Red Hat Summit

Connect

Your AI, Your Data, Your Rules

Navigating the Future of Machine Learning with Control and Compliance

Michael Bang
Associate Principal Solutions Architect

Henrik Løvborg
Tech Sales Leader Denmark

Michael Bang

Associate Principal Solutions Architect
Red Hat

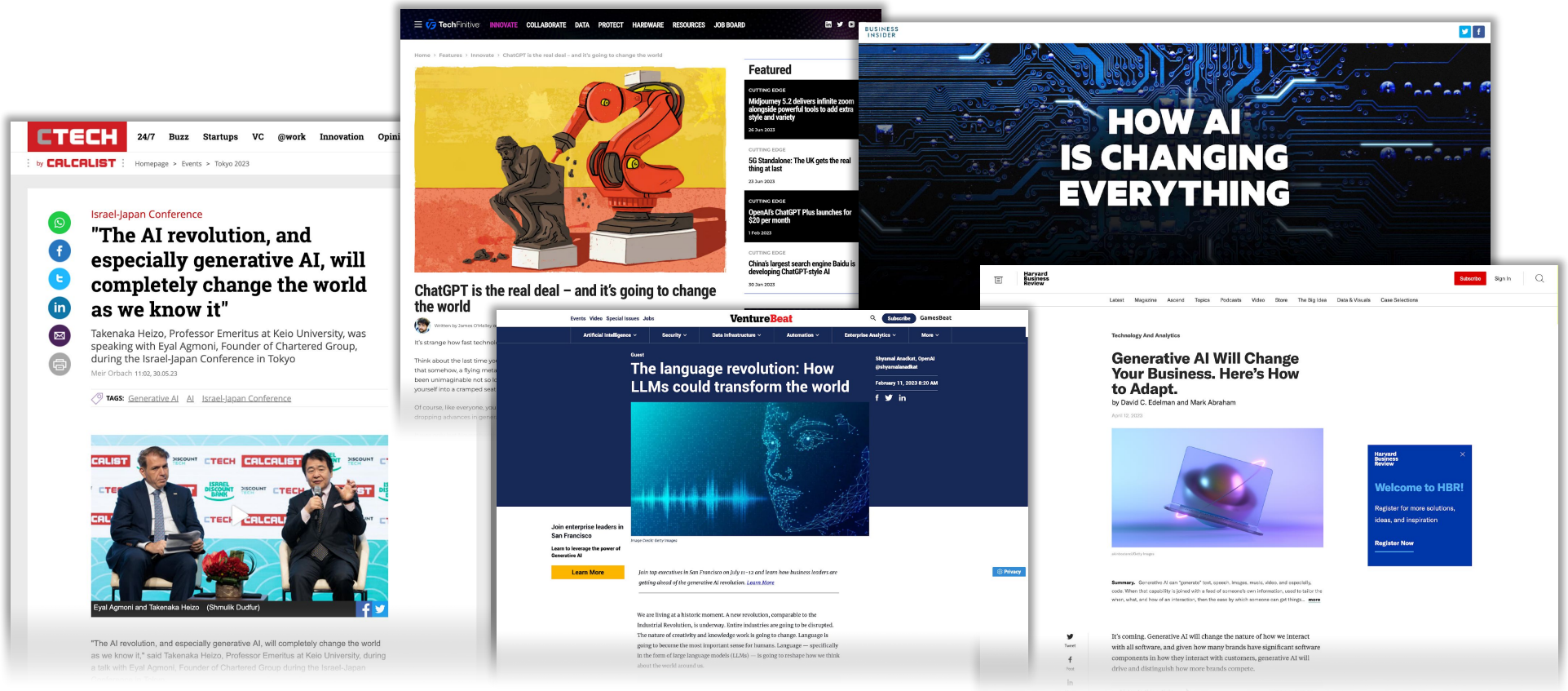
Henrik Løvborg

Tech Sales Leader Denmark
Red Hat

Navigating The AI Era

The World Changed in November 2022

ChatGPT woke the world up to the power of generative AI



Red Hat

We need AI

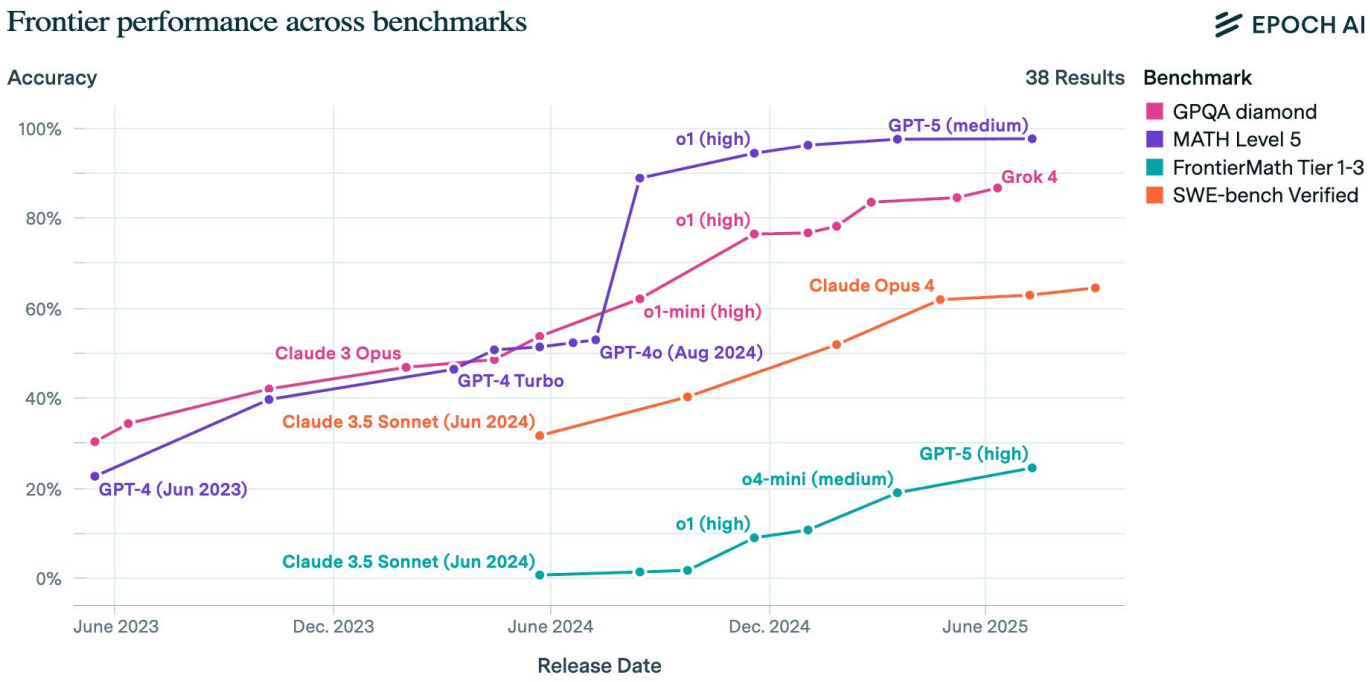
The CLOUD Act: Unveiling European Powerlessness

Emmanuelle Mignon
Responsable du département Public Réglementaire Environnement chez August Debouzy

Your AI conversations are a secret new treasure trove for marketers

The Register | 15 hours ago

ai-pocalypse Profound is a startup that promises to help companies understand how they appear in AI responses to customer queries. But one expert in the field thinks the AI analytics startup has been sucking up information on users' AI conversations without proper consent.



Red Hat

Regulating AI: The EU-AI Act

Unregulated, irresponsible or abusive use of AI could lead to negative consequences for individuals or the society, create public opposition and **hinder AI innovation in the EU**.

The EU is committed to strive for a balanced approach to AI

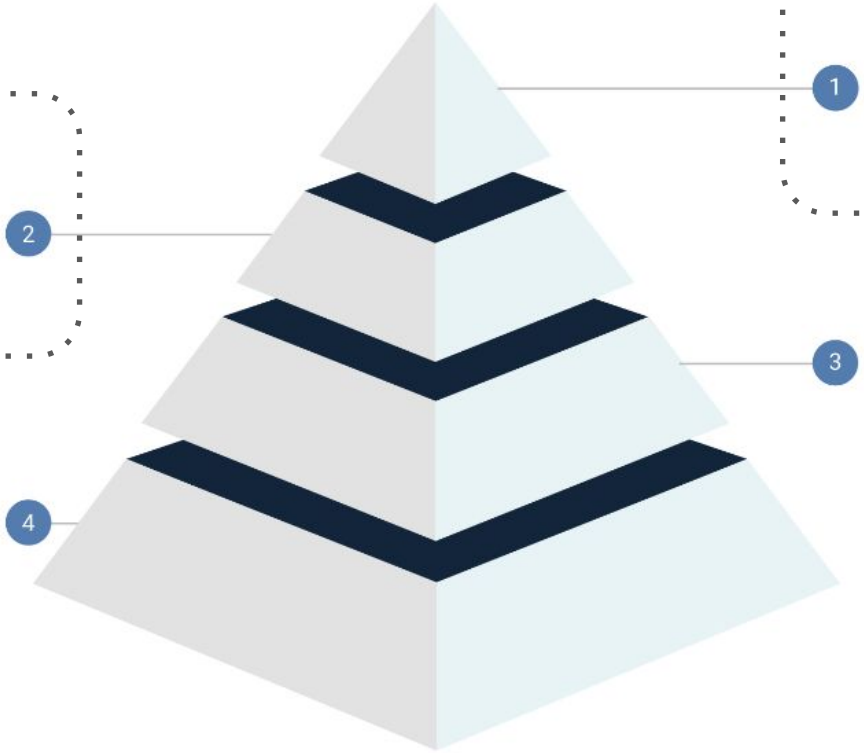
- Lawful → accurateness
- Ethical → transparency
- Robust → fairness
- → no (unintended) bias
- → security

EU AI Act Requirements:

Explainability, Documentation,
Process & Data Governance,
Human Oversight,
Risk Management, Auditability.

There are some exceptions for AI systems released under **Open Source** licenses.

banned:

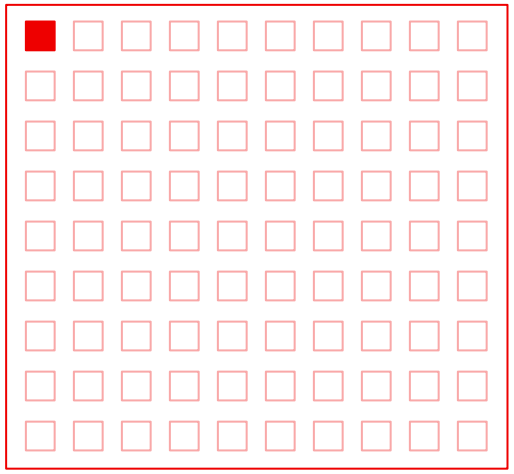


https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Red Hat

Enterprises need models aligned to their private data

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

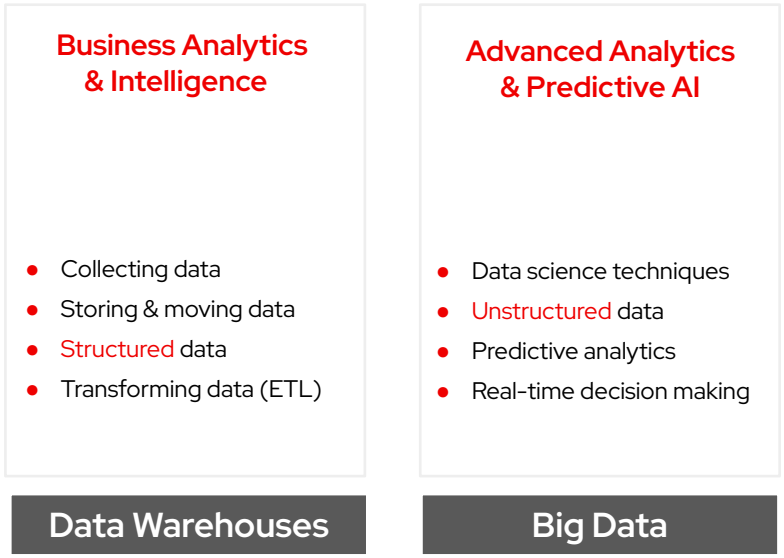


Less than 1% of all enterprise data
is represented in foundation models

1. Bloated models
2. Non enterprise data
3. Contextual uninteresting data

Data driven AI

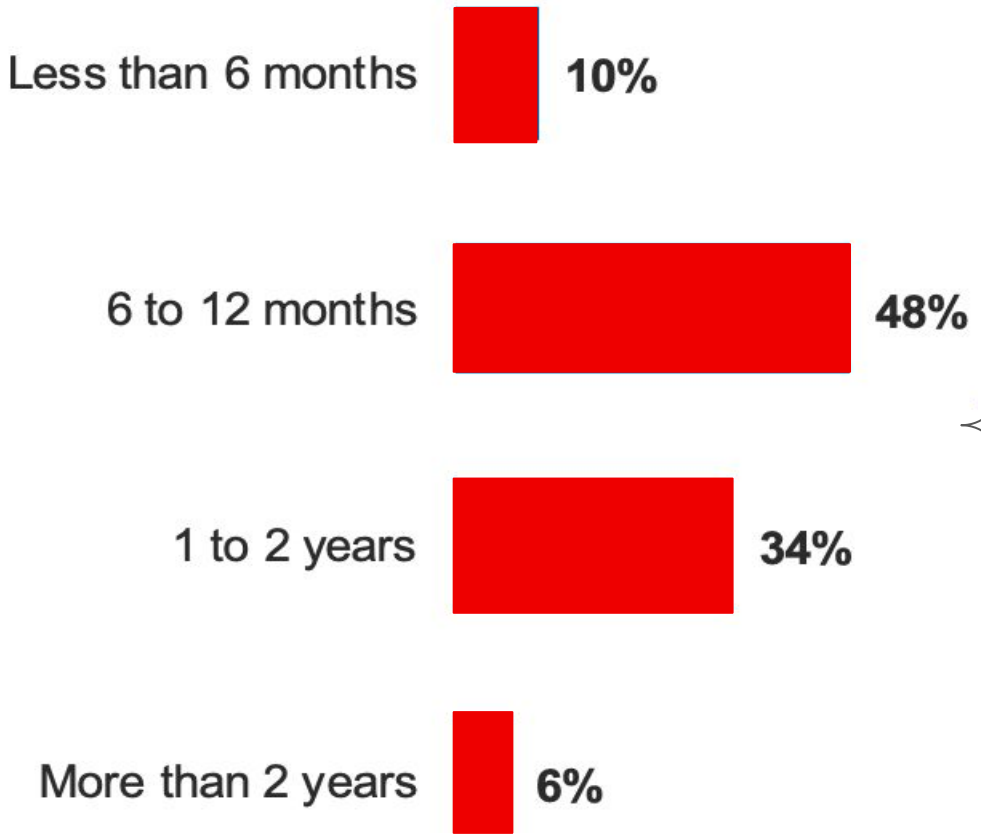
Data is the fuel for AI



- Most data is created and kept on-premise
- Where do we train and host models ?
- (Do) we have enough data ?!

Red Hat

Getting AI into production is hard



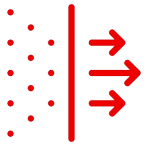
>80% of IT Decision Makers report that it takes 6 months to 2 years for their company to move from pilot to full production

Source: ROI of AI, IBM in collaboration with Lopez Research, December 2024

Generative AI customer adoption challenges

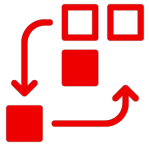
Cost

Generative AI frontier model services are cost prohibitive at scale for most enterprise customer use cases.



Complexity

Tuning models with private enterprise data for customer use cases is too complex for non-data scientists.



Flexibility

Enterprise AI use cases span data center, cloud & edge and can't be constrained to a single public cloud service.

What then?

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 Google aws IBM

Physical

Virtual

Private Cloud

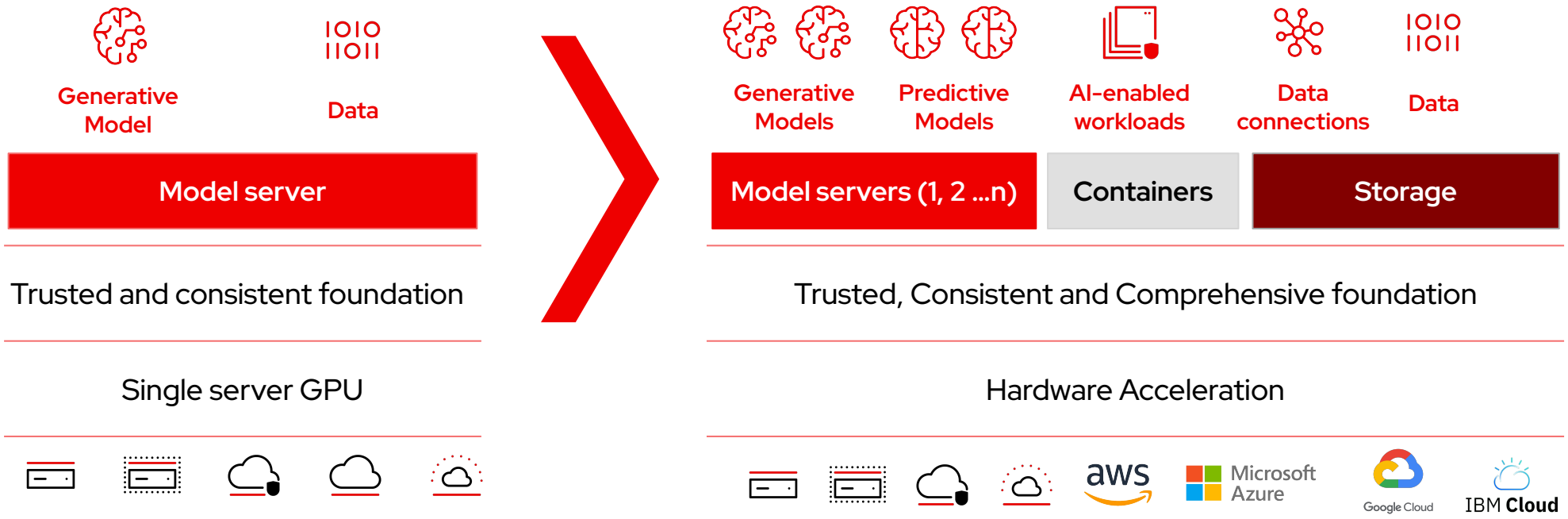
Public Cloud

Edge

* NVIDIA, AMD, Intel, Google TPU supported in Red Hat AI. AWS Inferentia/Neuron IBM AIU are on our roadmap

Red Hat AI supports each stage of the AI adoption journey

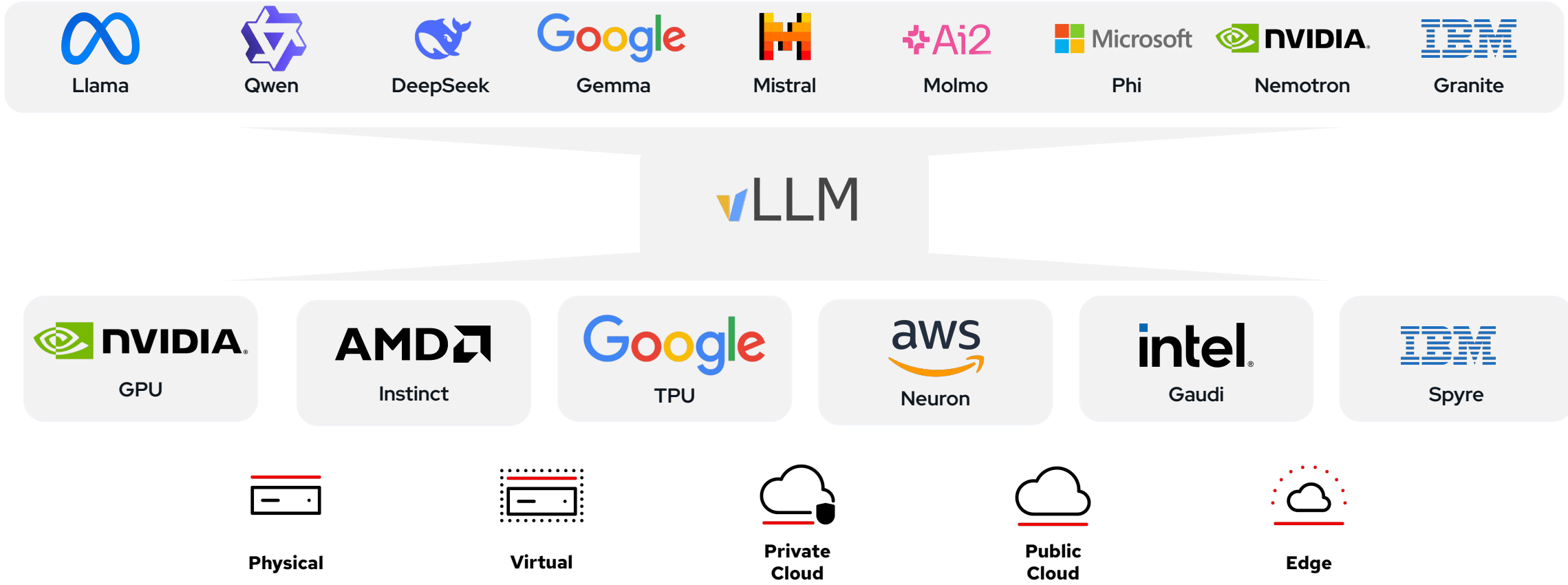
From single server deployments to highly scaled-out platform architectures



Red Hat

Red Hat AI the inference engine for the hybrid cloud

vLLM supports the key models on the key hardware accelerators



Red Hat

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

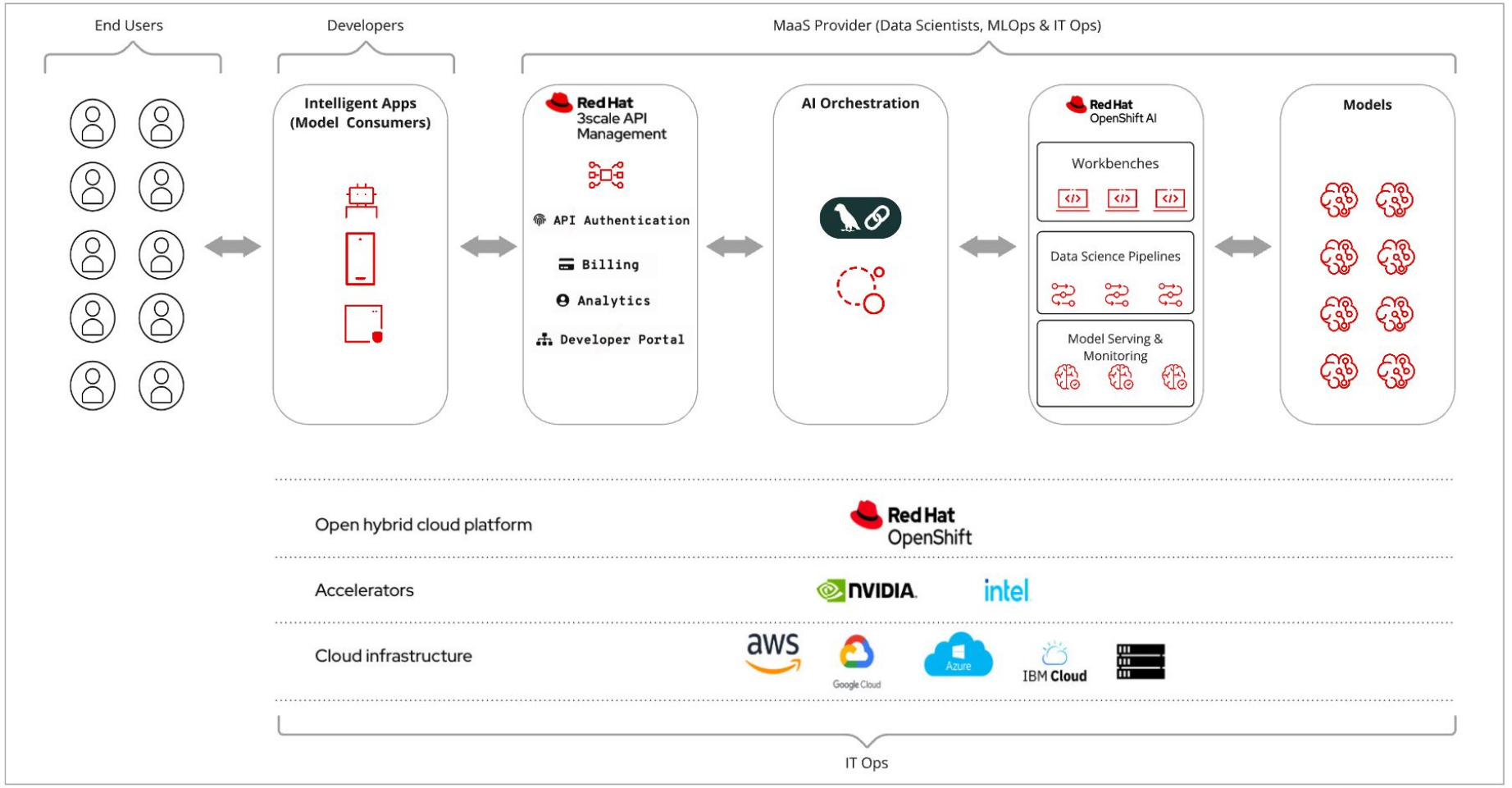
Validated models

- 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

Model as a Service - MaaS



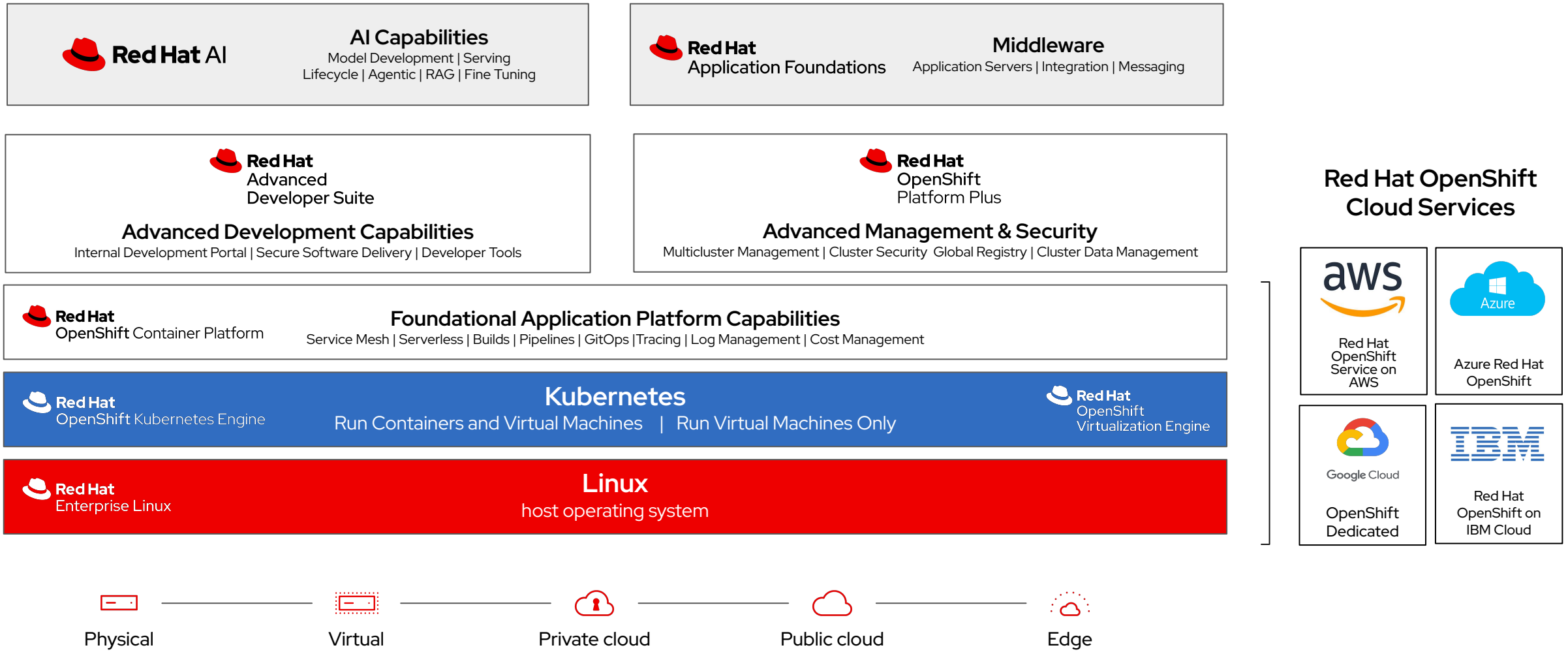
TrustyAI 

Validating our models

For compliance and efficiency

- **Bias Monitoring:** Ensure that your models are providing consistent results across all demographic groupings.
- **Drift Monitoring:** Verify that your real-world inbound data match the data that your model was trained on.
- **Explainability:** Produce human-readable explanations and justifications of model behavior.
- **Guardrails:** Moderate interaction between generative models and users
- **LM-Eval:** Evaluate language models over a variety of tasks and benchmarks , such as logical reasoning or toxic language production

Red Hat OpenShift and Open Hybrid Cloud



Red Hat

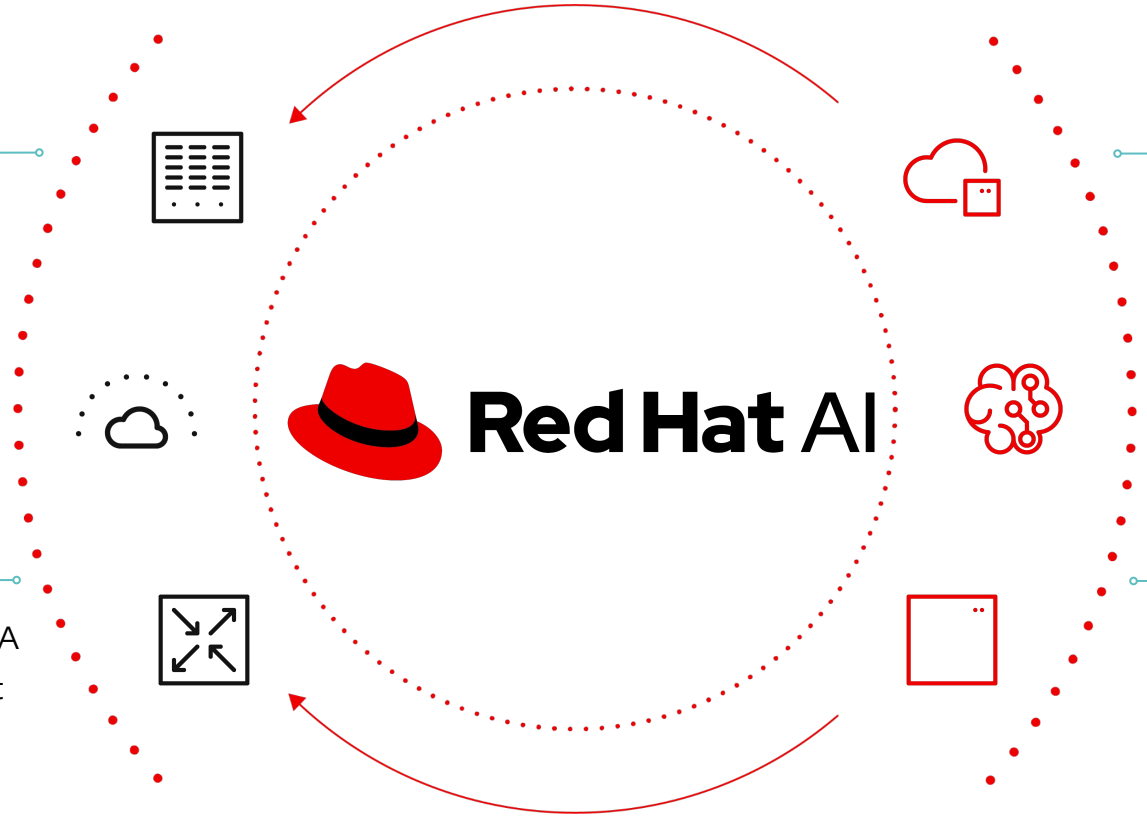
Red Hat AI Announcements

Fast and Efficient Inference

- Introducing Red Hat AI Inference Server
- LLM Compressor tools
- Announcing llm-d for inference at scale

Connecting Data to Models

- Red Hat AI InstructLab on IBM Cloud now GA
- InstructLab distributed training in OpenShift AI with Kubeflow Training Operator
- InstructLab multi-language capabilities



AI Platform

- Red Hat AI Validated models
- Model catalog and feature store in OpenShift AI (Tech Preview)
- RHEL AI now available in Google Cloud Marketplace

Agentic AI

- Llama Stack (Dev Preview)
- Model Context Protocol (Dev Preview)

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

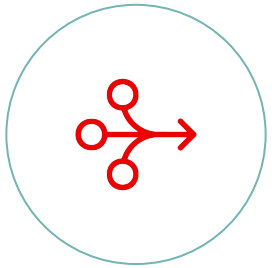
Red Hat

The value of Red Hat OpenShift AI

What differentiates us?

Simplify AI adoption

Promotes freedom of choice and access to latest innovation on AI/ML technologies



Drive AI/ML operational consistency

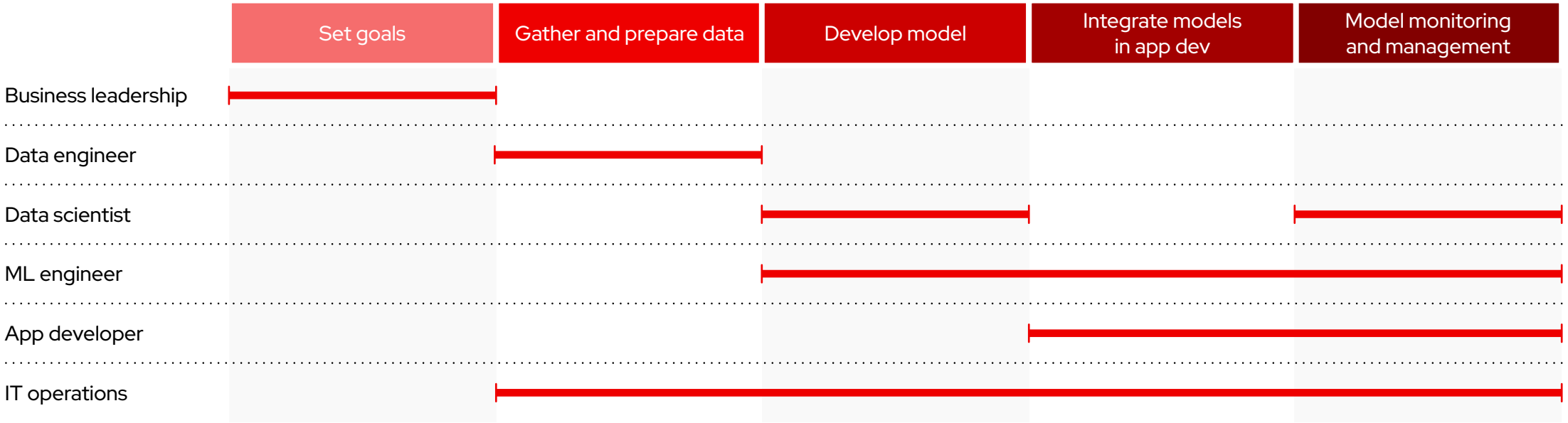
Streamline the process of moving models from experiments to production

Gain hybrid cloud flexibility

Deploy models in containerized format across on-prem, clouds and edge, including disconnected environments

Operationalizing AI/ML requires collaboration

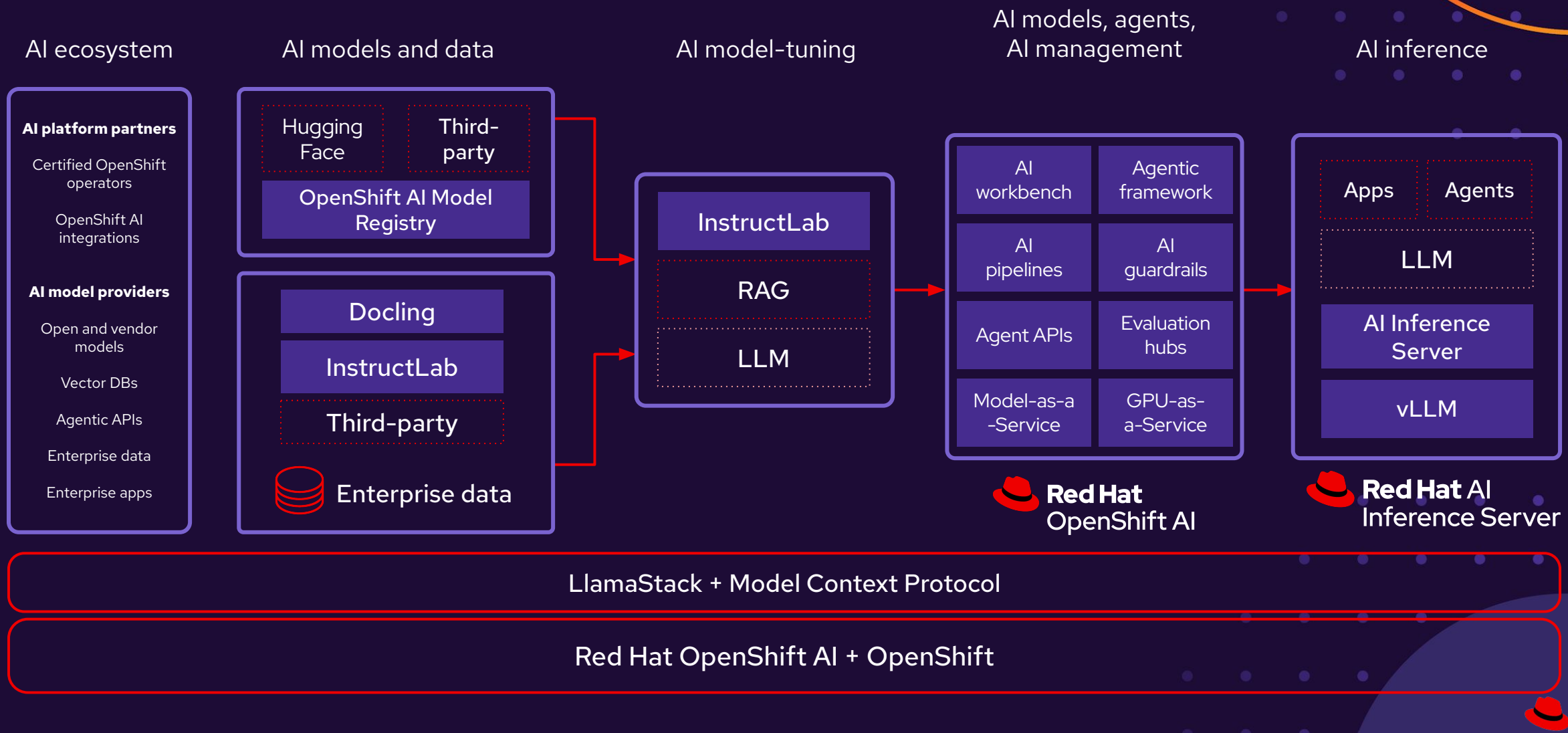
Every member of your team plays a critical role in a complex process



Red Hat

Red Hat AI

Open, Agentic, Enterprise-Ready



OpenShift AI
Architecture, Deployment, and Real-World Usage

Speaker Image

Lars Kromann

Container Service Architect/Solution specialist
JN Data

Speaker Image

Martin Mogensen

AI Specialist at AI Innovation Lab
JN Data

14:00 - Room A

Red Hat Summit

Connect

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos


twitter.com/RedHat