Red Hat AI

Generative AI in action

Considerations and practical applications with Red Hat AI

Contents

Generative AI:

Transforming markets and industries

Artificial intelligence (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$632 billion at a compound annual growth rate (CAGR) of 29.0% for the 2023-2028 period.[1]

Generative AI (gen AI) is a key motivator of this growth, with an expected worldwide spending CAGR of 59.2% 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 creates this new content or data based on patterns learned from existing information. It can generate text, images, code, sounds, or other media that resemble its training data, providing innovative solutions for content creation and personalization. As a result, gen AI is reshaping human-machine collaboration, inspiring new approaches to problem solving, and delivering significant business gains across industries.

According to IDC, worldwide investment in gen AI solutions is expected to exceed

US$202 billion

in 2028 at a CAGR of 59.2% for the 2023-2028 period.[1]

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

- Improve employee productivity.
- Increase customer satisfaction.
- Reduce operational costs.

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

1 IDC. "IDC FutureScape: Worldwide Artificial Intelligence and Automation 2025 Predictions." 28 October 2024. Doc #US51666724.

Choose the right AI strategy for your organization

Like any large IT or business initiative, creating a strategy for how your organization will implement AI is crucial for success.

There are 2 paths enterprises can take for their AI strategy: 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. These services offer access to frontier models via application programming interfaces (APIs). This allows your organization to integrate AI models into your applications without hosting the model yourself. Some private commercial offerings also let you fine-tune the provided models or deploy models in a dedicated or more controlled environment.

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. Even so, it typically involves higher up-front investments and ongoing operational effort and maintenance costs than a cloud-based AI service.

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 graphic processing units (GPUs).
- Access to an application platform with advanced AI tools and serving mechanisms.
- A governance solution for compliance and responsible AI use.

This approach gives you more control over your AI solutions, so it can be 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.

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 don't want to manage AI infrastructure in-house	Organizations that want more control and customization in their AI solutions

Considerations for evaluating AI solutions

Be sure to consider transparency, efficiency, and relevance when evaluating AI solutions and strategies.

Ensure transparency in AI solutions

AI solutions that provide transparency, accountability, and explainability while ensuring data privacy, security, and regulatory compliance are 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. Apply techniques like quantization and distillation to reduce hardware dependency, lower infrastructure costs, and decrease your overall environmental impact.

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. Tools that let you tune models with business-specific data provide enhanced context, resulting in more accurate and relevant responses.

Maximize the value of your AI investments

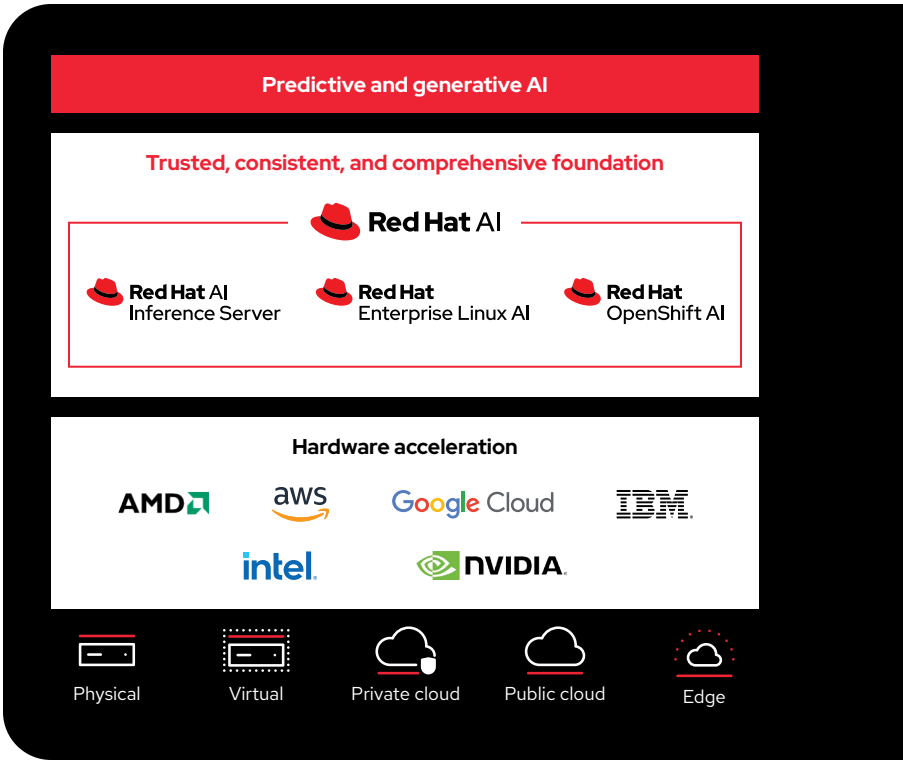
1. **Align your AI initiatives with business goals.** Ensure your chosen solution directly supports your strategic objectives like differentiation or efficiency.
2. **Optimize your total cost of ownership (TCO).** Consider maintenance, infrastructure, and talent expenses in addition to the upfront costs of your solution.
3. **Prioritize adoption and usability.** Choose a solution that balances adoption speed with features that your teams can actually use to be productive.
4. **Take advantage of centralized AI services.** Avoid duplicated efforts and optimize GPU use by designing and delivering scalable Models-as-a-Service (MaaS) that all teams can use.
5. **Continuously measure and adapt.** Track your return on investment (ROI) by measuring factors like cost savings, efficiency gains, and revenue gains, and adapt your approach accordingly.

Speed time to value with Red Hat AI

Red Hat® AI is a portfolio of products and services that accelerates the development and deployment of AI solutions across hybrid cloud environments. With a focus on simplifying AI adoption, the portfolio makes advanced AI technologies more accessible across your entire organization.

By balancing flexibility with consistency, Red Hat AI helps your teams deploy and manage both predictive and gen AI models wherever it makes the most sense for your workload and overall strategy. The portfolio supports all stages of your AI adoption journey—from single-server deployments to highly distributed, scalable platform architectures—so you can start small and expand according to your needs and plans. Support for a variety of hardware accelerators, original equipment manufacturers (OEMs), and cloud providers ensures a stable, optimized, and high-performance environment for your AI workloads. Plus, you can deploy your AI applications and services across diverse environments, including on-site infrastructure, public and private cloud resources, and edge locations.

The Red Hat AI portfolio includes Red Hat Enterprise Linux® AI for individual Linux server environments, Red Hat OpenShift® AI for distributed Kubernetes platforms, and Red Hat AI Inference Server for optimized inference of LLMs. These solutions deliver open source technologies and purpose-built SLMs, providing access to the latest AI tools while helping to address the high costs often associated with gen AI. In fact, the included Granite family models—distributed under the Apache 2.0 license with transparency into training data sets—helps you get started with gen AI in less time using smaller, efficient models that reduce operational costs without compromising performance. 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. Finally, the **Red Hat AI partner ecosystem** helps you speed innovation with a range of tested, supported, and certified products and services that address both business and technical challenges.

Streamline AI model delivery

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 reducing IT dependencies. With access to a catalog of pre-optimized and open source Granite family models, you can efficiently customize solutions to meet specific use cases. The platforms simplify 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, offering reliable day-to-day support through governance, monitoring, security, machine learning operations (MLOps), and large language model operations (LLMOps) services. And support for air-gapped deployments on-site or in private cloud instances reduces the risk of exposing sensitive data.

Reduce complexity and cost with SLMs

With a smaller size that requires less compute resources, data, and energy than LLMs, SLMs are efficient, cost-effective gen AI models for many applications. Included with Red Hat AI products, the purpose-built, open source Granite family models help you control AI costs and get started more simply.

We also provide tools for fine-tuning models using your own enterprise data in a security-focused manner, so you can ensure your models are accurate and relevant for your use without unneeded complexity or cost.

Read the e-book to learn more about the benefits of open source SLMs.

Red Hat AI benefits

Increased efficiency

Access to Granite family models and a catalog of pre-optimized open source models helps to boost AI deployment and operational efficiency. These models require fewer computational resources to tune while delivering faster inference, helping you reduce hardware dependency and minimize costs.

Ease and accessibility

AI tools for all roles—from developers to data scientists to AI engineers—speeds model development and customization. By simplifying environment setup and streamlining hardware allocation for model training and tuning, Red Hat AI makes enterprise AI accessible across your organization.

Deployment flexibility

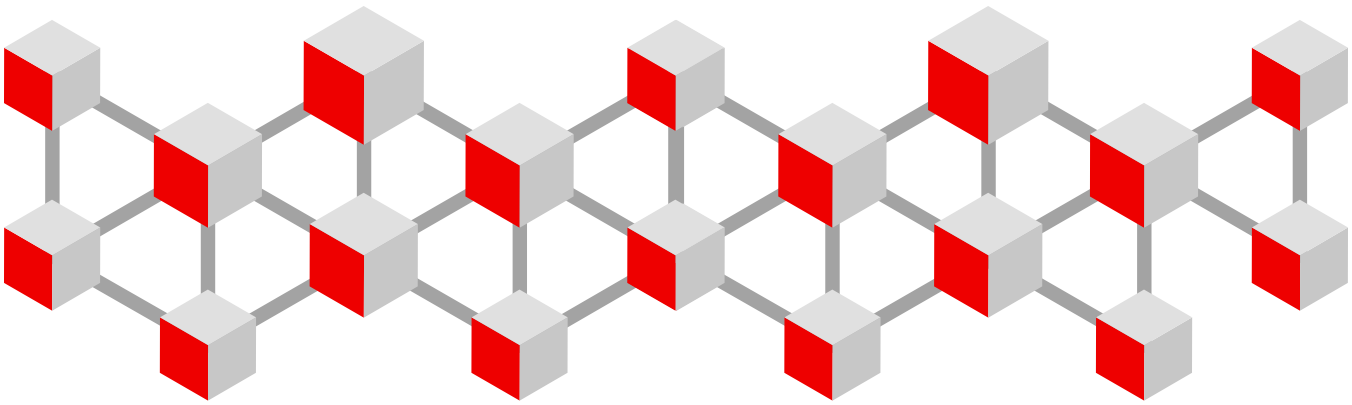
A consistent experience across hybrid cloud environments gives you the flexibility to choose where to train, tune, deploy, and run your models and gen AI applications. This helps you comply with data constraints, protect privacy, and maintain security while controlling AI infrastructure costs.

Get started with practical applications and use cases

You can use the Red Hat AI portfolio to implement a range of AI use cases to address many business challenges. The consistent user experience allows stakeholders from AI developers to data scientists to IT operations teams to more simply develop and deploy AI solutions across hybrid cloud environments.

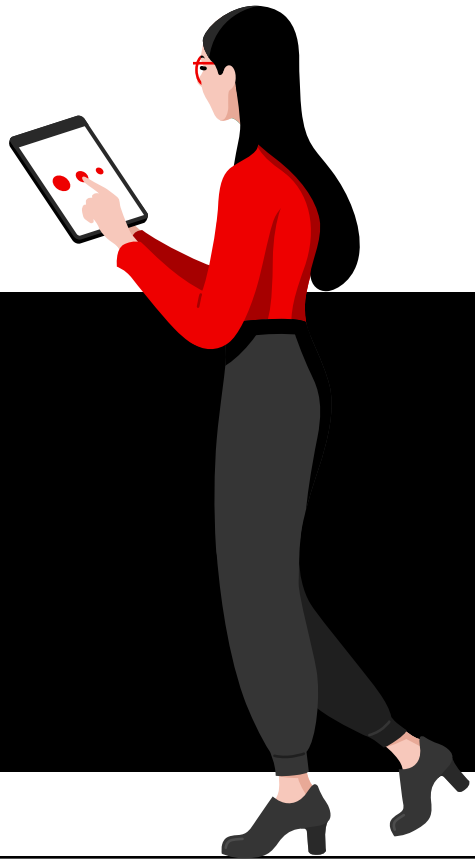
Common business use cases addressed by Red Hat AI

- 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. With greater insights, teams can maximize revenue, optimize operations, and enhance customer experiences and employee productivity.



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 AI basics to hands-on tool overviews. Complete a path to earn a certificate and boost your AI skills.

Start learning

> 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

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 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. Clalit is also using this platform to build learning processes and algorithms to identify new trends, patient and disease behavior patterns, and more.

Read the success story

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 to improve productivity and time to market for customer loan identification and fraud detection. As a key improvement, Intertech adopted Red Hat AI for its self-service capabilities and 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.

> 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

Read the success story

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.

> 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

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

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

Customer support and content creation

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

We use Red Hat AI within our own organization to increase the efficiency and scalability of customer and technical support services for our 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 our customers and support associates. These tools improve self-service, increase efficiency, and help bring about a faster response to support cases. For example, we increased the availability of knowledge content and minimized repetitive tasks for IT support associates that handle 30,000 new cases each month. And our AI-powered initiatives have shown the cost-saving potential of these solutions: We 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

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 for supporting employees in their daily work by providing instant answers to work-related questions and helping them respond with more accuracy to citizen inquiries and requests. With OpenShift AI on Red Hat OpenShift, the city can innovate faster, provide new services and functionality to the public, and maintain frequent release cycles.

Read the success story

Start your gen AI journey with Red Hat today

Build AI solutions for your world.

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 own 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 business goals.

Learn more about the Red Hat AI portfolio

Explore product information, key features and benefits, and access no-cost trials and developer sandboxes for Red Hat Enterprise Linux AI and Red Hat OpenShift AI.

Learn more

Discover AI resources and knowledge for your entire organization

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

Build your skills

0425_KVM

Red Hat