Red Hat

Gen AI in action

Considerations and practical applications with Red Hat AI

Table of contents

Worldwide spending on AI solutions

According to the IDC Market Forecast, worldwide investment in gen AI solutions is expected to be US$608.7 billion in 2029 at a compound annual growth rate (CAGR) of 56% for the 2024-2029 period.[2]

Introduction

Gen AI: Transforming markets and industries

AI continues to be a major area of innovation and investment for enterprise organizations worldwide. In fact, IDC expects worldwide spending on AI solutions to grow to US$1,262 billion at a compound annual growth rate (CAGR) of 31.9% for the 2024-2029 period.[1]

Gen AI in particular is a key motivator of this growth, with an expected worldwide spending CAGR of 56% for the same time period.1 Gen AI is a powerful tool for organizations that want to create innovative products, optimize processes, and gain competitive advantages in rapidly changing markets. Based on advancements in deep learning and neural networks, it goes beyond predictive AI capabilities by not only processing data, but generating new, original content. Gen AI can respond to prompts with text, images, code, sounds, or other media derived from its training data, which makes innovative solutions for content creation and personalization possible. As a result, gen AI is reshaping human-machine collaboration, inspiring new approaches to problem solving, and delivering significant business gains across industries.

Gen AI applications can deliver a variety of benefits for enterprise organizations:

- Improving employee productivity
- Increasing customer satisfaction
- Reducing operational costs

This e-book reviews key strategies and considerations for choosing AI solutions. You'll learn about the benefits of selecting a solution that balances ready-to-use and custom development approaches, and see some common use cases for getting started with gen AI in your enterprise. Discover how you can build a foundation for gen AI innovation.

Chapter 1

Choose the right AI strategy for your organization

Gen AI is a digital transformation project, and for it to succeed, it needs an implementation strategy.

There are 2 paths your enterprise can take for its AI strategy: you can adopt a cloud-based AI service or build and host an AI platform yourself. These options require different levels of technical involvement and operational effort, and offer different levels of customization and control.

Cloud-based AI services

Cloud-based AI services are provided by a third-party vendor as a paid and managed solution, and offer access to frontier models via application programming interfaces (APIs). These services allow your organization to integrate AI models into your applications without hosting the model yourself. Some private commercial offerings also let you fine-tune the models they provide, or deploy models in a dedicated or more controlled environment.

1 IDC Market Forecast. "Worldwide Artificial Intelligence IT Spending Forecast, 2025–2029." Doc #US53688725, August 2025.

Because this approach gives you ready-to-use AI solutions with minimal interaction with the model itself, it can be more straightforward and cost-effective for organizations that do not want to deal with the complexities of AI infrastructure management, have smaller operations teams, or are adopting AI at a lesser scale.

Self-hosted AI platforms

Building and hosting an AI platform yourself provides more choice and control over your models and environment. You can select the hardware, software, models, applications, and deployment location that best fit your organization's requirements. For example, you can choose to host your models and applications in public clouds, private clouds, on-site datacenters, or edge locations. This approach also gives you more opportunities to customize your models and applications, more control over your data, and less dependence on third-party providers. However, it typically involves higher up-front investments and ongoing operational effort and maintenance costs than a cloud-based AI service.

Model-as-a-Service

If you want to build a self-hosted platform but offer your users an experience similar to what they would get from a cloud-based AI service, you can implement the solution pattern for Model-as-a-Service (MaaS) as an internal cloud utility. By purchasing a dedicated AI platform, you can set up your IT team to act as an internal provider, offering developers an easy-to-use experience similar to public cloud offerings. This centralized approach allows developers to access curated models on demand without managing graphic processing units (GPUs) and other complex infrastructure.

These platforms go beyond just serving models. They support the entire AI lifecycle, from initial development and fine-tuning to the deployment and long-term management of predictive, generative, and agentic AI. This strategy combines the use of a cloud-like consumption model with the deep control, security focus, and governance required to scale diverse AI initiatives across the entire enterprise.

To build and host an AI platform, you need:

- Access to foundation models for your use case. Examples include large language models (LLMs), code models, small language models (SLMs), open source models, and multimodal models.
- Access to hardware acceleration capabilities like GPUs.
- Access to an application platform with advanced AI tools and serving mechanisms.
- A governance solution for compliance and responsible AI use.

Whether you use a MaaS strategy or not, building a self-hosted AI platform gives you more control over your AI solutions. That makes it an obvious choice for organizations that operate in highly regulated industries, plan to use sensitive data and intellectual property (IP) within their AI solutions, or have larger operations teams that can handle the complexities of building, running, and maintaining AI infrastructure.

Maximize the value of your AI investments

1. **Align your AI initiatives with business goals.** Align your chosen solution directly with your strategic objectives, and continuously measure return on investment (ROI) to track outcomes.
2. **Optimize your total cost of ownership (TCO).** Consider maintenance, infrastructure, and talent expenses in addition to upfront costs.
3. **Prioritize adoption and usability.** Choose a solution that balances adoption speed with features that can make your teams more productive.
4. **Take advantage of centralized AI services.** Avoid duplicated efforts and optimize GPU use by designing and delivering a scalable MaaS that all teams can use.

Table 1. Comparing approaches to AI strategies

	Cloud-based AI services	Self-hosted AI platforms
Deployment	+ Faster deployment with ready-to-use solutions	- Slower deployment with more planning required
Costs	+ Lower up-front costs - Potential hidden costs, especially at scale and with customization	- Higher up-front costs + No hidden costs
Data privacy and security	- Higher data privacy, security, and IP risks with less control	+ Increased data privacy and security when deployed on site
Solution customization	- Limited ability to customize - Vendor lock-in and dependency	+ Complete ability to customize + Low vendor dependency
Skills requirements	+ Minimal skills needed, as hardware, models, and support are included	- AI infrastructure architecture and operational skills required
Best for	Organizations that do not want to manage AI infrastructure in-house	Organizations that want more control and customization in their AI solutions

Considerations for evaluating AI solutions

When you're choosing an IT strategy for your organization, you need to keep transparency, efficiency, and relevance in mind. The way you think about these considerations will shift depending on your chosen path; for instance, a self-hosted strategy makes it possible for your internal teams to provide deep transparency, while with a cloud-based approach you will need to rigorously vet your vendors.

Ensure transparency in AI solutions

AI solutions should provide transparency, accountability, and explainability while ensuring data privacy, security, and regulatory compliance. That is critical for building trust, mitigating risks, and remaining competitive. Look for vendors that clearly disclose model architectures, training data, and performance metrics, and offer accountability mechanisms and explanations for AI-generated outputs.

Optimize infrastructure and cost efficiency

Scalable, low-cost infrastructure solutions that support model optimization, distributed training, and efficient hardware configurations can help you minimize operational expenses, enhance performance, and adapt quickly to changing demands.

While self-hosted platforms involve higher up-front infrastructure costs, you can mitigate these by using model optimization techniques like quantization and distillation. These allow you to reduce hardware dependency, lower infrastructure costs, and decrease your overall environmental impact.

Unlock the power of agentic AI with Red Hat technologies

While gen AI started the conversation, agentic AI is where the work gets done. By connecting LLMs to external tools, these systems move beyond models that talk to agents that act—autonomously orchestrating complex tasks to achieve defined goals.

In the enterprise, this means completing tedious workflows in a fraction of the time. Simplify and accelerate your journey to successful adoption with the trusted foundation of Red Hat AI.

Read the e-book[2] to learn more about agentic AI.

See how Red Hat AI can help your organization

Red Hat offers a broad selection of learning materials and tools to help you get started with AI.

Explore our AI learning paths, designed for business leaders and technology learners. Our step-by-step courses cover everything from AI basics to hands-on tool overviews.

Complete a path to earn a certificate and boost your AI skills.

Start learning

Explore industry-specific applications

Gen AI solutions can be applied to a wide variety of uses. Find solutions that include large libraries of industry-specific AI use cases and offer prebuilt templates for applications like recommendation engines and client support to speed time to market. You can provide enhanced context with tools that let you tune models with business-specific data, resulting in more accurate and relevant responses.

This is where a MaaS strategy can shine: by providing your developers with prebuilt templates and tailored models, you can achieve the accelerated time to market of a cloud service while maintaining the deep control of a self-hosted environment.

Chapter 2

Speed time to value with Red Hat AI

Red Hat® AI is a platform that accelerates the development and deployment of AI solutions across hybrid cloud environments. As we discussed in Chapter 1, the self-hosted AI path offers the most control, but the operational effort required to build it from scratch is a barrier for many organizations. Red Hat AI closes that gap, providing a preintegrated and comprehensive platform that can be deployed in a self-hosted environment.

Red Hat AI meets customers where they are—whether starting out or scaling to a full enterprise architecture—while supporting the deployment of any model on any hardware accelerator. Plus, you can deploy your AI applications and services across diverse environments, including on-site infrastructure, public and private cloud resources, and edge locations.

Red Hat AI empowers your teams to build predictive and gen AI models using your confidential enterprise data. The portfolio includes essential tools, GPU support, and self-service on-demand environments, increasing agility and abstracting away complexities for your developers.

With access to a catalog of optimized and validated third-party open models, you can efficiently customize solutions to meet specific use cases. The platform simplifies integration of applications and AI models by centralizing model, application, and code management. Designed for enterprise-grade production workflows, Red Hat AI prioritizes security, cost optimization, and operational efficiency. It offers reliable day-to-day support through governance, monitoring, security, machine learning operations (MLOps), and large language model operations (LLMOps) services. And customers seeking to reduce the risk of exposing sensitive data can take advantage of Red Hat AI's support for air-gapped deployments on site or in private cloud instances.

Red Hat AI includes Red Hat AI Enterprise, for enterprises looking to deploy and scale efficiently and anywhere, Red Hat AI Inference, for optimized inference of large language models (LLMs), Red Hat OpenShift® AI, for distributed Kubernetes platform environments, and Red Hat Enterprise Linux® AI, for individual Linux server environments. These solutions combine the power of open source technologies with state of the art open source models, helping organizations accelerate the pace of discovery and democratize access to cutting-edge tools and technologies. Access to the latest innovations is complemented by Red Hat's AI partner ecosystem, which offers an array of partner products and services that are tested, supported, and certified to perform with our technologies and help customers solve their business and technical challenges.

2 Red Hat e-book. "Unlock the power of agentic AI with Red Hat technologies." 8 Oct. 2025.

Red Hat AI benefits

Flexible and efficient inferencing

Red Hat AI optimizes model inference across hybrid cloud environments, allowing organizations to deploy their preferred models faster and more cost-effectively. Its runtime, vLLM, maximizes throughput and minimizes latency, while LLM Compressor uses advanced quantization techniques to improve inference speeds without sacrificing prediction accuracy. According to Forrester research, "Together, these components create a high-performance environment for serving models across a wide range of applications, and have helped interviewed enterprises double their GPU capacity."[3]

Connecting models to data

The platform simplifies the integration of private enterprise data, providing standardized tools for preparing and ingesting information to ground AI outputs in specific business knowledge. Organizations can boost model performance and relevance through multiple customization techniques, including fine-tuning, retrieval-augmented generation (RAG), and prompt tuning. "This foundation ensures that models are aligned with enterprise scenarios and deliver tangible business outcomes. Interviewed enterprises have increased productivity with a 60% reduction in model training time," according to Forrester research.[3]

Agentic AI innovation

Moving beyond simple chat interfaces, Red Hat AI accelerates the deployment and management of agentic AI on a scalable and flexible platform. It provides the infrastructure needed to manage and orchestrate autonomous agents that can reason through complex goals, such as onboarding a new client, by breaking them into executable tasks like updating databases or sending emails. This transition from chat to act is supported by managed workflows that maintain a human in the loop for high-stakes approvals and safety.

According to Forrester, "Interviewed enterprises have been able to accelerate 400 AI projects by year 3 via 75% faster MLOps provisioning."[3]

AI at scale across hybrid cloud environments

Red Hat AI allows organizations to manage and monitor the entire lifecycle of both predictive and generative AI models at scale, from single-server deployments to highly scaled-out distributed platforms. The platform ensures operational consistency and risk mitigation by providing a tested, supported AI platform powered by Red Hat OpenShift. Through centralized monitoring, teams can track bias, data drift, and anomaly detection to ensure models remain reliable, fair, and compliant with corporate governance standards. "Interviewed enterprises have seen up to 2% annual profit growth by year 3 from faster AI productization."[3]

3 Forrester Consulting. "The Total Economic Impact™ Of Red Hat AI," a commissioned study conducted by Forrester Consulting on behalf of Red Hat, February 2026. Results are for a composite organization based on interviewed customers.

Success in action with Clalit Health Services

"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."

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

Chapter 3

Get started with practical applications and use cases

You can use Red Hat AI to build applications that support a range of AI use cases and address many business challenges. Thanks to Red Hat AI's consistent user experience, stakeholders from AI developers to data scientists to IT operations teams can more simply develop and deploy AI solutions across hybrid cloud environments.

Common business use cases addressed by Red Hat AI include:

- **Natural language processing**
- **Content creation**
- **Knowledge bases**
- **Digital assistants**
- **Media creation**
- **Service personalization**
- **Recommendation engines**
- **Data analytics**
- **Cybersecurity**
- **Chatbots**
- **Task and workflow automation**
- **Sentiment analysis**
- **Computer vision**
- **Software development**

AI- and data-driven business operations

AI models can handle the massive amount and breadth of data that organizations collect to help them make more informed business decisions. Teams gain keener insights into the data that underlies the business, which can help them maximize revenue, optimize operations, and enhance customer experiences and employee productivity.

Recommendation engines

AI recommendation engines assess current situations against historical data to identify common factors and provide guidance. They can be used in many industries to deliver real-time suggestions for action.

Clalit Health Services recently established an advanced AI platform based on Red Hat AI to process historical medical data and train a LLM to identify at-risk patients who would benefit from preventive care and medication. The solution then provides recommendations on courses of action for patient treatment through a chatbot-like experience. Clalit is using the same platform to build learning processes and algorithms to identify new trends, patient and disease behavior patterns, and more.

Read the success story[4]

4 Red Hat case study. "Clalit is gaining insights from massive health data pools." 28 May 2024.

Success in action at Denizbank

"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

Success in action with AGESIC

"AGESIC is using OpenShift and OpenShift AI to combine best practices in architecture and software development with governance processes."

Gabriel Hernandez
Director of IT and Operations, AGESIC

Automated, self-service AI workflows

AI model and application development can be complicated. Automated AI pipelines and self-service operations can streamline this process while improving security and compliance.

Data scientists working at DenizBank wanted to convert its existing workflow into a less manual process with a more standardized approach. The bank's IT subsidiary, Intertech, provided a model development environment with automated pipelines and standards. Intertech helped DenizBank improve productivity and speed time to market for customer loan identification and fraud detection applications. As a key improvement, Intertech adopted Red Hat AI for its work with DenizBank, citing its self-service capabilities and its capacity to scale model serving and improve operational efficiency. The bank's more than 100 data scientists can now focus on building models that are more robust and secure than ever.

Read the success story[5]

Automated service ticket routing

Organizations in the public and private sectors use ticketing systems to serve citizens, customers, and employees. AI-based assessment can help them rapidly route incoming tickets to the right teams. And some tickets can even be automatically handled to accelerate resolution and user satisfaction.

Uruguay's Agency for Electronic Government and Information and Knowledge Society (AGESIC) adopted Red Hat AI to extend, scale, and standardize AI across government agencies. This solution empowers AGESIC to build, train, tune, and deploy models efficiently, fostering closer collaboration between data scientists, developers, and IT operations. For example, AGESIC built and deployed a series of models to automatically classify and route 2,000 citizen claims per month to the right team, reducing routing time from 1 hour to only seconds.

Read the press release[6]

Customer support and content creation

High-quality customer support is critical for delivering high-value user experiences. AI can help teams improve troubleshooting, summarize information and tickets, and create tailored content based on existing documentation.

Red Hat uses Red Hat AI internally[7] to increase the efficiency and scalability of customer and technical support services for its customer base. The Experience Engineering team at Red Hat developed, tested, and deployed 4 solutions powered by AI, all with the goal of simplifying IT support for customers and support associates. These tools improve the self-service experience, increase efficiency, and help accelerate responses to support cases. For example, Red Hat increased the availability of knowledge content and minimized repetitive tasks for IT support associates who handle 30,000 new cases each month. And Red Hat's AI-powered initiatives have shown the cost-saving potential of these solutions: they "saved an estimated US$1.5 million in support costs in only 10 months, with a projected savings of more than US$5 million overall."

Read the success story

5 Red Hat case study. "DenizBank empowers its data scientists." 16 Jan. 2025.
6 Red Hat press release. "Red Hat Helps AGESIC Scale AI Innovation Across Uruguay." 7 May 2024.
7 Red Hat case study. "Red Hat saves $5 million in IT support costs with AI augmentation." 17 Dec. 2024.

Success in action with Red Hat

By deploying 4 AI-powered support solutions, Red Hat saved an estimated US$1.5 million in just 10 months.

"AI augmentation doesn't just improve efficiency; it also enhances content creation and may contribute to job satisfaction."

Mandy Elliott
Senior Director AI and Data Experience Engineering, Red Hat

Success in action with Turkish Airlines

"Our AI projects are targeted to create over US$100 million in financial impact by boosting revenue, decreasing operational costs, and increasing efficiency."

Emre Yavuz
Head of Data and AI, Turkish Technology—IT subsidiary of Turkish Airlines

Virtual assistants and chatbots

AI-based chatbots and assistants continue to improve in response quality and accuracy. They often serve as the interaction point for advanced AI solutions and can be applied across industries in a multitude of use cases from customer service to information delivery to content creation.

The City of Vienna wanted to improve employees' productivity and satisfaction. The city developed a virtual assistant that supports employees in their daily work by providing instant answers to work-related questions, helping them respond more accurately to citizen inquiries and requests. With OpenShift AI, the city can innovate at their pace, provide new services and functionality to the public, and maintain frequent release cycles.

Read the success story[8]

From predictive to generative AI

Turkish Airlines aimed to gain competitive advantages over its rivals by starting a data-driven transformation program. The company wanted its data scientists to develop the latest AI models to take advantage of data, and hoped its entire business would become AI-driven. The airline's IT subsidiary developed a scalable and cloud-ready infrastructure with Red Hat AI. The operations team can now create workspaces in minutes rather than days, and data scientists can deploy AI models themselves, halving deployment times. The airline can address use cases like asset management, aircraft health monitoring, offer generation, seat pricing, traveler behavior, ground time prediction, and more.

Read the success story[9]

8 Red Hat success story. "City of Vienna increases efficiency of public services with innovative AI solutions." accessed 23 April 2026.
9 Red Hat case study. "Turkish Airlines innovates with Red Hat OpenShift AI." 3 June 2025.

Conclusion

Start your gen AI journey with Red Hat today

Build AI solutions for your world

Your organization's goal is to deliver AI value with the resources you have, the insights you own, and the freedom you need. Red Hat AI accelerates time to market and reduces the operational cost of delivering AI solutions across your hybrid cloud environment. Efficiently tune small, fit-for-purpose models with your enterprise data and gain the flexibility to deploy wherever the data resides. Manage and monitor the lifecycles of AI models at scale and focus on innovating with AI to reach your team's business goals.

Get started by learning more about Red Hat AI

Explore product information, key features and benefits, and access no-cost trials for Red Hat AI.

Learn more

Discover AI resources and knowledge for the entire organization

Build your AI skills with hands-on resources for practitioners, or strengthen your decision making with AI knowledge resources. Red Hat provides demos, guides, and case studies to help you get started.

Build your skills

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

facebook.com/redhat
x.com/RedHat
linkedin.com/company/red-hat

redhat.com
#3644407_0526