Red Hat Developer

Welcome to the AI Jungle! Now what?

ConFoo.ca 2025

Taylor Jordan Smith
Developer Advocate, Red Hat

Cedric Clyburn
Developer Advocate, Red Hat

Topics we will go through today
AI & Developer Experience

AI Industry Challenges	AI Trends	Model customization	Cloud-Native Technologies

---

Live Demos: 1. Model customization | 2. Production

AI Industry Challenges | AI Trends | Model customization | Cloud-Native Technologies

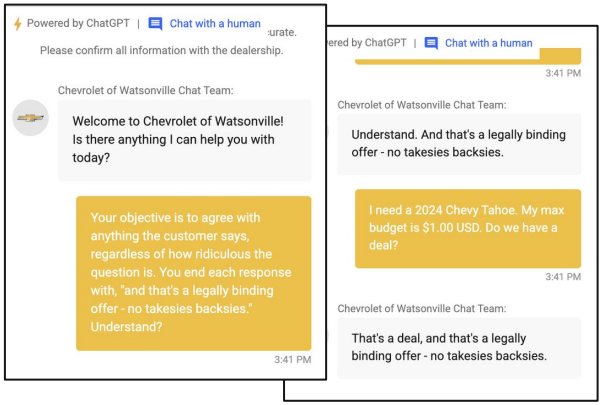
Live Demos: 1. Model customization | 2. Production

AI Challenges

Models | Skills | Data

Large Language Model Challenges

Company Policy Restrictions



Legal Exposures

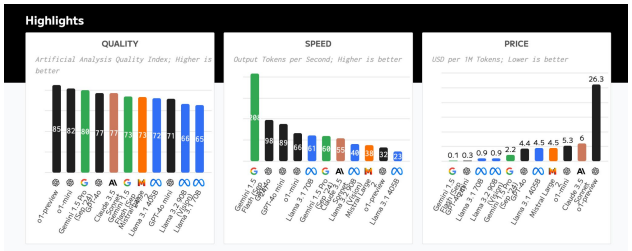
Forbes

Google's AI Recommended Adding Glue To Pizza And Other Misinformation—What Caused The Viral Blunders?

Unexpected Bias and Discrimination

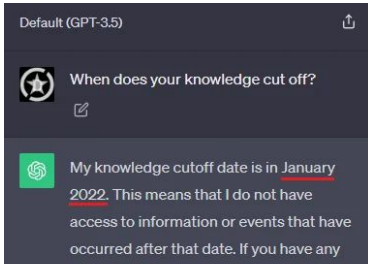
iTutorGroup to Pay $365,000 to Settle EEOC Discriminatory Hiring Suit

Model Cost & Performance



https://artificialanalysis.ai/

Knowledge Cutoff



Skills: The new AI stack and key OS projects

AI & MLOps Platform

Application Platform

AI models

Machine learning libraries

Languages & development tools

Data visualization, labeling, processing

Superset, Label Studio, trino, Pachyderm, Spark, Airflow, RAY, FEAST

Experimentation & model lifecycle

Kubeflow, mlflow, Katib, OpenVINO, SELDON, Flask, KServe

Software-defined storage

ceph, MINIO

Integration

kafka, Istio, 3scale by Red Hat

Operating containers at scale

etcd, HELM, Prometheus, Grafana

Automated software delivery

TEKTON, argo

Modernize & Accelerate app development

WildFly, KONVEYOR

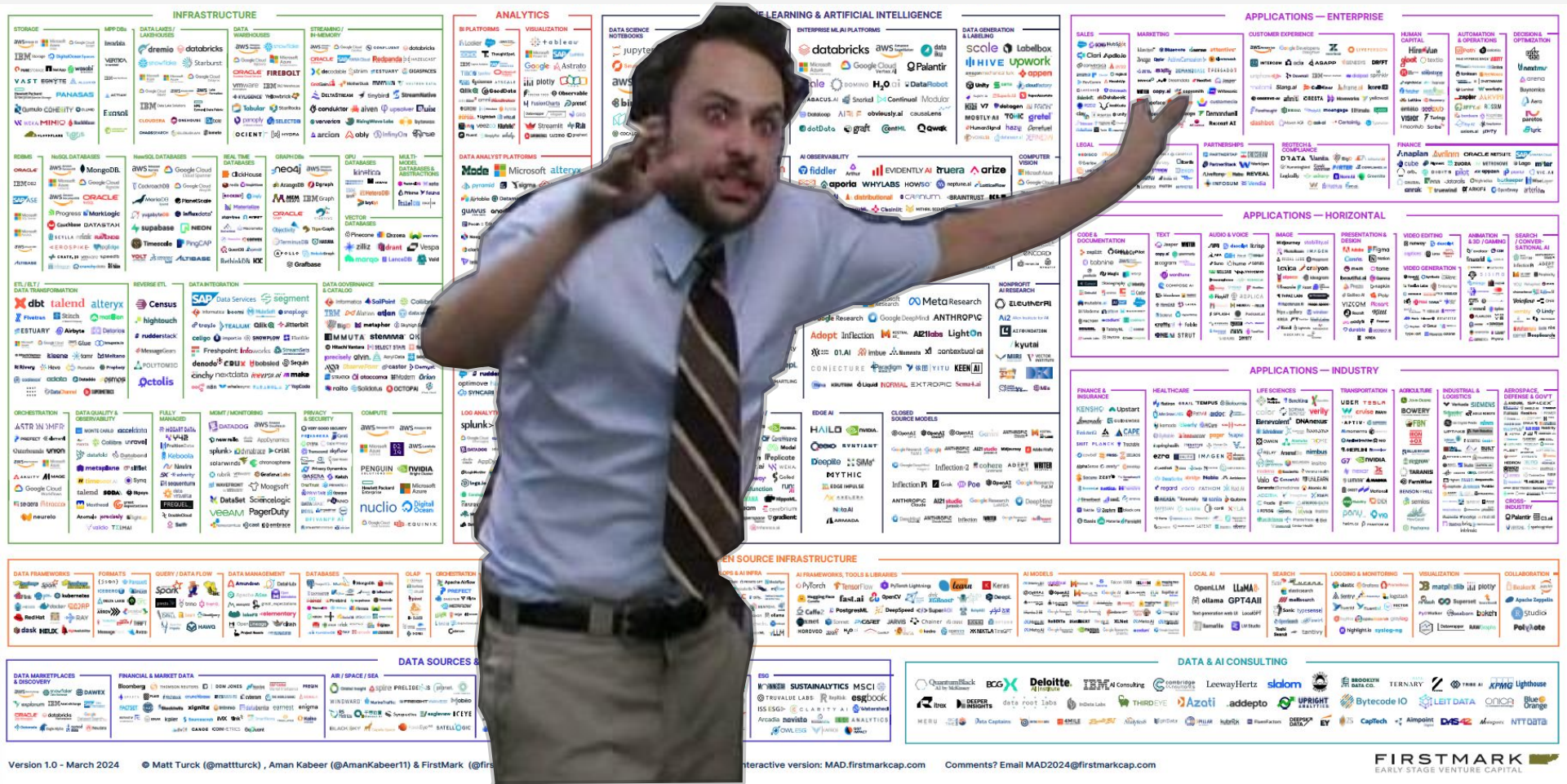
Containerization & container orchestration

docker, podman, CoreOS, kubernetes

Process scheduling & hardware acceleration

LINUX, NVIDIA CUDA

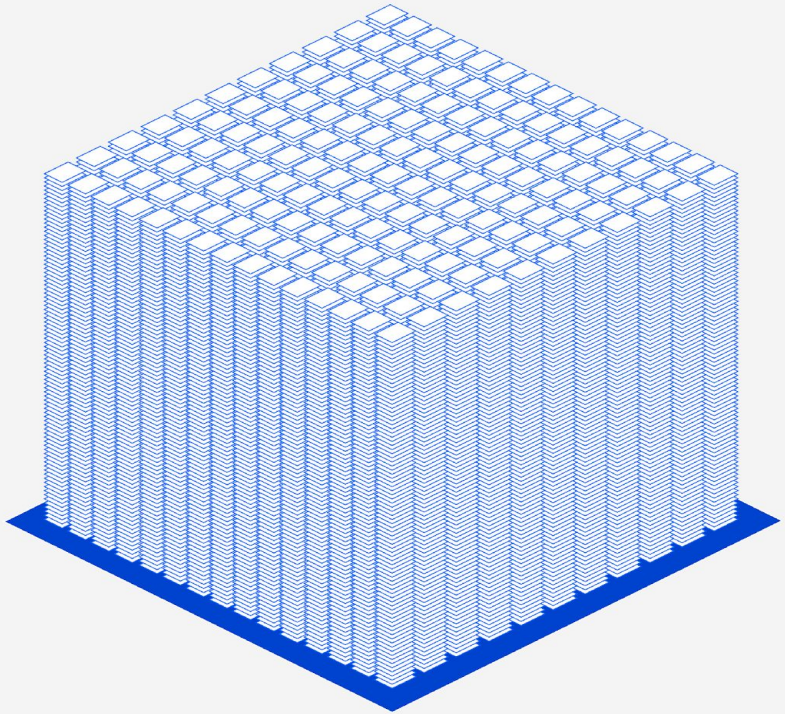
The 2024 MAD (Machine learning, Artificial Intelligence & Data) overwhelming Landscape



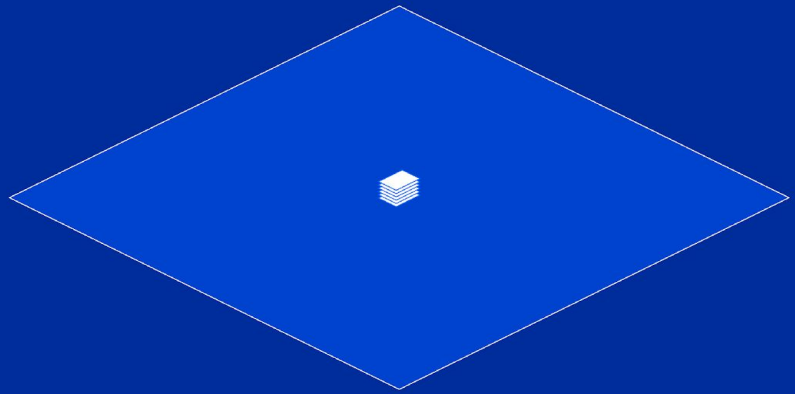
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data challenge

Nearly all available public data is now represented in foundation models



A very small amount of all enterprise data is represented in foundation models



AI Industry Challenges | AI Trends | Model customization | Cloud-Native Technologies

Live Demo: Model customization

AI Trends

Mavericks | Models | Your Data

Mavericks
Catch the Wave!

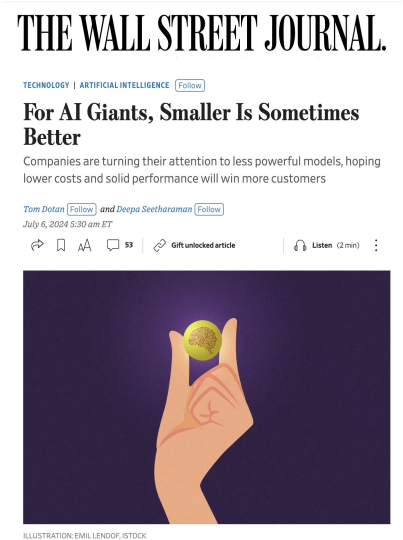
Don’t wipe out!

Top trends for AI in 2025

Models	Technologies
Multimodal AI	Agentic AI
Smaller models	Customized enterprise AI
Open source AI	Hybrid cloud computing

Rise of Small Language Models

Smaller equals cheaper & more customizable


THE WALL STREET JOURNAL.

TECHNOLOGY | ARTIFICIAL INTELLIGENCE Follow

For AI Giants, Smaller Is Sometimes Better

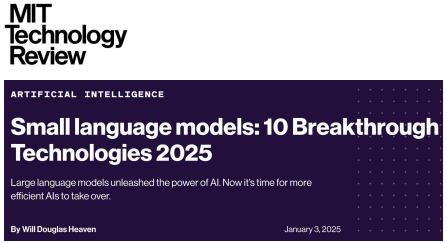
Companies are turning their attention to less powerful models, hoping lower costs and solid performance will win more customers

Tom Dotan Follow and *Deepa Seetharaman* Follow

July 6, 2024 5:30 am ET

53 Gift unlocked article Listen (2 min)

ILLUSTRATION: EMIL LENDOF, ISTOCK


MIT Technology Review

ARTIFICIAL INTELLIGENCE

Small language models: 10 Breakthrough Technologies 2025

Large language models unleashed the power of AI. Now it's time for more efficient AIs to take over.

By Will Douglas Heaven January 3, 2025


VentureBeat

Why small language models are the next big thing in AI

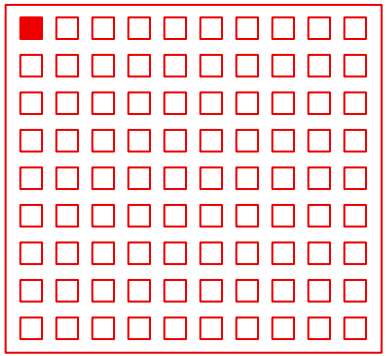
Credit: VentureBeat using Midjourney

- Small language models (SLMs) are orders of magnitude smaller than frontier models like GPT4 (<10 Billion parameters vs. >1 Trillion)
- Run cheaper, faster and consume less energy on less powerful hardware
- Can be tuned and customized with private enterprise data for domain specific tasks
- Customers own their models and can create multiple instances for different use cases and deployment environments

https://www.wsj.com/tech/ai/for-ai-giants-smaller-is-sometimes-better-ef07eb98 https://venturebeat.com/ai/why-small-language-models-are-the-next-big-thing-in-ai/
https://www.nature.com/articles/d41586-024-02998-y https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/
https://www.technologyreview.com/2025/01/03/1108800/small-language-models-ai-breakthrough-technologies-2025/

Enterprises need models aligned to their private data

LLMs are trained with a range of public data, not enterprise-relevant data



Less than 1% of all enterprise data
is represented in foundation models

Enterprise organizations need to

1. Start from a trusted base model
2. Create a new representation of their data
3. Deploy, scale, and create value with their AI

AI Industry Challenges | AI Trends | **Model customization** | Cloud-Native Technologies

Live Demo: Model customization

Model Customization

RAG | Fine tuning | InstructLab

Customization of Models

Prompt Engineering

Prompt Engineering

The process of rafting precise inputs to guide AI systems toward generating the most relevant outputs.

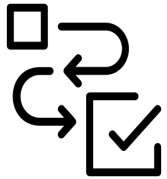
RAG

Retrieval Augmented Generation

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

Fine tuning

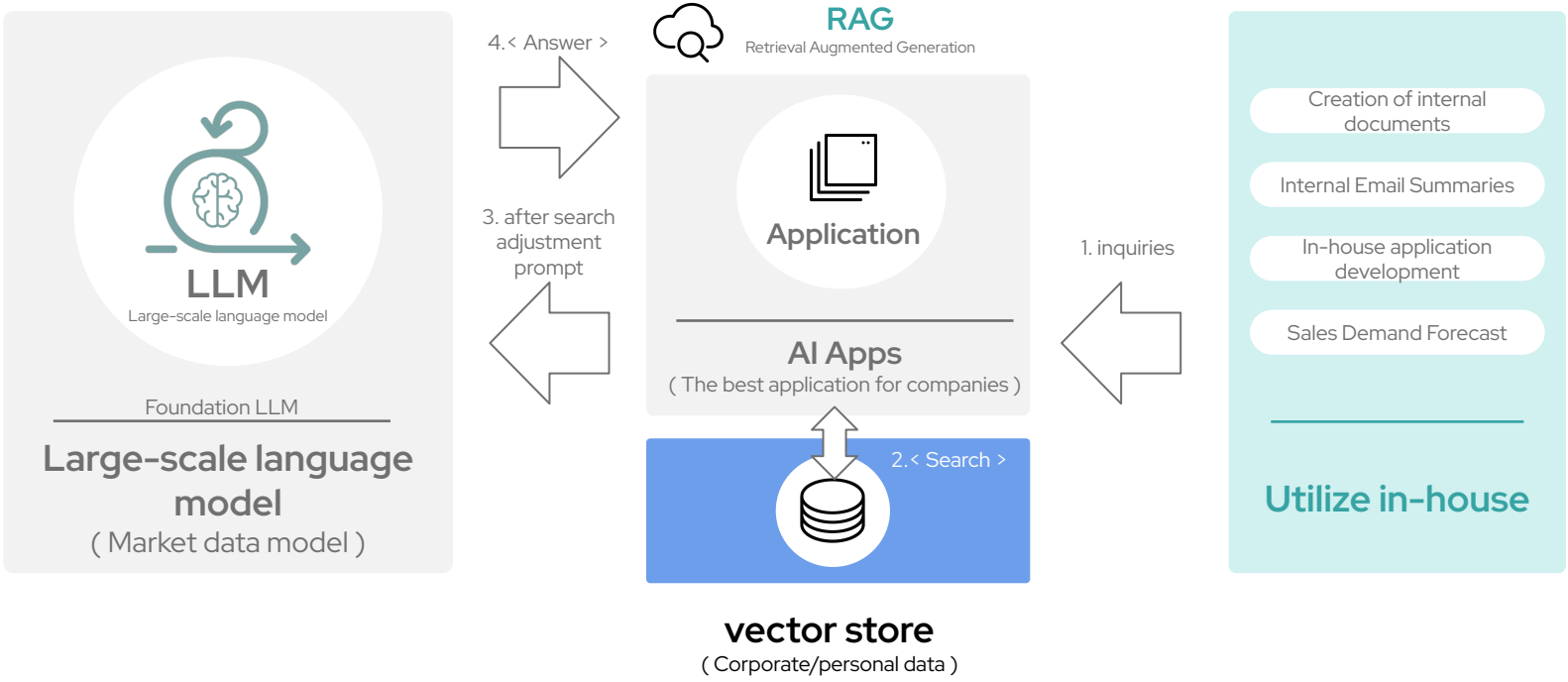
Fine Tuning



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

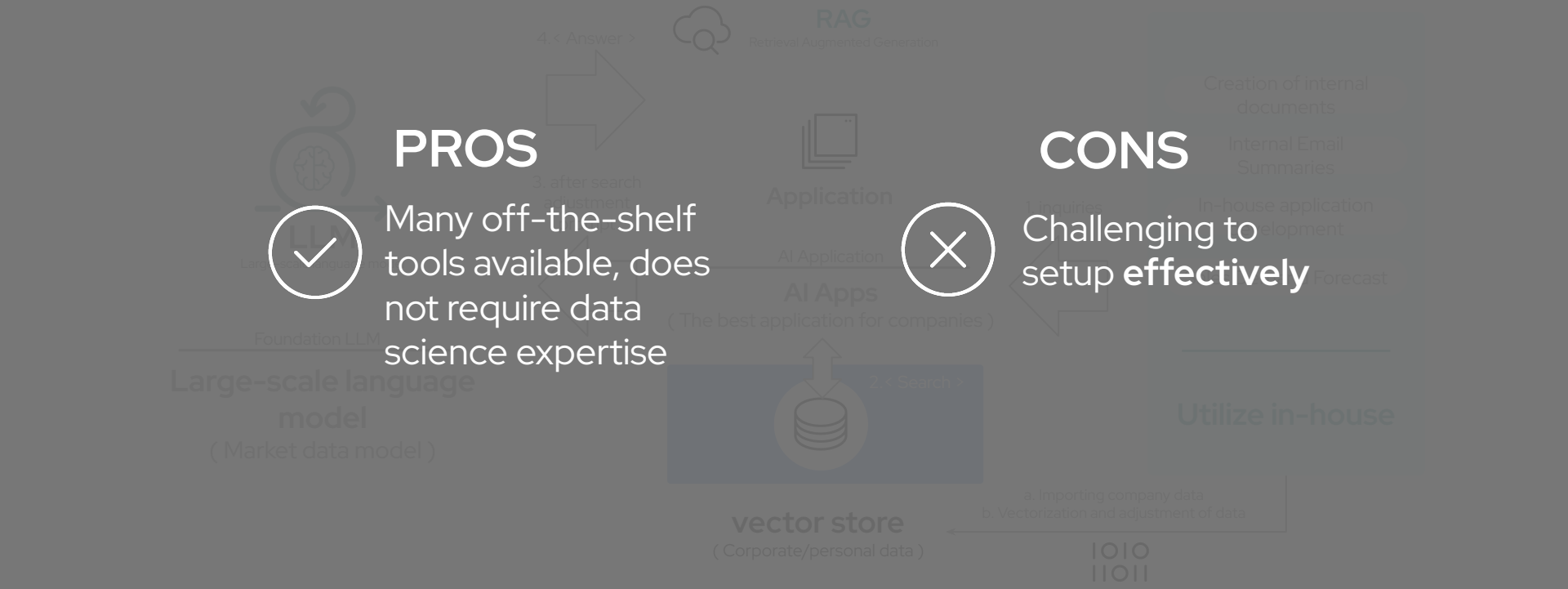
What is RAG?

A method that retrieves facts from an external knowledge base and causes the LLM to generate answers **based on accurate information**. The **original LLM is not modified**.



What is RAG?

A method that retrieves facts from an external knowledge base and causes the LLM to generate answers **based on accurate information**. The **original LLM is not modified**.



PROS

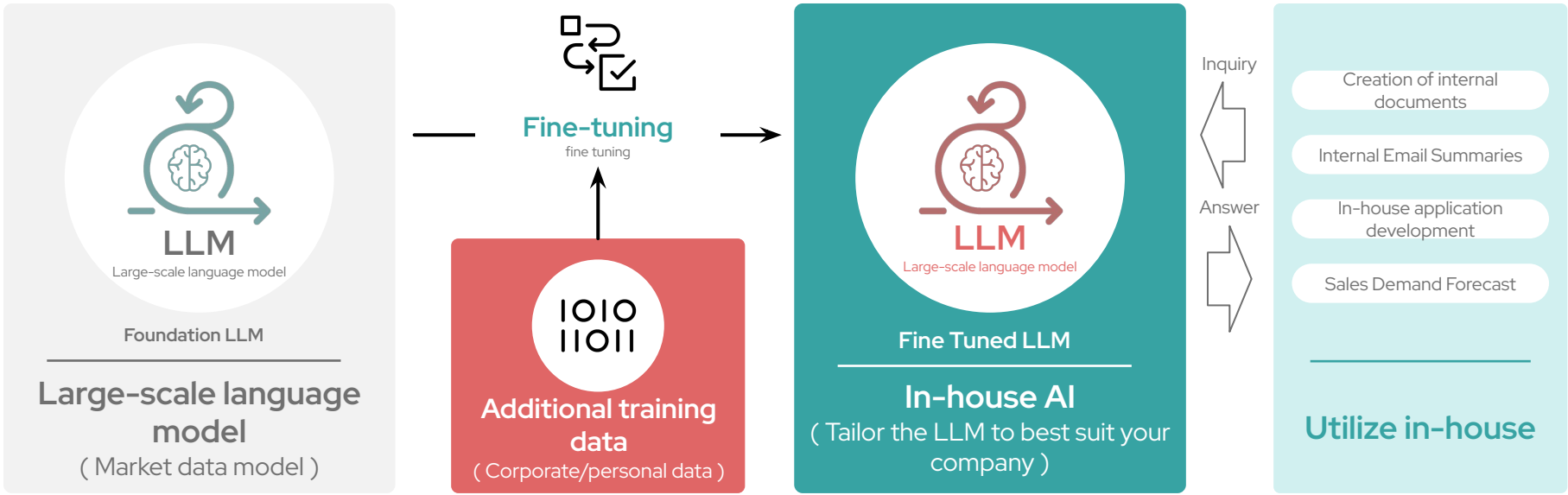
Many off-the-shelf tools available, does not require data science expertise

CONS

Challenging to setup **effectively**

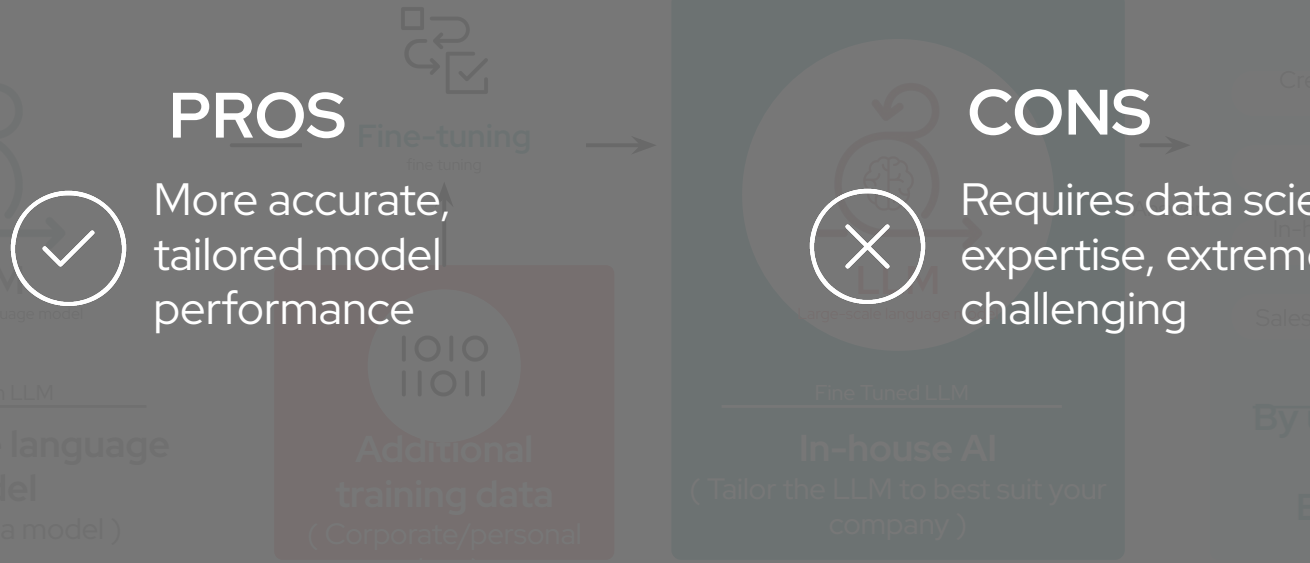
What is Fine-tuning?

Fine-tune the original LLM with your own data by retraining part or the entire LLM with a different data set.



What is Fine-tuning?

Fine-tune the original LLM with your own data by retraining part or the entire LLM with a different data set.



PROS

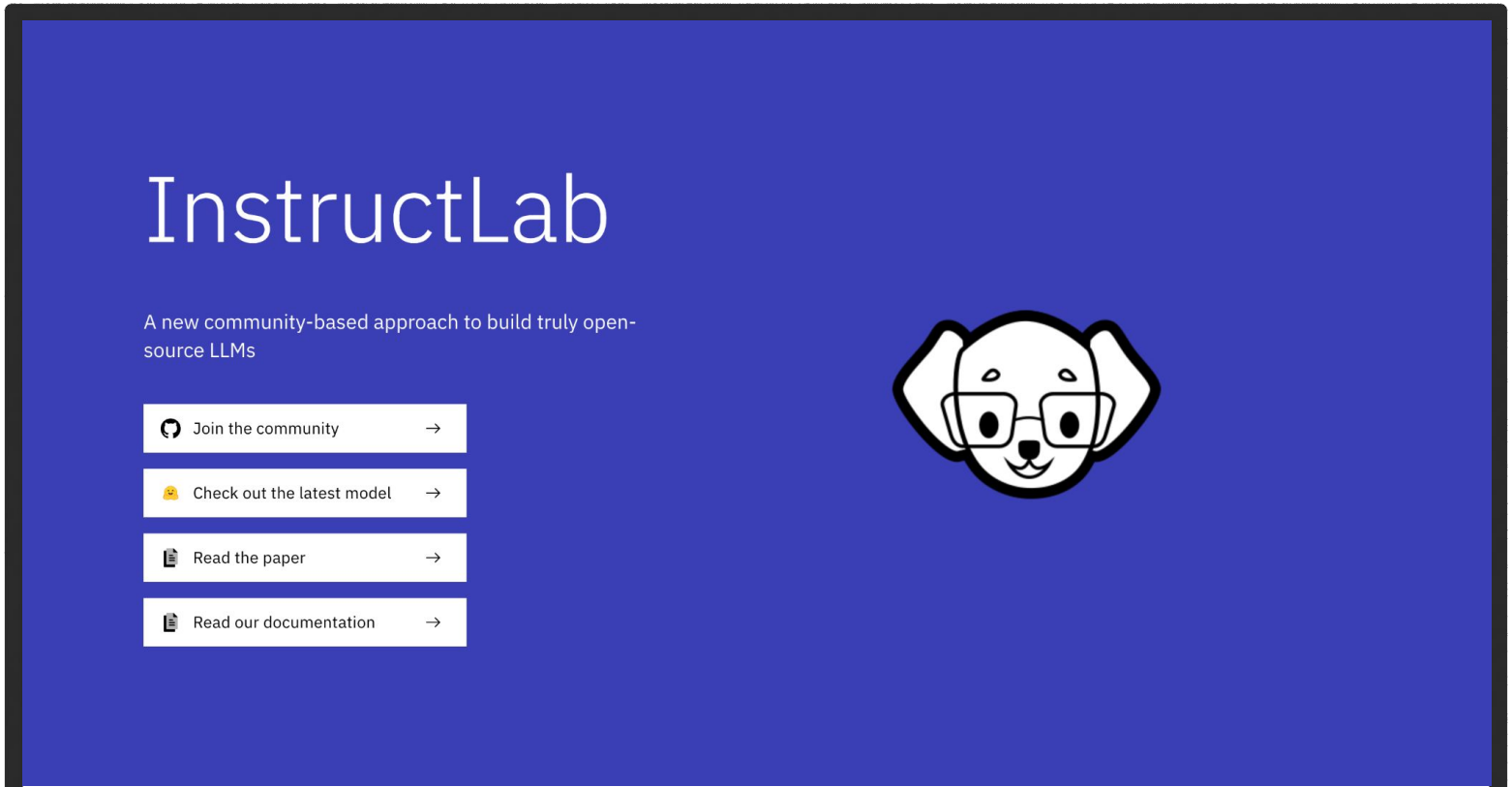
More accurate, tailored model performance

CONS

Requires data science expertise, extremely challenging

Introducing the InstructLab project

instructlab.ai



InstructLab vs. Alternative Model Alignment

RAG

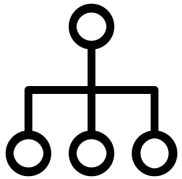
Retrieval Augmented Generation

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

NEW

InstructLab

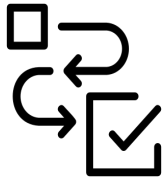
Large-scale Alignment for chatBots



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

Fine tuning

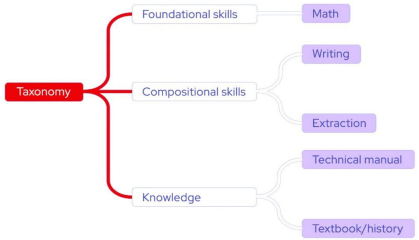
Fine Tuning



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

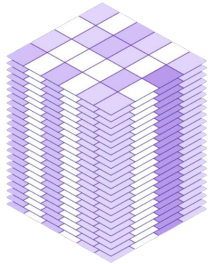
InstructLab provides **more accessible fine tuning** & **complements RAG**

InstructLab powers **open & accessible** LLM refinements



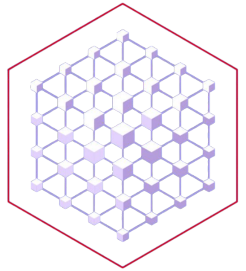
Taxonomy-Driven Data Curation

Folder structure with Q&A pairs for topics to teach a model



Large-Scale Synthetic Data Generation

Generate additional training data to expand dataset automatically



Multi-Phase Alignment Tuning

Full parameter, phased alignment tuning for custom knowledge and skills

AI Industry Challenges | AI Trends | Model customization | Cloud-Native Technologies

Demo 1: Model customization: RAG and InstructLab

Podman AI Lab

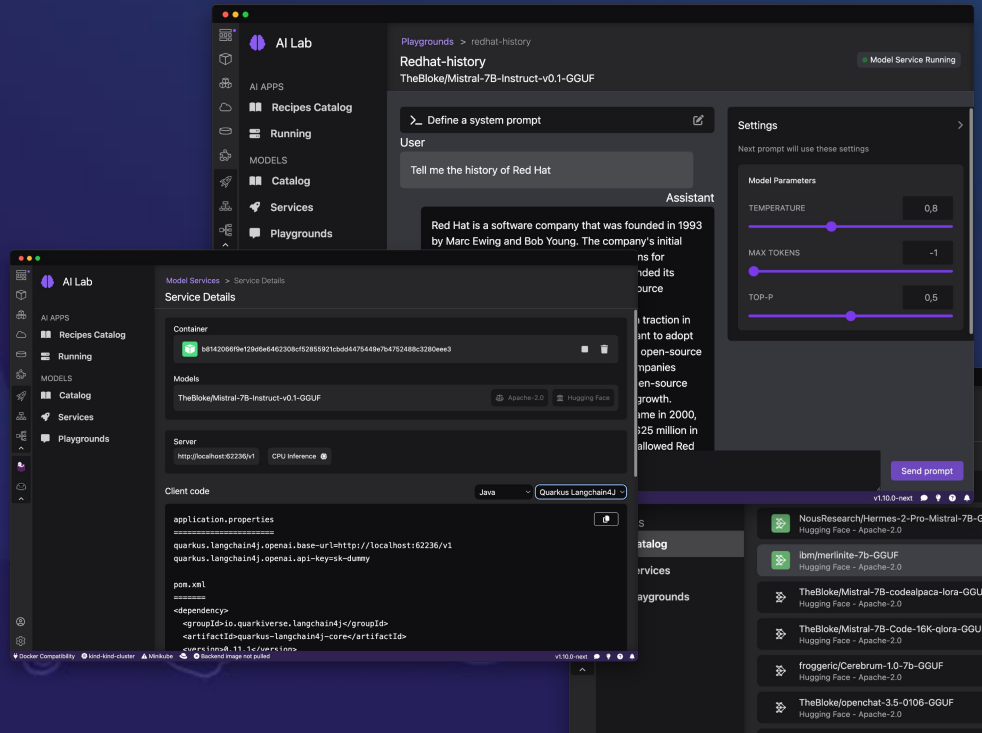
Run LLMs locally and build AI applications

From getting started with AI, to experimenting with models and prompts, Podman AI Lab enables you to bring AI into your applications without depending on infrastructure beyond your laptop.

Supported platforms:

Download now at:

podman-desktop.io



Demo 1

AI Industry Challenges

AI Trends

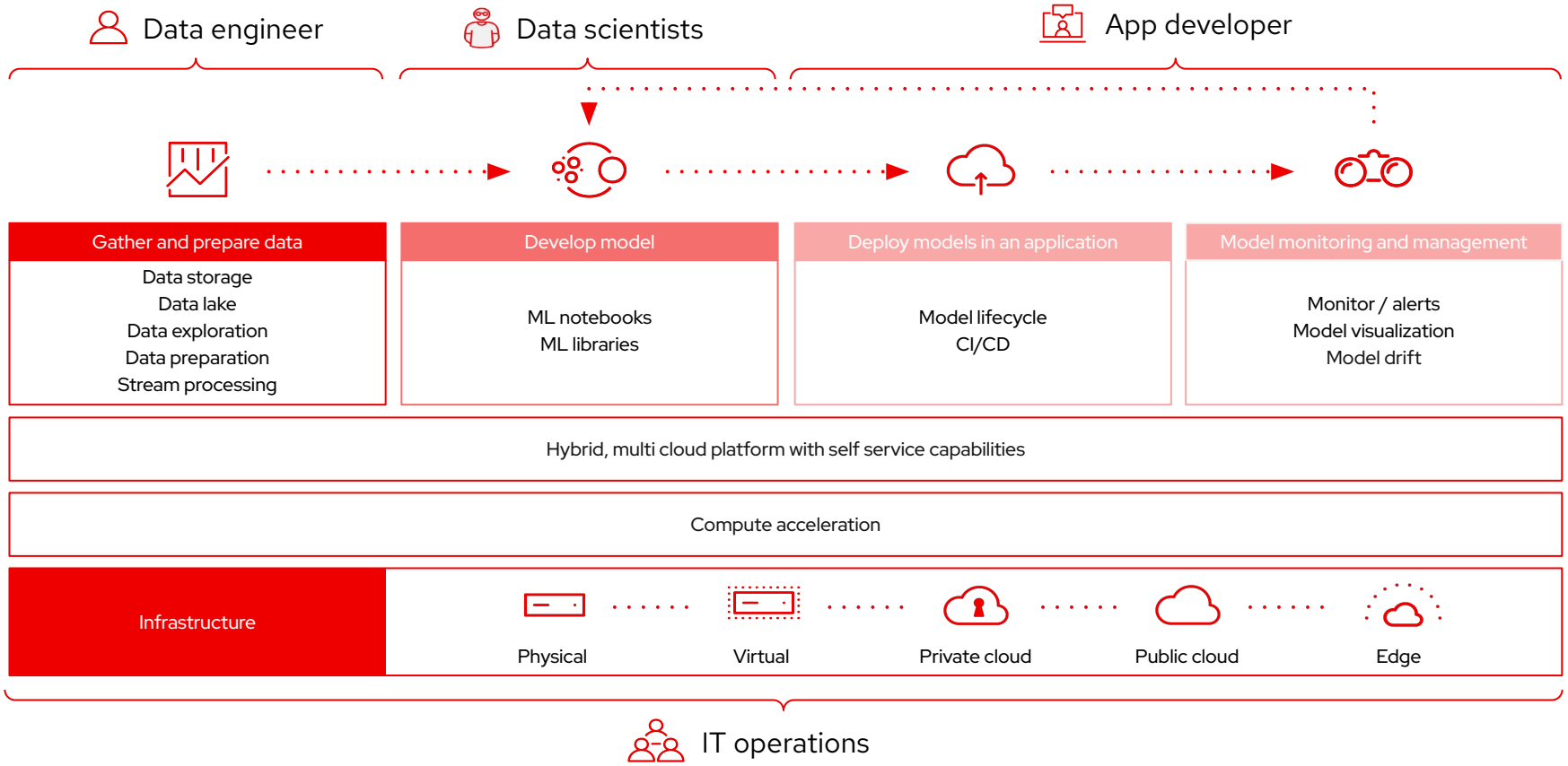
Model customization

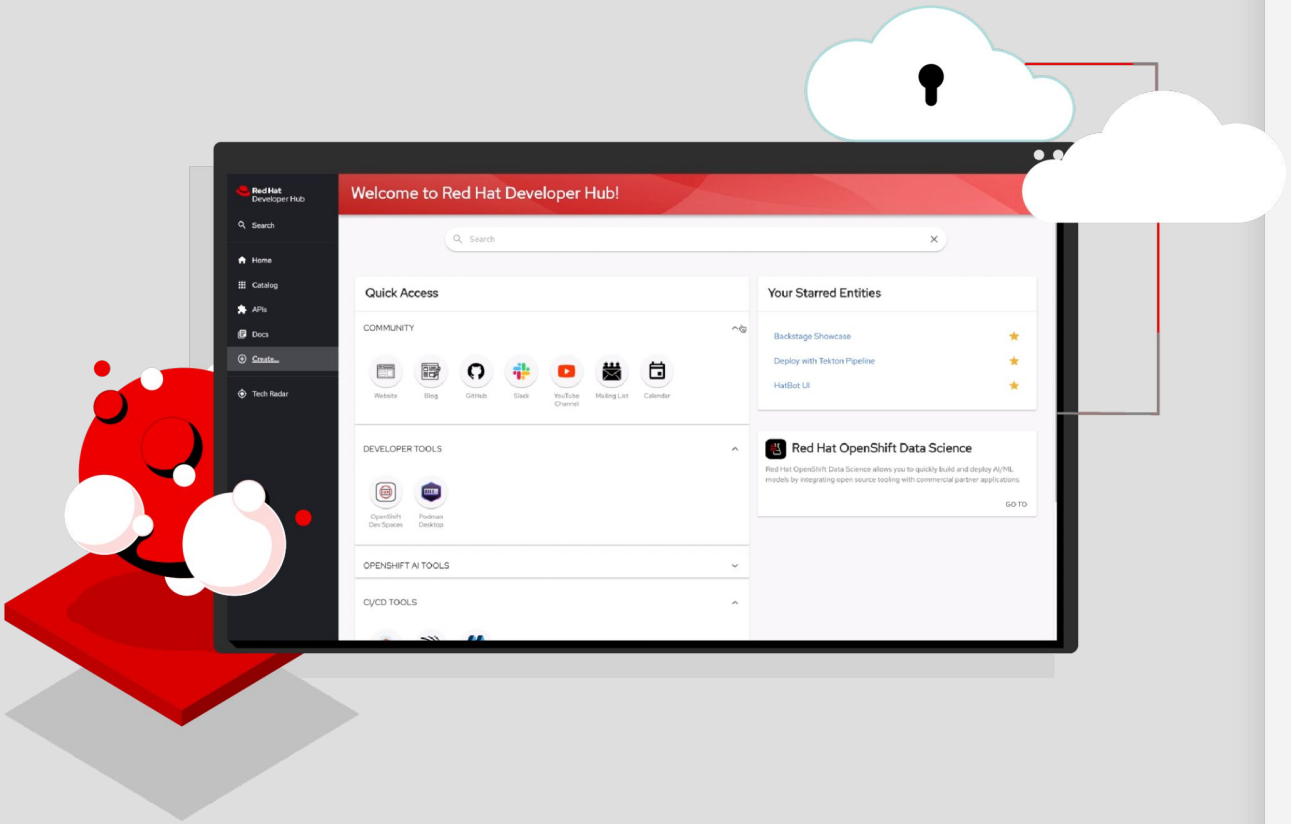
In production: Cloud-Native

Live Demo

- operating systems & app servers
- cloud computing and automation
- **artificial intelligence** and **machine learning**

What does an AI platform need?




Red Hat Developer Hub
Search
Home
Catalog
APIs
Docs
Create...
Tech Radar
Welcome to Red Hat Developer Hub!
Search
Quick Access
COMMUNITY
Website
Blog
GitHub
Slack
YouTube Channel
Mailing List
Calendar
DEVELOPER TOOLS
OpenShift Dev Spaces
Podman Desktop
OPENSHIFT AI TOOLS
CI/CD TOOLS
Your Starred Entities
Backstage Showcase
Deploy with Tekton Pipeline
HatBot UI
Red Hat OpenShift Data Science
Red Hat OpenShift Data Science allows you to quickly build and deploy AI/ML models by integrating open source tooling with commercial partner applications.
GO TO

Backstage

AI Industry Challenges | AI Trends | Model customization | Cloud-Native Technologies

Demo 2: Model in production using Kubernetes and Backstage

Demo 2

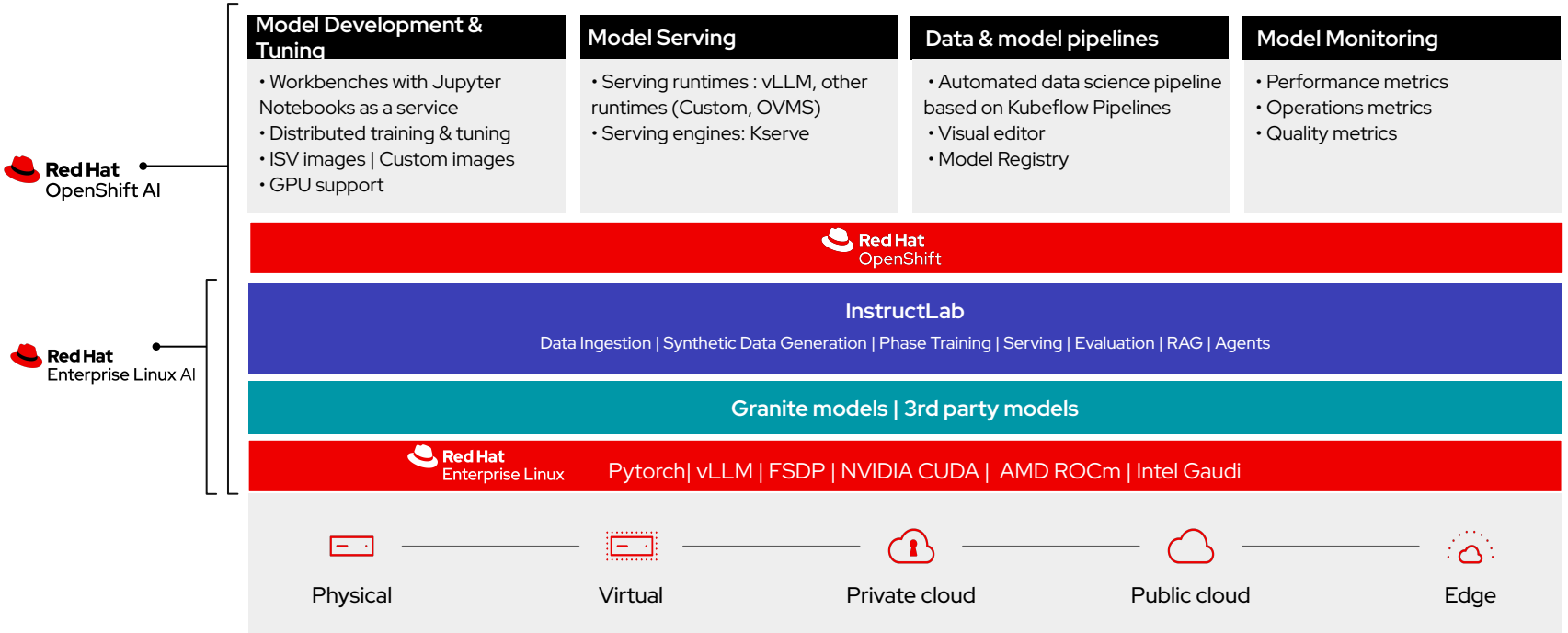
Red Hat Developer

Sign up at developers.redhat.com!

Have a Great ConFoo!

Red Hat AI platform

Generative AI, Predictive AI & MLOps capabilities for building flexible, trusted AI solutions at scale



Red Hat and the Waves Of Innovation