RHEL AI Foundation Model Platform for Generative AI

Solution Brief

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.

Red Hat Enterprise Linux AI allows portability across hybrid cloud environments, and makes it possible to then scale your AI workflows with Red Hat OpenShift® AI and to advance to IBM watsonx.ai with additional capabilities for enterprise AI development, data management, and model governance.

Key Benefits

Red Hat Enterprise Linux AI is a powerful tool that enables enterprises to harness the potential of artificial intelligence. It offers a comprehensive platform for developing, testing, and running generative AI foundation models, making it accessible to organizations at all stages of their AI journey. By leveraging open-source technologies and a community-driven approach, RHEL AI empowers businesses to innovate with trust and transparency, while reducing costs and removing barriers to entry.

One of its key advantages is its ability to facilitate collaboration between domain experts and data scientists, enabling the creation of purpose-built generative AI models tailored to specific business needs. RHEL AI's flexibility and scalability allow organizations to deploy and manage these models across hybrid cloud environments, ensuring seamless integration with existing infrastructure.

Furthermore, RHEL AI's focus on security and compliance provides a robust foundation for building and deploying AI solutions responsibly. By adhering to open-source principles and industry best practices, RHEL AI helps organizations mitigate risks and protect sensitive data.

Overall, RHEL AI offers a compelling solution for enterprises seeking to embrace the power of AI. Its ease of use, scalability, and focus on security make it a valuable tool for driving innovation and achieving business objectives.

Below is a summary of the key benefits of RHEL AI:

- 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

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

Part number	Feature code
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

In 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.

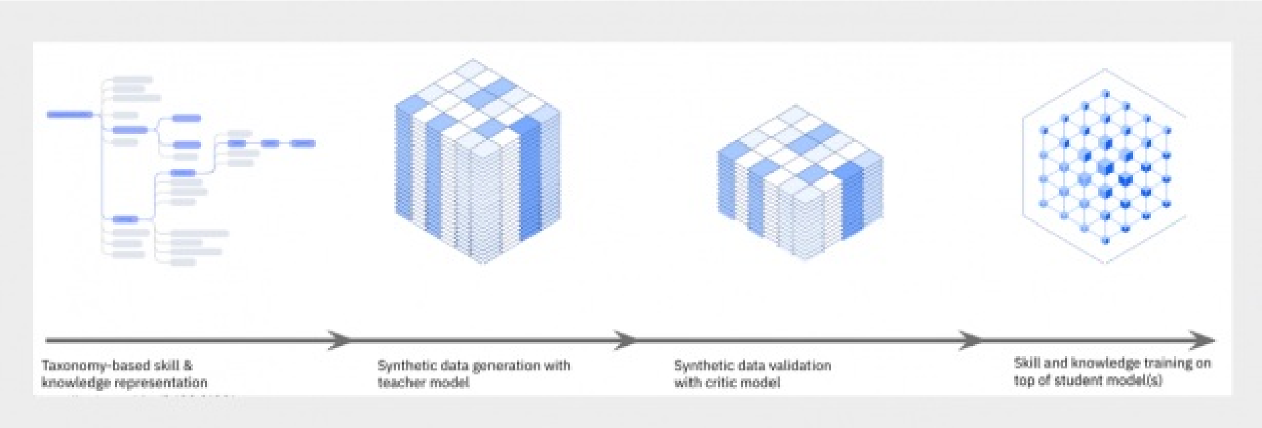
Taxonomy-based skill & knowledge representation
Synthetic data generation with teacher model
Synthetic data validation with critic model
Skill and knowledge training on top of student model(s)

Figure 1. InstructLab model fine-tuning workflow

InstructLab is an open source project for enhancing large language models (LLMs) used in generative artificial intelligence (gen AI) applications. Created by IBM and Red Hat, the InstructLab community project provides a cost-effective solution for improving the alignment of LLMs and opens the doors for those with minimal machine learning experience, skills and knowledge to contribute.

- Skills and knowledge contributions are placed into a taxonomy-based data repository.
- 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.
- The synthetic data output is reviewed, validated, and pruned by a critic model.
- 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.

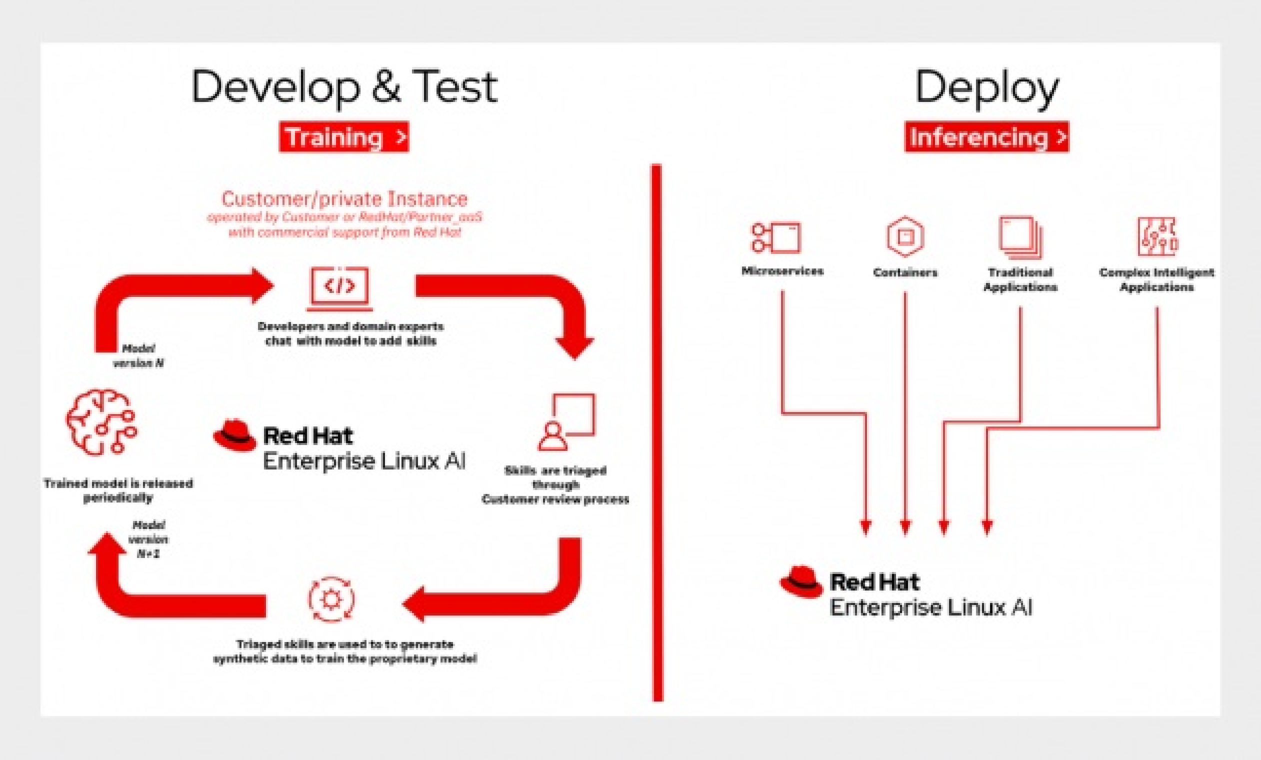
Develop & Test
Training >
Customer/private Instance
operated by Customer or RedHat/Partner_aaS
with commercial support from Red Hat
Developers and domain experts chat with model to add skills
Model version N
Red Hat Enterprise Linux AI
Trained model is released periodically
Skills are triaged through Customer review process
Model version N+1
Triaged skills are used to to generate synthetic data to train the proprietary model
Deploy
Inferencing >
Microservices
Containers
Traditional Applications
Complex Intelligent Applications
Red Hat Enterprise Linux AI

Figure 2. InstructLab approach used in the RHEL AI deployment

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.

Table 2. 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.
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.

Related product families

Product families related to this document are the following:

- Red Hat Alliance
- ThinkSystem SR675 V3 Server

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2111, was created or updated on December 11, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2111
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2111.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®

The following terms are trademarks of other companies:

AMD is a trademark of Advanced Micro Devices, Inc.

Intel® is a trademark of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Other company, product, or service names may be trademarks or service marks of others.