Red Hat

Get started with AI for enterprise organizations: A beginner's guide

Table of contents

AI terminology

Foundation model: A large AI model trained on broad data and designed as a general-purpose base that applications build upon, typically adapted or fine-tuned to perform many domain-specific tasks.

Frontier model: The most capable models available at any given time. This is a relative term that shifts as the field advances.

Generative AI (gen AI): AI models that learn from existing data and then generate new content, such as text, images, code, or audio, in response to prompts.

Large language model (LLM): A type of foundation model that is trained on large amounts of text (and often code) so it can understand and generate human-like language for many different purposes.

Small language model (SLM): A compact, efficient language model (typically 1–10 billion parameters) optimized for specific use cases and the tasks within them.

Introduction

Organizations are increasingly recognizing the opportunities artificial intelligence (AI) offers across all aspects of their business. From customer engagement, support, and sales to IT infrastructure, processes, code development, and solution delivery, AI is gaining momentum across all industries.

According to IDC's Worldwide Artificial Intelligence IT Spending Market Forecast, between 2025 and 2029, year-over-year spending on artificial intelligence (AI) initiatives will grow by 31.9%. This investment, driven by the growth of agentic AI applications and systems to manage agentic fleets, will reach $1.3 trillion in 2029.[1]

Amid 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 is moving far too fast for most organizations to increase their AI maturity at the same pace. 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 effect 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 adopting AI.

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.

Predictive AI helps organizations identify and connect patterns, historical events, and real-time data to predict future outcomes with extremely high accuracy. It can be applied to demand forecasting, preventive maintenance, and operational planning, helping organizations fix issues before they affect employees, customers, or company resources. Predictive AI is based on established data science and machine learning (ML) techniques, allowing AI to improve as more data is processed.

Generative AI (gen AI) lets organizations produce new content shaped by unique data. Gen AI applications typically rely on deep learning models, such as transformers, to generate new content, including text, images, and code. Popular large language models (LLMs), like OpenAI's family of GPT (generative pre-trained transformer) models, continue to advance rapidly, revolutionizing natural language processing and creative fields by generating human-like text and images. Gen AI is particularly useful for applications like chatbots, automated content generation, and creative tools. Multimodal AI systems may combine multiple generative AI models to deliver text and multimedia content more efficiently.

Other AI approaches, such as computer vision for object detection, image classification, and segmentation, are also widely used in enterprise settings, particularly in manufacturing and quality control.

Beyond these 2 categories, other AI approaches are widely used in enterprise settings. Computer vision, for example, enables object detection, image classification, and segmentation, which is particularly valuable in manufacturing and quality control. And popular LLMs like OpenAI's family of GPT (generative pre-trained transformer) models, among others, continue to advance rapidly, revolutionizing natural language processing and creative fields by generating human-like text and images.

1 IDC Press Release. "Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and $1.3 Trillion in 2029, According to IDC." 26 Aug 2025.

What are the benefits of AI implementations?

Consider the following challenges your business may be facing. How could you apply AI to help your organization?

- **Data volume.** Organizations often struggle to manage and derive insights from the vast amounts of information they collect. AI can quickly process, summarize, and analyze even extremely large data-sets, uncovering valuable insights and trends that would be difficult to identify manually.
- **Operational inefficiency.** Inefficient processes and bottlenecks hinder productivity. Automation using AI helps streamline operations, reducing errors and improving process efficiency. This could include simple applications such as automatically generating meeting notes with clear next steps, or accelerating the creation of graphics and video for websites or social media, to more complex tasks such as helping software engineers write and test code.
- **Customer expectations.** Customers expect personalized, efficient 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 more quickly to market changes, maintain a competitive edge, and even 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 existed—at least conceptually—since the 1950s. Like most scientific or technological evolutions, AI development has plateaued, jumped forward, then plateaued again—the cycle of hype, disillusion, and then more realistic progress played out over decades.

Today, AI has reached an inflection point. What was once experimental is now production-ready, and what was once theoretical is now transforming how enterprises operate.

The catalyst? A 2017 research paper from Google, "Attention Is All You Need," introduced the transformer architecture: the foundation upon which today's large language models (LLMs) are built. Transformers allowed AI systems to process vast sets of data and understand context in ways previous approaches couldn't match. Combined with advances in computational power and storage capacity, this breakthrough delivered the ability to generate human-like text, realistic images, and working software code.

But this is just the starting point. Continued work to develop models, perfect inference techniques, and integrate different AI approaches has brought us to systems that can perform complex reasoning and problem-solving, plan and execute actions with autonomy, and learn from their interactions. For enterprises, this opens the door to sophisticated automation and creative solutions that weren't possible even a few years ago.

This guide will help you better understand the models delivering modern innovation, assess your organization's readiness, choose the right approach, and build a practical roadmap to adopt and scale AI with confidence.

The types of AI models accelerating innovation

Large language models (LLMs) and image-generation models are among those delivering the recent explosive growth of generative AI.

LLMs (such as those from OpenAI, Anthropic, and Meta) are pretrained on massive datasets to process and generate natural language, making them invaluable for customer support automation, marketing copy generation, and more. Image generation models (such as Stable Diffusion, Midjourney, and DALL-E), on the other hand, create visuals from text prompts, fueling innovation in entertainment, marketing, and beyond.

The rise of reasoning and planning models

A new class of reasoning models emerged in 2025, fundamentally changing how AI approaches complex problems. These models use reinforcement learning to develop chain-of-thought reasoning, self-verification, and error correction capabilities, forming the basis for agentic AI workflows.

Enterprises are adopting a multimodel approach, using multiple specialized models rather than one monolithic system, with large-scale reasoning models for complex planning tasks (5–10% of queries) while routing simpler requests to models with 7–13 billion parameters. This pattern can deliver significant savings on inference costs and the computational expense of processing each query while maintaining quality.

AI models process text in chunks called tokens, roughly equivalent to a word or part of a word. Tokens represent the currency of AI: Pricing for AI services is typically calculated per million tokens processed, making token cost a key consideration for enterprise organizations deploying AI at scale. For example, a customer service implementation might route 80% of simple queries to smaller, more cost-efficient models, reserving resource-intensive reasoning models for complex issues that justify the higher per-token cost.

Emerging trends to consider

Businesses increasingly use multiple AI models in concert. Modern models also now support function calling, allowing them to interact with external tools, application programming interfaces (APIs), and databases, transforming them from text generators to action-takers.

One example of this trend is Llama Stack, Meta's open-source AI runtime environment launched in 2024, which standardises how organisations build and deploy multimodel AI systems. Think of it as Kubernetes for AI agents: just as Kubernetes orchestrates containers, Llama Stack orchestrates agents and their providers, offering common APIs for inference, retrieval-augmented generation (RAG), agents, tools, and safety that work consistently across development and production environments.

Open source: A foundation for AI innovation

Red Hat's AI strategy is deeply rooted in open source, helping enterprises to advance both predictive and 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.

Take control of LLMs with open source

While gen AI is changing nearly every aspect of business, from how software is made to how we communicate, it's not uncommon for the models (LLMs and others) used as part of a gen AI capability to be tightly controlled by the service provider. This means it isn't going to be easy for an enterprise to evaluate the capabilities of a gen AI service without specialized skills and significant investments of both money and time.

Without visibility into the datasets that created the model or an understanding of how the model uses that data, organizations are exposed to potential risks related to AI-generated content. What if your code-generation model is trained on copyrighted source code? Does any code generated by that model now also belong to the owner of that copyrighted code?

Many questions like these have not been fully answered, but understanding the consequences is imperative. Enterprises are turning to AI to ensure they have access to and control over their data, and can understand how it will be handled and used.

Red Hat has always believed in the power of open source to propel innovation, and a transparent approach to software development that gives customers control over the choices they make. That same philosophy now extends to AI. Our approach centers on that transparency and choice, and provides the stability and proactive support enterprise organizations need, no matter which model or models they deploy.

Choosing the right model is hard. With thousands of options available, how do you know which will perform reliably in production?

Chapter 2

Choosing the right AI model

A single application might use several types of AI: a predictive model to forecast demand, a generative model to draft responses, and an image recognition model to process uploads. Each brings different capabilities, costs, and infrastructure requirements.

AI models are mature enough that organizations can use common capabilities out of the box, including text generation, image recognition, voice-to-text, and image segmentation. The availability of these prebuilt functions can significantly accelerate AI development projects.

The key is matching models to your specific use case rather than chasing the latest release. Consider the problem you're solving, the accuracy you need, and the resources you have available.

Foundation models are trained on large sets of data and offer impressive flexibility across many tasks. But that comes with trade-offs: larger resource requirements, higher costs, and increased operational complexity. They're powerful, but not always the right fit.

Small language models (SLMs) offer a compelling middle ground. These models deliver strong performance on targeted tasks while requiring significantly fewer resources than their larger counterparts. Model providers typically offer models with parameter counts ranging from 2 billion to more than 70 billion, allowing you to match model size to task complexity. For many enterprise use cases, an SLM can perform just as well as a large foundation model at a fraction of the cost.

The trend in enterprise AI is clear: The right model is the smallest one that meets your accuracy requirements and has been optimized for your infrastructure.

Training Hub provides an accessible entry point for model customization. This open source library provides algorithm-level abstractions for modern post-training techniques, from supervised fine-tuning to reinforcement learning.

By pulling together proven community implementations, Training Hub lets developers apply advanced training methods without needing deep expertise in each technique. Training Hub can lower the barrier for adding enterprise knowledge to existing models, so your domain experts can contribute without being AI specialists.

Learn more about Training Hub.

Model building versus model tuning

Building an AI model from scratch can be a significant undertaking. It involves gathering and preparing large datasets relevant to your organization's business challenge, choosing an appropriate algorithm, and training it on your data. This process requires considerable computational power and expertise, making it time-consuming and resource-intensive. The model must be maintained over time and updated as the source data evolves. While this provides a custom solution, it might not always be the most efficient path.

Fine-tuning your model

An alternative approach is to tune a foundation model, adapting it to your organization's specific requirements. A common technique is transfer learning, which involves using a model trained on a large dataset and retraining it on a smaller, domain-specific dataset. The model retains the general knowledge it learned during its initial training, while adapting to the nuances of your specific data. This improves the model's ability to understand industry-specific terminology and your organization's brand voice, making responses less generic and more tailored to your business. It also allows you to start small and grow your AI implementation over time.

Starting from either a foundation model or an SLM can accelerate AI adoption. Fine-tuning adapts the model to perform specific tasks exceptionally well using your own data. If your use case is well-defined, such as classifying support tickets or extracting data from invoices, a fine-tuned model can deliver higher accuracy than a general-purpose model while remaining efficient to run. For organizations with specific business requirements, data privacy concerns, or a desire for greater control over model behavior, fine-tuning and hosting your own model can go a long way towards addressing those challenges.

Techniques for customizing model behavior

Fine-tuning isn't the only way to adapt AI models to your needs, and it's not always the right starting point. Several complementary techniques can enhance model performance, often working together to deliver the best results.

RAG connects models to external knowledge sources, typically vector (or graph-vector) databases, that provide relevant context at query time. Rather than relying solely on what a model learned during training, RAG retrieves up-to-date facts and feeds them into the prompt. This is particularly valuable when accuracy matters and information changes frequently—think product documentation, policy details, or customer records.

It is worth noting that RAG and fine-tuning solve different problems. RAG gives models access to current information they weren't trained on. Fine-tuning changes how a model behaves, reasons, or responds. Many production systems use both a fine-tuned model specialized to your domain and RAG to access the latest data.

Retrieval-augmented fine-tuning (RAFT) bridges these approaches. RAFT trains models to work effectively with retrieved documents by fine-tuning them on question-answer pairs that include both relevant and irrelevant context—mimicking real-world RAG scenarios. The result is a model that's better at identifying useful information, ignoring distractors, and reasoning over retrieved content. If your use case relies heavily on RAG, RAFT can significantly improve answer quality by teaching the model how to use retrieved context more effectively.

Agentic AI has emerged as a powerful approach for connecting models to enterprise systems. Rather than a single model answering in isolation, agentic architectures orchestrate multiple AI agents that can

The infrastructure supporting your AI model is just as critical 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 for handling parallel processing tasks, making them ideal for training deep learning models that require processing large amounts of data simultaneously.

Tensor Processing Unit (TPU)

Accelerators specifically for machine learning workloads, optimized for the tensor operations common in neural networks.

Neural Processing Unit (NPU)

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

query databases, call APIs, search internal knowledge bases, and take actions based on results. This moves AI from answering questions to completing tasks, filing tickets, updating records, or coordinating workflows across systems. For enterprises looking to deliver measurable value from AI investments, agentic approaches are rapidly becoming essential.

Prompt engineering and system prompts shape model behavior through carefully crafted instructions. This ranges from providing the AI with 2 to 10 examples (known as "few-shot" learning, as compared to zero-shot or one-shot approaches) to sophisticated system prompts that define persona, constraints, and output format (many-shot). When combined with RAG, the retrieved context becomes part of the prompt, guiding the model toward accurate, grounded responses.

In practice, these techniques layer on top of one another. A well-designed enterprise AI application might:

- Use a fine-tuned small language model for domain understanding and tone.
- Enhance with RAG for access to current data.
- Orchestrate using an agentic framework for multistep tasks.

Use additional guidance from carefully crafted prompts to ensure consistent behavior.

The role of hybrid cloud in AI adoption for enterprise

Hybrid cloud environments play a critical role in AI adoption. A hybrid cloud environment combines on-premise infrastructure with public and private cloud resources, offering flexibility in how and where you deploy and manage AI workloads. You might train models using powerful cloud-based graphics processing units (GPUs) and then deploy them on-premise or in a private cloud for security or compliance reasons. The key consideration is ensuring that the tooling and platform you choose work across environments.

Red Hat's open hybrid cloud approach helps organizations integrate AI to improve consistency, scalability, and flexibility. This allows you to manage AI workloads across multiple clouds, optimize data placement, and implement smooth migration, making it more efficient to adopt AI at scale.

Sovereign AI: Control in an uncertain landscape

As AI is woven into the fabric of enterprise operations, control and visibility over AI systems becomes business-critical. Sovereign AI addresses this by keeping data, models, and inference within your control, whether that means on-premise, in a specific region, or within a trusted cloud environment.

Several factors are promoting enterprise interest in sovereign AI:

Regulatory compliance: Key provisions of the EU Artificial Intelligence Act came into effect in 2025, joining Europe's General Data Protection Regulation (GDPR), the US Department of Health and Human Services' Health Insurance Portability and Accountability Act (HIPAA), and sector-specific regulations that dictate how AI models must be governed, audited, and documented. Running models within controlled environments simplifies compliance.

Data residency: Sensitive data must remain within particular jurisdictions. Sovereign deployment ensures your data never leaves approved boundaries.

Cost predictability: Cloud-based AI services can change pricing or deprecate models with little notice, forcing expensive re-evaluation of connected applications. Self-hosted models eliminate per-token API charges and protect against vendor pricing shifts.

Vendor independence: When a cloud provider replaces or retires a model, every application built on it must be reworked. Sovereign deployment of open source models removes this dependency.

Performance: Edge and on-premise deployments deliver low-latency inference for real-time applications without round-trip calls to external APIs.

Why open source is a strategic advantage for AI

Sovereign AI considerations and concerns are pushing open source technology to the forefront of enterprise strategy. Open source methods provide transparency into training data and model behavior, portability across infrastructure, and freedom from single-vendor lock-in. When AI underpins critical business processes, the ability to inspect, modify, and control your models isn't optional—it's essential.

Small language models (SLMs) make sovereign AI practical. Models like IBM Granite, Mistral, and Meta's Llama deliver strong capabilities in compact, efficient forms that run on enterprise infrastructure. "Gartner® predicts that small, task-specific AI models will be used at least three times more than general-purpose LLMs by 2027."[2]

By understanding how models, data, infrastructure, and sovereignty intersect, you can navigate the complexities of AI adoption while maintaining the control your business requires.

Chapter 3

What you need to get started

As with any new technology, AI adoption comes with challenges. This chapter helps you assess whether your organization is ready for AI and provides a practical roadmap for getting started.

Assessing your organization's readiness

Before diving into implementation, consider whether your organization is positioned to support AI initiatives. This initial review isn't about having everything in place; it's about understanding where you are, what challenges you need to overcome, and what gaps need attention.

Strategic alignment: Do your AI ambitions support broader business goals? AI projects that lack a clear connection to strategic objectives struggle to protect sustained investment and executive sponsorship.

Infrastructure capacity: Can your current environment support AI workloads? This includes computing resources, storage, network capabilities, and the flexibility to scale as requirements grow.

Skills and expertise: What AI capabilities exist within your organization? Understanding your current skills helps identify where training, hiring, or external support may be needed.

2 Gartner Press Release. "Gartner Predicts by 2027, Organizations Will Use Small, Task-Specific AI Models Three Times More Than General-Purpose Large Language Models." 9 Apr 2025.

AI adoption depends on collaboration

Building a team with a range of capabilities is key for successful gen AI projects.[3]

Business leaders represent the people who use or are affected by the solution.

AI specialists tune, maintain, and update gen AI models, optimizing performance for specific use cases.

AI and data engineers prepare training data, build RAG pipelines, implement fine-tuning workflows, and integrate models into production systems.

Ethics and compliance officers ensure that gen AI initiatives comply with regulations.

IT operations specialists integrate solutions with existing infrastructure, enforce security policies, and deliver AI playgrounds and Models-as-a-Service capabilities that enable rapid learning and testing across the organization.

Data maturity: What data does your organization hold, and how accessible is it? Enterprise data is often fragmented across systems, formats, and teams. A realistic view of your data landscape helps identify which AI opportunities are feasible.

A practical roadmap for AI adoption

The speed and scale of AI adoption depend on many factors, but starting small and growing incrementally remains a proven approach. The following steps provide a practical path from initial concept through to production deployment.

Step 1: Define the problem and success criteria

Start with a specific business problem, not a technology. The best starting points are use cases with clear outcomes, available data, and engaged stakeholders. Define success criteria in advance: accuracy thresholds, response quality, latency requirements, or business metrics. Without clear criteria, you won't know whether your implementation is working. Avoid the temptation to pursue AI for its own sake; focus on problems where AI offers a genuine advantage over existing approaches.

Step 2: Assemble your team

AI adoption depends on collaboration across disciplines. Form a cross-functional team early, bringing together:

- Business stakeholders who understand the problem and will use the solution.
- Data engineers who can prepare data and build pipelines.
- AI/ML practitioners who select, customize, and evaluate models.
- IT operations teams that integrate solutions with existing infrastructure and enforce security policies.
- Ethics and compliance specialists who ensure initiatives meet regulatory requirements.
- Development teams and communities should be involved from the start. A cross-functional approach ensures AI investments deliver measurable outcomes.

Domain expertise is particularly valuable. People who understand your business processes well can often identify the highest-impact opportunities and catch issues that pure technologists might miss.

Step 3: Evaluate your data

With a defined use case, assess whether you have the data to support it. Evaluate completeness, accuracy, and relevance for the specific problem you are solving. Identify where data lives, who owns it, and what is required to access it. If critical data doesn't exist or can't be obtained, revisit the use case before investing further.

Step 4: Select your model

Choose AI models based on your identified use case and success criteria. Consider capability, cost, scalability, and compatibility with your infrastructure. Multimodel architectures are now standard practice; different models excel at different tasks, and production applications often combine several.

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

What is inference?

AI inference is the last step in the machine learning process. It is when connections are made between what the model has learned from training and the patterns it sees in the real world. This is how AI applications draw their conclusions—whether it is being used to provide customer-service support, predict car crashes, or identify anomalies in medical data.

Explore AI use cases.

Providing an AI playground where developers and stakeholders can explore and experiment with different models can significantly accelerate model selection. Red Hat® AI includes a gen AI studio for precisely this purpose, allowing rapid prototyping and evaluation before committing to production.

As mentioned earlier, the right model is often the smallest one that meets your accuracy requirements. A smaller model fine-tuned on your specific data, such as customer service transcripts or product documentation, can deliver faster response times, lower costs, and higher accuracy for your use case than a larger general-purpose model.

Step 5: Connect models to your data

Much enterprise knowledge lives in documents scattered across the organization: PDFs, wikis, support tickets, and internal documentation. Connecting models to this knowledge is often a more efficient path to value.

Retrieval-augmented generation (RAG) lets models access your data at query time without retraining. The key is effective document ingestion: Converting unstructured content into formats that models can use. Docling, an open source framework included in Red Hat AI, handles this complexity by extracting text, tables, and structure from PDFs and other documents.

Start with RAG to ground models in your enterprise knowledge in less time. As you learn which information matters most, you can selectively fine-tune models for even better performance.

Step 6: Customize if needed

When RAG alone isn't sufficient, fine-tuning adapts models to your specific domain and requirements. This involves training the model on your data to improve accuracy, adjust tone, or teach specialized knowledge.

Red Hat AI offers training capabilities that simplify this process, providing accessible workflows for model customization. Synthetic data generation enables models to generate training examples from your documents and guidelines. By incorporating synthetic data generation into the workflow, you can build high-quality training datasets even when real-world examples are limited or sensitive.

Consider RAFT (retrieval-augmented fine-tuning) if your application relies heavily on RAG. RAFT trains models to work more effectively with retrieved content, improving their ability to identify relevant information and reason over documents.

Step 7: Optimize for production

Moving from proof-of-concept to production introduces new challenges around cost, latency, and scale. Optimization ensures your models perform efficiently under real-world conditions, where every millisecond of latency and every dollar of compute cost matters.

Model quantization reduces size and accelerates inference by using lower-precision numerical formats. By representing model weights with fewer bits (for example, 8-bit or 4-bit integers instead of 16-bit floating point), quantized models require less memory and process in less time. Red Hat's benchmarks of over half a million evaluations found that 8-bit quantization delivers approximately 1.8x performance speedup with full accuracy recovery, while 4-bit quantization achieves 2.4x speedup for latency-sensitive applications.

Inference servers matter as much as the models themselves. Purpose-built serving infrastructure, such as Red Hat AI Inference Server (based on vLLM), maximizes throughput through techniques such as continuous batching, paged attention, and optimized GPU use. For workloads that exceed a single server's capacity, distributed inference spreads computation across multiple machines, supporting larger models and higher throughput.

Red Hat AI offers a growing collection of validated, compressed versions of the most popular open source models, along with LLM Compressor tools for quantizing and optimizing your own models. Our validated models are benchmarked across hundreds of scenarios, so you can deploy with confidence in both performance and accuracy.

As we mentioned in Step 4: The right model is the smallest one that meets your accuracy requirements. A well-tuned, quantized model will almost always outperform an oversized model in production economics.

Step 8: Deploy with guardrails and monitoring

Production AI requires ongoing oversight. Deploy models with appropriate guardrails: content filters, output validation, and safety boundaries that reflect your policies and risk tolerance.

Implement drift monitoring to track model behavior over time, including changes in accuracy, response quality, and adherence to safety guidelines. Models can degrade as the world changes around them; monitoring catches this before users do.

Step 9: Iterate and scale

AI adoption is rarely a single project. Use learnings from initial deployments to refine your approach, expand to new use cases, and build internal capabilities. Each iteration builds organizational confidence and expertise.

If your in-house AI expertise is still developing, external guidance can accelerate progress. Red Hat Consulting brings experience from hundreds of AI implementations, helping you navigate architecture decisions, avoid common pitfalls, and build internal capabilities.

Chapter 4

Adopt and scale with Red Hat

The speed of AI development means that AI models that were state-of-the-art 6 months ago are routinely surpassed, and new techniques emerge constantly. This rapid, iterative development process, in many ways, mirrors the development model Red Hat was built on.

Red Hat is an enterprise organization, with an open source development model: a modular, flexible foundation that evolves with the community. By building on innovative open source AI projects—rather than building novel code behind closed doors—Red Hat AI stays current with the latest innovations while providing the enterprise support, transparency, and stability your organization requires.

Red Hat AI delivers a unified platform for enterprise AI deployment, expanding from traditional infrastructure to address AI-specific challenges: fragmented tools, unpredictable costs, and workload portability. Red Hat AI provides a portfolio of products and services needed to build, deploy, monitor, and use AI models and AI-powered applications, regardless of where you are on your AI adoption journey.

Delivering business outcomes with Red Hat AI

Red Hat AI delivers trust, choice, and consistency across hybrid cloud environments, accelerating the development and delivery of AI solutions and addressing the challenges AI adopters face. Choosing Red Hat AI allows organizations to:

- **Optimize with efficient, cost-effective inference**, and do more with the resources you have. Red Hat AI optimizes model inference across hybrid cloud environments, delivering faster deployments at a lower cost. With vLLM at its core, the platform maximizes throughput and minimizes latency. Combined with LLM Compressor tools and a growing collection of validated and compressed models, organizations can reduce compute costs while maintaining accuracy.

- **Connect models to your data**, so that AI intelligent applications understand your organization's domain and deliver accurate, relevant responses. Gen AI models are only as useful as their understanding of your business. Red Hat AI provides a simplified, consistent experience for integrating your private data with AI models through RAG, fine-tuning with Training Hub, and built-in data ingestion capabilities.

- **Accelerate agentic AI innovation**, and bring experiments to production with confidence. Red Hat AI provides a flexible, scalable foundation for building and deploying agentic AI workflows with Llama Stack API as a unified entry point, dedicated dashboard experiences, and Model Context Protocol (MCP) support for connecting models to external tools and data sources.

- **Deploy AI across hybrid cloud environments**, and scale efficiently while maintaining control. Deploy AI workloads wherever your business needs them: on-premise, in cloud environments, or at the edge of your network. Red Hat AI provides the flexibility to choose a model and a hardware accelerator, with consistent tooling and governance. It also offers centralized model registries, enhanced observability, and intelligent GPU management.

- **An AI platform, built on open source and backed by enterprise support**, that offers innovation, choice, and confidence. The platform includes a fully indemnified IBM Granite family of open source models built specifically for enterprise use, with 24x7 production support and extended lifecycle coverage.

Red Hat's extensive partner ecosystem further strengthens your AI capabilities. NVIDIA continues to partner with Red Hat to provide a complete enterprise platform optimized for AI workloads. Red Hat AI extends accelerator support beyond NVIDIA to include AMD GPUs, Intel Gaudi, IBM Spyre, and Google Tensor Processing Units. This gives customers the flexibility to choose the hardware that best fits their workloads, budget, and strategic direction.

Red Hat Consulting can help accelerate your journey, whether you're pursuing a pilot project or building a foundation for rapid adoption.

Red Hat's complete technology portfolio, proven expertise, and strategic partnerships can help you develop and deploy AI models and applications, as well as the services and training you need to achieve your AI goals.

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

Accelerate your AI adoption with an 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 from Red Hat

Learn more about Red Hat AI

Schedule a complimentary discovery session

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

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

redhat.com
#3235271_0326