Red Hat

TechTalks

Red Hat AI 101

Build, Train & Run AI on Your Terms

Red Hat

TechTalks

Red Hat AI 101

Adnan Drina

Principal Solution Architect at Red Hat NL

adnan.drina@redhat.com
linkedin.com/in/adnandrina/

Introducing Red Hat AI

Any model. Any accelerator. Any cloud.

Red Hat AI

- **Red Hat AI** Inference Server
- **Red Hat** Enterprise Linux AI
- **Red Hat** OpenShift AI

Trusted, Consistent and Comprehensive foundation

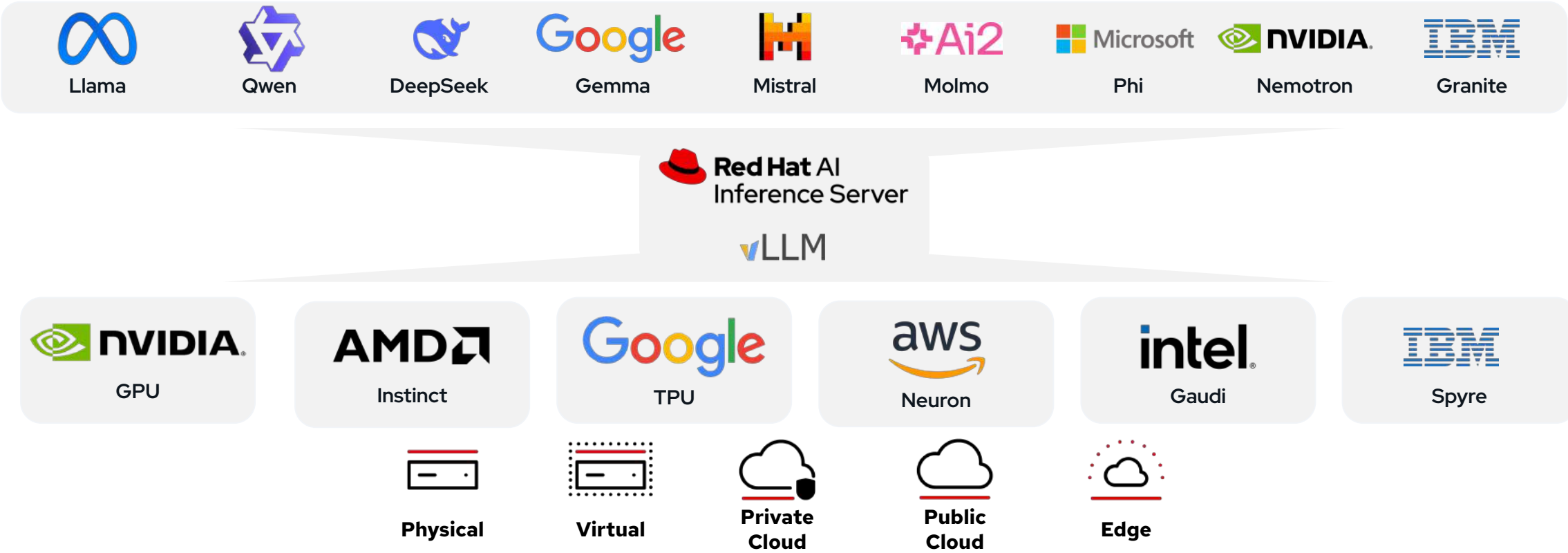
NVIDIA AMD intel Hardware Acceleration aws Google IBM

Physical | **Virtual** | **Private Cloud** | **Public Cloud** | **Edge**

Red Hat

Introducing Red Hat AI Inference Server

vLLM connects model creators to accelerated hardware providers

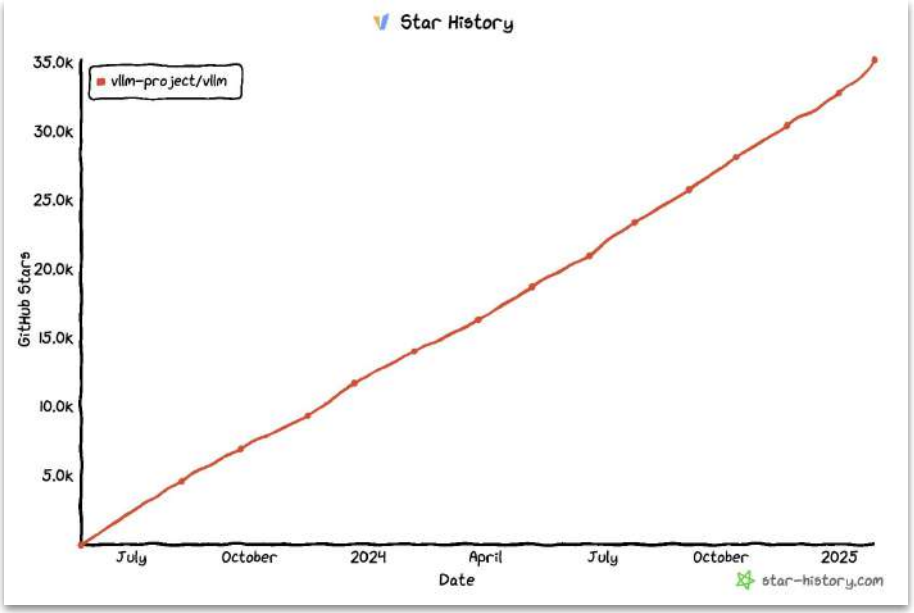


Single platform to run any model, on any accelerator, on any cloud

Red Hat

vLLM

High-Performance, Open Source Inference Engine for LLMs



Llama | Granite | Gemma | Qwen | DeepSeek | Mistral

NVIDIA GPU | AMD Instinct | intel. Gaudi | Google TPU | aws Neuron

vLLM is a high-performance and memory-efficient inference engine for serving large language models (LLMs). Originally developed at UC Berkeley, vLLM has evolved into a community-driven project with contributions from both academia and industry.

Technical Value

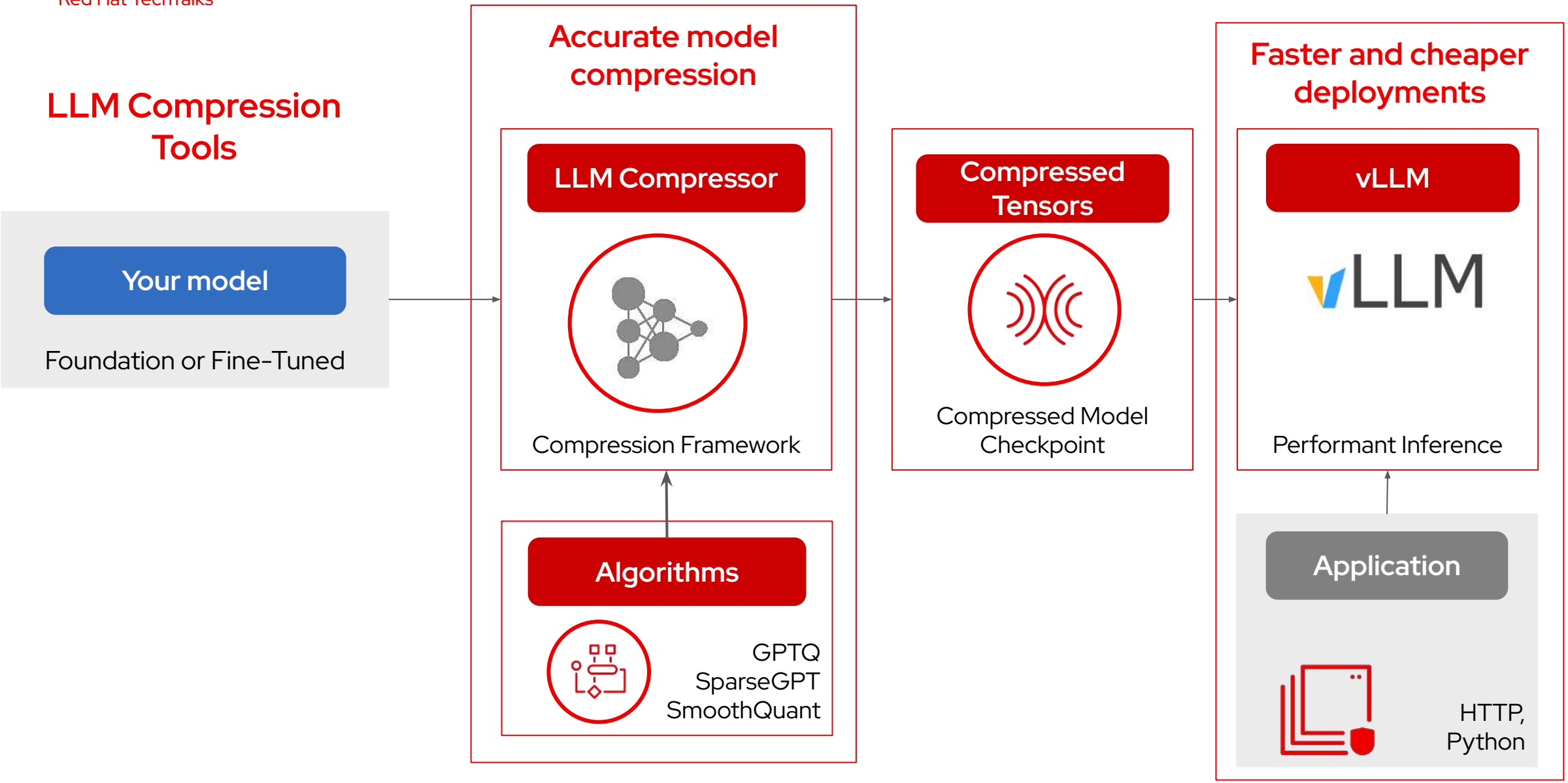
- Fast inference via **PagedAttention** & **Continuous Batching**
- Supports **quantization** (INT4, INT8, FP8) & **sparsity**
- Scalable: **multi-GPU**, distributed, **OpenAI-compatible**
- **Hardware-agnostic** (NVIDIA, AMD, Intel, AWS Inferentia, etc.)

Business Value

- Scales across cloud, edge, and on-prem
- Streamlines deployment & ops
- Speeds up time-to-market
- **Reduces infra & compute costs**

Red Hat

LLM Compression Tools



Introducing Red Hat AI repository on Hugging Face

A collection of third-party validated and optimized large language models

Broad Collection of models

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

Validated models

- Tested using realistic scenarios
- Assessed for performance across a range of hardware
- Done using GuideLLM benchmarking and LM Eval Harness

Optimized models

- Compressed for speed and efficiency
- Designed to run faster, use fewer resources, maintain accuracy
- Done using LLM Compressor with latest algorithms

Hosted on the Red Hat AI repository on Hugging Face

Cut GPU costs with inference optimized models.

Red Hat Enterprise Linux AI

Foundation Model Platform

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

Red Hat AI Inference Server

Optimize model inference across the hybrid cloud to create faster, more cost-effective model deployment and have access to repository of pre-optimized models

InstructLab model alignment tools

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

Granite family models

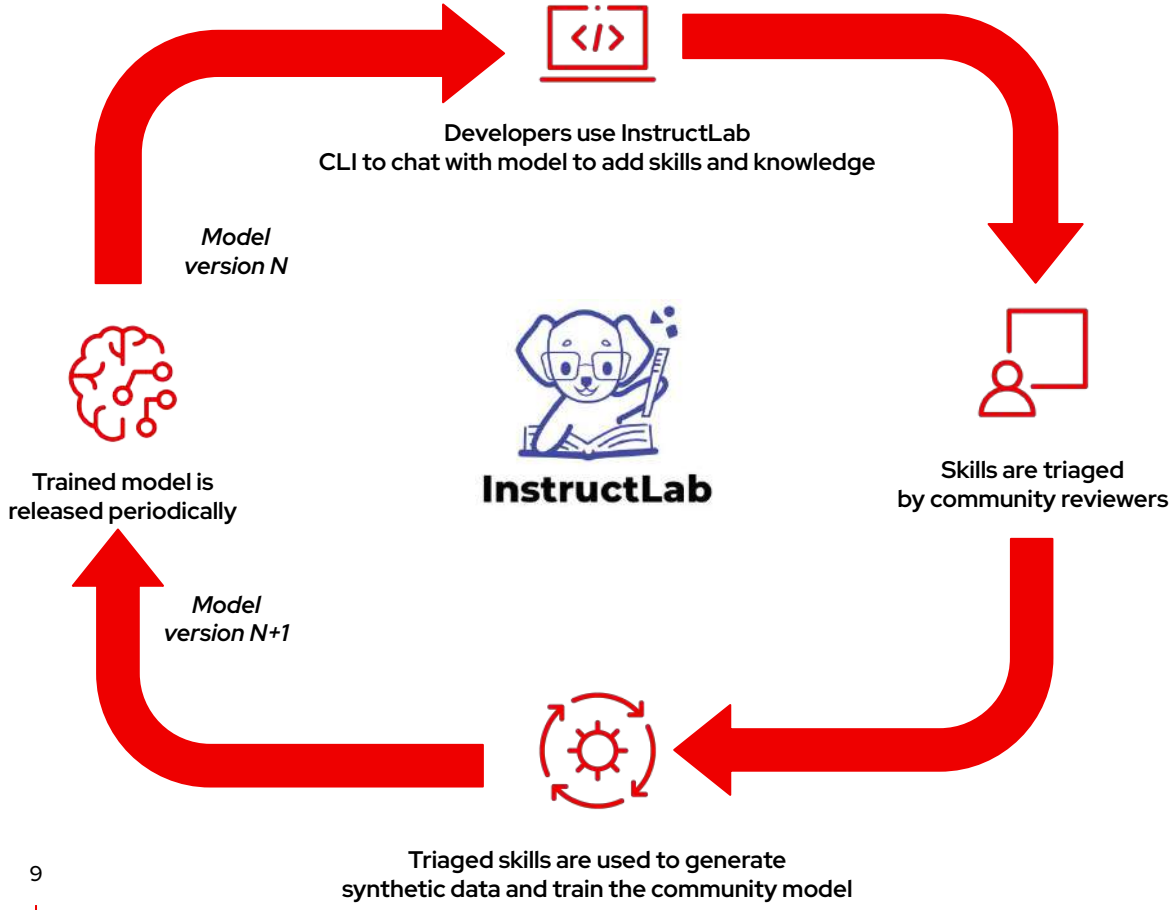
Open source-licensed LLMs, distributed under the Apache-2.0 license, with complete transparency on training datasets and model IP indemnification.

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).

InstructLab

Fine-tune LLMs with *your* data



InstructLab is an open-source framework for customizing large language models (LLMs) by adding new knowledge and skills at a fraction of the cost. Developed by Red Hat and IBM Research, it promotes collaborative and cost-effective AI development.

Technical Value

- Adds **new knowledge** to LLMs without full retraining
- Combines minimal human input with **synthetic data**
- **Model-agnostic**: works with LLaMA, Mistral, Granite, etc.
- Easy contribution via **CLI and YAML** – no ML expertise needed

Business Value

- Lowers fine-tuning cost & compute demands
- Speeds up delivery of domain-specific AI
- Broadens access to AI development
- Promotes open source collaboration & innovation

Red Hat

What If A Business Analyst Could Train Your AI?

Product Data Sheet: Wonderful Widget

Product Overview

The **Wonderful Widget** is a cutting-edge, multi-functional tool designed to revolutionize the way you approach everyday tasks. Engineered with precision and innovation, the Wonderful Widget combines versatility, durability, and ease of use into a compact and stylish design. Whether you're a professional or a DIY enthusiast, this tool is your pocket companion for a wide range of applications.

Key Features

- **Multi-Functionality**: The Wonderful Widget integrates 10 essential tools into one sleek device, including a screwdriver, pliers, bottle opener, wire cutter, and more. It's like having a toolbox in your pocket!
- **Compact & Portable**: With a foldable design, the Wonderful Widget easily fits into your pocket, glove box, or backpack. It's lightweight yet robust, making it perfect for on-the-go use.
- **Durability**: Constructed from high-grade stainless steel, the Wonderful Widget is built to withstand tough conditions. It's resistant to rust, corrosion, and wear, ensuring long-lasting performance.
- **Ergonomic Design**: The Wonderful Widget is designed with user comfort in mind. Its ergonomic handle ensures a firm, comfortable grip, reducing hand fatigue during extended use.
- **Quick Access**: With a smart locking mechanism, you can swiftly and safely access the tool you need. The Wonderful Widget's intuitive design allows for easy, one-handed operation.
- **Versatile Applications**: Whether you're fixing a bike, assembling furniture, or opening a bottle of your favorite beverage, the Wonderful Widget is up to the task.

Specifications

- **Material**: High-grade stainless steel
- **Dimensions (Closed)**: 4.5 x 1.5 x 1 inches
- **Weight**: 6.5 ounces
- **Tools Included**:
  - Flathead screwdriver
  - Phillips screwdriver
  - Pliers
  - Wire cutter
  - Bottle opener

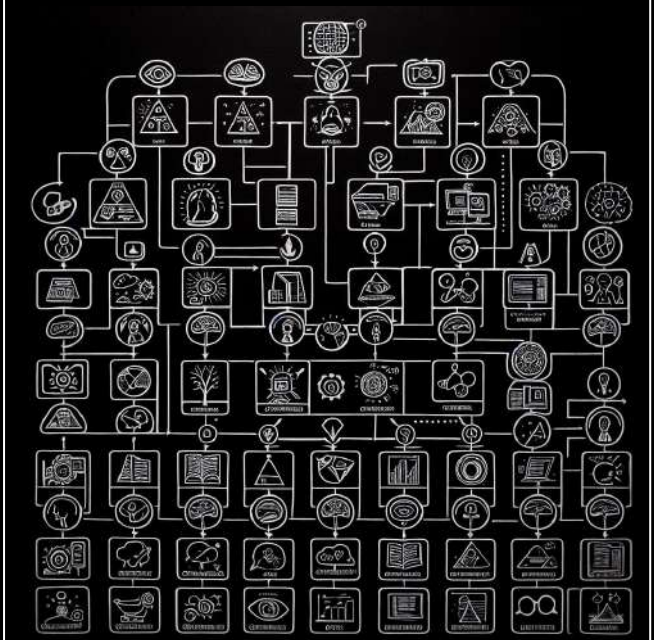
```
- context: |
    - Durability: Constructed from high-grade stainless steel, the Wonderful Widget is built to withstand tough
      conditions. It's resistant to rust, corrosion, and wear, ensuring long-lasting performance.
  questions_and_answers:
    - question: |
        - What material is the Wonderful Widget made from?
    - answer: |
        - The Wonderful Widget is made of high precision stainless steel
    - question: |
        - Is the Wonderful Widget suitable for outdoor use?
    - answer: |
        - The Wonderful Widget is built to withstand tough conditions. It is resistant to rust, corrosion, and
          wear.
    - question: |
        - Will the Wonderful Widget rust or corrode?
    - answer: |
        - No. The Wonderful Widget is made from high precision stainless steel and is resistant to rust and
          corrosion.

- context: |
    - Ergonomic Design: The Wonderful Widget is designed with user comfort in mind. Its ergonomic handle
      ensures a firm, comfortable grip, reducing hand fatigue during extended use.
```

Red Hat

InstructLAB

Add Skills/Knowledge to your taxonomy



Synthetic Data Generation
($ ilab data generate)

mixtral-8x7b-instruct-v0-1
(teacher model)

Training
($ ilab model train)

mixtral-8x7b-instruct-v0-1
(critic model)

Evaluation
($ ilab model evaluate)

Prometheus-8x7b-v2-0
(judge model)

Red Hat OpenShift AI

Integrated AI platform

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

Model development

Bring your own models or customize Granite models to your use case with your data. Supports integration of multiple AI/ML libraries, frameworks, and runtimes.

Model serving and monitoring

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

Lifecycle management

Expand DevOps practices to MLOps to manage the entire AI/ML lifecycle.

Resource optimization and management

Scale to meet workload demands of gen AI and predictive models. Share resources, projects, and models across environments.

Responsible AI and Governance

Built-in safeguards for fairness, safety, and regulatory compliance

AI model monitoring

Monitors tabular model inferences with customizable metrics for **bias** (outcome disparities) and **drift** (deployment vs. training data differences)

LLM Evaluation

Perform a huge variety of evaluation tasks over LLMs to understand and quantify their knowledge, capabilities, and behaviors

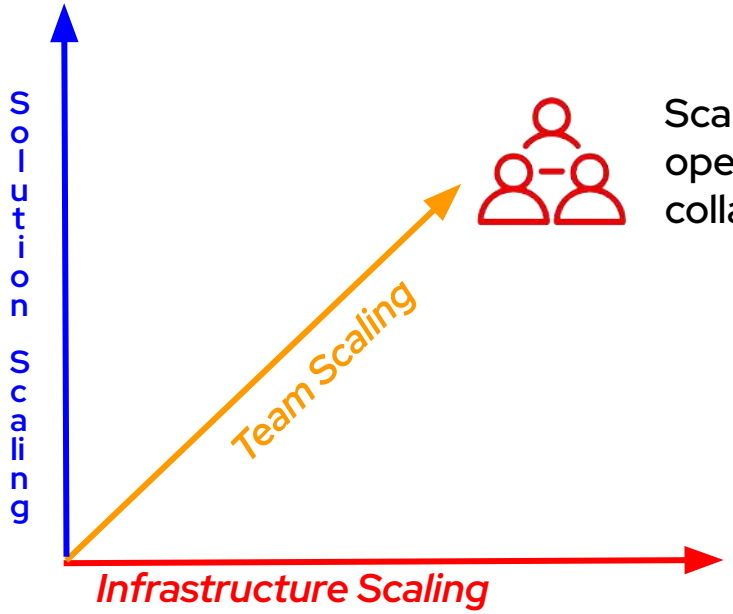
LLM Guardrails

Customizable guardrails framework to moderate interactions between users and generative AI models, ensuring secure, compliant, and efficient operations

Responsible AI and Governance

Built-in safeguards for fairness, safety, and regulatory compliance

Scale by integrating various AI models and techniques to leverage your data, foster innovation, and **differentiate yourself from the competition**.



Scale your AI projects across data scientists, operations, and developer teams, enabling collaboration and **accelerating time to value**.

Scale your infrastructure to meet user demand by easily adding GPUs, CPUs, memory, and other resources while **controlling costs**.

Red Hat

Full Lifecycle Control with OpenShift AI

Automate, govern, and scale your AI workflows from development to production

Gather and prepare data	Develop model	Deploy models in an application	Model monitoring and management
Data engineers	Data scientists	Data scientists | App developers	Data scientists
Data storage Data lake Data exploration Data preparation Stream processing	ML notebooks ML libraries	Model life cycle Continuous integration, continuous delivery (CI/CD) Collaboration	Monitoring and alerts Model visualization Model drift

Red Hat OpenShift AI and Red Hat OpenShift	Compute acceleration | Easy-to-manage, large ecosystem, and self-service capabilities
Infrastructure	

Red Hat

Red Hat OpenShift AI

Dashboard

- Accelerator Profiles
- Data Science Projects
- Admin Features
- Model Registry

Accelerators

- NVIDIA CUDA
- AMD ROCm
- Intel Gaudi

Model Development, Training & Tuning

Notebooks

- Minimal Python
- PyTorch
- VS Code
- RStudio*
- TensorFlow
- JupyterLab
- TrustyAI

CodeFlare SDK | Custom (BoY)

Kubeflow Notebooks

Distributed Workloads

- Kubeflow Training Operator
- KubeRay

Training Runtimes

- Ray
- PyTorch
- ISV, Custom (BoY)

Pipelines

- Kubeflow Pipelines

Kueue

Model Serving

Serving Engines

- KServe
- ModelMesh

Serving Runtimes

- vLLM, TGIS
- OpenVINO
- Custom

Models

- Granite Models
- Ecosystem models
- Performance metrics
- Operations metrics
- Quality metrics

OpenShift Operators

- OpenShift GitOps (ArgoCD)
- OpenShift Pipelines (Tekton)
- Authorino
- OpenShift Service Mesh (Istio)
- OpenShift Serverless (Knative)
- OpenShift Monitoring (Prometheus)

Red Hat OpenShift

Red Hat AI

Align Red Hat AI Capabilities to Customer Challenges

Red Hat AI
Inference Server

Model Inference

- De Facto Standard Inference Engine with leading performance
- Inference Efficiency through performant Quantization
- Support and curation of leading Foundation Models

Red Hat
Enterprise Linux AI

Model Customization

- Flexible Model Selection and Performance/Cost: LLM vs SLM
- Configurable Model Alignment on Enterprise Data
- Protect Sensitive Data and Own the Model Inputs and Weights

Red Hat
OpenShift AI

Operate Models at Scale

- Hybrid platform across compute as well as accelerators
- Enterprise class distributed platform and operations
- Trusted and Tested ML & LLM Ops Platform

Red Hat

TechTalks

Thank you for joining!

Red Hat the Netherlands