# Build on an integrated AI platform with Red Hat AI Enterprise

## The rapid rise of AI

AI has already seen widespread adoption that has resulted in:

- 55% of organizations using gen AI.[1]
- Over 50% of organizations conducting proofs of concept or implementing selected use cases for agentic AI.[2]
- 30% of AI-enabled applications expected to be using agentic AI by the end of 2026.[2]

## Evolving AI opportunities creates increased complexity

AI has given enterprises an opportunity to redefine how they operate and compete. It has quickly become a key piece of today's business landscape, but despite increasing AI adoption, many enterprises face challenges that obstruct its full value.

This includes rising model costs, complex customization needs, strict deployment constraints, and keeping up with the accelerated pace of innovation, among others.

As enterprises begin to focus more on gen AI innovation, agentic AI workflows, and sovereign AI, those challenges are only becoming more complex to navigate. This requires a strategic shift to AI platforms that can help enterprises lower inference costs, streamline scaling and monitoring, build sovereign on-premise AI, and swiftly adapt to change.

## Address AI complexity with enterprise platforms and tools

Whether enterprises are at the beginning of their AI journey or scaling established AI initiatives across an organization, Red Hat® AI can help them accelerate AI innovation and reduce the operational cost of developing and delivering AI solutions across hybrid and multicloud environments.

Red Hat AI delivers cost-effective solutions with optimized models and efficient inference, streamlined integration with private data, and accelerated delivery of agentic AI on a scalable, flexible platform. It allows organizations to manage and monitor the lifecycle of both predictive and gen AI models at scale, from single-server deployments to highly scaled-out distributed platforms.

Built on proven open source technologies, Red Hat AI provides support for a wide range of leading open source AI models to help enterprises accelerate innovation and democratize access to valuable tools and technologies.

This is complemented by an AI partner ecosystem that offers tested and certified products and services focused on performance, stability, and accelerator support across infrastructures to help enterprises solve key business and technical challenges with AI.

1 IDC Tech Buyer Presentation. "Worldwide Generative AI Industry Use Case Early Adoption Trends, 2025: Executive Summary." Document #US53280825, Apr. 2025. (purchase required)

2 IDC Survey. "Agentic AI Impact on Digital Infrastructure Strategies." Document #US53418526, Oct. 2025. (purchase required)

Red Hat AI delivers:

- Complete AI model lifecycle monitoring and management.
- The flexibility to use any AI model on any hardware accelerator.
- A validated collection of optimized AI models.
- Flexible and cost-effective inference across hybrid or multicloud environments.
- Trusted integration into enterprises' private data estates.
- Streamlined delivery of agentic AI workflows.
- The ability to scale as needed with efficient use of resources.

Red Hat AI Enterprise is a unified AI platform that is offered as part of the Red Hat AI portfolio to help enterprises solve the challenges of adopting AI today.

## 6 reasons to choose Red Hat AI Enterprise as your AI platform

Red Hat AI Enterprise provides all of the tools and capabilities needed to support enterprises at any stage of their journey. It allows them to deploy and manage efficient and cost-effective AI models, agents, and AI-powered applications across hybrid and multicloud environments on a unified AI platform.

This platform offers enterprises a ready-to-use development environment and key AI capabilities. This includes model tuning, high-performance inference, agentic AI workflow management, and the flexibility to support any model, use any hardware, and deploy anywhere to meet evolving data location requirements.

### Streamline AI deployments with flexible and efficient inferencing

As AI innovation continues to accelerate, the complexity of these deployments has also increased, which can result in inefficient resource usage, heightened operational costs, and lowered quality of results.

Red Hat AI Enterprise offers optimized model inference across diverse environments, including on premise, hybrid and multicloud environments, or at the edge of a network. This helps enterprises streamline model deployments while using fewer resources, lowering inference costs, and maintaining accuracy.

This is made possible through an optimized runtime powered by vLLM—an inference server that accelerates the output of gen AI-powered applications by making better use of the GPU memory to optimize throughput and latency. It also includes advanced compression techniques that help optimize for cost by lowering the size and computational requirements of a model.

When an organization is ready to scale—moving from a few models to dozens and from dozens of users to hundreds—the platform offers the option to use llm-d. This technology provides a distributed inference framework for cost-effective, predictable performance at scale for gen AI models by building on vLLM's

core innovations and delivering improved performance as volume increases.

**Improve AI model accuracy with private, enterprise data**

Despite the capabilities of publicly available gen AI models, enterprises need AI models that are trained on their own unique data to address domain-specific use cases.

Red Hat AI Enterprise delivers a streamlined and consistent experience for enterprises to support efficient model customization and help AI engineers and data scientists improve the accuracy and relevance of model responses. It allows enterprises to connect models to organizational data for Retrieval Augmented Generation (RAG) use cases.

This platform also supports fine-tuning and continual and reinforcement learning, and delivers a modular, flexible approach for document processing and parsing, synthetic data generation, and model evaluation and tasks.

**Accelerate innovation with agentic AI workflows**

AI technology has evolved at a rapid pace in a relatively short period of time. Starting with predictive AI opening up many new use cases, followed by the popularization of gen AI models, and now with the focus on agentic AI workflows to deliver the next evolution that will further extend AI strategies and bring even more business value to enterprises.

Red Hat AI Enterprise provides an agile, stable platform to help enterprises successfully deliver value with agentic AI. It achieves this with a unified application programming interface (API) layer, out-of-the-box components that streamline agentic AI workflow assembly for enterprise use cases, dedicated user experiences, and a flexible, scalable foundation that supports the deployment and management of agentic AI systems.

The platform also provides support for the Model Context Protocol (MCP), which is a crucial component for agentic deployment, acting as a standardized translator between a wide range of tools and capabilities and large language models (LLMs).

**Scale AI across the hybrid cloud with flexibility and consistency**

Organizations across many industries have been able to prove the value of AI in initial test use cases. Yet, too many of them still struggle to scale that value across an enterprise while also maintaining cost efficiency and adherence to regulatory compliance, privacy, and security. This becomes especially difficult when that includes diverse IT environments, such as hybrid and multicloud environments, or AI deployments at the edge of a network.

Red Hat AI Enterprise provides the control and consistency needed to train, tune, deploy, manage, and scale predictive AI and gen AI models wherever it makes the most sense for an organization's AI workload strategy, across any hardware or IT environment.

This platform offers flexibility by supporting any choice of AI model on any combination of hardware, original equipment manufacturers (OEMs), cloud providers, and datacenters. Combining this with a comprehensive suite of observability and monitoring capabilities, enterprises are able to optimize performance, control costs, and maintain governance and focus on security as they scale to meet

evolving business needs.

**Maintain data sovereignty with on-premise AI deployments**

Government regulations on the use of AI and how data is stored and processed continue to be put into place around the world. This has led many enterprises to explore how they can build on-premise AI deployments to maintain sovereignty and privacy.

By storing and processing their sensitive, private data on premise, enterprises can not only keep themselves in adherence with the many evolving regulations, but also safeguard the privacy of that data with a smaller attack surface and an increased focus on security.

Red Hat provides a comprehensive and layered approach to AI security, allowing organizations to confidently embrace AI innovation. Red Hat AI Enterprise offers this security approach on a trusted open hybrid cloud foundation to help address the unique security and safety challenges of the AI lifecycle.

It facilitates true sovereignty by supporting any model, any hardware, and security-focussed GPU sharing across all cloud and edge environments, including air-gapped environments. Red Hat works closely with sovereign cloud providers to allow organizations to deploy AI in private, security-focussed cloud environments.

Red Hat AI Enterprise also includes tools that support auditable trust via model explainability, fairness, and output policy enforcement, as well as lifecycle traceability with reproducible pipelines and audit readiness.

**Build AI on a trusted Kubernetes platform**

Red Hat AI Enterprise delivers this complete set of AI capabilities as part of an integrated, ready-to-use, end-to-end AI stack built on the trusted foundation of Kubernetes.

This provides enterprises who already use Kubernetes and containerization with a familiar and consistent experience across on-premise datacenters, hybrid or multicloud environments, and AI deployments at the edge. Enterprises can build, train, deploy, and manage modern, cloud-native, AI-powered applications on a single, centralized platform using the same tools, frameworks, and skills their teams already know.

**Learn more about the value of Red Hat AI Enterprise**

Read more about how Red Hat AI Enterprise helps organizations deploy efficient and cost-effective AI across the hybrid cloud with a unified platform that supports every stage of the AI journey.

**About Red Hat**

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.


| North America | Europe, Middle East, and Africa | Asia Pacific | Latin America |
|---|---|---|---|
| 1 888 REDHAT1<br>www.redhat.com | 00800 7334 2835<br>europe@redhat.com | +65 6490 4200<br>apac@redhat.com | +54 11 4329 7300<br>info-latam@redhat.com |

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat