Red Hat Summit

Connect

GenAI mit Parasol AI Studio

Hands-On Day Darmstadt 18.11.2025

Jochen Cordes
Solution Architect
EMEA TelCo CoE
Red Hat

Alessandro Arrichiello
Solution Architect
EMEA TelCo CoE
Red Hat

Jochen Cordes

Solution Architect
EMEA TelCo CoE
Red Hat
https://www.linkedin.com/in/jochencordes/

Alessandro Arrichiello

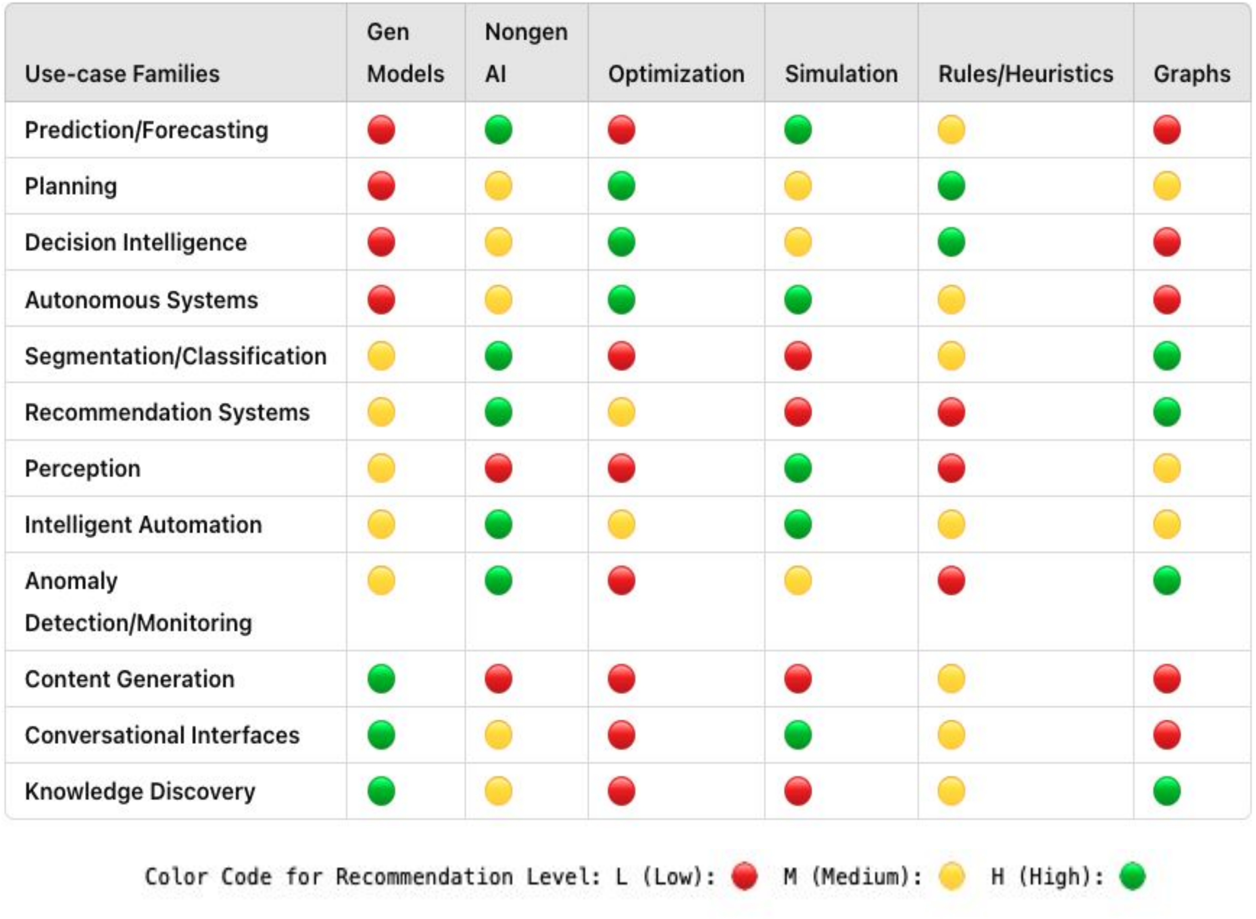
Solution Architect
EMEA TelCo CoE
Red Hat

Any Model
Any Accelerator
Any Cloud

Why Any Model?

The Myth of Universal Solutions

Model choice depends on use-case, precision, performance and resource constraints



Use-case Families	Gen Models	Nongen AI	Optimization	Simulation	Rules/Heuristics	Graphs
Prediction/Forecasting	🔴	🟢	🔴	🟢	🟡	🔴
Planning	🔴	🟡	🟢	🟡	🟢	🟡
Decision Intelligence	🔴	🟡	🟢	🟡	🟢	🔴
Autonomous Systems	🔴	🟡	🟢	🟢	🟡	🔴
Segmentation/Classification	🟡	🟢	🔴	🔴	🟡	🟢
Recommendation Systems	🟡	🟢	🟡	🔴	🔴	🟢
Perception	🟡	🔴	🔴	🟢	🔴	🟡
Intelligent Automation	🟡	🟢	🟡	🟢	🟡	🟡
Anomaly Detection/Monitoring	🟡	🟢	🔴	🟡	🔴	🟢
Content Generation	🟢	🔴	🔴	🔴	🟡	🔴
Conversational Interfaces	🟢	🟡	🔴	🟢	🟡	🔴
Knowledge Discovery	🟢	🟡	🔴	🔴	🟡	🟢

Color Code for Recommendation Level: L (Low): 🔴 M (Medium): 🟡 H (High): 🟢

Avoid Hype-Driven Adoption

- Using GenAI for unsuitable cases can lead to high failure rates. Evaluate feasibility and appropriateness for each use case.

Focus on Alternative AI Techniques

- Established techniques like ML, optimization, simulation, and rule-based systems may be more suitable and reliable.

Combine AI Techniques for Robust Solutions

- Combining GenAI with other AI techniques can mitigate limitations like inaccuracies. Use GenAI for interfaces and rule-based systems for decision-making.

GenAI's Limitations in Specific Use Cases

- GenAI isn't ideal for prediction, planning, decision intelligence, or autonomous systems. It's better for content generation, conversational interfaces, and knowledge discovery.

Manage GenAI-Specific Risks

- Consider risks like output unreliability, data privacy, IP issues, cybersecurity, and regulatory compliance. Evaluate these risks for each use case.

Why Any Accelerator?

The Advantages of Having a Choice

Matching hardware to model needs enables faster AI operationalization and cost savings

Performance

Different AI models perform better on specific accelerators

Cost Efficiency

Selecting the most suitable accelerator avoids over-provisioning or under-utilizing resources leading to significant cost savings

Innovation

As the AI landscape is evolving rapidly the ability to choose accelerators allows for adapting new hardware innovation

Operationalize

Operationalize AI faster by matching the hardware to the specific needs of models

Flexibility Meets Hybrid Cloud

Aligning models, accelerators and cloud to your needs

Llama | Qwen | DeepSeek | Google Gemma | Mistral | Ai2 Molmo | Microsoft Phi | NVIDIA Nemotron | IBM Granite

I need to run this …

⊗

NVIDIA GPU | AMD Instinct | Google TPU | aws Neuron | intel Gaudi | IBM Spyre

⊗

here

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

because of

- **Compliance**
- **Availability**
- **Price**
- **Use-Case**
- **…**

A flexible AI platform for Any Model, Any Accelerator and 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 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 Inference Server

Gen AI model inference

- Packaging: Linux container
- Red Hat vLLM inference server
- Validated & optimized model repository
- LLM Compressor tool
- Certified: RHEL/RHEL AI and OpenShift/OpenShift AI
- 3rd Party Support Policy: Non-Red Hat Linux & Kubernetes platforms

I need Gen AI model Inference on RHEL/Linux or OpenShift/Kube

Red Hat Enterprise Linux AI

AI model inference & training

- Packaging: Linux server appliance
- Granite family models
- InstructLab model alignment
- Optimized RHEL image with integrated accelerators
- **Includes Red Hat AI Inference Server**

I need an integrated AI Linux server appliance for inference & training

Red Hat OpenShift AI

AI model inference, training & LLMOps

- Packaging: Kubernetes distributed cluster
- Supports Gen AI & Predictive AI
- Distributed Training, Tuning & Inference in OpenShift Kubernetes
- LLMOps & MLOps / Day 2 Mgt
- **Includes RHEL AI**
- **Includes Red Hat AI Inference Server**

I need a complete distributed AI platform for inference, training and LLMOps

Flexible and Efficient Inference

- GA distributed inference (**llm-d**)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Models to Data

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Continual Post Training Algorithm
- Feature Store GA

Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

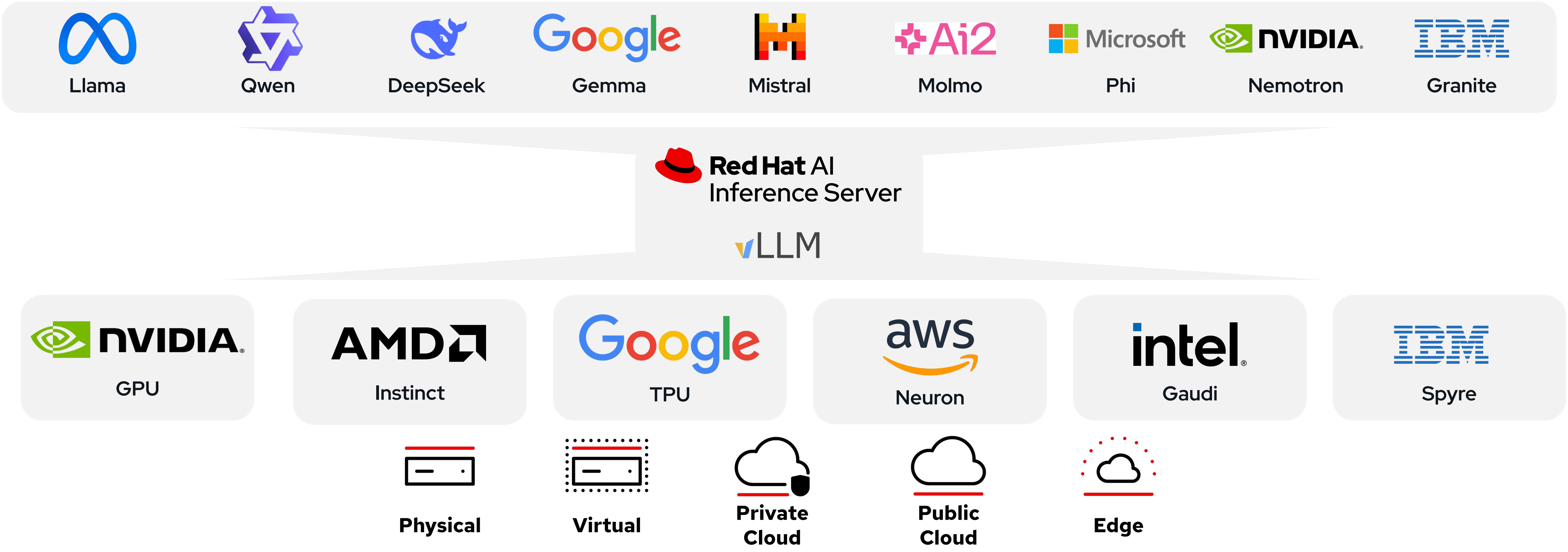
AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integration
- GPU as a Service enhancements

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

Optimized Inference

vLLM connects model creators to accelerated hardware providers



Llama

Qwen

DeepSeek

Gemma

Mistral

Molmo

Phi

Nemotron

Granite

Red Hat AI Inference Server

vLLM

NVIDIA GPU

AMD Instinct

Google TPU

aws Neuron

intel Gaudi

IBM Spyre

Physical

Virtual

Private Cloud

Public Cloud

Edge

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

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

Red Hat AI tooling for model optimization

Optimize and validate your choice of model

Inference benchmarks with GuideLLM

Tool for evaluating LLM performance to guarantee efficient, scalable, and affordable inference serving.

Accuracy evaluation with LM-eval-harness

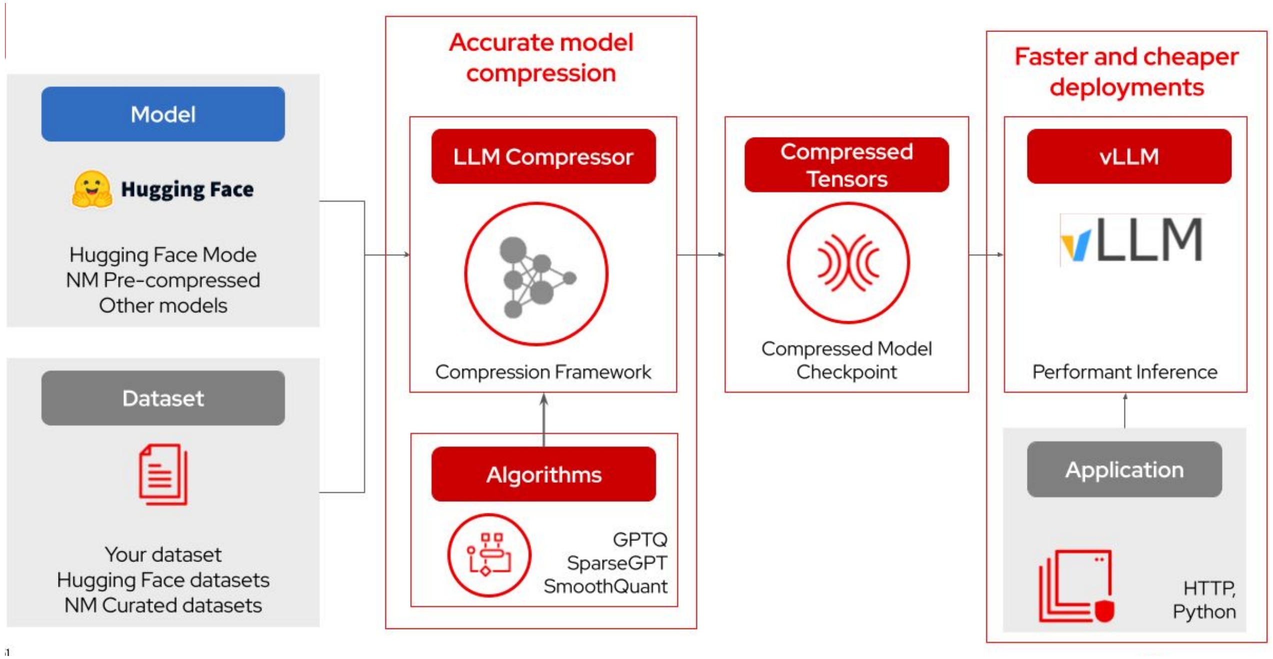
A unified framework for evaluating the accuracy of LLMs across a variety of tasks and benchmarks.

LLM Compression tools

Framework for reducing the size and computational requirements of a LLMs while preserving accuracy

Receive tailored capacity planning guidance from our experts

Optimization Support with vLLM-Compressor framework



- Comprehensive set of algorithms in unified interface
  - GPTQ, AWQ, SmoothQuant
  - FP8, INT8, INT4, MxFp4 and W4A8
- Seamless integration w/ HF AutoModel
- Safetensors-based checkpoint format compatible with vLLM
- Large model support via HF accelerate

Optimize fine-tune models for inference

Why Use Red Hat AI Compressed Models?

Delivering near-baseline accuracy and reliability through rigorous engineering and evaluation

Exceptional Quality and Accuracy

- Achieve **near-perfect (~99%) accuracy** recovery compared to the original, uncompressed baseline.
- Derived from **intensive hyperparameter tuning**, not a simple quantization run.

Rigorously Evaluated and Reliable

- **Evaluated on diverse, rigorous benchmarks** (Arena-Hard, etc.) to ensure baseline performance.
- **Extensive testing** provides a trustworthy and reliable model for end-users.

Competitive Differentiation

- Our **comprehensive compression tuning** is resource-intensive, requiring multiple runs for proper recovery.
- This **commitment to quality at scale** provides is key to the unique value Red Hat provides.

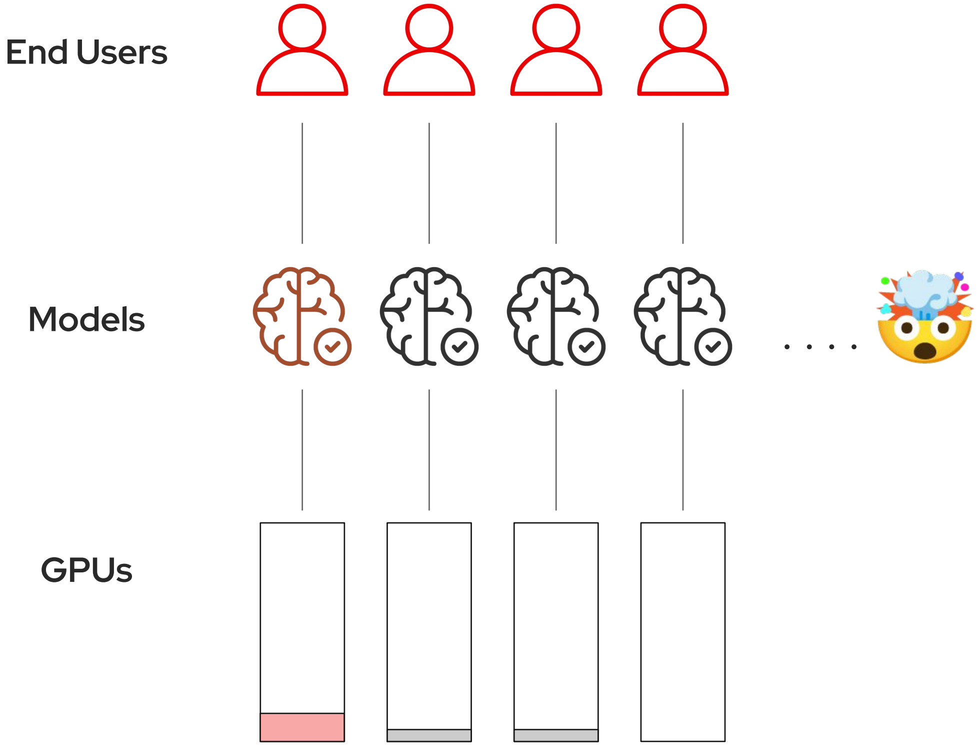
Example: [Llama-3.3-70B-Instruct-quantized.w4a16]() compressed for **75% reduced disk size and GPU memory**

Model-as-a-Service

Resources Optimization on Shared Inference Servers

CPUs & **especially GPUs**

Infrastructure as a Service can be costly



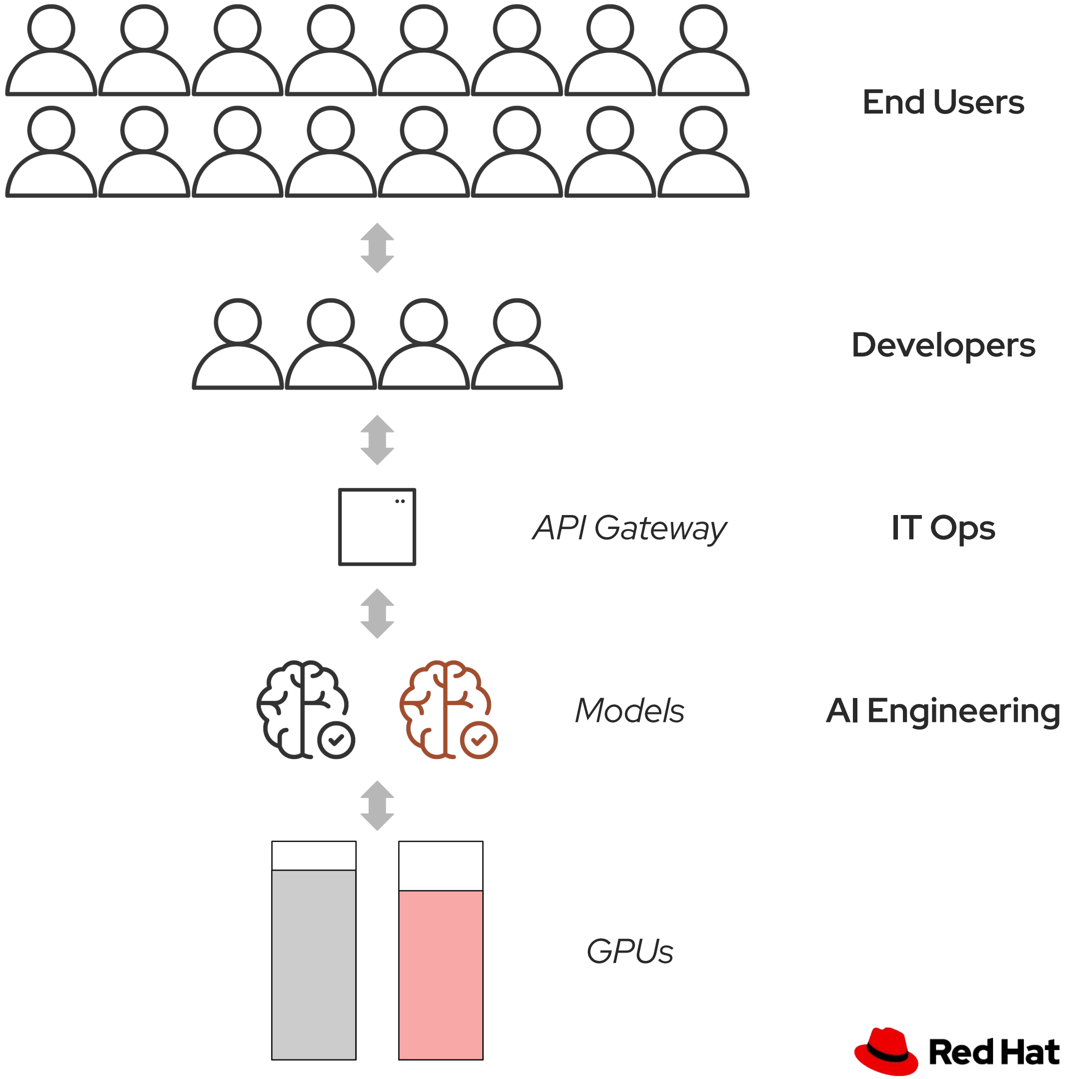
Self-Service is good for plentiful resources & small teams

- Throwing GPUs at the problem is risky
- Few people know how to use them correctly
- Leads to duplication and underutilization
- Leads to high costs
- Most people want an LLM endpoint, not a GPU

Models as a Service

Offering AI **models as *the* service** to a larger audience

- IT serves common models centrally
  - Generative AI focus, applicable to any model
  - Centralized pool of hardware
  - Platform Engineering for AI
  - AI management (versioning, regression testing, etc)
- Models available through API Gateway
- Developers consume models, build AI applications
  - For end users (private assistants, etc)
  - To improve products or services through AI
- Shared Resources business model keeps costs down



Hosted AI services are **not** the only option

Gemini

OpenAI

ANTHROP\C

Become the **Private AI Provider**

LLaMA by Meta

IBM Granite

MISTRAL AI_

Risks & Challenges:

- Costs at scale
- Data privacy and security policies
- IP leakage

Models-as-a-Service Benefits:

- Cost effective & optimize performance
- Easy to use
- Consistent with data & security requirements

Parasol AI Studio: an example use case

AI Applications

Anything LLM

- Granite 3.3 8B Instruct LLM
- Weaviate Vector Database
- Chatbot
- Document embeddings

Continue

- VS Code with continue.dev plugin
- Granite 3.1 8B Code Instruct LLM

Docling

- Transformation of
  - Spreadsheet
  - PDF

  to Markdown

Stable Diffusion

- Stable Diffusion LLM
- Generation of images

Models-as-a-Service

Red Hat 3scale API Management

AI Platform

Red Hat OpenShift AI + Red Hat OpenShift

NVIDIA

aws

Connection Details

Wifi:

Red Hat Summit: Connect 2025

Password:

redhat_2025

red.ht/3JtNcSx

Replace **studentX** with your actual assigned user!

Jetzt Session bewerten!

Einfach QR-Code scannen,
Session aus der Liste wählen
und bewerten. **Vielen Dank!**

red.ht/rhsc-darmstadt-feedback

Red Hat

Open Tech Quest

Solve technical challenges around
Ansible, OpenShift & RHEL as a team

Nicht vergessen!

Schließt euch zu Teams von rund 5 Personen zusammen, legt eure jeweiligen Benutzeraccounts an und startet dann ab 14:45 Uhr gemeinsam durch!

red.ht/otq

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat