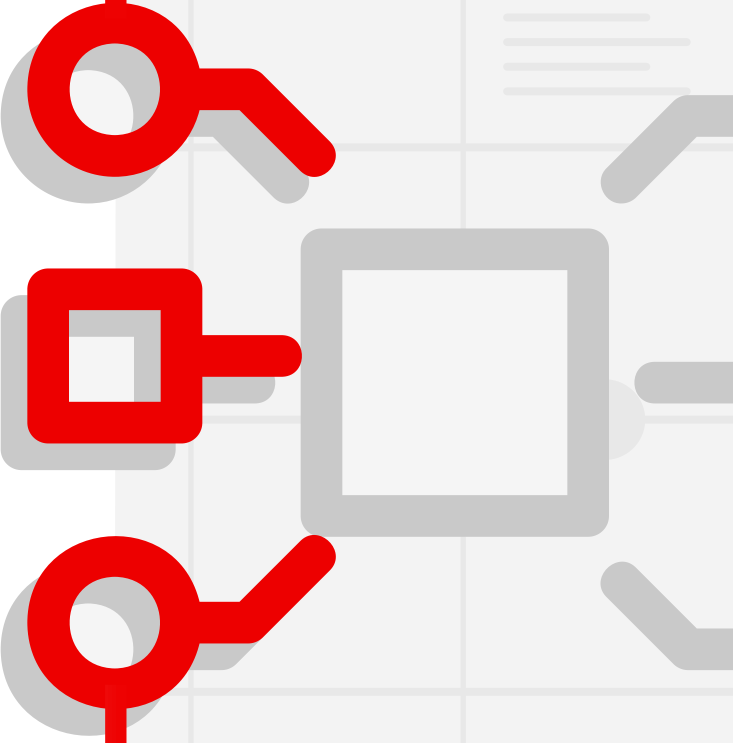
Get started with AI for enterprise: A beginner’s guide

Red Hat

Table of contents

Introduction

Organizations are increasingly recognizing the opportunity that artificial intelligence (AI) presents for all aspects of their business.



From customer engagement, support, and sales, through to IT infrastructure, processes, code development, and solution delivery, AI is expanding in use cases and gaining momentum across all industries.

According to IDC, the market is expected to exceed US$423 billion by 2027 with a 5-year compound annual growth rate (CAGR) of 26.9%, with many businesses focusing their AI initiatives on improving operational efficiency, customer experience, and productivity.[1]

Amidst this rapid evolution, leaders are under pressure to identify, choose, build, and deliver AI solutions that will give their organizations a competitive advantage. But the rate of AI innovation and the ability of most organizations to increase their AI maturity are moving at different paces. This makes it challenging to unlock the full value of AI and, in many cases, creates more questions than answers.

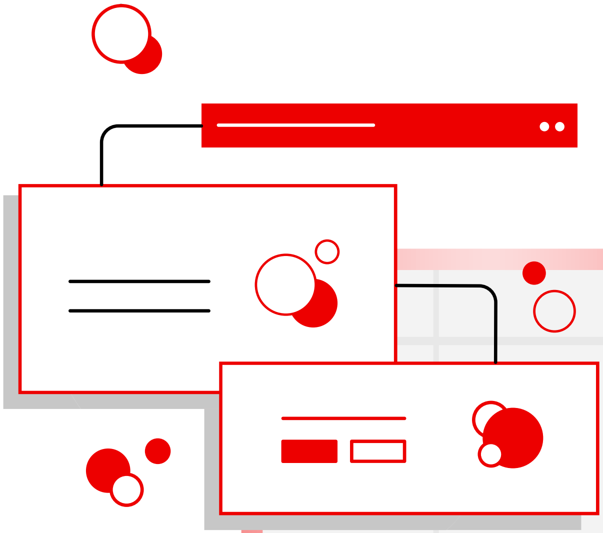
Whether you're just getting started with your AI journey, looking to understand more about the impact AI will have on your business, or figuring out how to scale existing AI implementations, this e-book aims to answer many of the questions about AI today.

[1] IDC FutureScape Webcast. "Worldwide Artificial Intelligence and Automation 2024 Predictions." Document #US51901124, March 2024.

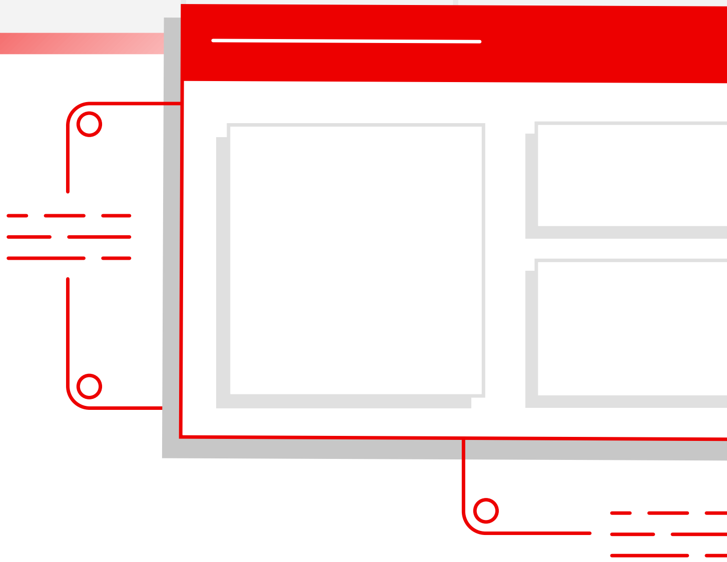
What are the types of AI?

To make the most of AI, get to know everything about it—including 2 of the most prominent types being used by organizations today.

Predictive AI: Using historical data, predictive AI helps organizations identify patterns and make informed decisions about the future. Predictive models power applications like demand forecasting, predictive maintenance, and operational planning. Predicative AI is based on well-established data science and machine learning (ML) techniques, allowing AI to improve as more data is processed.

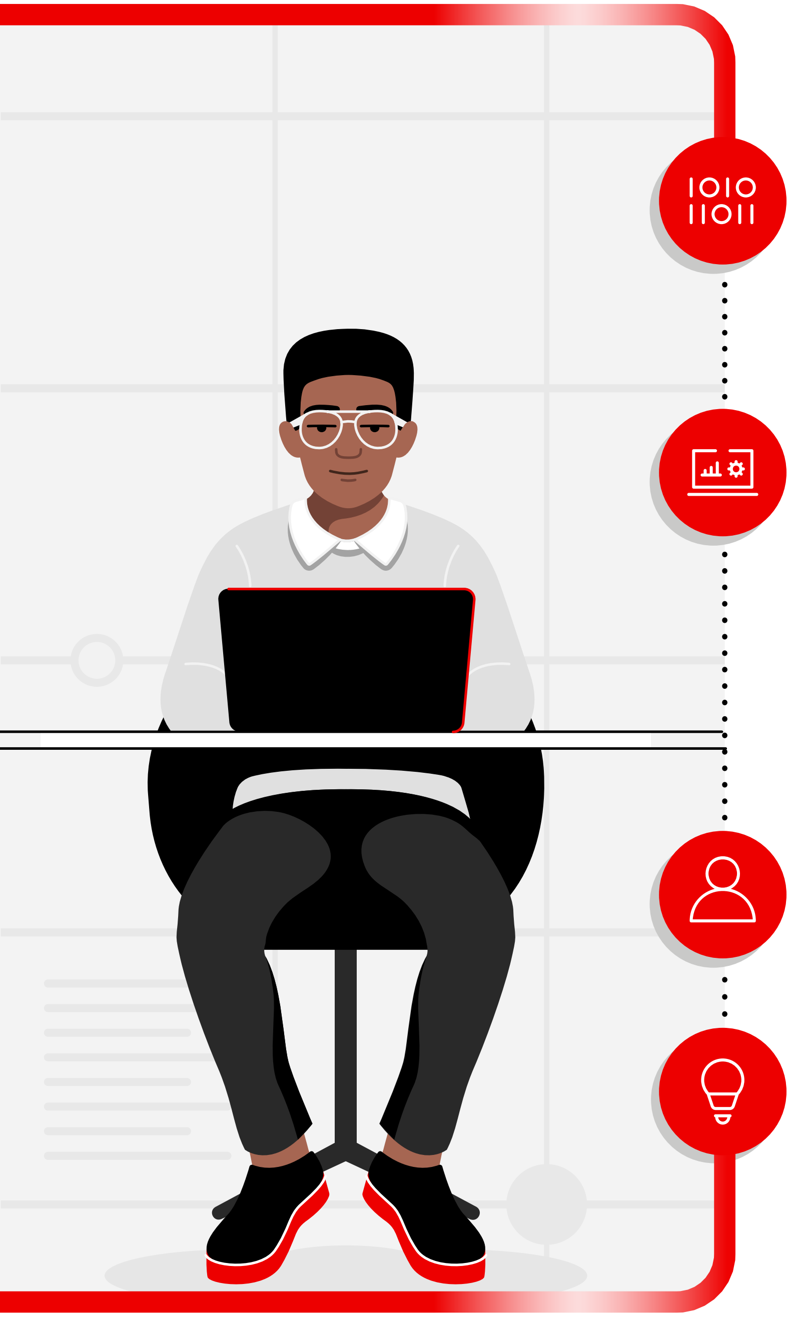


Generative AI (gen AI): Generative AI, powered by deep learning models like transformers, can create new content such as text, images, and code. It's particularly useful for applications like chatbots, automated content generation, and creative tools. Models like generative pretrained transformers (GPTs) have revolutionized natural language processing and creative fields by producing human-like text and images.



What are the benefits of AI implementations?

The full capabilities of AI are still yet to be discovered, but understanding how this rapidly evolving technology is already benefiting organizations of all sizes in various industries is a good way to determine where to integrate AI into your organization.



Consider the following benefits of AI and how they might help your organization:

Data volume. With the exponential growth of data, organizations often struggle to manage and derive insights from the vast amounts of information they collect. AI can process and analyze large datasets quickly, uncovering valuable insights and trends that would be difficult to identify manually.

Operational inefficiency. Many organizations understand that inefficient processes and bottlenecks can hinder productivity, and as a result, more time and effort is required to remove these hurdles. Automation that uses AI helps to streamline operations, which reduces errors and improves process efficiency. For example, this could include simple applications such as automatically generating meeting notes with action items and clear next steps, or accelerating the creation of graphics and video creation for websites or social media.

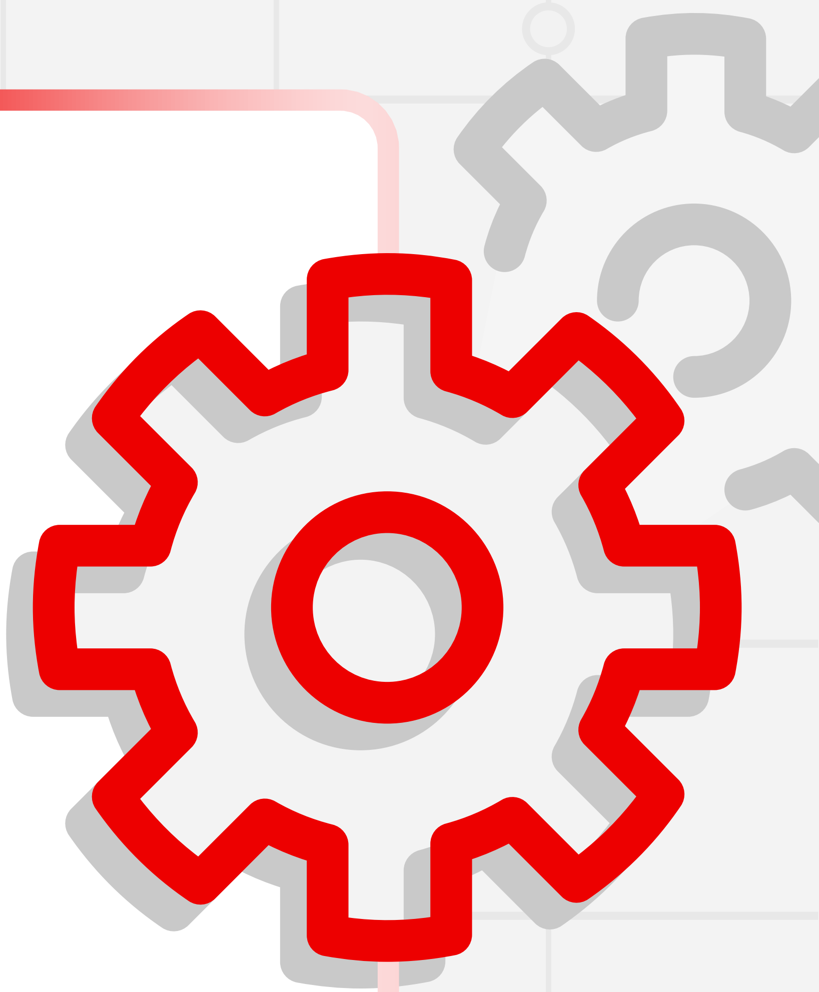
Customer expectations. Customers expect personalized and hassle-free experiences. By analyzing customer data and providing tailored recommendations and bespoke interactions, AI can enhance customer service and personalization.

Market competitiveness. Staying competitive in a rapidly evolving market requires continuous innovation. AI can help organizations adapt quickly to market changes and maintain a competitive edge. Gen AI can even be used to help refine your approach when used as a thought partner for leadership or when preparing for an important meeting.

Chapter 1

The rise of AI

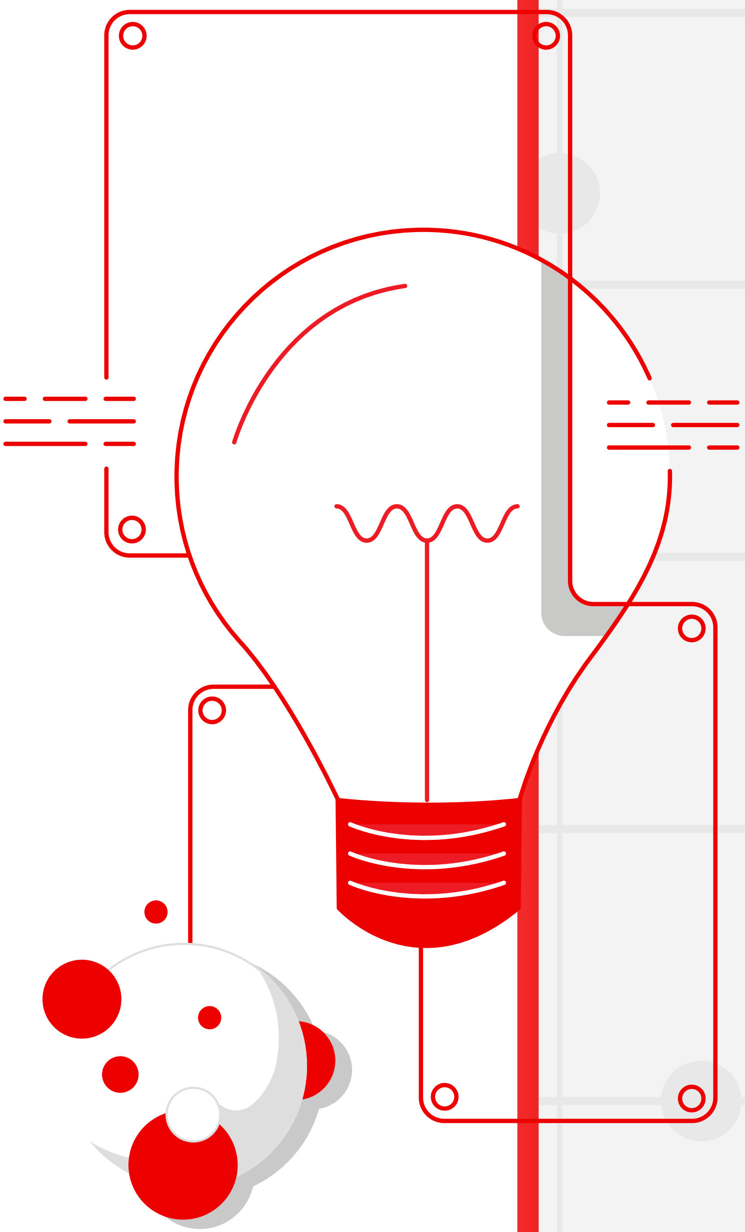
AI has been evolving for decades, powering advances in industries such as healthcare, finance, and manufacturing.



However, the recent rise of gen AI has captured attention due to its ability to create human-like text, realistic images, and even software code. Unlike traditional AI that automates tasks or analyzes data, gen AI opens the door for creative problem-solving and advanced content creation.

The types of AI models accelerating innovation

Large language models (LLMs) and stable diffusion models are among the AI models responsible for the explosive growth of gen AI. LLMs, like GPT, are pretrained on massive datasets and can understand and generate natural language, making them invaluable for customer support automation, marketing copy generation, and more. Stable diffusion models, on the other hand, create hyper-realistic images, fueling innovation in entertainment, marketing, and beyond.



Emerging trends to consider

Businesses are increasingly exploring multimodal AI, which combines text, image, and data processing capabilities into a single model, offering more versatile solutions. Staying ahead of these trends is key to taking advantage of the full potential of AI in enterprise settings.

Open source: A foundation for AI innovation

Red Hat's AI strategy is deeply rooted in open source, helping enterprises to advance gen AI with transparency, trust, and lower costs. By using Red Hat's open hybrid cloud platforms, organizations can innovate freely while maintaining control over their AI solutions.

Learn more about LLMs and how they work

Take control of LLMs with open source

While gen AI is creating change in nearly all aspects of business, from how software is made to how we communicate, it's not uncommon for the models (LLM and other kinds) used as part of a gen AI capability to be tightly controlled by the provider of the service. This means it isn't easy for an enterprise to evaluate the capabilities of a gen AI service without specialized skills and sometimes high costs (both monetary and time).

Lacking visibility into the data sets that created the model or details on how the model uses data exposes enterprises to potential risks in terms of AI-generated content. What if a code generation model was trained on copyright source code? Is any code generated by that model also part of that copyrighted code? Many questions such as this have not been fully answered but understanding the consequences can be significant, enterprises are turning to open source AI.

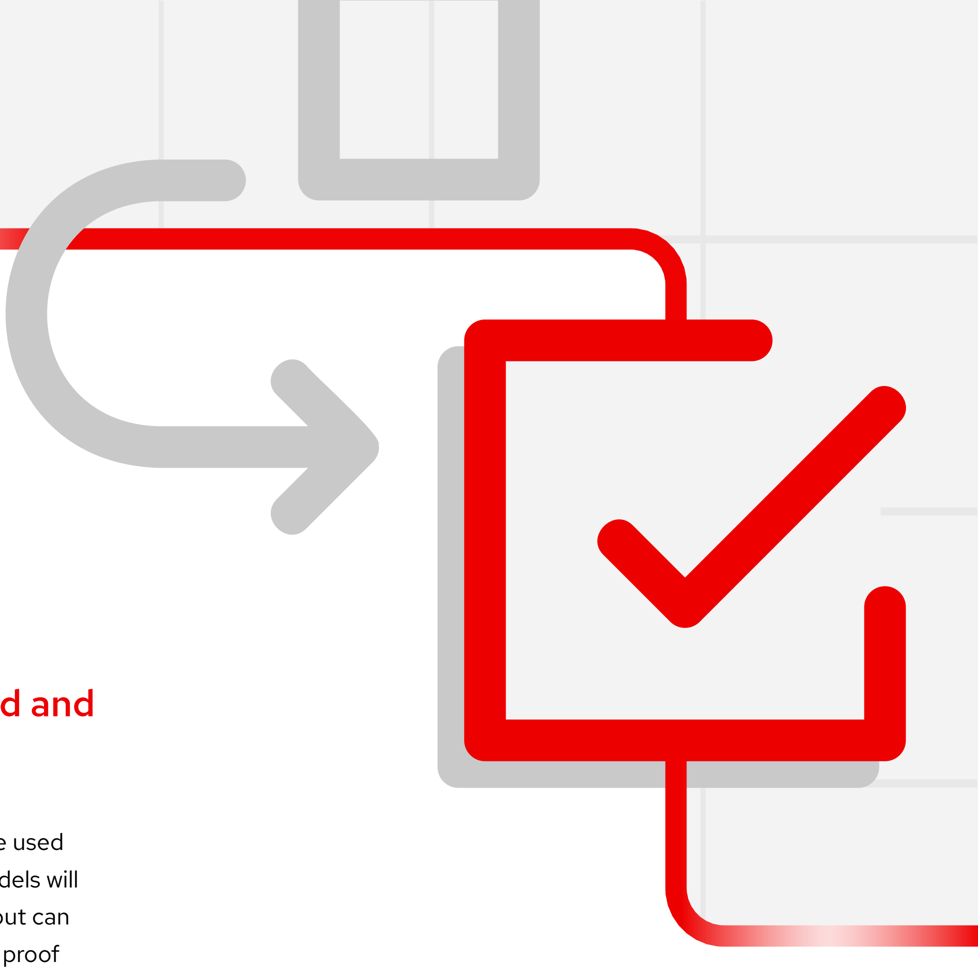
Red Hat's approach to AI is rooted in open source, as is our support for open source models such as the IBM Granite family of foundation models.

Red Hat's AI solutions even contribute directly to AI model development with InstructLab, a community-led solution for enhancing LLM capabilities.

Explore InstructLab on GitHub

Chapter 2

Choosing the right AI model



Different AI models can be used and applied to diverse use cases.

Predictive AI and generative AI models could all be used in a single application or service. Each of these models will incur different costs and provide unique benefits, but can help organizations reduce time to market for initial proof of concept. Image segmentation, voice-to-text, and image recognition models are common and highly capable examples but the important part is to evaluate what is best for your use case.

Foundations models, trained on vast amounts of data, provide great flexibility in their capabilities, however their large size can inflate cost, boost management demands, and increase complexity, which means they may not be suitable for all use cases.

Smaller, fine-tuned models (that are still part of the gen AI family) can be a better solution in cases where tuning an existing model to suit your requirements is preferred. When deciding on a model, you might opt for a prebuilt option, which is readily available and can be easily integrated in your systems. A common example of this kind of model is a LLM, which is a powerful tool that has already been trained on vast amounts of data.

However, if you have specific business requirements, data privacy concerns, or a desire for greater control over the model's behavior, there may be a need to build and self-host a custom model.

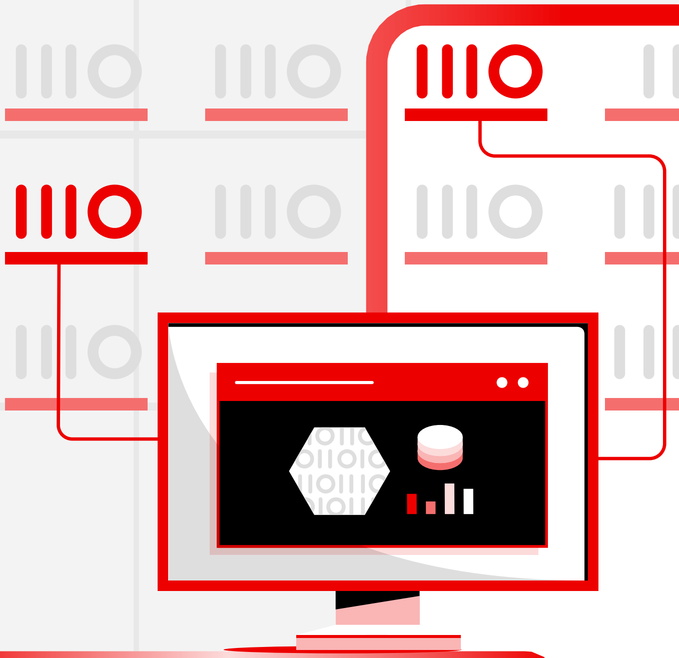
Model building versus model tuning

Building an AI model from scratch can be a huge undertaking. You will need to gather and prepare large datasets relevant to your organization's business use case. Then, you must choose an appropriate algorithm and train it on your data. This process requires significant computational power and expertise, making it a time-consuming and resource-intensive endeavor. While building a traditional or foundational model can provide a custom solution, it might not always be the most efficient path.

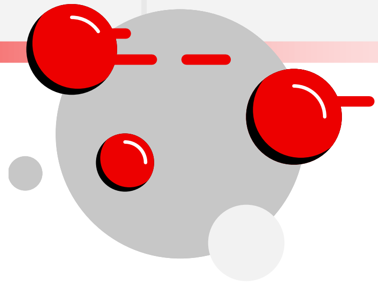
On the other hand, tuning a foundation model involves adapting a pretrained model to your specific requirements. A common approach is transfer learning, which involves using a model trained on a large dataset and retraining it on a smaller, domain-specific dataset. This method allows the model to retain the general knowledge it learned during its initial training while adapting to the nuances of your specific data.

Fine tuning your model

Another approach is fine tuning, where you adjust the model's parameters to improve performance on your specific task. A model's parameter refers to the variables of a selected model that can be estimated by fitting the given data to the model. Fine tuning might involve changing the learning rate, modifying the model's architecture, or training certain layers of the model more intensively than others. These techniques help enhance the model's knowledge, making it more effective for your particular use case.



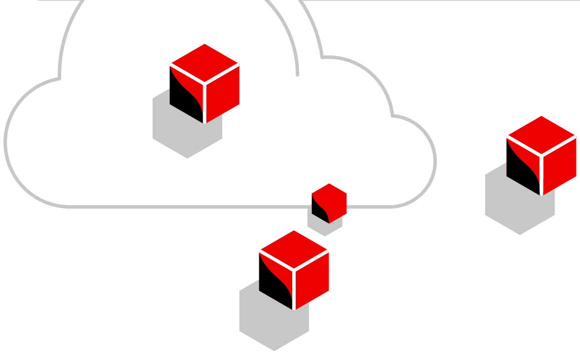
InstructLab takes a fine-tuning approach, with the goal of reducing the prerequisite AI knowledge and simplifying the addition of enterprise knowledge to existing Granite foundational models.



Alternatives to fine tuning models

Researchers are investigating ways to tune foundation models for greater speed and efficiency. Some common techniques are retrieval-augmented generation (RAG); this is a technique for retrieving facts from an external source in which additional knowledge (context) has been encoded.

RAG relies on the use of 1 or more external databases (vector databases) which feed additional context to the question being asked of the gen AI model. Another emerging approach is Agentic AI systems which combine a number of gen AI agents together to query outside systems for knowledge—like internal databases, corporate intranets, or the internet—to provide gen AI models with the most accurate, up-to-date information.



The final example is prompt tuning, where AI models receive cues or front-end prompts, including extra words or AI-generated numbers, that guide models toward a desired decision. The result from a RAG query would make up additional context for a prompt, with prompt tuning and RAG working together. The combination of fine tuning an existing LLM and the use of RAG techniques and prompt tuning allows organizations with limited data to tailor a foundation model to a narrow task.

The infrastructure supporting your AI model is just as important as the model itself. Different tasks require different types of hardware.

Central processing unit (CPU)

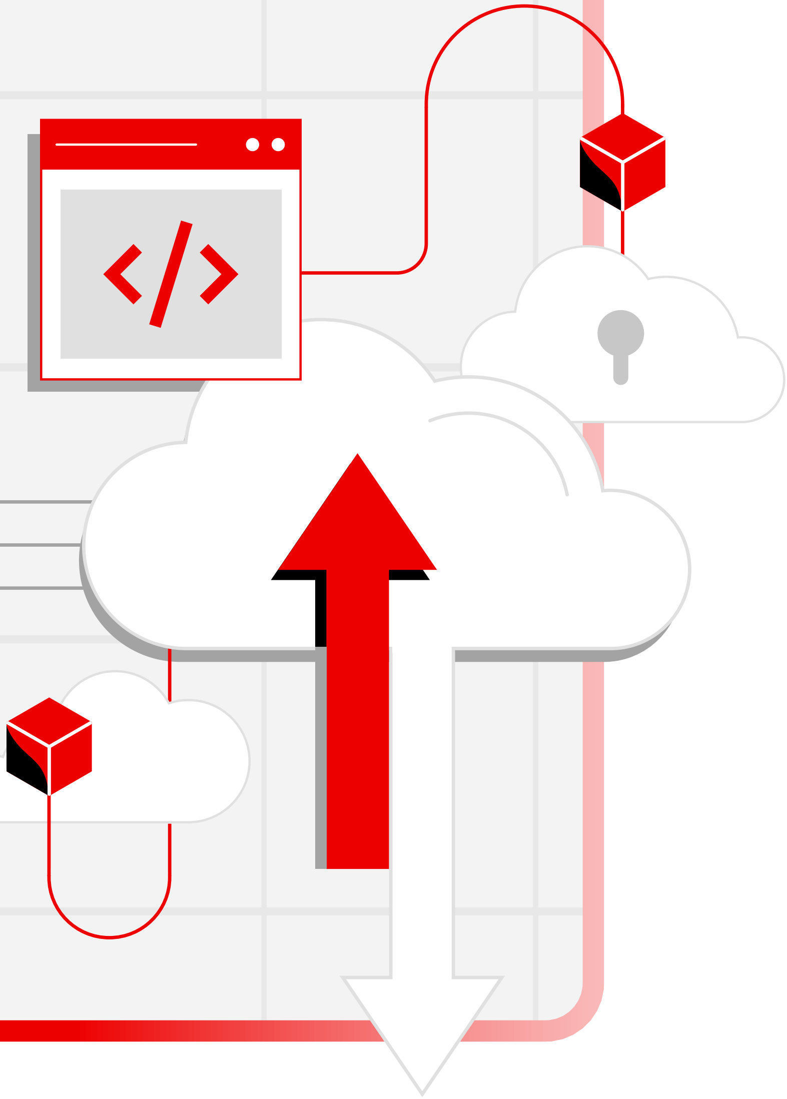
Traditional processors that handle general computing tasks. They're versatile but may not be efficient for large-scale AI workloads.

Graphics processing unit (GPU)

Specialized processors designed to handle parallel processing tasks, making them ideal for training deep learning models that require processing large amounts of data simultaneously.

Neural processing unit (NPU)

A newer type of processor designed specifically for AI tasks, offering even greater efficiency and speed for certain types of models.



The role of hybrid cloud in AI adoption for enterprise

Hybrid cloud environments play a critical role in AI adoption. A hybrid cloud combines on-premise infrastructure with public and private cloud resources, offering flexibility in how and where you deploy and manage AI workloads. For example, you might train your AI models using powerful cloud-based GPUs and then deploy them on premise or in a private cloud for security or compliance reasons. Therefore, a key consideration when taking advantage of the hybrid cloud approach is consistency of tooling and the platform that you choose.

Red Hat's open hybrid cloud approach helps organizations to integrate AI across different environments, improving consistency, scalability, and flexibility. This approach allows you to manage your AI workloads across multiple cloud environments, on premise, or at the edge of the network, optimize data placement, and promote smooth data migration, making it easier to adopt AI at an enterprise scale.

By understanding AI models, data, and infrastructure, you can better navigate the complexities of AI adoption and use its full potential.

What you need to get started

As with the adoption of any new technology, there are challenges associated with AI that an organization must overcome to be successful.

Use the following considerations to assess your organization's readiness and identify the areas where you may need to focus to accelerate your AI adoption.

Evaluate data quality and availability. Access to high-quality, relevant data is essential for AI. Data quality is critical for training accurate AI models, so it's important to assess the completeness, accuracy, and relevance of your data.

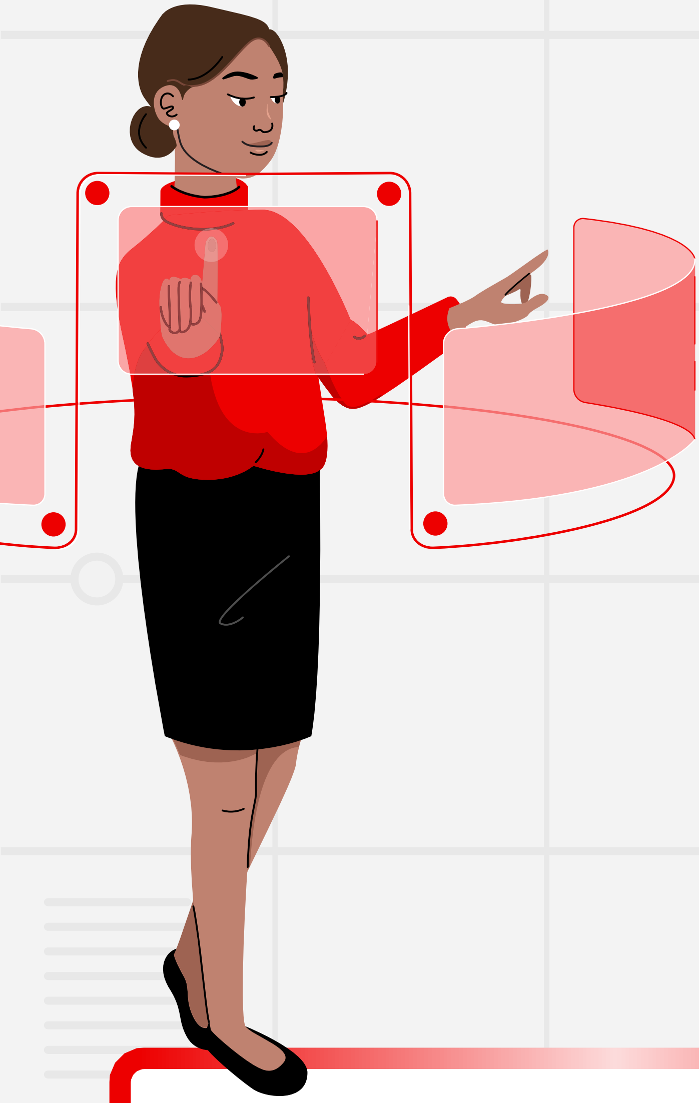
Assess technological infrastructure. Determine if your current infrastructure can support AI workloads. This includes evaluating the availability of high-performance computing resources, storage solutions, automation, and network capabilities.

Identify where skills are needed. Evaluate the availability of AI expertise within your organization. Assessing the current skill set and identifying where training or specialized skills may be needed.

Review strategic alignment. Make sure that your AI initiatives align with your business goals and strategies. AI projects should support your organization's overall strategic objectives and deliver measurable business value.

How to get started with AI

The speed and scale of AI adoption within an organization depends on many factors, but starting small and growing incrementally is often a good approach for almost any technology modernization project.



Here are 8 steps to help your organization get started and advance your journey to AI adoption:

1

Evaluate abilities and goals

Begin by evaluating your organization's current capabilities, infrastructure, and strategic goals. Determine if AI aligns with your broader objectives and identify the potential areas where AI can add value. This initial assessment will help set a clear direction for your AI adoption journey.

2

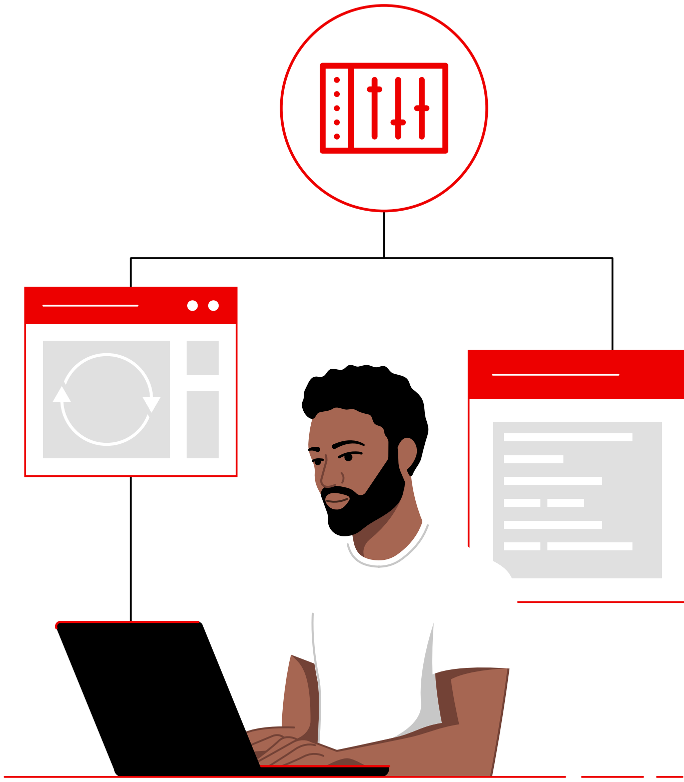
Identify use cases and AI teams

Identify opportunities within your organization that AI can address. Form a dedicated AI application team, including cross-functional members such as developers, domain experts, data scientists, and IT specialists, to lead the initiative. A well-defined use case will guide your AI adoption efforts and focus resources.

3

Model selection

Choose the appropriate AI model based on your identified use case. Whether it's a LLM for gen AI or a predictive model for data analysis, make sure the model's capabilities align with the goals of your use case. Consider factors like the model's complexity, scalability, and compatibility with your existing systems.



4

Testing and validation loops

Define clear success criteria for your AI implementation, such as performance metrics, accuracy rates, or business objectives. Establish testing and validation loops to continuously assess your model's effectiveness. Regular feedback from these loops will help fine tune the model and keep your AI journey on track.

5

Model tuning

Customize your selected model using your organization's data. This tuning process involves feeding the model relevant data to improve its accuracy and relevance to your specific business' use case. Fine tuning makes sure that the model adapts to your organization's unique context and needs.

6

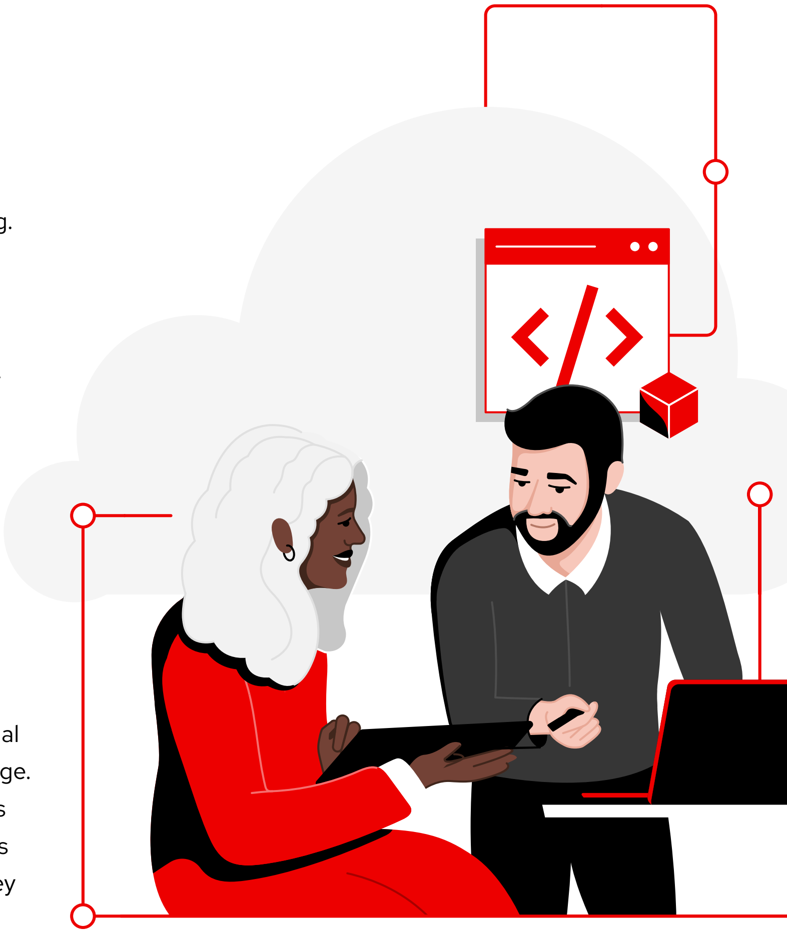
Synthetic data training

Consider using synthetic data to further train and enhance your AI models. This approach, using methods such as LLM teacher and LLM student, allows you to generate high-quality training data when real data is scarce or sensitive. Synthetic data can help improve the model's robustness and performance without compromising privacy.

7

Monitor drift

Drift monitoring provides general, content-based monitoring. Rather than structured configuration management, drift monitoring tracks changes to content on the local file system. Drift monitoring can help you detect and address any deviations or declines in the model's accuracy over time. Continuous monitoring ensures the model remains effective and relevant as conditions change.



8

Engaging expert help

If your in-house AI expertise is still growing, engaging external experts such as Red Hat® Consulting can be a huge advantage. Red Hat experts can guide you through the complex aspects of AI adoption, provide valuable insights, and give you access to training. Red Hat Consulting can accelerate your AI journey and improve the likelihood of success.

AI adoption depends on collaboration

Building a team with a range of capabilities is key for successful gen AI projects.[2]

- **Business leaders** represent the people who use or are affected by the solution.
- **AI specialists** tune, maintain, and update gen AI models.
- **Data scientists** preprocess and provide correct, unbiased training data for models.
- **Ethics and compliance officers** ensure that gen AI initiatives comply with regulations.
- **IT operations specialists** integrate solutions with existing infrastructure and enforce security policies.
- **Development teams and communities** need to be involved at the beginning to collaborate, create, share, and improve open source tools, frameworks, and best practices for AI adoption. This will ensure AI uses are connected to business value.

[2] Kearney. "Standing up tiger teams to tackle generative AI complexity," 15 Nov. 2023.

Chapter 4

Adopt and scale with Red Hat

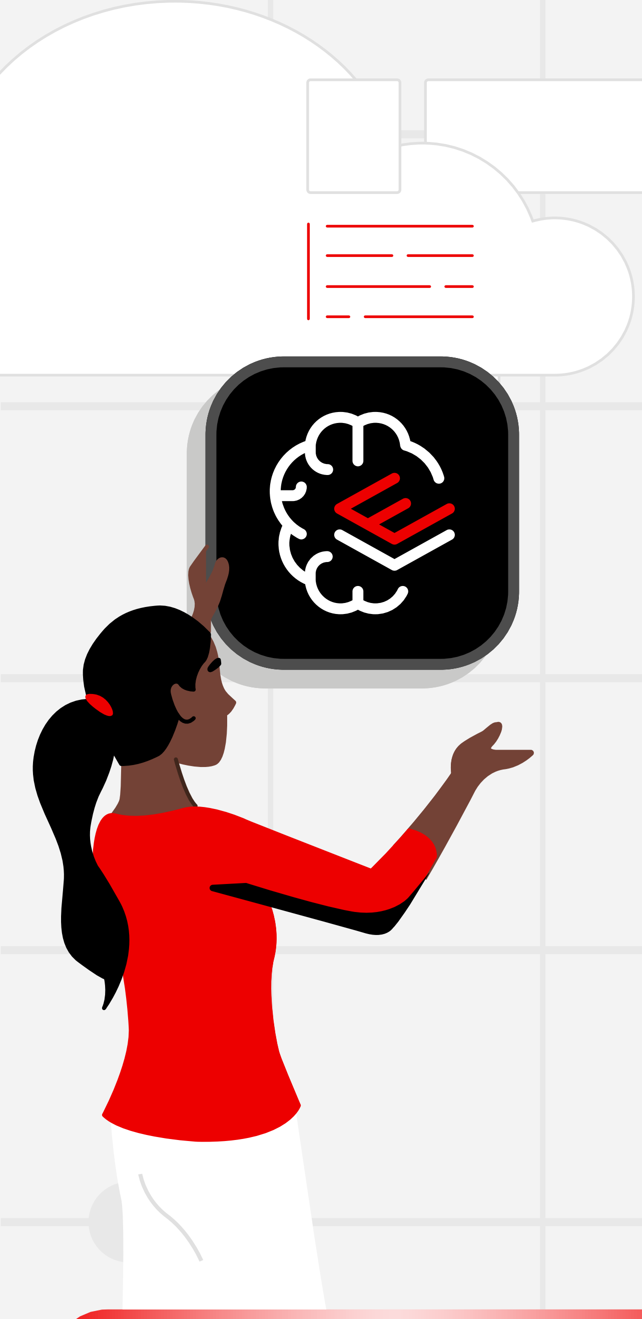
Red Hat AI delivers trust, choice, and consistency across the hybrid cloud to accelerate enterprise adoption of AI.

The Red Hat AI portfolio includes Red Hat Enterprise Linux® AI, for individual Linux server environments, and Red Hat OpenShift® AI, for distributed Kubernetes platform environments, and provides integrated machine learning operations capabilities. Both solutions are powered by open source technologies and models, helping organizations stay at the forefront of AI innovation, accelerate the pace of discovery, and democratize access to cutting-edge tools and technologies.

Red Hat's extensive partner ecosystem can further strengthen your AI capabilities. For example, NVIDIA, the major AI company known for popularizing the GPU, continues to partner with Red Hat to unlock the power of AI by providing an end-to-end enterprise platform optimized for AI workloads. NVIDIA helps enterprise customers adopt GPU-accelerated computing for AI and high-performance computing applications.

> **"Red Hat and NVIDIA have a long history of close collaboration, and Red Hat Enterprise Linux AI demonstrates our shared focus on bringing full-stack computing and software to the developers and researchers building the next wave of AI technology and applications."[3]**
>
> **Justin Boitano**, Vice President, Enterprise Products, NVIDIA

[3] Red Hat press release. "*Red Hat Delivers Accessible, Open Source Generative AI Innovation with Red Hat Enterprise Linux AI,*" 7 May 2024.



A closer look at Red Hat Enterprise Linux AI

Red Hat Enterprise Linux AI comprises 4 foundational components:

1 **Open Granite models**

Red Hat Enterprise Linux AI includes open source Granite models that are fully supported by Red Hat. These flexible models allow you to create custom language models and use them publicly or privately.

2 **InstructLab model alignment**

InstructLab is an open source project led by Red Hat and IBM. It customizes AI models with specific knowledge and generates synthetic data for training. As a command-line tool that integrates with a git repository, users can add skills and train models easily.

3 **Optimized bootable Red Hat Enterprise Linux for Granite models and InstructLab**

Granite models and InstructLab tools run on a specialized Red Hat Enterprise Linux image optimized for AI, which is compatible with virtually all hardware and cloud environments. This setup allows for efficient performance with high-end GPUs, necessary for quick training and model deployment.

4 **Enterprise support and indemnification**

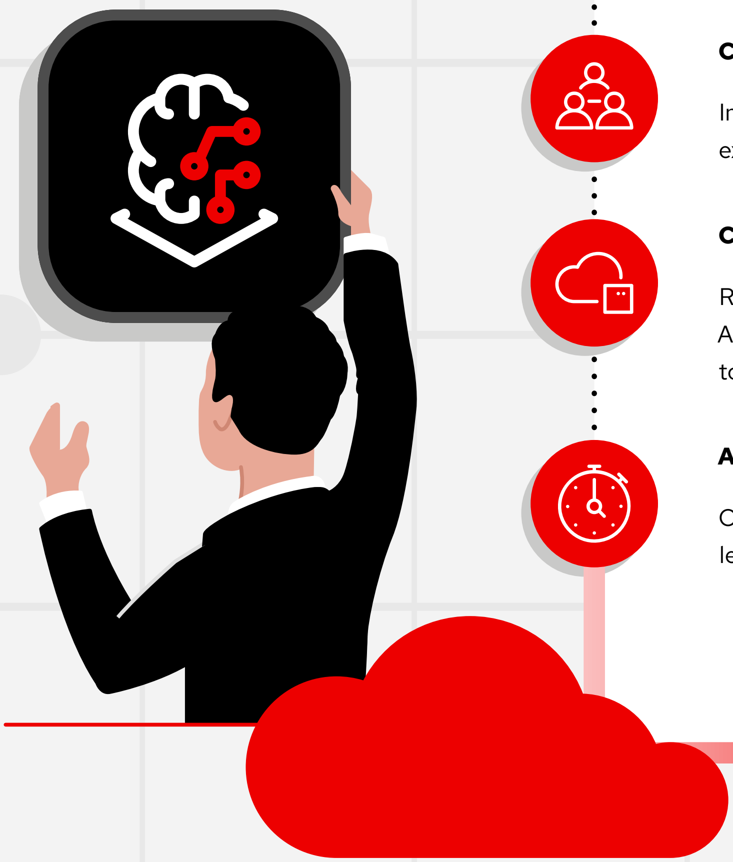
Red Hat Enterprise Linux AI subscriptions include enterprise support, a complete product lifecycle starting with the Granite 7B model and software, and IP indemnification by Red Hat.

Red Hat Enterprise Linux AI helps bring gen AI applications to life

For organizations just getting started with gen AI, Red Hat Enterprise Linux AI provides ready-to-use LLMs and code language models in a single server development and inference environment.

This provides a unified environment with models and tools, making it easy to get started with gen AI and customize models using your business data, without the need for extensive AI expertise or infrastructure.

Fully supported and indemnified by Red Hat, Red Hat Enterprise Linux AI reduces risk. It also provides a simplified approach to gen AI designed to be more accessible to developers and domain experts, allowing them to collaborate and accelerate the time it takes to see business results.



Why Red Hat Enterprise Linux AI?

LLMs for the enterprise

Open source-licensed IBM Granite LLMs are included under the Apache-2.0 license, and fully supported and indemnified by Red Hat.

Community collaboration

InstructLab makes it possible to simplify gen AI model experimentation and alignment tuning.

Cloud-native scalability

Red Hat Enterprise Linux image mode lets you manage your AI platform as a container image, streamlining your approach to scaling.

Acceleration and AI tooling

Open source hardware accelerators, plus optimized deep learning features support accelerated results.

Scale with Red Hat OpenShift AI

Red Hat OpenShift AI provides an integrated MLOps platform for building, training, tuning, deploying, and monitoring AI-enabled applications and predictive and foundation models at scale across hybrid cloud environments.

Red Hat OpenShift AI builds on top of Red Hat OpenShift to deliver a consistent, streamlined, and automated experience when handling the workload and performance demands of AI/ML projects. MLOps practices can help organizations respond rapidly to AI innovations and deliver AI-enabled applications into production more quickly.

Experiment in the Red Hat OpenShift AI sandbox

With components curated from Open Data Hub and other open source projects, Red Hat OpenShift AI gives data scientists and developers a powerful open hybrid AI/ML platform for gathering insights from data and building AI-enabled applications.

Try it in our developer sandbox >

Why Red Hat OpenShift AI?

Scale model serving

Models can be served for integration into intelligent applications on premise, in the public cloud, or at the edge. These models can be rebuilt, redeployed, and monitored based on changes to the source notebook.

Spend less time managing AI infrastructure

Provide your teams with on-demand access to resources, so they can focus on exploring data and building applications that add value to your organization.

Tested and supported AI/ML tooling

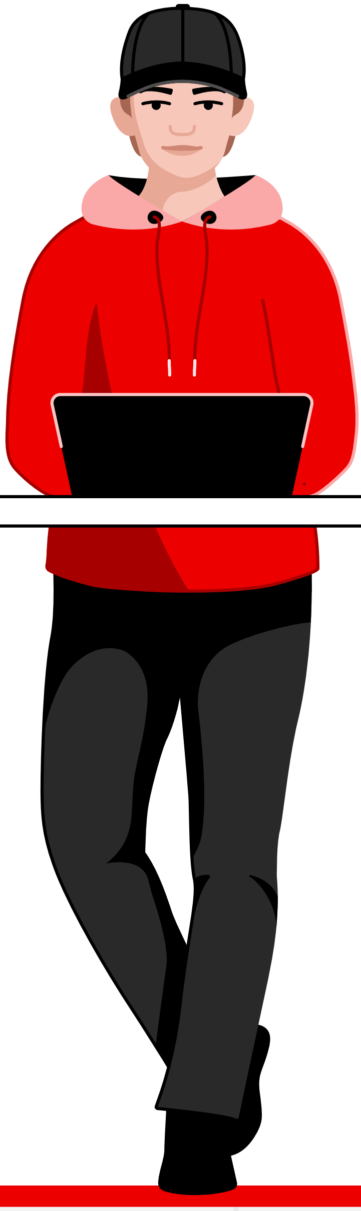
Red Hat tracks, integrates, tests, and supports common AI/ML tooling and model serving on Red Hat OpenShift application platform.

Flexibility across the hybrid cloud

Offered as either self-managed software or as a fully managed cloud service on top of Red Hat OpenShift, Red Hat OpenShift AI provides a security-focused and flexible platform that gives you the choice of where you develop and deploy your models–whether on premise, the public cloud, or even at the edge.

Operate using our best practices

Red Hat Consulting provides services that allow you to install, configure, and use Red Hat OpenShift AI to its fullest extent. Whether you're pursuing an Red Hat OpenShift AI pilot experience or need guidance on building your MLOps foundation, Red Hat Consulting will provide support and mentorship.



Red Hat provides a complete technology portfolio, proven expertise, and strategic partnerships to help you achieve your gen AI goals. Gain a foundation for developing and deploying gen AI models and applications, as well as services and training for rapid adoption.

Learn more

Ready to take the next step on your AI adoption journey?

Accelerate your AI adoption with Red Hat's open hybrid cloud strategy, giving you the flexibility to run your AI applications anywhere you need them.

Jumpstart your AI/ML projects with Red Hat expertise, consulting, and training services to help your organization get where you want to be with AI.

Learn about AI/ML services: red.ht/aiml-consulting

Schedule a complimentary discovery session: redhat.com/consulting

Learn more about Red Hat AI

Read more about Red Hat Enterprise Linux AI

Discover how you can scale with Red Hat OpenShift AI