Red Hat

Foundation Model Platform for Generative AI

Red Hat Enterprise Linux AI

Foundation model platform for generative AI

Key benefits

- Let users update and enhance large language models (LLMs) with InstructLab

- Align LLMs with proprietary data, safely and securely, to tailor the models to your business requirements

- Get started quickly with generative artificial intelligence (gen AI) and deliver results with a trusted, security-focused Red Hat® Enterprise Linux® platform

- Packaged as a bootable Red Hat Enterprise Linux container image for installation and updates

Product overview

Red Hat Enterprise Linux AI (RHEL AI) is an enterprise-grade gen AI foundation model platform to develop, test, and deploy LLMs for gen AI business use cases.

RHEL AI brings together:

- The Granite family of open source LLMs.

- InstructLab model alignment tooling, which provides a community-driven approach to LLM fine-tuning.

- A bootable image of Red Hat Enterprise Linux, along with gen AI libraries and dependencies such as PyTorch and AI accelerator driver software for NVIDIA, Intel, and AMD.

- Enterprise-level technical support and model intellectual property indemnification provided by Red Hat.

- RHEL AI gives you the trusted Red Hat Enterprise Linux platform and adds the necessary components for you to begin your gen AI journey and see results.

The future of AI is open and transparent

RHEL AI includes a subset of the open source Granite language and code models that are fully indemnified by Red Hat. The open source Granite models provide organizations cost- and performance-optimized models that align with a wide variety of gen AI use cases. The Granite models were released under Apache 2.0 license. In addition to the models being open source, the datasets used for training the models are also transparent and open.

Table 1. Granite models in RHEL AI

IBM Granite Language models	Granite-7B-Starter
	Granite-7B-RedHat-Lab
IBM Code models	Granite-8B-Code-Instruct
	Granite-8B-Code-Base

Accessible gen AI model training for faster time to value

MIn addition to open source Granite models, RHEL AI also includes InstructLab model alignment tooling, based on the Large scale Alignment for chatBots (LAB) technique. InstructLab allows teams within organizations to efficiently contribute skills and knowledge to LLMs, customizing these models for the specific needs of their business.

f facebook.com/redhatinc
X @RedHat
in linkedin.com/company/red-hat

redhat.com   Datasheet   Foundation model platform for generative AI

Red Hat Enterprise Linux AI

Foundation model platform for generative AI

Key benefits

▸ Let users update and enhance large language models (LLMs) with InstructLab

▸ Align LLMs with proprietary data, safely and securely, to tailor the models to your business requirements

▸ Get started quickly with generative artificial intelligence (gen AI) and deliver results with a trusted, security-focused Red Hat® Enterprise Linux® platform

▸ Packaged as a bootable Red Hat Enterprise Linux container image for installation and updates

Product overview

Red Hat Enterprise Linux AI (RHEL AI) is an enterprise-grade gen AI foundation model platform to develop, test, and deploy LLMs for gen AI business use cases.

RHEL AI brings together:

▸ The Granite family of open source LLMs.

▸ InstructLab model alignment tooling, which provides a community-driven approach to LLM fine-tuning.

▸ A bootable image of Red Hat Enterprise Linux, along with gen AI libraries and dependencies such as PyTorch and AI accelerator driver software for NVIDIA, Intel, and AMD.

▸ Enterprise-level technical support and model intellectual property indemnification provided by Red Hat.

▸ RHEL AI gives you the trusted Red Hat Enterprise Linux platform and adds the necessary components for you to begin your gen AI journey and see results.

The future of AI is open and transparent

RHEL AI includes a subset of the open source Granite language and code models that are fully indemnified by Red Hat. The open source Granite models provide organizations cost- and performance-optimized models that align with a wide variety of gen AI use cases. The Granite models were released under Apache 2.0 license. In addition to the models being open source, the datasets used for training the models are also transparent and open.
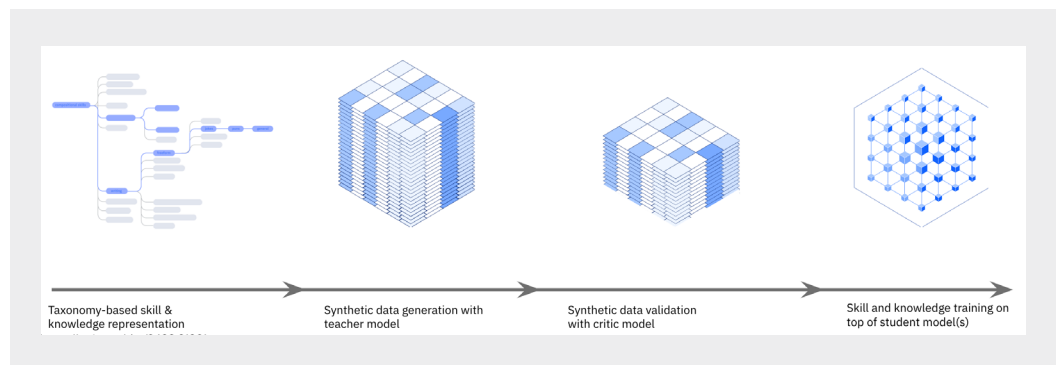
Table 1. Granite models in RHEL AI

IBM Granite Language models	Granite-7B-Starter
	Granite-7B-RedHat-Lab
IBM Code models	Granite-8B-Code-Instruct
	Granite-8B-Code-Base

Accessible gen AI model training for faster time to value

MIn addition to open source Granite models, RHEL AI also includes InstructLab model alignment tooling, based on the Large scale Alignment for chatBots (LAB) technique. InstructLab allows teams within organizations to efficiently contribute skills and knowledge to LLMs, customizing these models for the specific needs of their business.

▸ Skill: A capability domain intended to teach a model how to do something. Skills are classified into two categories.

   ▸ Compositional skills

      ▸ Let AI models perform specific tasks or functions.

      ▸ Are either grounded (includes context) or ungrounded (does not include context).

         ▸ A grounded example is adding a skill to provide a model the ability to read a markdown-formatted table.

         ▸ An ungrounded example is adding a skill to teach the model how to rhyme.

   ▸ Foundation skills

      ▸ Skills like math, reasoning, and coding.

▸ Knowledge: Data and facts that provide a model with additional data and information to answer questions with greater accuracy.

Taxonomy-based skill & knowledge representation   Synthetic data generation with teacher model   Synthetic data validation with critic model   Skill and knowledge training on top of student model(s)

The above image outlines the InstructLab model fine-tuning workflow.

1. Skills and knowledge contributions are placed into a taxonomy-based data repository.

2. A significant quantity of synthetic data is generated, using the taxonomy data, in order to produce a large enough dataset to successfully update and change an LLM.

3. The synthetic data output is reviewed, validated, and pruned by a critic model.

4. The model is trained with synthetic data rooted in human-generated manual input.

InstructLab is accessible to developers and domain experts who may lack the necessary data science expertise normally required to fine-tune LLMs. The InstructLab methodology allows teams to add data, or skills especially suited to business use case requirements, to their chosen model for training in a collaborative manner allowing for quicker time to value.

Train and deploy anywhere

RHEL AI helps organizations accelerate the process of going from proof of concept to production server-based deployments by providing all the tools needed and the ability to train, tune, and deploy these models where the data lives, anywhere across the hybrid cloud. The deployed models can then be used by various services and applications within your company.
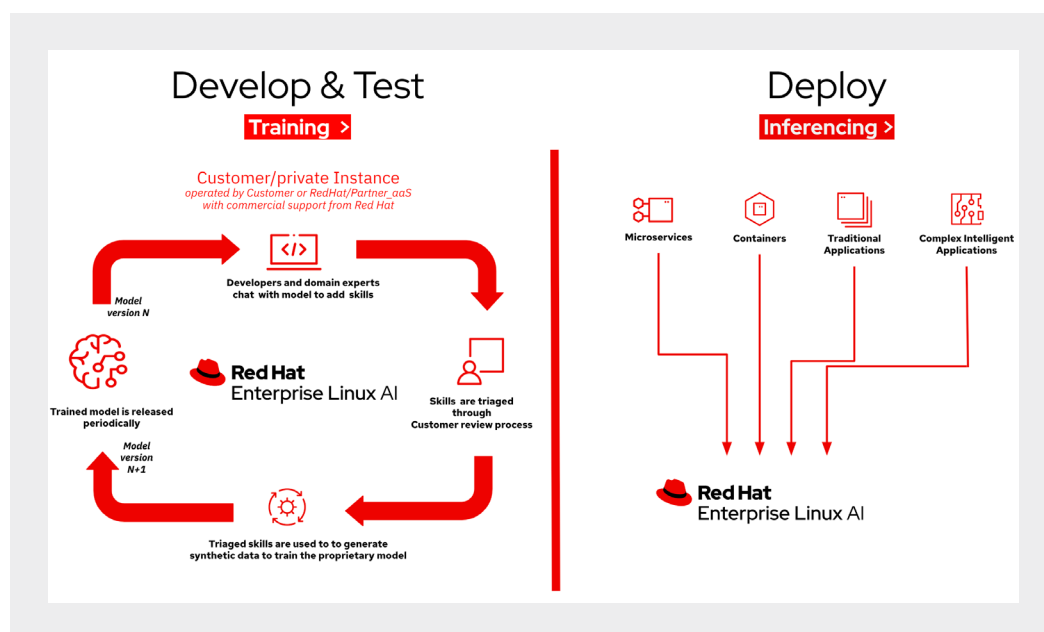
When organizations are ready, RHEL AI also provides an on-ramp to Red Hat OpenShift® AI, for training, tuning, and serving these models at scale across a distributed cluster environment using the same Granite models and InstructLab approach used in the RHEL AI deployment.

Features and benefits

Features	Benefits
Fully supported Granite language and code models, open sourced under the Apache 2.0 license	Open source and transparent LLMs, along with openly accessible training data, enhance data transparency and address ethical concerns about data content and sources, ultimately reducing overall business risk.

Red Hat Enterprise Linux AI

Features	Benefits
Model IP indemnification for Granite models	Indemnification for the Granite models within RHEL AI reflects the strong confidence Red Hat and IBM have in the rigorous development and testing of these models. This indemnification provides customers with enhanced assurance, empowering them to use the Granite models with greater trust and confidence in Red Hat's commitment to their success.
InstructLab LLM alignment tooling for scalable and accessible model fine-tuning	InstructLab provides an accessible method to fine-tune LLMs, lessening the need for deep data science expertise and enabling various roles within your organization to contribute. This allows your business to adopt gen AI, accelerating your time to value and maximizing your return on investment.
Optimized, bootable model runtime instances	RHEL AI is delivered as a bootable container image, a deployment method called image mode for Red Hat Enterprise Linux. This technology reduces installation, configuration, and update complexity, allowing for a simple setup and change management process.
Gen AI package dependencies and software drivers for AI hardware	Begin gen AI right away with a comprehensive set of tools, including essential packages and drivers like PyTorch, vLLM, and NVIDIA drivers, ensuring you're equipped to tackle your gen AI business use cases from day one.

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.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1 888 REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com

redhat.com
#1384734_0924

Red Hat

Thank you for downloading this Red Hat resource! Carahsoft is the Master GSA and SLSA Dealer and Distributor for Red Hat Enterprise Open Source solutions available via GSA, SLSA, ITES-SW2, The Quilt and other contract vehicles.

To learn how to take the next step toward acquiring Red Hat's solutions, please check out the following resources and information:

For additional resources:
carah.io/RedHatResources

For upcoming events:
carah.io/RedHatEvents

For additional Red Hat solutions:
carah.io/RedHatSolutions

For additional Open Source solutions:
carah.io/OpenSource

To set up a meeting:
redhat@carahsoft.com
877-RHAT-GOV

To purchase, check out the contract vehicles available for procurement:
carah.io/RedHatContracts