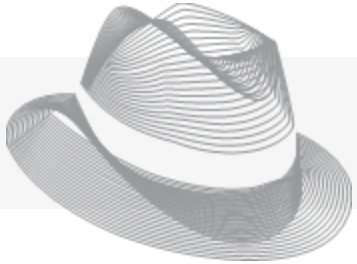
INTER COMPUTER GROUP

PRECISION. RELIABILITY. STABILITY.

Red Hat

RED HAT ENTERPRISE LINUX AI OVERVIEW

Date



RED HAT'S AI PORTFOLIO STRATEGY

TRUST | CHOICE | CONSISTENCY

AI models

RHEL AI

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

AI platform

OpenShift AI

Development | Serving | Monitoring & Lifecycle | MLOps | Resource Management

AI enabled portfolio

Lightspeed portfolio

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

AI workload support

Optimize AI workloads

Deployment & Run | Compliance | Certification | Models | Open Source Ecosystem

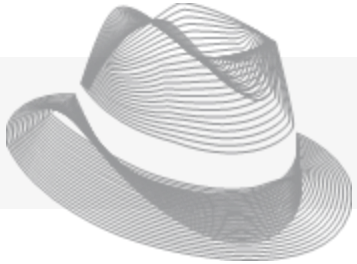
Open Hybrid Cloud Platforms

Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform

Acceleration | Performance | Scale | Automation | Observability | Security | Developer Productivity | App Connectivity | Secure Supply Chain

Partner Ecosystem

Hardware | Accelerators | Delivery



OVERVIEW OF RED HAT ENTERPRISE LINUX AI

Red Hat
Enterprise Linux AI

FOUNDATION MODEL PLATFORM

Seamlessly develop, test, and run Granite family large language models (LLMs) for enterprise applications.

Granite family models

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

InstructLab model alignment tools

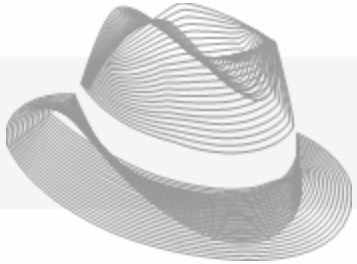
Scalable, cost-effective solution for enhancing LLM capabilities and making AI model development open and accessible to all users.

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).

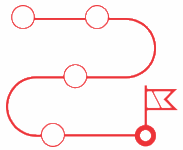
Enterprise support, lifecycle & indemnification

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



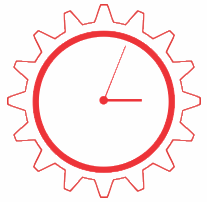
RED HAT ENTERPRISE LINUX AI

Gen AI adoption challenges



EFFICIENCY

Large, proprietary gen AI models are expensive to run and difficult to train/tune

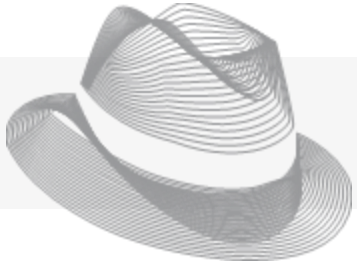


ACCESSIBILITY

Aligning models to enterprise requirements is difficult for non-data scientists

FLEXIBILITY

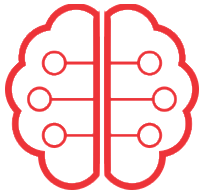
Training, tuning and serving models everywhere your data lives can be a challenge



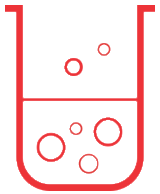
RED HAT ENTERPRISE LINUX AI

Rhel AI benefits

Streamline adoption of generative AI



Unlock the power of efficient, open source gen AI models with Granite



Enable users to easily add their skills & knowledge to models with InstructLab

Train and tune models to serve across hybrid cloud environment

OVERVIEW OF RED HAT ENTERPRISE LINUX AI

IBM GRANITE MODEL FAMILY

Released under the Apache 2 license

IBM Granite Language models

English Base
Granite-7B-Base

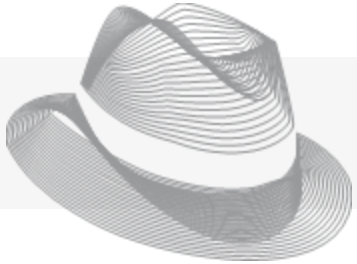
English Instruction-tuned
Granite-7B-Instruct

IBM Granite Code models

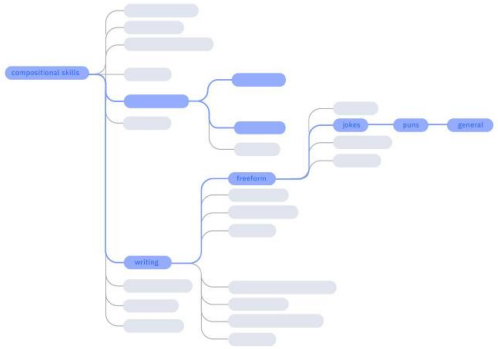
Base
Granite-34B-Code-Base
Granite-20B-Code-Base
Granite-8B-Code-Base
Granite-3B-Code-Base

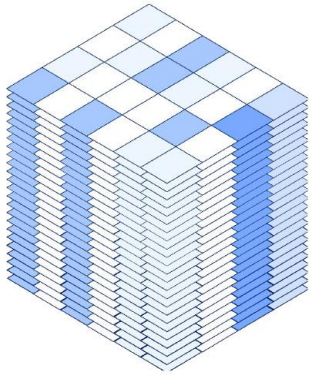
Instruction-tuned
Granite-34B-Code-Instruct
Granite-20B-Code-Instruct
Granite-8B-Code-Instruct
Granite-3B-Code-Instruct

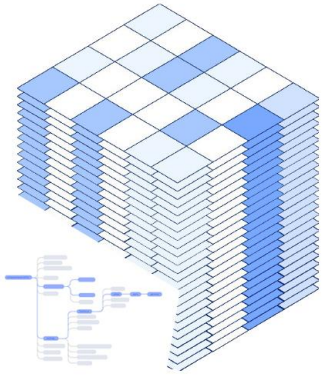
Granite

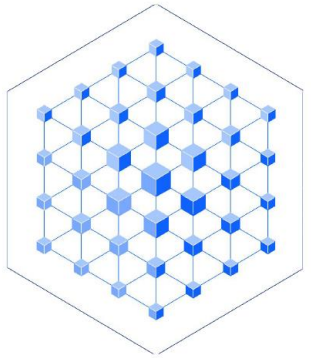


LAB (Large-scale Alignment for ChatBots) METHOD









Taxonomy-based skill & knowledge representation

Synthetic data generation with teacher model

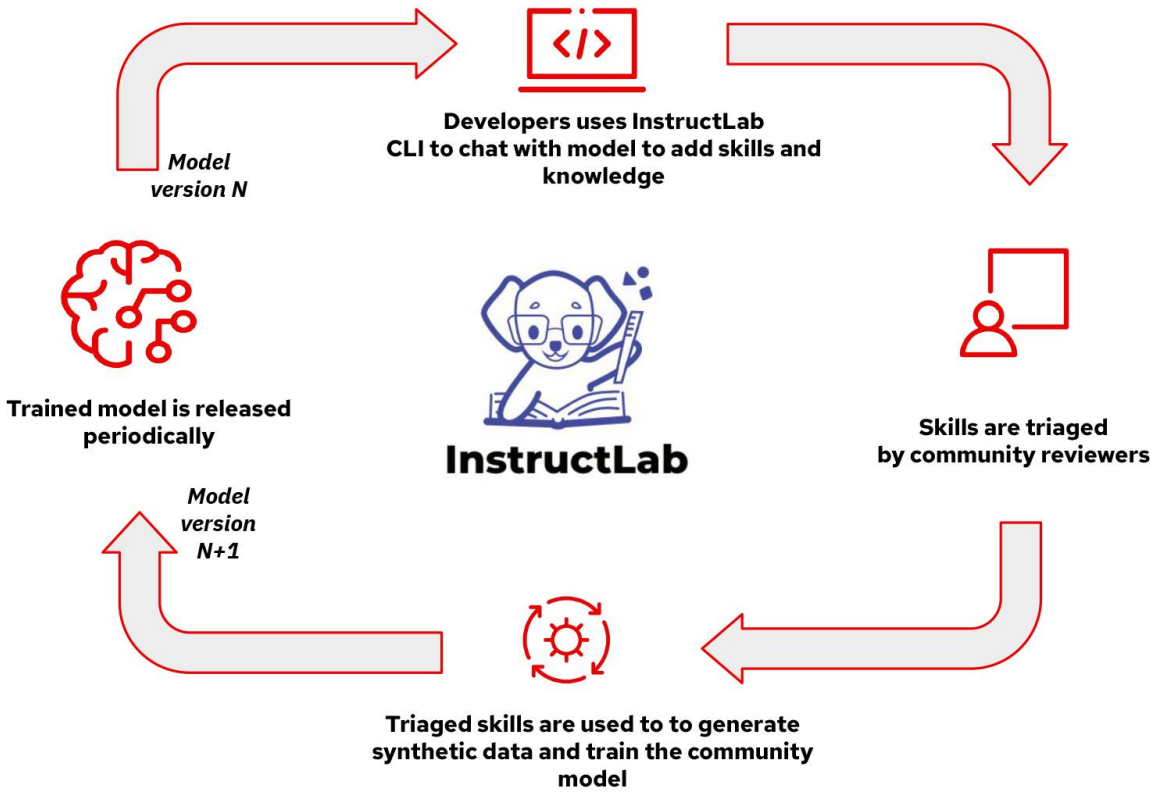
Synthetic data validation with critic model

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

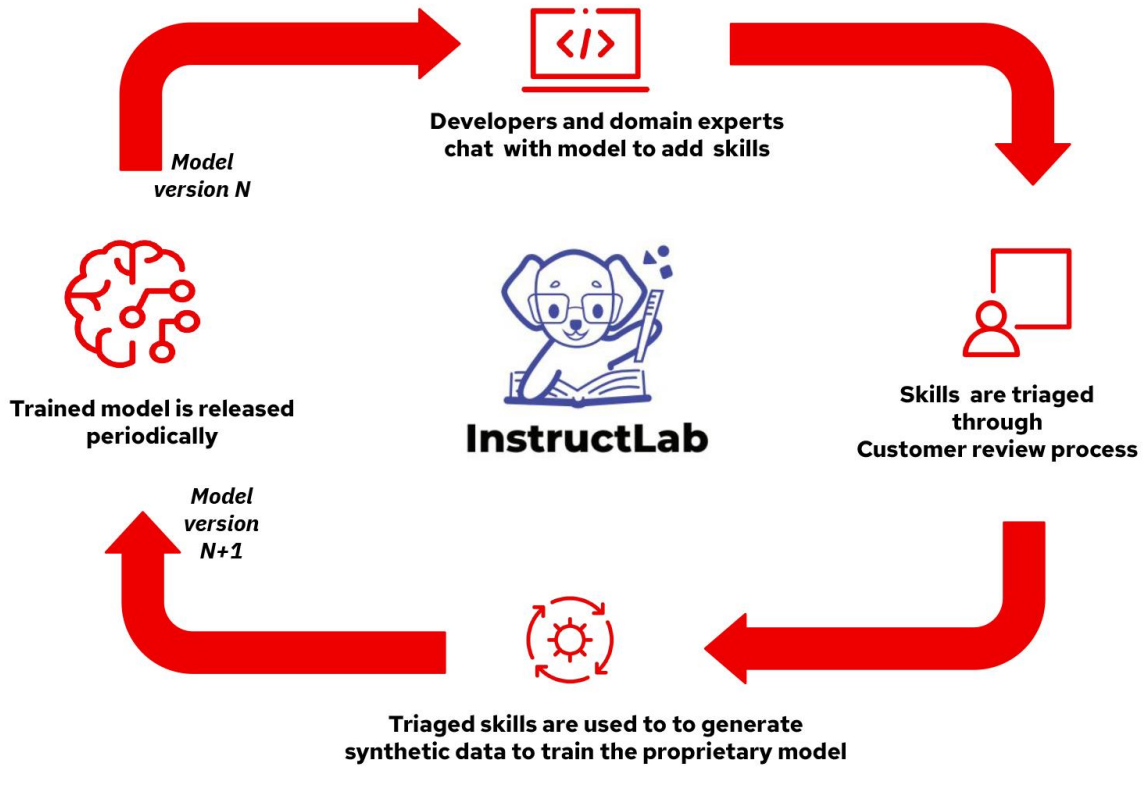
https://arxiv.org/abs/2403.01081

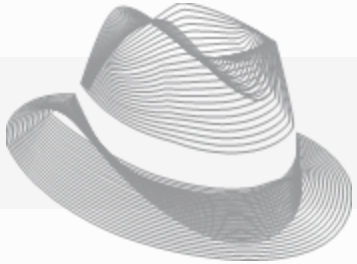
InstructLab TOOLING

COMMUNITY INSTANCE
operated by Red Hat with support
from IBM Research team



CUSTOMER/PRIVATE INSTANCE
operated by Customer or RedHat/Partner_aaS
with commercial support from Red Hat





OVERVIEW OF RED HAT OPENSHIFT AI

Red Hat
OpenShift AI

INTEGRATED MLOps PLATFORM

Create and deliver GenAI and predictive models at scale across hybrid cloud environments.

Model development
Provides flexibility and composability by supporting multiple AI/ML libraries, frameworks, and runtimes.

Model serving and monitoring
Deploy models across any OpenShift footprint and centrally monitor their performance.

Lifecycle management
Expands DevOps practices to MLOps to manage the entire AI/ML lifecycle.

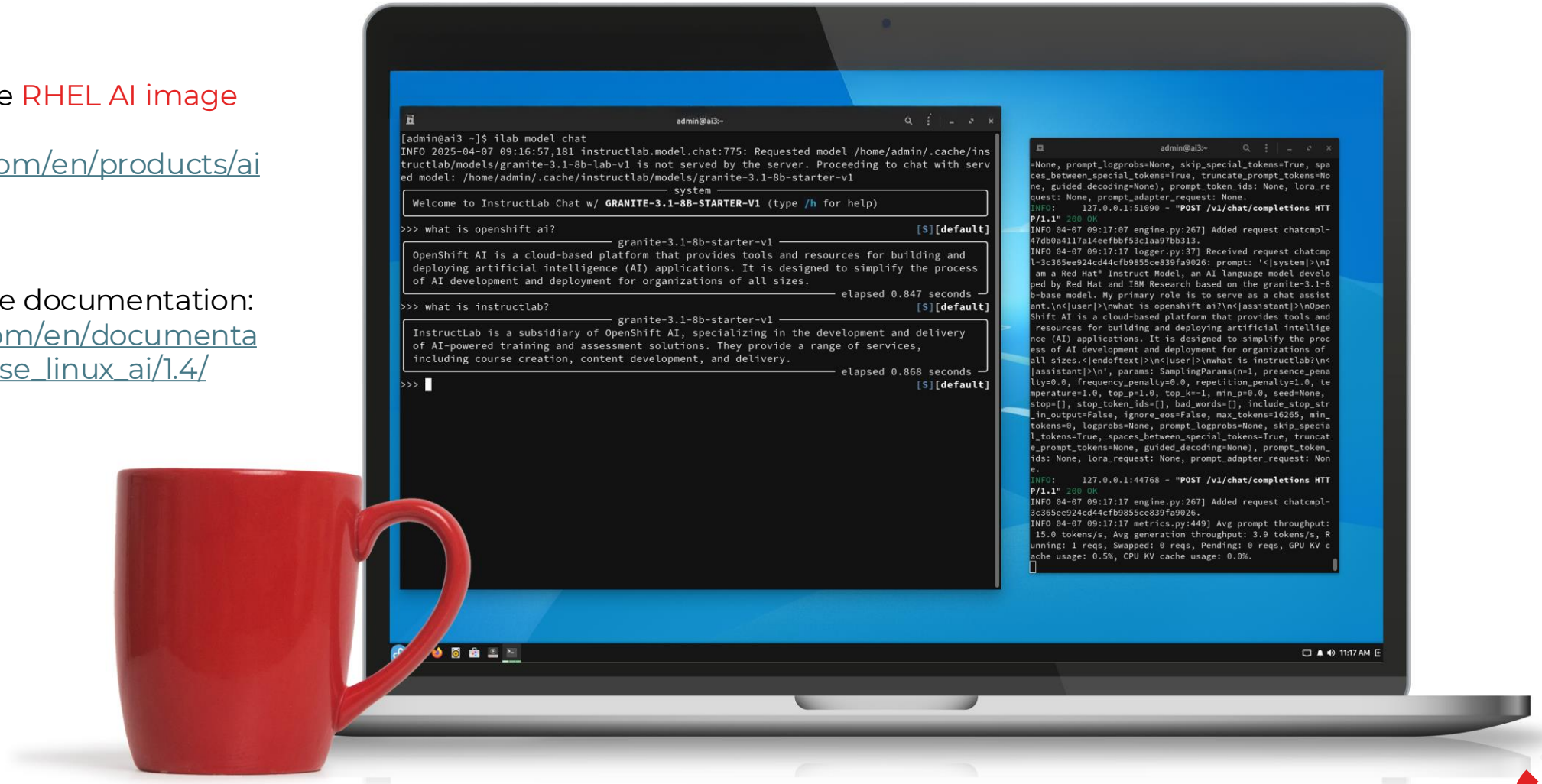
Resource optimization and management
Scales to meet the workload demands of foundation models and traditional machine learning.

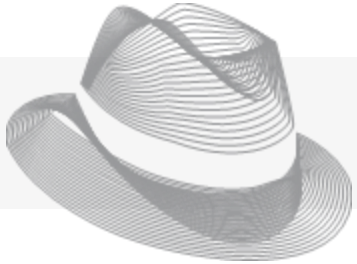
Available as
- Fully managed cloud service
- Traditional software product on-site or in the cloud!

RED HAT ENTERPRISE LINUX AI

You can download the RHEL AI image from this link:
https://www.redhat.com/en/products/ai/enterprise-linux-ai

The correct link for the documentation:
https://docs.redhat.com/en/documentation/red_hat_enterprise_linux_ai/1.4/





RED HAT AI PORTFOLIO

InstructLab

Red Hat
Enterprise Linux AI

Red Hat
OpenShift AI

Open Source

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

Laptop / desktop

Small Scale

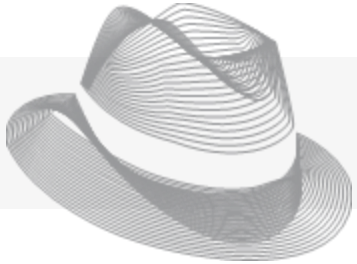
Production-grade model training using full synthetic data generation, teacher and critic models. CLI tooling with building blocks.

Server / VM

Large Scale

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

Cluster



TUNING

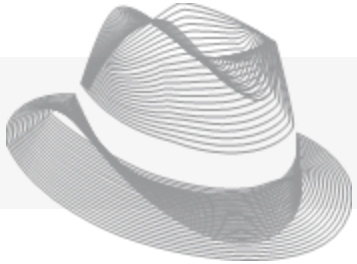
Customize smaller, fit-for-purpose models using InstructLab to build an efficient, cost-effective solution.

InstructLab methodology reduces the complexity of customizing models with enterprise private data at a fraction of the cost.

Models

Enterprise private data

Tuning tools



TUNING

InstructLab vs. Alternative Model Alignment Approaches

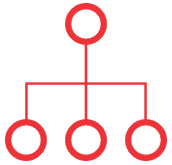
RAG

Retrieval Augmented Generation

Enhance Gen AI model-generated text by retrieving relevant information from external sources, improving accuracy and depth of model's responses.

InstructLab

Large-scale Alignment for chatBots



Leverage a taxonomy-guided synthetic data generation process and a multi-phase tuning framework to improve model performance.

NEW

Fine Tuning

Fine Tuning

Adjust a pre-trained model on specific tasks or data, improving its performance and accuracy for specialized applications without full retraining.

InstructLab provides **more accessible fine tuning** & **complements RAG** (RAFT pattern)

You can download the RHEL AI image from this link:

Lea Vass Boros, technical sales
M: +36 20 542 5111
E: lea.vass@intercomputer.hu

INTER-COMPUTER-INFORMATIKA ZRT.
www.intercomputer.hu