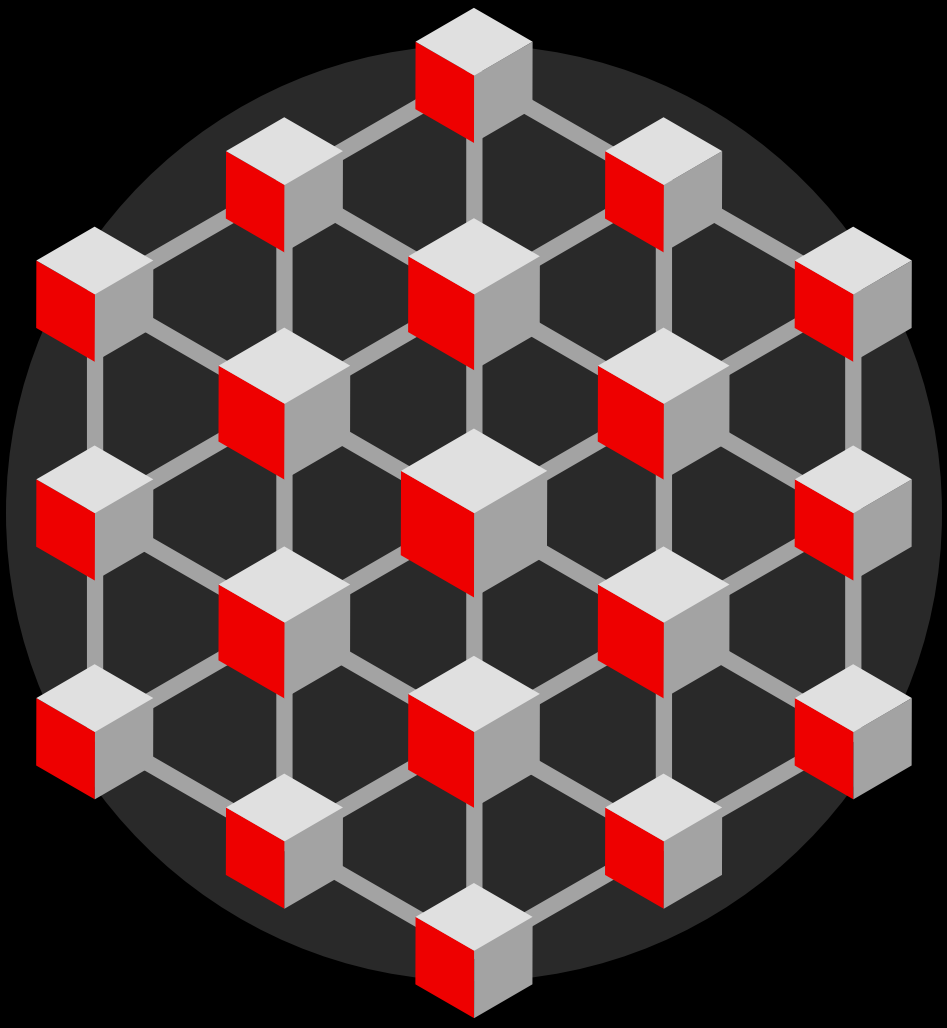
Maximize AI innovation with open source models



Red Hat

Contents

Simplify generative AI across your enterprise

Across industries, generative artificial intelligence (AI) continues to transform how organizations manage operations, make decisions, and deliver innovative solutions.

From personalized customer experiences and automated content creation to enhanced decision support, generative AI helps businesses streamline processes, increase productivity, and explore new possibilities. By investing in innovative AI solutions, enterprise organizations can strengthen their ability to navigate complexity, respond to shifting demands, and achieve strategic goals.

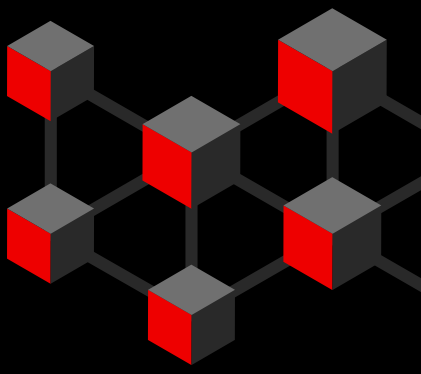
However, for many enterprises, realizing the full potential of generative AI technologies presents challenges. Proprietary large language models (LLMs) come with high license costs and require extensive hardware resources for both training and inferencing, making them expensive to deploy and maintain within enterprise IT environments. Training and tuning LLMs for specific customer and industry use cases can be difficult when developers and organizational subject matter experts—who fully understand the application and business requirements—lack the necessary specialized data science expertise. And proprietary, confidential data stored across complex hybrid cloud environments—including private datacenters, public cloud resources, and edge deployments—can make it hard to train generative AI models effectively, especially when regulatory requirements limit data movement between locations.

of organizations believe that generative AI will disrupt their business within the next 18 months.[1]

Red Hat offers a simplified approach that makes generative AI accessible to developers and domain experts across enterprise organizations. With Red Hat® AI, you can efficiently train and run generative AI models on production server deployments, so your teams can progress rapidly along their AI journeys without unnecessary complexity.

1 IDC eBook. "Are You Ready for AI Everywhere?" 2024.

Speed innovation with task-specific small language models



LLMs offer natural language understanding and generation capabilities. However, substantial computational costs for resource-intensive training and inference can make them impractical for many business applications. Fortunately for most enterprise organizations, these broad capabilities are unnecessary, as their needs are better served by more focused, task-specific **small language models (SLMs)**.

With a smaller size that requires significantly less compute resources, data, and energy, SLMs are efficient, cost-effective generative AI models for many enterprise applications. They offer faster model training, simplified fine-tuning, and a more sustainable approach to AI development. By streamlining deployment and improving the performance of AI-based applications, SLMs help reduce costs while delivering tailored, innovative solutions to meet your business goals.

SLMs offer several advantages for enterprise organizations. With a compact size and reduced data requirements, you can rapidly and efficiently customize SLMs for specific tasks or domains, resulting in more accurate and cost-effective generative AI models. By embedding organizational knowledge and expertise directly into SLM parameters during training, you can enhance model relevance, reduce retraining needs, and shorten the development timeline of your critical AI-based applications and services. Because you can efficiently deploy them in on-site enterprise IT or private cloud environments, SLMs make it easier to protect sensitive, proprietary training data while maintaining control over access, simplifying regulatory compliance, and reducing reliance on external providers. And in resource-constrained environments and edge deployments, SLMs allow real-time applications to run directly on user devices, streamlining development and eliminating the need for external cloud infrastructure.

Watch this video to learn more about the advantages of small language models.

Gain control with open source generative AI solutions

The use of commercial AI solutions often raises concerns about visibility, security, privacy, and safety, while uncertainty around the training data used and response accuracy can increase an organization's legal risk. In response to these challenges, LLMs are increasingly becoming open source, providing freedom, choice, and transparency that helps you better understand, customize, and trust the models you use.

Including both open source licensed software components and model weights—pretrained by a trusted provider—open source LLMs and SLMs let you contribute data and expertise to base models and training data sets in a transparent and reliable way. They also help avoid vendor lock-in, allowing you to retain full control over your AI solutions without being tied to proprietary systems. Because trusted providers disclose the origin of the data used to train open source models, you can evaluate model quality and ensure that no harmful or biased data is involved before adding your own proprietary and confidential data. Additionally, open governance practices help identify and mitigate biases during model development, fostering trust and fairness in AI systems.

Finally, with open source LLMs distributed under the Apache 2.0 license, you gain the flexibility to customize and control your models, ensuring you can adapt your innovative AI solutions to meet business requirements.

The power of open source in AI

Open source approaches are becoming an important factor in creating trust in generative AI technologies. Learn more about the impact of open source in AI:

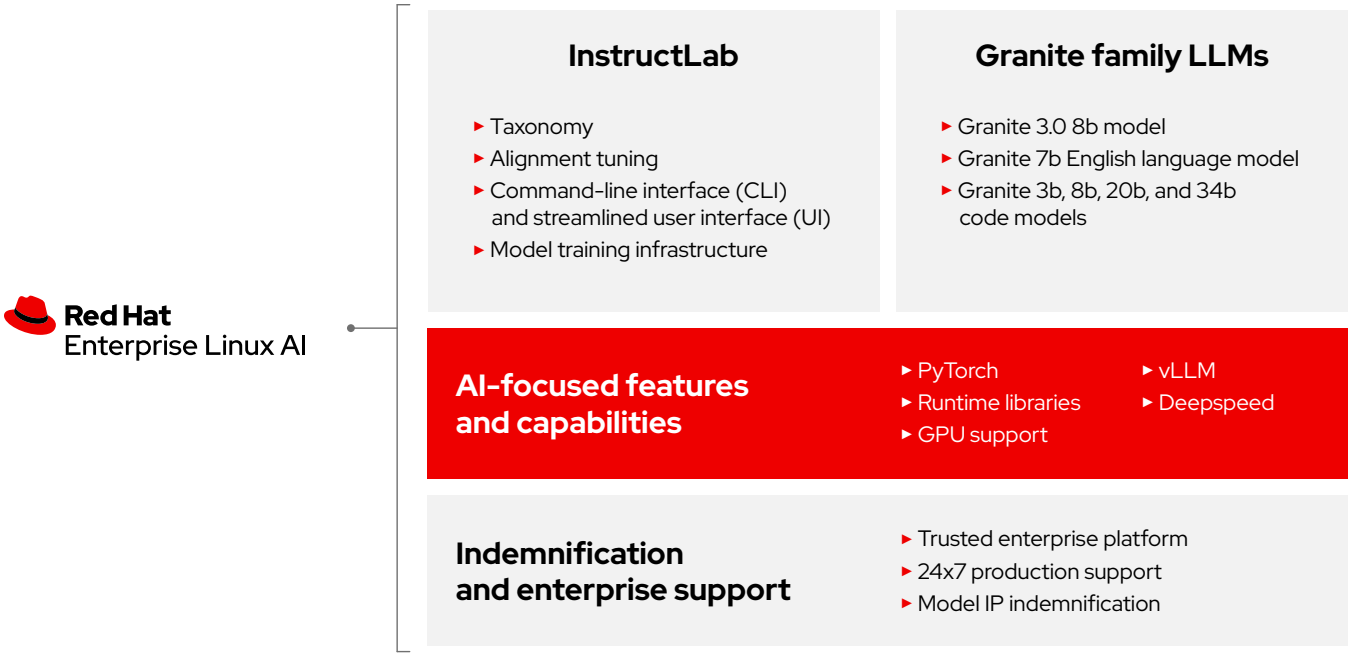
- **Read this blog** to see how open source communities are advancing AI technologies.
- **Read this blog** to understand why more organizations are gaining trust and confidence in open source AI technologies.
- **Watch this video** to find out how you can contribute to open source AI initiatives.

Streamline AI development with Red Hat Enterprise Linux AI

Delivering AI solutions that are both cost-effective and tailored to business needs can be a struggle for many organizations. Red Hat Enterprise Linux® AI offers a consistent, stable foundation model platform that speeds and simplifies development and deployment of innovative, enterprise-ready generative AI solutions.

With advanced features and capabilities for training, testing, and running task-specific SLMs, Red Hat Enterprise Linux AI empowers you to build cost-effective customized AI applications and services. A key platform component, the **Granite family** of open source licensed generative AI models—distributed under the Apache 2.0 license with transparency into training data sets—helps you get started with generative AI in less time using smaller, efficient models that reduce operational costs without compromising performance. Also included, **InstructLab** model alignment tools simplify fine-tuning processes, making AI accessible to developers and domain experts by aligning models with organizational data and broadening access to community-developed generative AI models.

Delivered as a **bootable image** with popular AI libraries like PyTorch and hardware-optimized accelerators from NVIDIA, Intel, and AMD, Red Hat Enterprise Linux AI simplifies integration of crucial AI technologies while streamlining model training and inference on production servers so that you can begin your generative AI projects efficiently. 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.



Adopt open source Granite models for enterprise AI

Successful AI applications and services require models that can balance transparency, costs, and performance while meeting the demands of enterprise applications. Developed by IBM, the Granite family of open source generative AI models addresses the needs of enterprise applications and real business scenarios. Optimized for both cost and performance, these models support a wide range of generative AI use cases involving both language and code. The open source Granite model series includes the Granite 3.0 8b model, Granite 7b English language models, and the Granite 3b, 8b, 20b, and 34b code models.

Watch this video to learn more about the open source Granite model series.

Granite models are distributed under the Apache License 2.0, providing full open source accessibility. These foundation AI models are built independently and are not derived from proprietary or non-open source models. Additionally, IBM discloses the data sets and methodologies used for training, offering enhanced transparency and confidence for your critical AI solutions. Read the **Granite 3.0 language models technical paper** for detailed insights into training data and practices that show IBM's commitment to openness and accountability.

Fully indemnified by Red Hat, the open source Granite language and code models offer an enterprise-ready solution for developing and deploying tailored AI applications. By combining IBM's innovative models with the reliability and support of Red Hat platforms, you can confidently adopt open source generative AI solutions to address business challenges while mitigating risk across your organization.

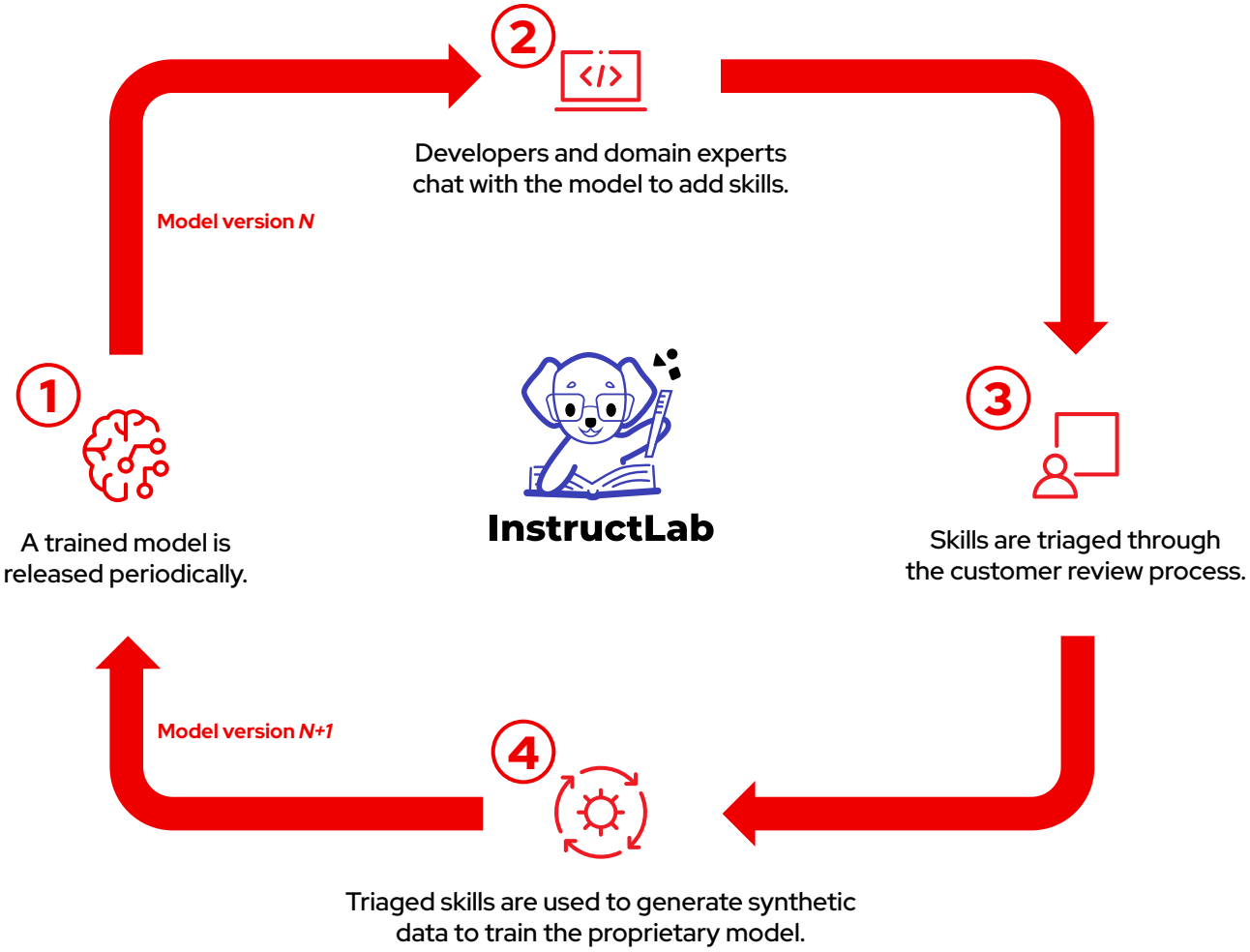
Democratize AI development with InstructLab

Aligning generative AI models to your business requirements can be a difficult, time-consuming task, especially when expertise and resources are limited. InstructLab model alignment tools let you customize generative AI models with specialized knowledge and skills. Based on **Large-Scale Alignment for ChatBots (LAB)** technology, InstructLab uses a taxonomy-driven approach to generate and train models with high-quality synthetic data. As a result, you can enhance AI models using much less human-generated information and fewer computing resources than typically required to retrain a model.

InstructLab simplifies training and fine-tuning, allowing a wider range of contributors—including developers and domain experts without extensive data science experience—to participate in generative AI model development. By empowering subject matter experts to participate directly in building innovative AI solutions, InstructLab fosters collaboration between technical teams and business stakeholders, ensuring AI applications and services are aligned with actual needs. This democratization of AI model alignment speeds time to value, helping your entire organization refine models to meet business goals more effectively. Accordingly, you can align generative AI models with your operational and strategic objectives and improve the accuracy and relevance of critical AI solutions. Through efficient and cost-effective integration of enterprise business knowledge, InstructLab helps you achieve faster results and a higher return on your AI investments.

Watch this video to learn about training generative AI models with InstructLab.

InstructLab supports an iterative approach to fine-tuning generative AI models. It encourages collaboration by allowing team members to contribute their expertise, while subject matter experts validate and approve input. This process helps ensure that your models remain transparent, relevant, and unbiased as they evolve.



Accelerate your AI journey with Red Hat

Deploy a trusted foundation for AI innovation.

No matter where you are in your AI journey, Red Hat Enterprise Linux AI provides the consistency, reliability, and innovation needed for success. With a trusted open source operating foundation and integrated AI technologies, you can get started on your AI projects more easily and move your ideas from concept to production in less time.

Learn more about Red Hat Enterprise Linux AI.

Try Red Hat Enterprise Linux AI free for 60 days

Access a no-cost, 60-day, self-supported subscription and resources to see how your organization can develop, train, test, and run Granite family generative AI models for enterprise applications.

Try it for free

Get started faster with expert guidance

Work with our experts to plan and build AI solutions that deliver real business results. Whether you're just starting with generative AI or want to expand your AI capabilities, Red Hat Services can help.

Learn more

0125_KVM

Red Hat