Red Hat AI

Manage AI in production with an enterprise-grade platform

Trusted, comprehensive, and consistent

Red Hat AI is a portfolio of products and services that accelerates time to market and reduces the operational cost of delivering AI solutions across hybrid cloud environments. It supports efficient tuning of small, fit-for-purpose models with enterprise-relevant data and provides the flexibility to deploy models wherever your data resides.

Advancements in AI are transforming business models

Rapid advancements in artificial intelligence (AI) are transforming industries and redefining traditional business models. As a result, enterprise IT organizations are under increased pressure to design, build, and deliver AI solutions that provide a competitive advantage. Although AI offers the potential to boost operational efficiency and productivity, many organizations have yet to fully integrate these technologies into their daily operations or realize their full business benefits.

Training and running large AI models at scale can be costly, particularly as organizations deploy multiple models and AI-enabled applications into production. Customizing these models with business-specific data requires significant compute resources, extensive datasets, and specialized expertise, all of which can lead to greater expenses. Investments in specialized hardware and qualified professionals to manage these models can further increase costs and complicate scaling of critical AI services. Additionally, managing model lifecycles and serving models for inference can also result in substantial, recurring expenses, especially for large AI models.

Aligning AI models with your specific business requirements and proprietary, confidential data can be challenging if your training and tuning processes are primarily designed for data scientists. Subject matter experts and developers often possess insights into enterprise data, applications, and business needs that are essential for effectively training and fine-tuning AI models. Without the right toolset, organizations can miss these valuable perspectives, resulting in AI models that are misaligned with business objectives and fail to deliver competitive advantages.

Data that is inaccessible or scattered across multiple locations can complicate model training, tuning, and inference. The location of models and data can influence decisions related to hardware availability, data privacy, and governance. Safeguarding proprietary data and minimizing the costs of complex AI infrastructure can be challenging without the flexibility to deploy AI solutions across distributed environments.

Red Hat offers comprehensive, trusted technologies that speed development and delivery of innovative AI solutions across hybrid cloud environments.

Streamline and speed AI operations

Red Hat® AI is a portfolio of products and services that accelerates time to market and reduces the operational cost of delivering AI solutions across hybrid cloud environments. The portfolio supports all stages of your AI adoption journey—from single-server deployments to highly distributed, scalable platform architectures. Designed to simplify AI adoption, Red Hat AI makes advanced technologies more

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

accessible across your entire organization. Integrate and manage both predictive and generative AI (gen AI) models at scale with increased security. Take advantage of support for a variety of hardware accelerators, original equipment manufacturers (OEMs), and cloud providers to ensure a stable, optimized, and high-performance environment. And deploy your critical AI applications and services across diverse environments, including on-site infrastructure and public cloud resources.

The Red Hat AI portfolio includes Red Hat Enterprise Linux® AI for individual Linux server environments, Red Hat OpenShift® AI for scalable, distributed Kubernetes platforms, and Red Hat AI Inference Server for optimized inference of large language models (LLMs). These solutions deliver open source technologies and models, providing access to the latest AI tools curated and integrated together for your entire organization. And the Red Hat AI partner ecosystem helps you speed innovation with a range of tested, supported, and validated products and services that address both business and technical challenges.

Increase efficiency with optimized models for any AI solution

Choosing the right model can significantly impact the efficiency of your AI solutions. Red Hat AI helps you deliver efficient, cost-effective, high-performance models fine-tuned with enterprise-relevant data. Red Hat AI increases efficiency by reducing development and deployment costs through smaller, optimized, open source language models. The portfolio helps you build predictive models, tune gen AI models, and deploy a combination of both across hybrid cloud environments.

Red Hat AI includes several key features for increasing efficiency:

- **Granite family models.** Red Hat AI provides access to Granite family models that are designed for enterprise use and smaller in size. These efficient small language models (SLMs) require fewer compute resources and deliver faster inference compared to LLMs. These models are available under the Apache 2.0 license with full support and indemnification from Red Hat.
- **Catalog of pre-optimized models.** Red Hat AI includes access to a collection of maintained, pre-optimized gen AI models that are ready for inference deployments. These models are efficient, scalable and performant while maintaining accuracy.
- **Third-party model validation program.** Red Hat AI offers a validation framework for third-party models, so you can validate, test, deploy, and manage third-party gen AI models on the platform, with the option to bring your own models from external libraries and open source communities. Inference performance benchmarking and accuracy evaluations give you the flexibility, confidence, and predictability to run these models using virtual large language model (vLLM) servers.
- **LLM compressor.** Red Hat AI lets you create compressed, accurate versions of open source models for faster inference with vLLM. The LLM compressor applies the latest compression best practices like quantization and sparsification.

Red Hat AI lets you develop predictive machine learning (ML) models or start with pretrained gen AI models, so you can experiment with your data for regression, classification, and decision-making applications. By combining predictive and gen AI, you can develop AI-enabled applications and agentic systems that can predict future outcomes, generate optimal responses, and automate workflows.

Proprietary LLMs excel in general-purpose applications, but they are not always the best fit for enterprise artificial intelligence solutions. Their significant computational requirements can limit flexibility and increase costs and operational complexity.

SLMs, particularly those based on open source principles, offer an alternative for organizations looking to develop customized AI solutions, maintain control over data, and manage costs effectively.

Read the checklist to learn more about the benefits of SLMs.

Achieve your goals with Red Hat AI

You can use Red Hat AI for a wide range of use cases:

- Build, migrate, and run ML and predictive AI models.
- Build, deliver, and run gen AI applications.
- Tailor AI solutions with relevant enterprise data.
- Deploy private AI solutions in on-site or air-gapped environments.
- Operationalize and automate model lifecycles via MLOps and DevOps.
- Build multi-architecture AI deployments.

Simplify and speed model customization

Gen AI models are typically trained on generic data, which may not provide the business-specific context that your organization needs for accurate responses and meaningful insights. Red Hat AI lets your developers, data scientists, and domain experts solve unique business challenges by customizing gen AI models with private, enterprise-specific data or deploying ready-to-use solutions.

Through a consistent, simplified AI tooling experience, Red Hat AI supports users with varying levels of AI expertise in refining smaller language models using enterprise-relevant data. The portfolio includes access to the InstructLab model alignment tool, providing an efficient and cost-effective way to fine-tune models. This tool allows data scientists and AI engineers to refine models more effectively. InstructLab embeds domain-specific knowledge directly into models using a taxonomy-driven synthetic data generation process and a multiphase tuning workflow that improves results when proprietary data is unavailable or insufficient.

Red Hat AI also supports retrieval-augmented generation (RAG), allowing you to enhance model outputs and accuracy without retraining by querying external data sources at inference time. The combination of InstructLab and RAG improves alignment with enterprise data while offering the speed, flexibility, and simplicity of RAG. This combined approach is often referred to as the retrieval augmented fine tuning (RAFT) pattern.

For data scientists, the platform offers advanced AI capabilities to build predictive models, fine-tune LLMs, and run distributed workloads efficiently. And you can benefit from self-service access to development environments that include AI tooling for model customization and open source frameworks.

Gain the flexibility to deploy AI solutions anywhere

Red Hat AI provides the flexibility to train, tune, deploy, and run gen AI models and applications wherever it best aligns with business needs. This approach helps you meet data privacy, security, and compliance requirements while optimizing hardware infrastructure costs.

With a focus on enterprise-grade AI workloads, Red Hat AI delivers a trusted, consistent, and comprehensive platform for managing AI in production. The platform simplifies model integration into both new and existing applications, while unifying the management of models, applications, and code into a single location. As a result, you can deploy and manage both predictive and gen AI models across diverse environments—both on-site and in the cloud—with operational consistency, stability, and flexibility.

Red Hat AI prioritizes security, cost optimization, and operational efficiency to support enterprise AI strategies. It offers optimized inference and serving runtimes—like vLLM—to improve the efficiency of LLMs at inference time. A range of deployment options across different hardware accelerators, cloud providers, and OEM server environments provides the flexibility you need to balance cloud spend, data storage, and graphics processing unit (GPU) availability.

Effective AI implementation also requires streamlined model lifecycle management. Red Hat AI simplifies this process with robust machine learning operations (MLOps) and large language model operations (LLMOps) capabilities—including enhanced automation, monitoring, governance, resource allocation, and security. The platform abstracts the complexity of provisioning development environments and managing hardware acceleration for training and tuning, allowing you to focus on AI innovation rather than infrastructure challenges. And for organizations with strict data security requirements, Red Hat AI supports on-site and air-gapped deployments, reducing the risk of exposing sensitive data.

Trust Red Hat AI for every stage of your journey

Red Hat AI delivers capabilities and services that support every stage of AI adoption, from initial single-server deployments to highly scaled, distributed AI platform architectures. For organizations that only need to support AI inference, Red Hat AI Inference Server optimizes model inference across hybrid cloud and edge environments for faster, more flexible, and cost-effective model deployments.

Red Hat Enterprise Linux AI lets you quickly get started with gen AI on physical servers or virtual machines. It provides the 3 essential components for building gen AI solutions: models, tooling, and GPU support.

And for enterprises running increasing numbers of AI workloads in production, Red Hat OpenShift AI offers a scalable solution for deploying predictive and gen AI applications and models across hybrid cloud environments. With integrated MLOps capabilities integrated across distributed Kubernetes platforms, it improves security, reliability, and scalability, ensuring critical AI workloads run efficiently and consistently.

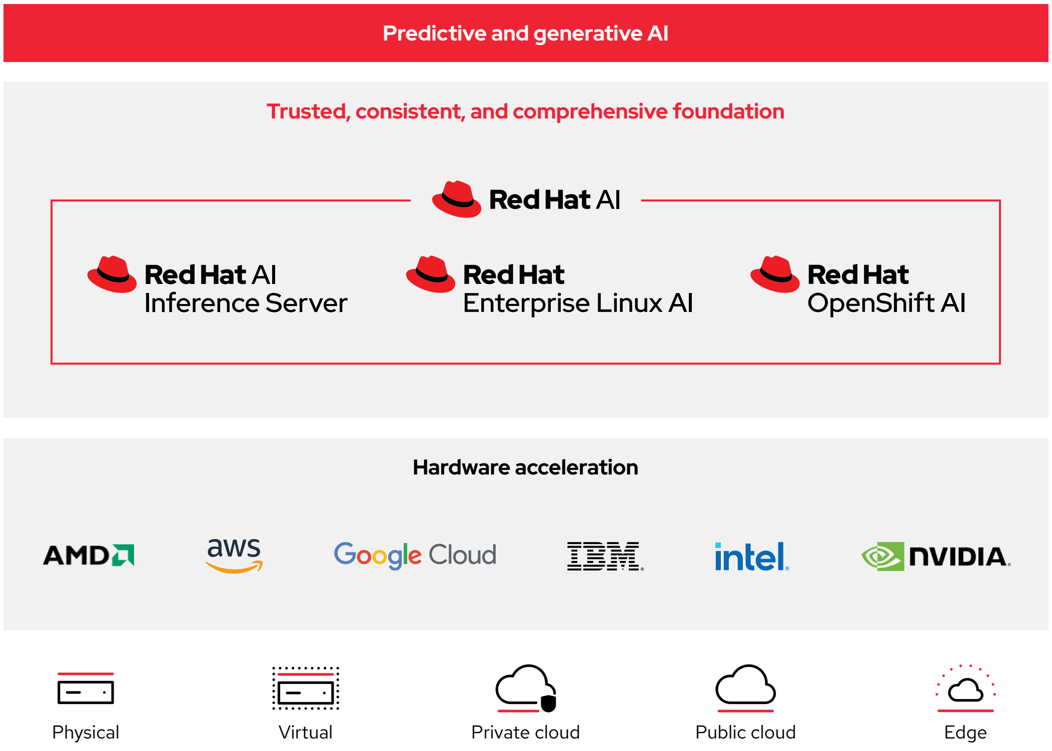
Predictive and generative AI
Trusted, consistent, and comprehensive foundation
Red Hat AI
Red Hat AI Inference Server
Red Hat Enterprise Linux AI
Red Hat OpenShift AI
Hardware acceleration
AMD
aws
Google Cloud
IBM
intel
NVIDIA
Physical
Virtual
Private cloud
Public cloud
Edge

Figure 1. Red Hat AI is a portfolio of products and services that support a range of innovative AI solutions.

Start inferencing with Red Hat AI Inference Server

Red Hat AI Inference Server provides rapid, cost-effective inferencing at scale. With the vLLM inference runtime at its core, the solution gives your organization a unified, high-performance platform to run your choice of models across hardware accelerators, Kubernetes and Linux environments, and IT infrastructure from on-site datacenters and clouds to edge deployments.

Red Hat AI Inference Server includes access to a set of optimization capabilities—including a model repository and LLM compressor—to help models run faster, use fewer resources, and reduce inference costs. The model repository offers a collection of validated and preoptimized models, ensuring rapid deployment with benchmarked performance. The LLM compressor helps reduce the size of models and improve inferencing speeds using advanced quantization techniques, all while maintaining accuracy. Together, these components support accurate and cost-effective inferencing across a wide range of applications.

Develop, test, and run AI models with Red Hat Enterprise Linux AI

Red Hat Enterprise Linux AI is a foundation model platform for developing, testing, and running AI models to power enterprise applications. The included Granite family of open source licensed gen AI models provides a trusted foundation that lets you build innovative gen AI applications in less time. Released under the Apache 2.0 License and fully supported and indemnified by Red Hat, these enterprise-grade language and code models offer cost- and performance-optimized solutions with full transparency into training datasets. You can also deploy your own AI models on the platform while ensuring flexibility, security, and compliance.

Included with Red Hat Enterprise Linux AI, InstructLab model alignment tools simplify fine-tuning processes, making AI more accessible across your organization. Built on the large-scale alignment for chatbots (LAB) methodology, InstructLab follows a community-driven approach to model development. It lets you align models with organizational data and broaden access to community-developed models so that you can tailor AI solutions to your specific needs. InstructLab uses a taxonomy-driven synthetic data generation process and a multiphase tuning workflow to embed domain-specific knowledge directly into models. As a result, you can efficiently integrate business-specific skills and knowledge into generative AI models, improving their relevance and effectiveness for AI-based applications.

Built using Red Hat Enterprise Linux image mode, Red Hat Enterprise Linux AI comes as a bootable image with popular AI libraries—including PyTorch—and hardware-optimized accelerators from NVIDIA, Intel, and AMD. This solution simplifies integration of crucial AI technologies while streamlining model training and inference on production servers so that you can begin your AI projects efficiently. Optimized for individual server deployments, Red Hat Enterprise Linux AI bootable images provide a consistent, high-performance platform for innovative AI development across hybrid cloud environments.

Production technical support and model intellectual property (IP) indemnification help you mitigate risks while focusing on building, deploying, and managing innovative AI solutions with confidence, transparency, and cost efficiency. Red Hat offers comprehensive 24x7 enterprise production support, a trusted product distribution, extended model lifecycle management, and full model IP indemnification.

Red Hat Enterprise Linux AI empowers your teams to rapidly transition from proof of concept to production deployments with a comprehensive suite of tools to train, tune, and deploy AI models wherever your data resides. And when you're ready to scale your innovative AI solutions, Red Hat OpenShift AI lets you train, tune, and serve models across a distributed cluster environment using the same Granite models and InstructLab approach.

Deliver innovation at scale with Red Hat OpenShift AI

Red Hat OpenShift AI is an integrated AI platform for managing the lifecycles of predictive and gen AI models and delivering AI-based applications at scale across hybrid cloud environments. It unites data scientists and developers under IT oversight to develop, train, fine-tune, and manage models, speeding the journey from experimental AI applications to production-ready solutions. As a self-managed offering or fully managed cloud service, Red Hat OpenShift AI builds on the proven capabilities of Red Hat OpenShift to provide a trusted, consistent, and scalable environment for building, deploying, and monitoring AI/ML applications and models across on-site, public cloud, and edge environments. In partnership with our technology ecosystem, Red Hat OpenShift AI speeds innovation, improves operational consistency, and offers hybrid cloud flexibility—promoting transparency, freedom of choice, and responsible AI implementation.

> *"Combining the power of Red Hat OpenShift with Red Hat OpenShift AI gives AGESIC users the flexibility to explore opportunities for AI innovation across Uruguay while maintaining AGESIC's standards and processes for AI usage. Red Hat's work with AGESIC exemplifies our dedication to improving the user experience for our customers and their customers."*[1]
>
> **Steven Huels**
> Vice President and
> General Manager, AI
> Business Unit, Red Hat

Red Hat OpenShift AI empowers data scientists to create AI models, integrate their preferred foundation models, and customize models with enterprise relevant data. The platform includes a comprehensive suite of tools and environments—from Jupyter Notebooks and PyTorch to data science pipelines and enhanced monitoring and observability tools—to support AI development. It also provides access to Granite models and optimized open source models designed for enterprise use and ready for deployment. Simplified, self-service access to core AI/ML libraries, widely used frameworks, popular integrated development environments (IDEs), and a broad selection of predefined and customer-provided images and workbenches boost productivity throughout the model development process. And the InstructLab alignment tool allows users to customize models with enterprise-relevant data to deliver domain-specific solutions.

Red Hat OpenShift AI streamlines the creation and execution of continuous integration and continuous deployment (CI/CD) pipelines that simplify management of AI model and application lifecycles. It automates data science workflows during model development, while allowing application developers and IT operations teams to deploy models rapidly and efficiently using standard DevOps techniques. The platform features a visual editor that simplifies design and automation of data science pipelines and experiments, supporting efficient data exploration, model training, validation, and storage.

Enhanced model serving capabilities in Red Hat OpenShift AI simplify and streamline AI model delivery for inference. Support for model servers and runtimes like KServe, vLLM, and Text Generation Inference Server (TGIS) allow for flexible deployment across production environments. For example, vLLM delivers high throughput performance for gen AI inferencing and scales from a single GPU to multi-GPU distributed systems.

Red Hat OpenShift AI gives you flexibility in AI development and deployment through cross-platform support for hybrid and multicloud environments. It supports a wide range of accelerators—including those from NVIDIA, AMD, Intel, Google Cloud, and Amazon Web Services (AWS)—and runs across major original equipment manufacturer (OEM) servers from partners like Dell Technologies, Lenovo, HPE, and Cisco. With this flexibility, you can adapt your AI strategy over time, moving operations to cloud or edge environments as needs change. You can also train and deploy models and AI-enabled applications in the appropriate environment—including air-gapped and disconnected environments—to meet relevant regulatory, security, and data requirements.

[1] Red Hat press release. "Red Hat Helps AGESIC Scale AI Innovation Across Uruguay," 7 May 2024.

"Adding Red Hat OpenShift AI to our arsenal has given our data scientists greater autonomy and better standards."[2]

Ömer Uyar
CEO, Intertech

Distributed workload capabilities in Red Hat OpenShift AI increase the efficiency of data processing and model training, tuning, and serving across multiple cluster nodes. The platform optimizes job execution by prioritizing workloads, automating resource scaling—including hardware accelerators—and maximizing node utilization. These capabilities are essential for handling foundation models with varying data volumes, training durations, model sizes, and compute requirements.

Finally, Red Hat OpenShift AI offers monitoring and observability capabilities and tools to help data scientists determine whether models are fair and unbiased both based on training data and during deployment. These built-in features provide real-time insight into key performance and operational metrics, allowing both data scientists and IT operations teams to proactively identify and address potential issues.

Explore a comprehensive partner ecosystem

Red Hat's partner ecosystem offers tools, services, and solutions designed for a wide range of AI use cases. Partners—including independent software vendors (ISVs), global systems integrators (GSIs), and cloud service providers—collaborate with Red Hat to integrate and certify advanced AI/ML technologies with Red Hat AI. These extensive relationships let you explore, select, and implement the most suitable technologies for your innovative AI solutions, ensuring flexibility and efficiency throughout your AI model and application lifecycles.

Technology partners

With technology solutions from partners like AMD, IBM, Intel, NVIDIA, and Starburst, you can build, deploy, and manage AI-based applications with the security and support of the Red Hat ecosystem. These partnerships allow you to operationalize and scale AI with foundation models, gen AI capabilities, and traditional ML solutions. Streamline integration and production deployment, ensure performance at scale, and accelerate the delivery of AI-powered applications across datacenters, edge environments, and public clouds. Gain fast, efficient access to the data you need to support AI innovation.

Hardware partners

AI workloads demand significant compute power, and hardware partners like Dell and Lenovo provide the performance needed to run AI applications effectively. Red Hat collaborates with these partners to test and certify their hardware with Red Hat solutions, ensuring your AI-ready enterprise platform runs on a trusted, reliable foundation. These partnerships help you rapidly develop and deploy innovative AI applications at scale with integrated hardware and software solutions that support high-performance AI/ML workloads across environments.

Cloud partners

Cloud partners—including IBM, Amazon Web Services (AWS), Microsoft Azure, and Google Cloud—offer hybrid and multicloud solutions that streamline critical AI deployments. These partners provide tools that let you innovate faster, enhance customer experiences, and scale infrastructure to meet evolving demands. By taking advantage of open source innovation, they help modernize IT environments, strengthen security, and speed application development across hybrid and multicloud architectures. Together with our cloud partners, we can help you navigate cloud complexity, adopt cloud-native development, and stay competitive in fast-moving markets.

[2] Red Hat case study "DenizBank transforms AI operations and empowers innovation," Jan. 2025.

See success in action

Many customers are already experiencing the benefits of deploying Red Hat AI solutions.

AGESIC

Uruguay's Agency for Electronic Government and Information and Knowledge Society (AGESIC), leader of the nation's e-government strategy, recognized that integrating AI into government services was essential to meet the evolving needs of its citizens. To modernize its operations, AGESIC adopted Red Hat OpenShift AI to extend, scale, and standardize AI across government agencies while automating the development and lifecycle management of AI models. This solution empowers AGESIC to build, train, tune, and deploy models efficiently, fostering closer collaboration between data scientists, developers, and IT operations.

> *"AGESIC is using OpenShift and OpenShift AI to combine best practices in architecture and software development with governance processes. With the help of these technologies and associated methodologies, we have been able to advance rapidly and securely towards fulfilling that objective."*
>
> **Gabriel Hernandez**
> Director of IT and Operations, AGESIC

Key outcomes:

- Reduced ticket resolution times with an automated platform that cuts manual service ticket processing to less than 1% of previous effort
- Enhanced collaboration between data scientists, developers, and IT operations to speed development and integration of AI models
- Automated processes, streamlined workload management, and boosted overall efficiency with a containerized, centralized platform that unifies development and operations

Read the press release to learn more about AGESIC's experience.

DenizBank

Data scientists working at DenizBank, a prominent private bank in Türkiye and the 5th largest in the country, wanted to convert its existing workflow into a less manual process with a more standardized approach. The bank's IT subsidiary, Intertech, began a project to provide a model development environment with automated pipelines and standards to improve productivity and time to market. As a key improvement, Intertech adopted Red Hat OpenShift AI for its self-service capabilities and capacity to scale model serving and improve operational efficiency. Data scientists can now focus on building models that are more robust and secure than ever.

> *"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

Key outcomes:

- Provided more than 120 data scientists from different lines of business greater autonomy and more consistent standards
- Accelerated time to market while ensuring more robust and secure models with automated environment builds and self-service capabilities
- Optimized GPU usage with slicing to maximize resource utilization, increase flexibility, and allow more workloads to run simultaneously without the need for additional GPU hardware

Read the case study to learn more about DenizBank's experience.

Clalit Health Services

Clalit Health Services provides healthcare services for half of Israel's population across 14 general, mental health, geriatric, and children's hospitals. It also operates community clinics, dental clinics, imaging facilities, and a lifestyle program. Clalit recently established an advanced AI platform based on Red Hat OpenShift AI with support for hardware acceleration, model development, and production workloads. With this platform, Clalit is processing historical medical data and training a LLM to identify patients at risk for preventive care and medication. The solution then provides recommendations on courses of action for patient treatment through a chatbot-like experience.

Key outcomes:

- Automated provisioning of self-service access to development environments with data mining and data science tools
- Simplified management and allocation of GPU resources for data experimentation and model development
- Adheres to key cost and security requirements

Read the case study to learn more about Clalit's experience.

"Red Hat OpenShift AI ensures [our researchers and scientists] have the computing power they need to search text, search images, train models, and, in the future, process genomic data. We have found OpenShift AI to be beneficial in different ways, such as self-service to our data scientists and agility in the deployment of containers."

Eyal Dviri
Innovation Team Leader
in the Data Department,
Clalit Health Services

Red Hat

Red Hat wanted to increase the efficiency and scalability of customer and technical support services for our growing customer base with AI solutions. The Experience Engineering team at Red Hat started on a program—using Red Hat OpenShift AI and Red Hat Enterprise Linux AI—to develop, test, and deploy 4 solutions powered by AI, all with the aim of simplifying IT support for our customers and support associates. These tools improve self-service, increase efficiency, and help bring about a faster response to support cases.

Key outcomes:

- Delivered more than $5 million in cost avoidance, with estimated $1.5 million in just 10 months
- Increased availability of knowledge content and minimized repetitive tasks for IT support associates who handle 30,000 new cases each month
- Provided faster responses to customers with AI, enhancing overall user experiences

Read the case study to learn more about our IT support solution.

Learn more

No matter where you are in your AI journey, Red Hat AI can help you move forward.

Learn more about our solutions and how to get started.

> *"AI not only makes self-service more accessible but also ensures faster, more accurate responses when customers need support from Red Hat. With support managed consistently, our customers can stay focused on serving their own clients."*
>
> **Manikandan Sivanesan**
> AI Technical Strategy Lead, Experience Engineering, Red Hat

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	00800 7334 2835	+65 6490 4200	+54 11 4329 7300
www.redhat.com	europe@redhat.com	apac@redhat.com	info-latam@redhat.com

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