Red Hat Summit: Connect | Zurich, Switzerland

# Red Hat AI: Strategy and Roadmap

## Chris Wright

Chief Technology Officer and Senior Vice President,
Global Engineering
Red Hat

Red Hat

# Generative AI customer adoption challenges

## Model Cost

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

✓ Granite Models

## Alignment Complexity

Aligning models to enterprise data and use cases is difficult for non-data scientists.

✓ InstructLab

## Deployment Constraints

Tuning and serving models everywhere your data lives can be a challenge.

✓ Red Hat AI

Red Hat

# Red Hat AI platform

## Red Hat Enterprise Linux AI

**Foundation model platform for developing, testing, and running Granite family LLMs**

- Provides a simplified approach to get started with generative AI that includes open source models
- Makes AI accessible to developers and domain experts with little data science expertise
- Provides the ability to do training & inference on individual production server deployments

## Red Hat OpenShift AI

**Integrated MLOps platform for model lifecycle management at scale anywhere**

- Provides support for both generative and predictive AI models with a BYOM approach
- Includes distributed compute, collaborative workflows, model serving and monitoring
- Offers enterprise MLOps capabilities and the ability to scale across hybrid-clouds
- Includes Red Hat Enterprise Linux AI, including InstructLab and Granite models

# 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 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 AI portfolio

## InstructLab

### Open Source

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

Laptop / desktop

## 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.

Server / VM

## Red Hat OpenShift AI

### Large Scale

Production-grade model training as in RHEL AI, using full power of Kubernetes scaling, automation and MLOps services.

Cluster

# Neural Magic + Red Hat

**Neural Magic**

Neural Magic's expertise in software and algorithms that accelerate generative AI (gen AI) inference workloads.

vLLM LLM Compressor

nm-vllm DeepSparse

+

**Red Hat**

Red Hat's vision of high-performing, functional AI workloads that directly map to customer-specific use cases and data, anywhere and everywhere across the hybrid cloud.

Red Hat Enterprise Linux AI

InstructLab

Red Hat OpenShift AI

=

The potential to supercharge LLM deployments anywhere and everywhere across the hybrid cloud by providing a ready-made, highly-optimized, open inference stack.

# Red Hat AI real-world use cases

## Ticket classification and routing for citizen claims

Agesic, Uruguay's Agency for Electronic Government, improved citizen experience by expediting ticket classification, routing 2,000 citizen claims per month in seconds.

## AI models library for improved business operations

**A large global airline** produced multiple models targeting different use cases like crew planning, fuel optimization, and baggage handling optimization.

## Chat-bot experience for doctors to better service patients

Clalit Health improved patient care with a chatbot-like experience for pinpointing patients who need preventive medication or follow-ups.

## Identifying patients who are at-risk

The US Department of Veterans Affairs and partner, Guidehouse, optimized patient diagnosis at time of care by helping detect a high risk of suicide ideation.

# Q&A time –
# And thank you!