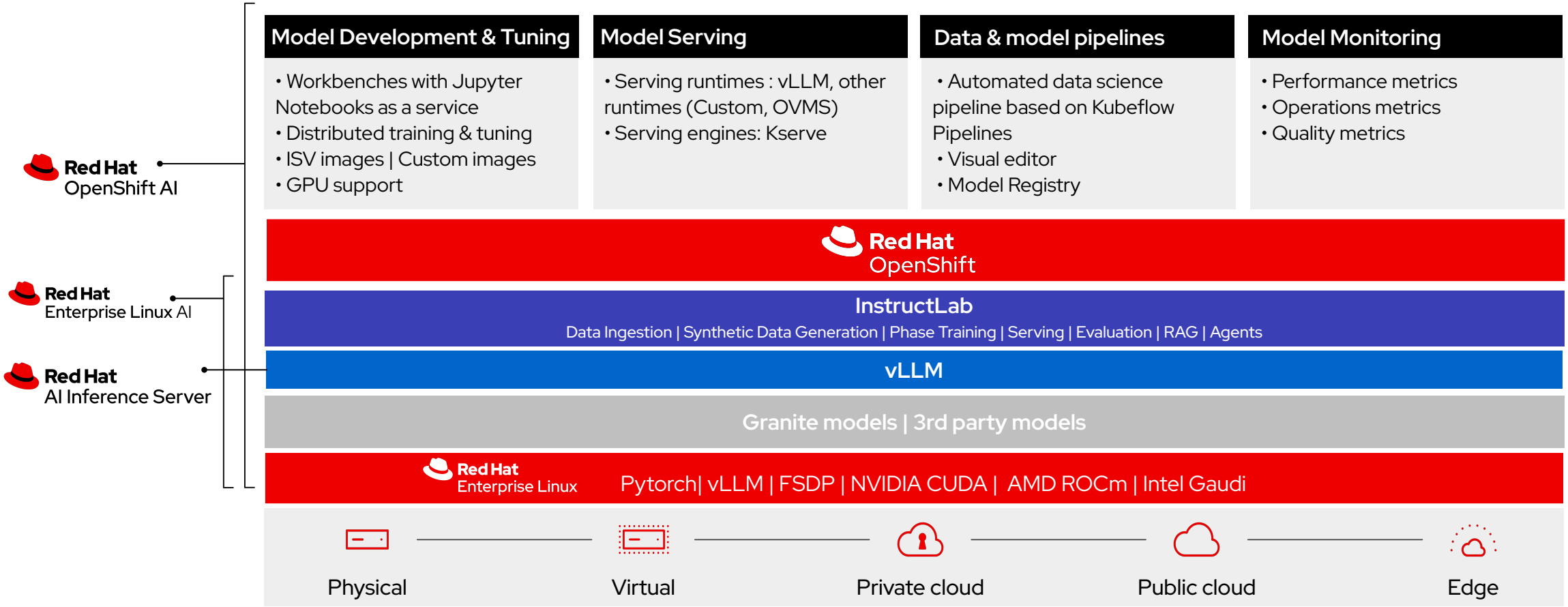
Contribution Trajectory



Red Hat

Red Hat AI platform

Generative AI, Predictive AI & MLOps capabilities for building flexible, trusted AI solutions at scale



Red Hat