What is RHEL AI? A Guide to the Open Source Way for AI

Positioning Information

Accelerating artificial intelligence and machine learning deployments

Artificial Intelligence (AI) is a powerful new tool in the IT industry but, as with most digital tooling, one size does not fit all. A general AI bot trained on random Internet content from 5 years ago can impress a casual user, but your organization likely needs current, specialized knowledge about developing (and possibly confidential) domains. You need an AI that's been fine-tuned for your organization, and there's no better way to get that than with Red Hat Enterprise Linux AI (RHEL AI) and InstructLab.

However, deploying these technologies can be complex. As data scientists work to build their models, they often face a lack of alignment between rapidly evolving tools. These discrepancies can hinder productivity and collaboration among data scientists, software developers, and IT operations. Scaling AI/ML deployments can be resource-constrained and administratively complex, requiring expensive graphics processing unit (GPU) resources for hardware acceleration and distributed workloads for Generative AI (gen AI.) Popular cloud platforms offer scalability and attractive toolsets but often lock users in, limiting architectural and deployment choices.

Thanks to the open source community, building an AI isn't as daunting as it may first seem. Multiple LLMs already exist, and thanks to InstructLab they're easy to modify for your organization's needs.

With RHEL AI plus rudimentary knowledge of Python and YAML, you can bring a customized AI to your organization. In fact, Red Hat Developers has released a new AI learning path that steps you through the process. The complementary course reveals everything you need to know about the data collection, pre-training, and fine-tuning steps, and enhancing an existing LLM. The learning path includes a step-by-step guide on how to initialize your Python environment, download and train a model, and how to make it available to your users.

Often, AI users want to experiment with large language models (LLMs) before developing and scaling out to a much larger production environment. However, experimenting with popular LLM-based applications like ChatGPT means sharing sensitive personal information, which often raises concerns about how user data is stored and used.

RHEL AI addresses this challenge by providing a low-cost, security-focused, single-server environment to experiment with large language models. RHEL makes it easy for users with little to no data science expertise to start developing and enhancing large language models without any data privacy or security concerns.

What is Red Hat Enterprise Linux AI and how does it work?

Red Hat Enterprise Linux AI is a foundation model platform designed to help you develop, test, and run the open source Granite AI models. RHEL AI is based on the open source InstructLab project and combines the Granite large language models (LLM) from IBM Research with InstructLab's model alignment tools. Based on the Large-scale Alignment for chatBots (LAB) methodology, RHEL AI can produce a bootableRHEL image so your AI deployment is as easy as booting up a container or virtual machine. Its core focus is to enable users to develop, test and run gen AI models to power enterprise applications more seamlessly.

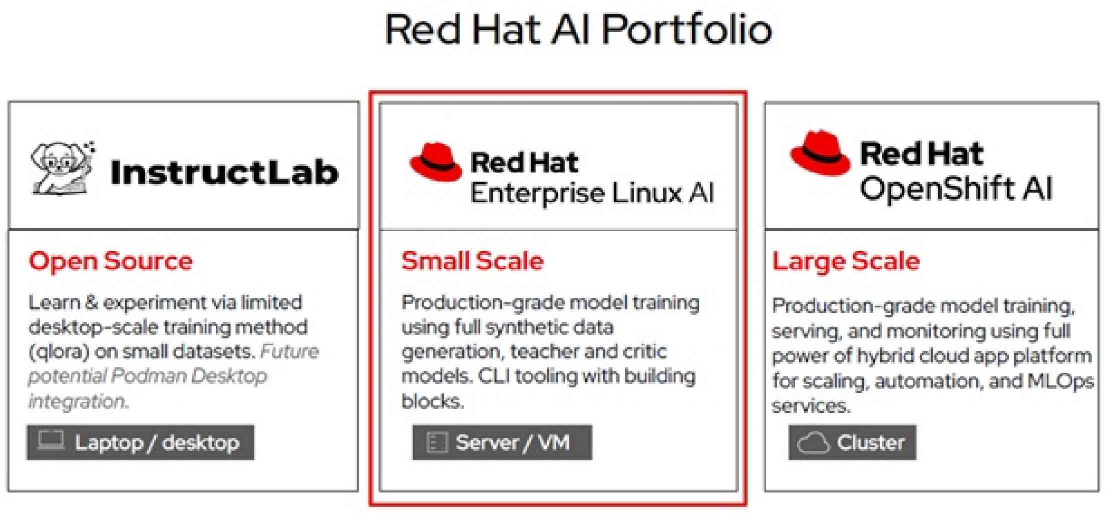
Red Hat Enterprise Linux AI allows portability across hybrid cloud environments and makes it possible to then scale your AI workflows with Red Hat OpenShift® AI and to advance to IBM watsonx.ai with additional capabilities for enterprise AI development, data management, and model governance.

For context, a foundation model is a type of machine learning (ML) model that is pre-trained to perform a range of tasks. Until recently, artificial intelligence (AI) systems were specialized tools, meaning that an ML model would be trained for a specific application or single use case. The term foundation model (also known as a base model) entered our lexicon when experts began noticing 2 trends within the field of machine learning:

- A small number of deep learning architectures were being used to achieve results for a wide variety of tasks.
- New concepts can emerge from an artificial intelligence (AI) model that were not originally intended in its training.

Foundation models have been programmed to function with a general contextual understanding of patterns, structures, and representations. This foundational comprehension of how to communicate and identify patterns creates a baseline of knowledge that can be further modified, or fine-tuned, to perform domain-specific tasks for just about any industry.

With the technological foundation of Linux, containers, and automation, Red Hat's open hybrid cloud strategy and AI portfolio gives you the flexibility to run your AI applications anywhere you need them.



Red Hat AI Portfolio

InstructLab

Open Source

Learn & experiment via limited desktop-scale training method (qlora) on small datasets. *Future potential Podman Desktop integration.*

Red Hat Enterprise Linux AI

Small Scale

Production-grade model training using full synthetic data generation, teacher and critic models. CLI tooling with building blocks.

Red Hat OpenShift AI

Large Scale

Production-grade model training, serving, and monitoring using full power of hybrid cloud app platform for scaling, automation, and MLOps services.

The launch of ChatGPT ignited tremendous interest in generative AI, and the pace of innovation has only accelerated since then. Enterprises have transitioned from initial evaluations of generative AI services to developing AI-enabled applications. A rapidly expanding ecosystem of open model options has fueled further AI innovation, demonstrating that there won't be a single dominant model. Customers will benefit from a diverse range of choices to meet specific requirements, all of which will be further accelerated by an open approach to innovation.

Although generative AI offers immense potential, the associated costs of acquiring, training, and fine-tuning large language models (LLMs) can be exorbitant, with some leading models costing nearly $200 million to train before launch. This doesn't include the cost of aligning the model with a specific organization's requirements or data, which typically requires data scientists or highly specialized developers. Regardless of the chosen model, alignment is necessary to adapt it to company-specific data and processes, making efficiency and agility crucial for AI in real-world production environments.

Red Hat predicts that over the next decade, smaller, more efficient, and purpose-built AI models will become a significant part of the enterprise IT stack, alongside cloud-native applications. However, to achieve this, generative AI must be more accessible and available, from its costs to its contributors to its deployment locations across the hybrid cloud. For decades, open source communities have helped address similar challenges for complex software problems through contributions from diverse user groups. A similar approach can lower the barriers to effectively adopting generative AI.

Gen AI is a catalyst for groundbreaking change, disrupting everything from how software is made to how we communicate. But frequently, the LLMs used for gen AI are tightly controlled, and cannot be evaluated or improved without specialized skills and high costs. The future shouldn't be in the hands of the few.

> With RHEL AI and its open source approach, you can encourage gen AI innovation with trust and transparency, while lowering costs and removing barriers to entry.

Using RHEL AI, companies will be able to **train and deploy generative AI anywhere across the hybrid cloud, close to where their data resides**. In addition, the platform provides an on-ramp to Red Hat's OpenShift AI platform for training, tuning, and serving generative AI models using the same tools and concepts.

What are large language models and how do they work?

LLMs acquire an understanding of language through a method known as unsupervised learning. This process involves providing a machine learning model with vast datasets of words and phrases, allowing it to learn by example. This unsupervised pretraining phase is fundamental to the development of LLMs like GPT-3 and BERT.

Even without explicit human instruction, the computer can extract information from the data, establish connections, and "learn" about language. As the model learns the patterns that govern word sequencing, it can predict how sentences should be structured based on probability. The outcome is a model capable of capturing intricate relationships between words and sentences.

LLMs require substantial computational resources due to their constant calculation of probabilities to find connections. One such resource is the graphics processing unit (GPU). A GPU is a specialized hardware component designed for complex parallel processing tasks, making it ideal for machine learning and deep learning models that require extensive calculations, like LLMs.

GPUs are crucial for accelerating the training and operation of transformers, a type of software architecture specifically designed for natural language processing (NLP) tasks that most LLMs implement. Transformers are fundamental building blocks for popular LLM foundation models like ChatGPT and BERT.

Transformer architecture enhances machine learning models by efficiently capturing contextual relationships and dependencies between elements in a sequence of data, such as words in a sentence. It uses self-attention mechanisms (parameters) to weigh the importance of different elements, improving understanding and performance. Parameters define boundaries, which are crucial for making sense of the vast amount of data processed by deep learning algorithms.

Transformer architecture incorporates millions or billions of parameters, allowing it to capture intricate language patterns and nuances. The term "large" in "large language model" refers to the substantial number of parameters required for an LLM to operate effectively.

Modern LLMs exhibit an unparalleled ability to understand and utilize language, surpassing what was previously conceivable from a personal computer. These machine learning models can generate text, summarize content, translate, rewrite, classify, categorize, analyze, and more.

This powerful toolset empowers humans to augment their creativity, enhance productivity, and tackle complex problems. Some of the most common applications of LLMs in a business setting include:

- **Automation and efficiency**: LLMs can help supplement or entirely take on the role of language-related tasks such as customer support, data analysis, and content generation. This automation can reduce operational costs while freeing up human resources for more strategic tasks.

- **Generating insight**: LLMs can quickly scan large volumes of text data, enabling businesses to better understand market trends and customer feedback by scraping sources like social media, reviews, and research papers, which can in turn help inform business decisions.
- **Creating a better customer experience**: LLMs help businesses deliver highly personalized content to their customers, driving engagement and improving the user experience. This may look like implementing a chatbot to provide round-the-clock customer support, tailoring marketing messages to specific user personas, or facilitating language translation and cross-cultural communication.

While there are many potential advantages to using an LLM in a business setting, there are also potential limitations to consider:

- **Cost:** LLMs require significant resources to develop, train, and deploy. This is why many LLMs are built from foundation models, which are pretrained with NLP abilities and provide a baseline understanding of language from which more complex LLMs can be built on top of. Open source and open source-licensed LLMs are free for use, making them ideal for organizations that otherwise wouldn't be able to afford to develop an LLM on their own.
- **Privacy and security:** LLMs require access to a lot of information, and sometimes that includes customer information or proprietary business data. This is something to be especially cautious about if the model is deployed or accessed by third-party providers.
- **Accuracy and bias:** If a deep learning model is trained on data that is statistically biased, or doesn't provide an accurate representation of the population, the output can be flawed. Unfortunately, existing human bias is often transferred to artificial intelligence, thus creating risk for discriminatory algorithms and bias outputs. As organizations continue to leverage AI for improved productivity and performance, it's critical that strategies are put in place to minimize bias. This begins with inclusive design processes and a more thoughtful consideration of representative diversity within the collected data.

IBM InstructLab and Granite Models are revolutionizing LLM training

IBM Granite is a series of decoder-only AI foundation models created by IBM. Initially intended for use in IBM's cloud-based data and generative AI platform Watsonx along with other models, IBM opened the source code of some code models. Granite models are trained on datasets curated from Internet, academic publishings, code datasets, legal and finance documents.

As generative AI shifts from experimentation to value creation, the training methods for foundation models are also evolving. Similar to how humans learn to learn more effectively with increased expertise, IBM Research teams, in collaboration with Red Hat counterparts, have begun to explore how generative AI models can learn more efficiently. Their recently launched InstructLab demonstrates significant acceleration in customizing foundation models for specific tasks.

InstructLab is an open-source project that aims to lower the cost of fine-tuning LLMs by enabling the ability to integrate changes to an LLM without the need to fully retrain the entire foundation model.

With its recently released family of Granite models, using InstructLab, IBM was able to demonstrate a 20% higher code generation score along with a reduction in the time it takes to achieve that quality.

Granite models are IBM's family of large language models (LLMs) designed to enhance the productivity of human programmers. These LLMs vary in parameter size and apply generative AI to multiple modalities, including language and code. Granite foundation models are being fine-tuned to create assistants that aid in translating code from legacy languages to current ones, debugging code, and writing new code based on plain English instructions. Given IBM's focus on enterprise-grade generative AI, Granite models have been trained on datasets encompassing not only code generation but also those related to academics, legal, and finance.

The large-scale training of new foundation models has significantly advanced generative AI and its potential applications for humanity. As foundation models are applied to real-world use cases and applications, especially within enterprises, it's crucial to build upon this impact. However, traditional training methods for these foundation models require substantial data center resources, leading to significant capital and operational costs. To fully realize the promise of generative AI, companies must rethink their model training processes. For widespread AI model deployment, fine-tuning techniques need to evolve to incorporate more domain-specific data at a lower cost. Based on the results demonstrated so far, IBM and Red Hat's InstructLab project appears to be making significant strides in this direction.

An open source approach to gen AI

RHEL AI aims to make generative AI more accessible, efficient, and flexible for CIOs and enterprise IT organizations across the hybrid cloud.

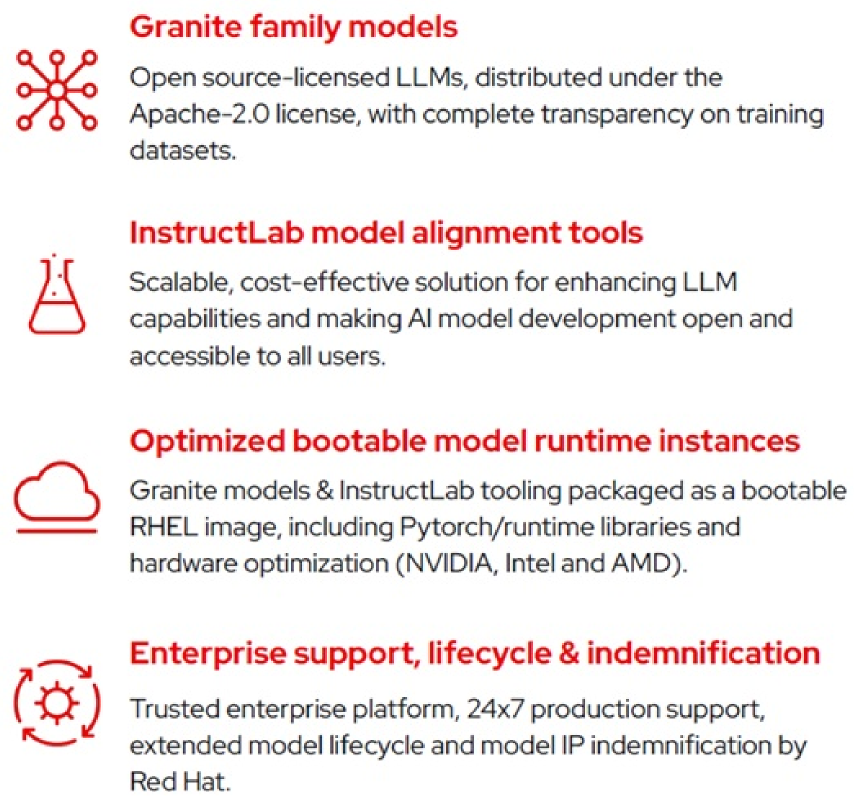
RHEL AI helps to accomplish this by:

- Empowering gen AI innovation with enterprise-grade, open source-licensed Granite models, and aligned with a wide variety of gen AI use cases.
- Streamlining alignment of gen AI models to business requirements with InstructLab tooling, making it possible for domain experts and developers within an organization to contribute unique skills and knowledge to their models even without extensive data science skills.

- Training and deploying gen AI anywhere across the hybrid cloud by providing all of the tools needed to tune and deploy models for production servers wherever associated data lives.

RHEL AI also provides a ready on-ramp to Red Hat OpenShift AI for training, tuning and serving these models at scale while using the same tooling and concepts.

Red Hat Enterprise Linux AI brings together:



Granite family models

Open source-licensed LLMs, distributed under the Apache-2.0 license, with complete transparency on training datasets.

InstructLab model alignment tools

Scalable, cost-effective solution for enhancing LLM capabilities and making AI model development open and accessible to all users.

Optimized bootable model runtime instances

Granite models & InstructLab tooling packaged as a bootable RHEL image, including Pytorch/runtime libraries and hardware optimization (NVIDIA, Intel and AMD).

Enterprise support, lifecycle & indemnification

Trusted enterprise platform, 24x7 production support, extended model lifecycle and model IP indemnification by Red Hat.

Open Granite family models

RHEL AI includes highly performant, open source licensed, collaboratively developed Granite language and code models from the InstructLab community, fully supported and indemnified by Red Hat. These Granite models are Apache 2 licensed and provide transparent access to data sources and model weights. In the future, Red Hat Enterprise Linux AI will also include additional Granite models including the Granite code model family.

InstructLab model alignment tools

LAB: Large-Scale Alignment for ChatBots is a novel approach to instruction alignment and fine-tuning of large language models with a taxonomy-driven approach leveraging high-quality synthetic data generation. In simpler terms, it allows for users to customize an LLM with domain-specific knowledge and skills. InstructLab then generates high-quality synthetic data that is used to train the LLM. A replay buffer is used to prevent forgetting.

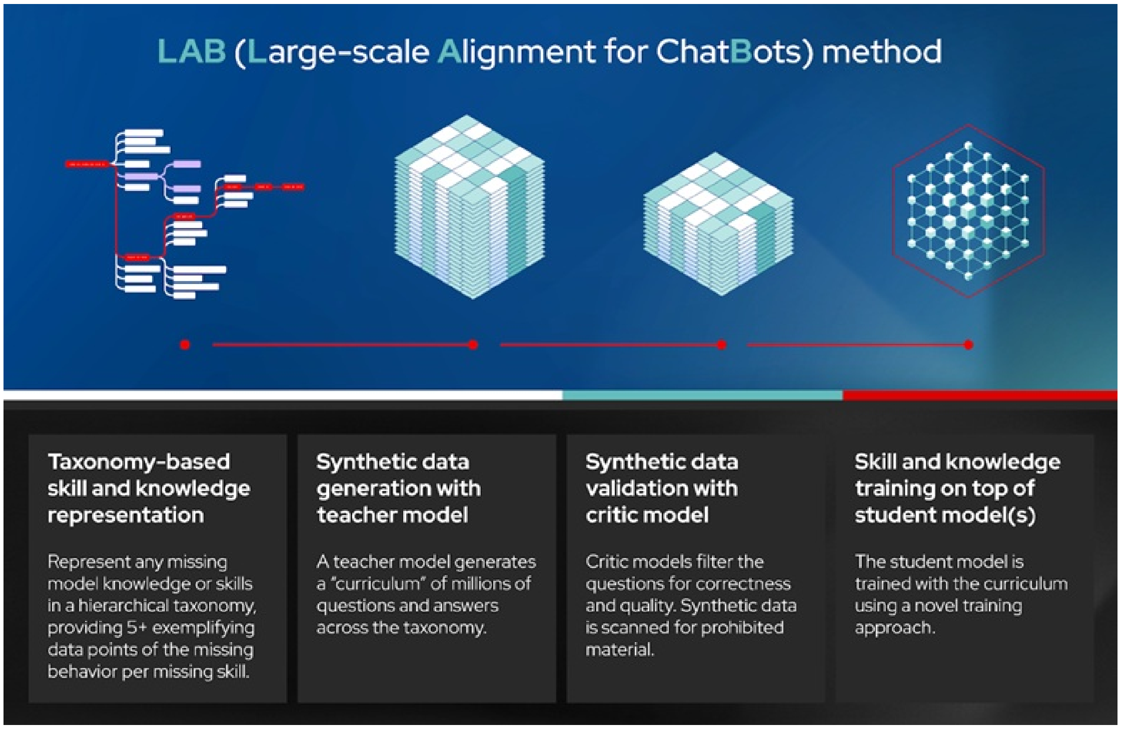
InstructLab is an open source project for enhancing large language models (LLMs) used in generative artificial intelligence (gen AI) applications. Created by IBM and Red Hat, the InstructLab community project provides a cost-effective solution for improving the alignment of LLMs and opens the doors for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

LLMs can drive a variety of useful applications, including chatbots and coding assistants. These LLMs can be proprietary (like OpenAI's GPT models and Anthropic's Claude models) or offer varying degrees of openness regarding pretraining data and usage restrictions (like Meta's Llama models, Mistral AI's Mistral models, and **IBM's Granite models**).

AI practitioners frequently need to adapt a pretrained LLM to meet specific business needs. However, there are limitations to how an LLM can be modified. InstructLab employs an approach that overcomes these limitations. It can enhance an LLM using significantly less human-generated information and fewer computing resources compared to traditional retraining methods. Additionally, it allows for continuous model improvement through upstream contributions.

The LAB technique contains four distinct steps (Figure 1):

- Taxonomy based skills and knowledge representation
- Synthetic data generation (SDG) with a teacher model
- Synthetic data validation with a critic model.
- Skills and knowledge training on top of the student model(s)



LAB (Large-scale Alignment for ChatBots) method

Taxonomy-based skill and knowledge representation

Represent any missing model knowledge or skills in a hierarchical taxonomy, providing 5+ exemplifying data points of the missing behavior per missing skill.

Synthetic data generation with teacher model

A teacher model generates a "curriculum" of millions of questions and answers across the taxonomy.

Synthetic data validation with critic model

Critic models filter the questions for correctness and quality. Synthetic data is scanned for prohibited material.

Skill and knowledge training on top of student model(s)

The student model is trained with the curriculum using a novel training approach.

Figure 1. InstructLab Technique

InstructLab is the name of the software that implements the LAB technique. It consists of a command-line interface that interacts with a local git repository of skills and knowledge, including new ones that the user has added, to generate synthetic data, run the training of the LLM, serve the trained model and chat with it.

Users can create their own custom LLM by training the base models with their own skills and knowledge. They can choose to either share the trained model and the added skills and knowledge with the community or keep them private.

Optimized bootable Red Hat Enterprise Linux for Granite models and InstructLab

The Granite models & InstructLab tooling are downloaded and deployed on a bootable RHEL image with an optimized software stack for popular hardware accelerators from vendors like AMD, Intel and NVIDIA. Furthermore, these RHEL AI images will boot and run across the Red Hat Certified Ecosystem including public clouds and AI-optimized servers from Lenovo.

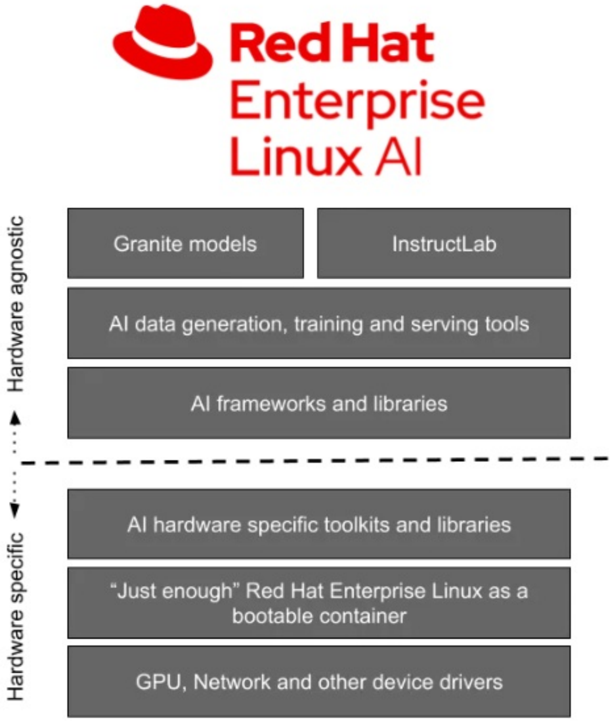
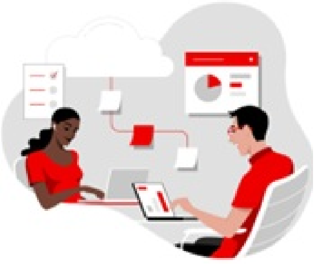

Red Hat Enterprise Linux AI
Hardware agnostic
Granite models
InstructLab
AI data generation, training and serving tools
AI frameworks and libraries
Hardware specific
AI hardware specific toolkits and libraries
"Just enough" Red Hat Enterprise Linux as a bootable container
GPU, Network and other device drivers

Figure 2. RHEL AI optimized stack

Enterprise support, lifecycle & indemnification

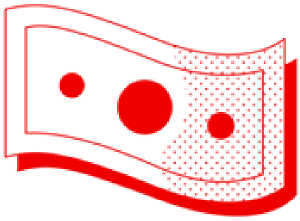
At general availability (GA), Red Hat Enterprise Linux AI Subscriptions will include enterprise support, a complete product life cycle starting with the Granite 7B model and software, and IP indemnification by Red Hat.

What problems does RHEL AI solve



Getting started with Generative AI is often very challenging. With InstructLab* and the Granite models, we make the ability to add knowledge and skills to the language model accessible to everyone. This allows companies to add their specific knowledge and skills to align the model to their specific use case.

Companies are not comfortable using large language models due to the unknown license of the outputted text. RHEL AI includes indemnification and the model itself is licensed under the Apache-2.0 Open Source license.



Getting started with Generative AI is very expensive. RHEL AI and InstructLab provides a way for people to get started in a cost effective way and scale out when needed.

Summary

The launch of ChatGPT ignited tremendous interest in generative AI, and the pace of innovation has only accelerated since then. Enterprises have transitioned from initial evaluations of generative AI services to developing AI-enabled applications. A rapidly expanding ecosystem of open model options has fueled further AI innovation, demonstrating that there won't be a single dominant model. Customers will benefit from a diverse range of choices to meet specific requirements, all of which will be further accelerated by an open approach to innovation.

Implementing an AI strategy involves more than simply choosing a model; technology organizations need the expertise to tailor a specific model to their use case and address the significant costs of AI implementation. The scarcity of data science skills is compounded by substantial financial requirements, including:

- Procuring AI infrastructure or consuming AI services
- The complex process of tuning AI models for specific business needs
- Integrating AI into enterprise applications
- Managing both the application and model lifecycle.

To truly lower the entry barriers for AI innovation, enterprises need to be able to expand the roster of who can work on AI initiatives while simultaneously getting these costs under control. With InstructLab alignment tools, Granite models and RHEL AI, Red Hat aims to apply the benefits of true open source projects - freely accessible and reusable, transparent and open to contributions - to GenAI in an effort to remove these obstacles.

Related product families

Product families related to this document are the following:

- Red Hat Alliance

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.

© Copyright Lenovo 2024. All rights reserved.

This document, LP2032, was created or updated on September 19, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2032
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2032.

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.