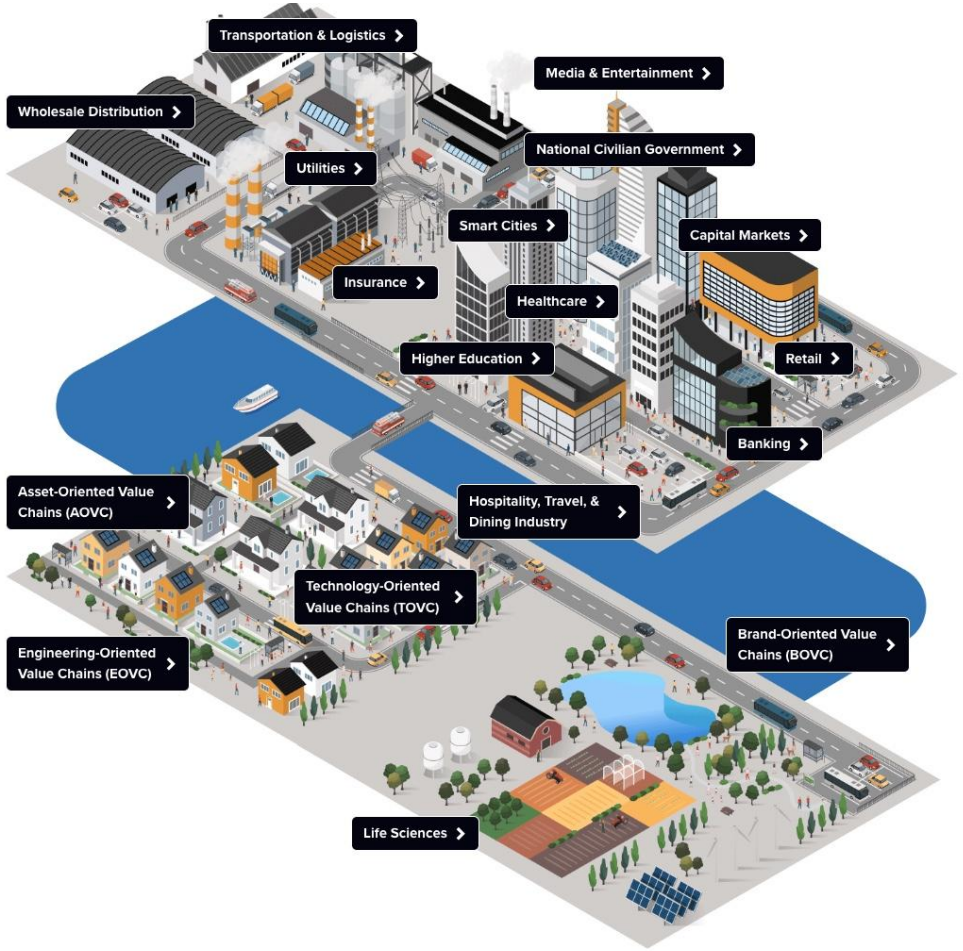
Red Hat AI: Our vision, strategy and roadmap

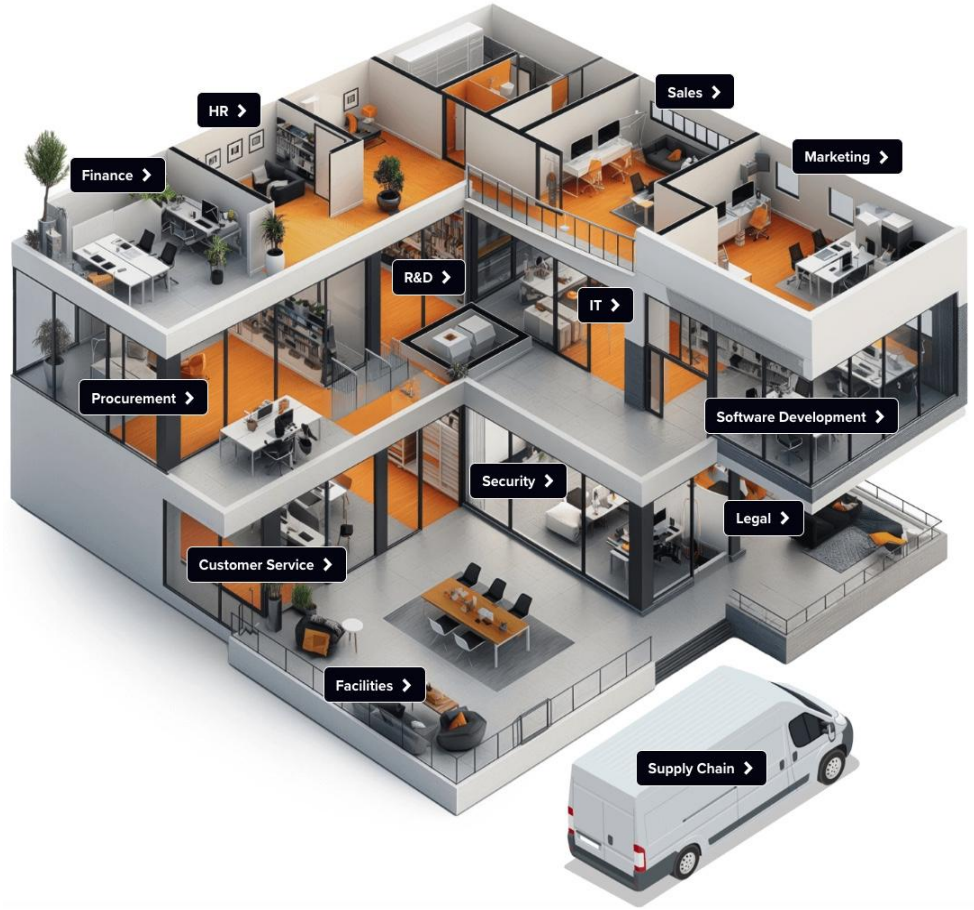
Red Hat

AI is a strategic enabler across industries

AI use cases that drive productivity and efficiency



Every vertical industry



Every business function

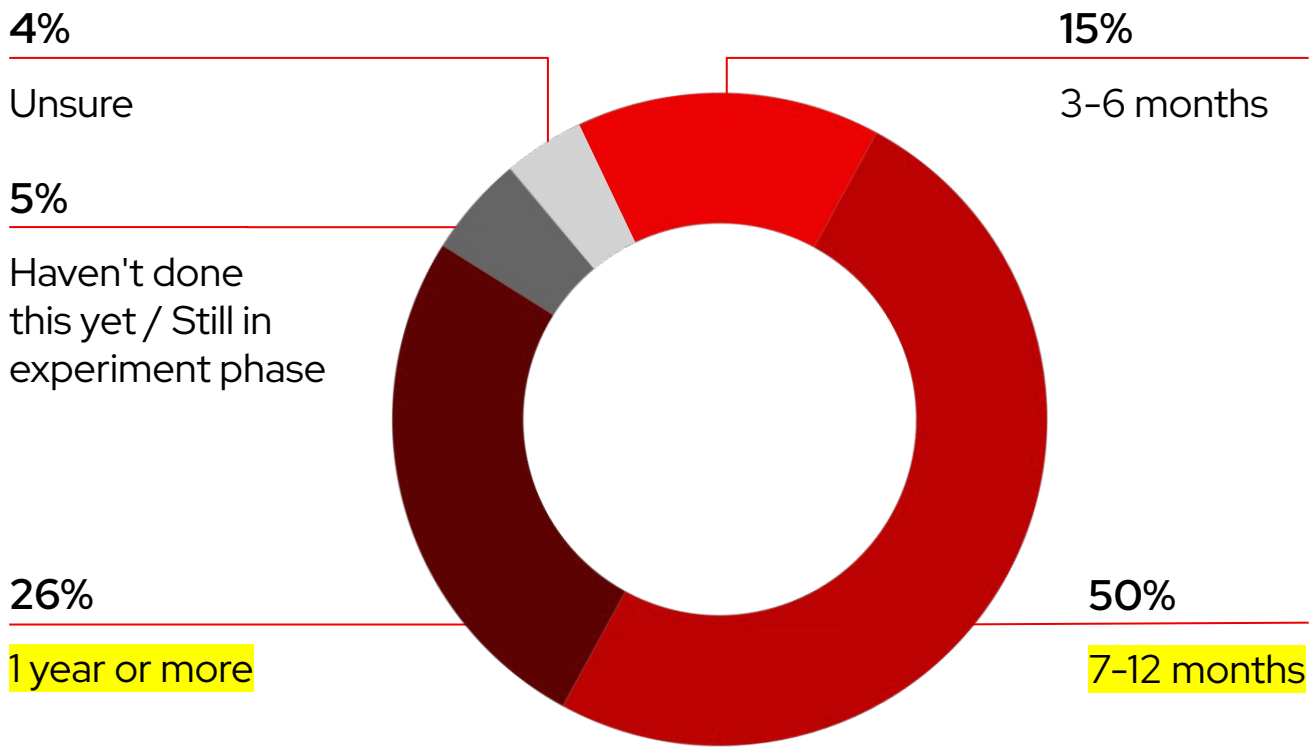
IDC: Use cases by industry or function

Red Hat

Operationalizing AI is still a challenging process

What is the average AI/ML timeline from idea to operationalizing the model?

Half of respondents (50%) say their average AI/ML timeline from idea to operationalizing the model is 7-12 months.



Source: Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

AI has undergone significant evolution

The evolution of AI: from Business Intelligence to Generative AI

- Predictive AI runs businesses **today**
- Foundation models provide a **shortcut** for realizing the value of AI

Business Analysis & Intelligence	Advanced Analytics & Predictive AI	Foundation Models & AI-enabled apps
• Collecting data • Storing & moving data • Transforming data	• Data science techniques • Predictive analytics • Real-time decision making	• Deep learning techniques • Model experimentation • Model tuning
Data warehouses	**Big data**	**Gen AI**

Why is **NOW** a good time to invest in AI?

Growing demand for AI solutions and services

The worldwide AI software market will grow to nearly $790 billion by 2026 (5 yr CAGR 18%)[1]

52%

of organizations cite 'lack of MLOps tools' as a challenge[2]

65%

of organizations are currently investing in generative AI[3]

Source

(1) IDC - Worldwide AI Software Forecast, 2023–2027 | (2) IDC Market Analysis Perspective: Worldwide Artificial Intelligence Software, 2023 | (3) IDC's AI Strategies View 2022, CY22

Generative AI Customer Adoption Challenges

Cost

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

Complexity

Integrating models with private enterprise data for customer use cases is too difficult.

Flexibility

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

Accelerate the development and delivery of AI solutions across hybrid-cloud environments

Increase efficiency with **fast, flexible and efficient inferencing**

Simplified and consistent experience for **connecting models to data**

Accelerate Agentic AI delivery and stay at the forefront of innovation

Flexibility and consistency when **scaling AI across the hybrid cloud**

Red Hat

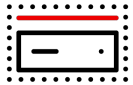
Red Hat AI Inference Server

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

For customers who need Gen AI model Inference on RHEL/Linux or OpenShift/Kube

Red Hat AI

Gen 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**

For customers who need an integrated Gen AI Linux server appliance for inference & training

Red Hat OpenShift AI

Gen AI model inference, training & LLMOps

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

For customers who need a complete distributed Gen AI platform for inference, training and LLMOps

Red Hat

Any model. Any accelerator. Any cloud.

Red Hat

Areas of focus for 2025 and Beyond

Models	Multi-lingual; Multimodal Reasoning models	Granite Models 3rd party open models	Model Validation Model Optimizations

Alignment to Enterprise AI	**Model Serving**	**AI Agents and Apps**
Data ingestion and chunking Synthetic Data Generation Training Evaluation RAG	Optimizations Service Level Objectives Massive scaling Model Hub and Registry	API: Inference, datasets, safety, vector_io, telemetry, Agents Tool/Function Calling

LLMOps **MLOps**	Observability: TFT, ITL, TPS, Metrics, Logs, Traces Costs: $/million i/o tokens	Security: Provenance, Signing, Encryption Governance: guardrails

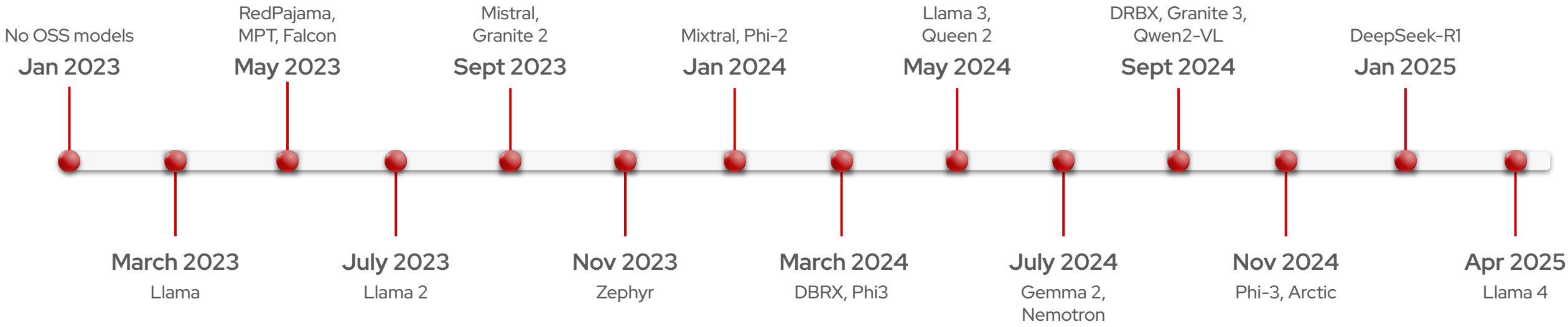
Hybrid Cloud	Hardware Accelerators, OEMs, Clouds
Open Source	Models, Tools, Frameworks
Partners	Model Providers, OEMs, CSPs, ISVs

Red Hat

Model Inference/Serving

The power of open

There has been an explosion of capability from open-source over the last 2 years



Red Hat

vLLM Inference Server - Connecting Models to the Hardware

vLLM is emerging as the defacto open platform for inferencing

Model Creators

Meta Llama

Mistral

Qwen

Microsoft Phi

NVIDIA Nemotron

snowflake Arctic

AI21 labs Jamba

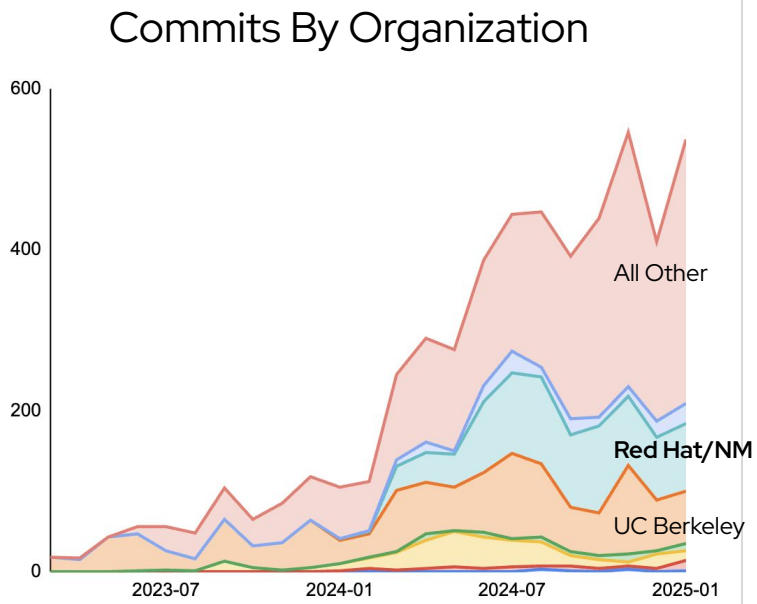
databricks DBRX

Hugging Face

Hardware Vendors

NVIDIA	Features for new HW
AMD	Choice for MI300X
intel	Gaudi enablement
Google	TPU enablement
aws	Neuron enablement
IBM	Spyre enablement

Contribution Trajectory



Interactive Demo: red.ht/vllm-interactive

vLLM is available as GA in all the Red Hat AI platforms

Red Hat

vLLM scheduler V1

Re-architect the "core" of vLLM
based on the lessons from V0

Unchanged	Changed
• **User-level APIs** • **Models** • GPU Kernels • Utility functions	• Scheduler • Memory Manager • Model Runner • API Server

→ **New implementation of speculative decoding!**
→ **V1 gives better performance across the board!**

vLLM scheduler V1 is available on RHAIIS 3.0, RHEL AI 1.5 and RHOAI 2.20

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

Interactive demo: red.ht/llm-compress-interactive

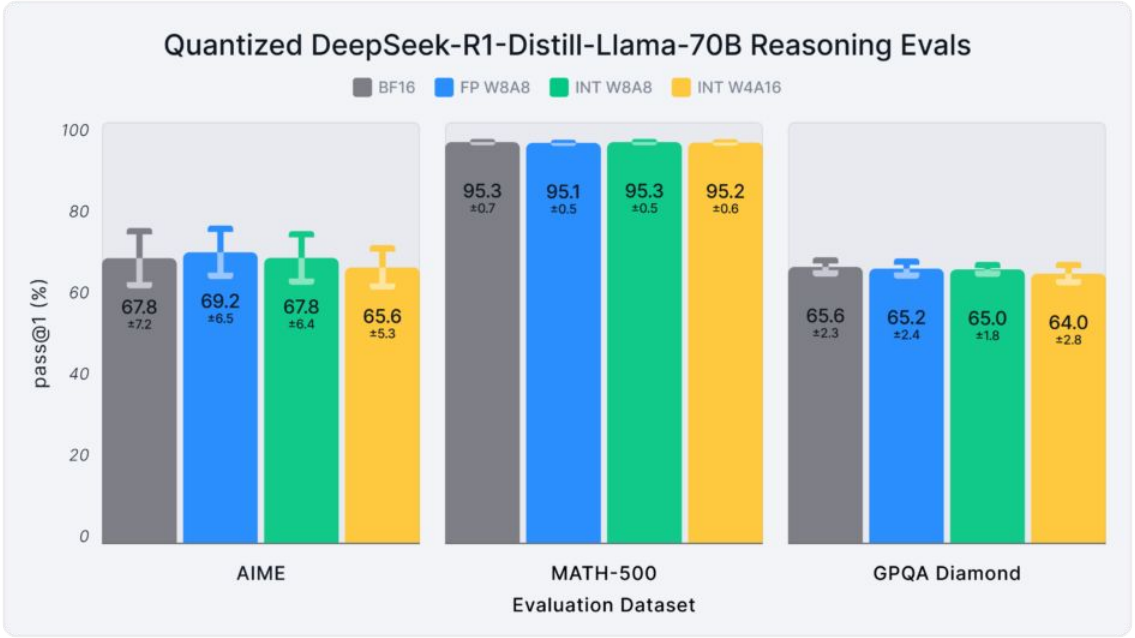
Llm-compressor and GuideLLM are Developer Preview; lm-evaluation-harness is GA

Red Hat

Compressed DeepSeek-R1 models

State-of-the-art, open-source quantized reasoning models built on the DeepSeek-R1-Distill

- **FP8 and INT8 quantized versions achieve near-perfect accuracy recovery** across all tested reasoning benchmarks and model sizes —except for the smallest INT8 1.5B model, which reaches 97%.
- **INT4 models recover 97%+ accuracy** for 7B and larger models, with the 1.5B model maintaining ~94%.
- With **vLLM 0.7.2,** deliver **4X better inference performance** across many common inference scenarios.
- **Reduce GPU requirements** by (e.g. 50% for INT8)



https://developers.redhat.com/articles/2025/03/03/deployment-ready-reasoning-quantized-deepseek-r1-models#

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

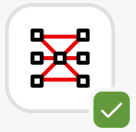
Phi

Molmo

Granite

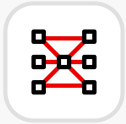
Nemotron

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

Access and Support (via 3rd party model support) is available to all Red Hat AI customers

Red Hat

Introducing the llm-d project

Industry leaders unite to power distributed, scalable gen AI inference

- **Inference at scale everywhere:** Gen AI will be as ubiquitous as Linux
- **Powered by Kubernetes & vLLM:** Unlocking efficient, scalable inference
- **Backed by industry leaders:** founded in collaboration with CoreWeave, Google, IBM Research, NVIDIA
  - Supported by AMD, Cisco, Intel, Lambda and Mistral AI

Red Hat

llm-d: distributed inference at scale

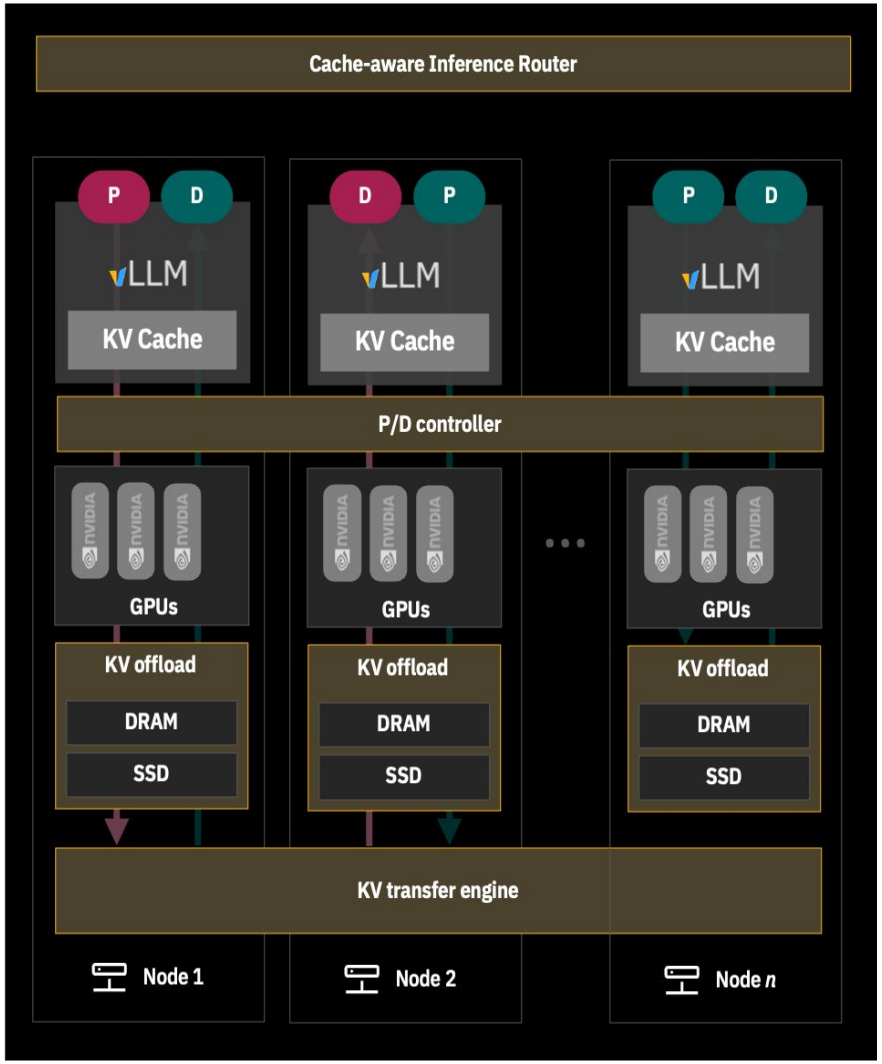
Flexible and Distributed architecture to meet SLOs efficiently

Core Features

- Prefill/decode disaggregation
- KV Cache distribution, offloading
- AI-aware router
- Operational telemetry for production
- vLLM and Kubernetes-based
- NIXL inference transfer library

Benefits

- Maximize Token Revenue
- Reduce Token Costs
- Boost Latency and Throughput
- Seamless Scaling



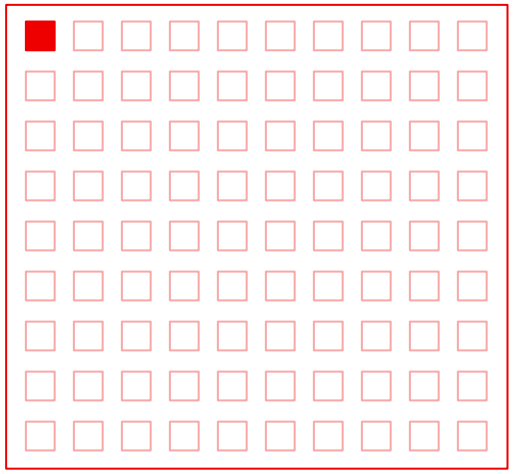
Red Hat

Model Customization / Alignment

Red Hat

Enterprises need models aligned to their private data

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



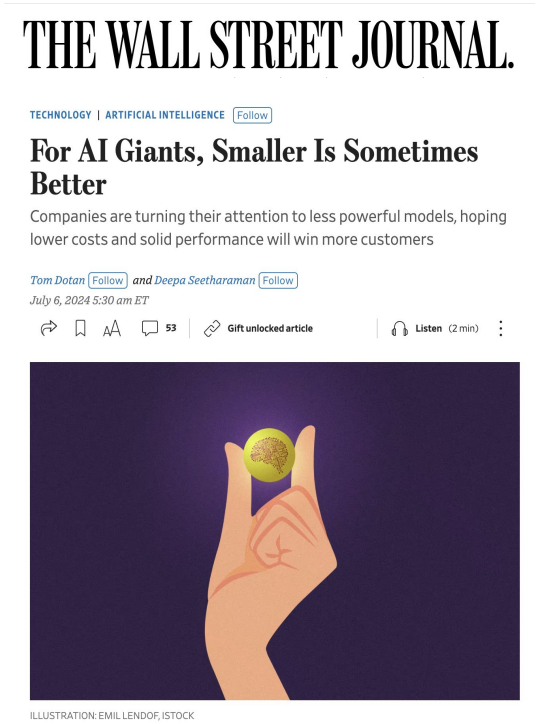
Less than 1% of all enterprise data is represented in foundation models

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

Open Source AI and Rise of Small Language Models

Smaller models are more efficient & customizable


THE WALL STREET JOURNAL.

TECHNOLOGY | ARTIFICIAL INTELLIGENCE Follow

For AI Giants, Smaller Is Sometimes Better

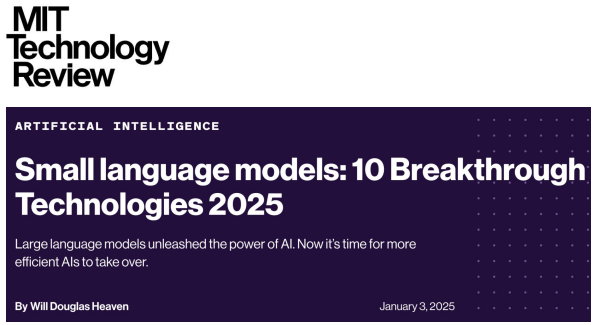
Companies are turning their attention to less powerful models, hoping lower costs and solid performance will win more customers

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

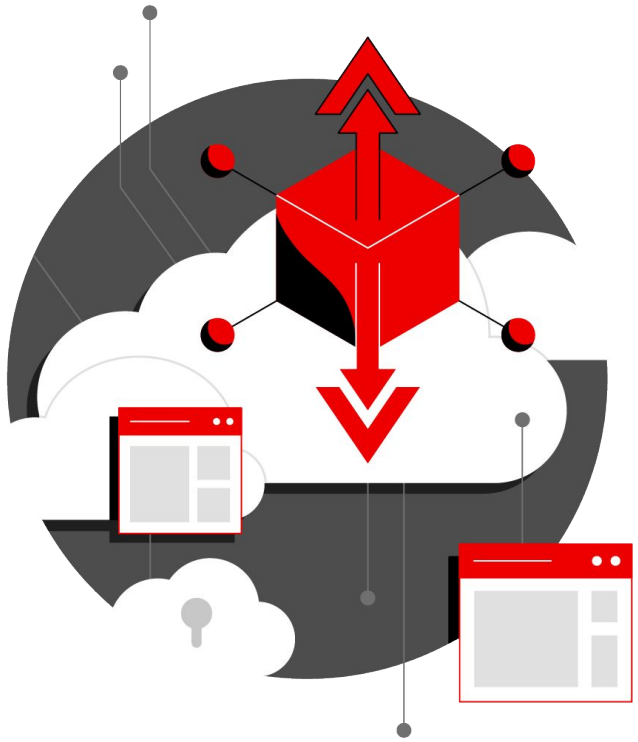
Credit: VentureBeat using Midjourney

- **Open source** AI models (Llama, Mistral, Granite, & more) are catching up to proprietary models
- **Small language models (SLMs)** are orders of magnitude smaller than frontier models like GPT4 (<10 Billion parameters vs. >1 Trillion)
- **Cost:** Run cheaper, faster and consume less energy on less powerful hardware
- **Customization:** Can be tuned & customized with private enterprise data for domain specific tasks
- **Control:** Customers own their own 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/

The value of open source and smaller language models

Smaller models are more efficient & customizable



- Open source AI models are catching up to proprietary models.
- Smaller language models, like **IBM Granite**, are orders of magnitude smaller than frontier models.
  - Models with less than 10 billion parameters **are cheaper and faster to run**, and consume less energy.
- These models can be **tuned and customized with private enterprise data** for domain specific tasks.
- **Customers own their own models** and can create multiple instances for different use cases and deployment environments.

Model Alignment Approaches

RAG and Fine Tuning increases accuracy and optimizes costs

RAG

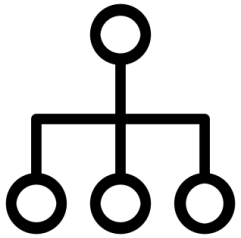
Retrieval Augmented Generation

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

NEW

InstructLab

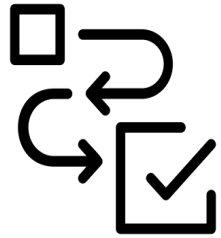
Large-scale Alignment for chatBots



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

Fine tuning

Fine Tuning



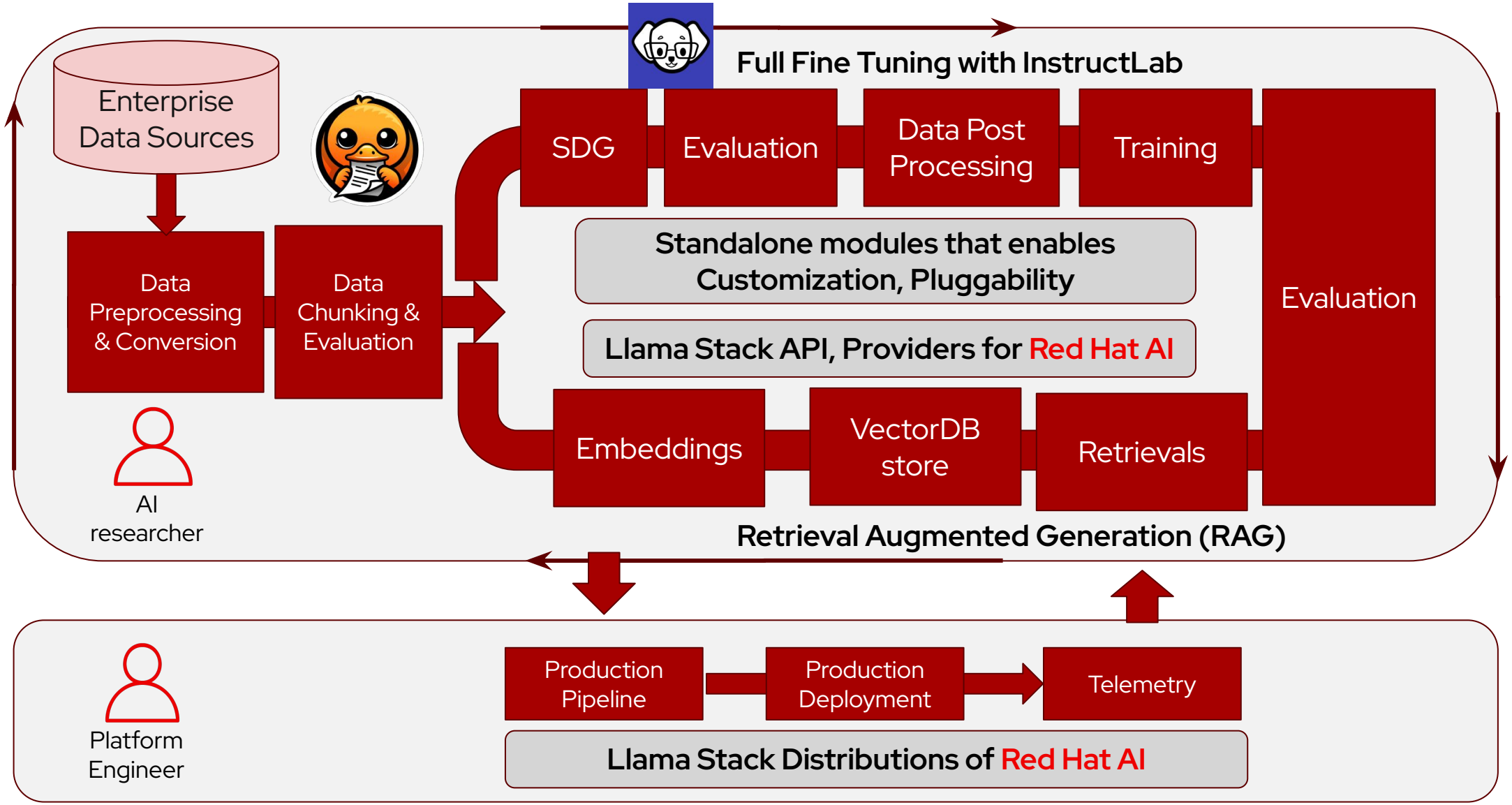
Adjust a pre-trained model on specific tasks or data, improving its performance and accuracy for specialized applications without full retraining.

InstructLab provides **more accessible fine tuning** & **complements RAG** (RAFT)

Red Hat

Aligning AI to the Enterprise

RAG and Fine Tuning with Red Hat AI



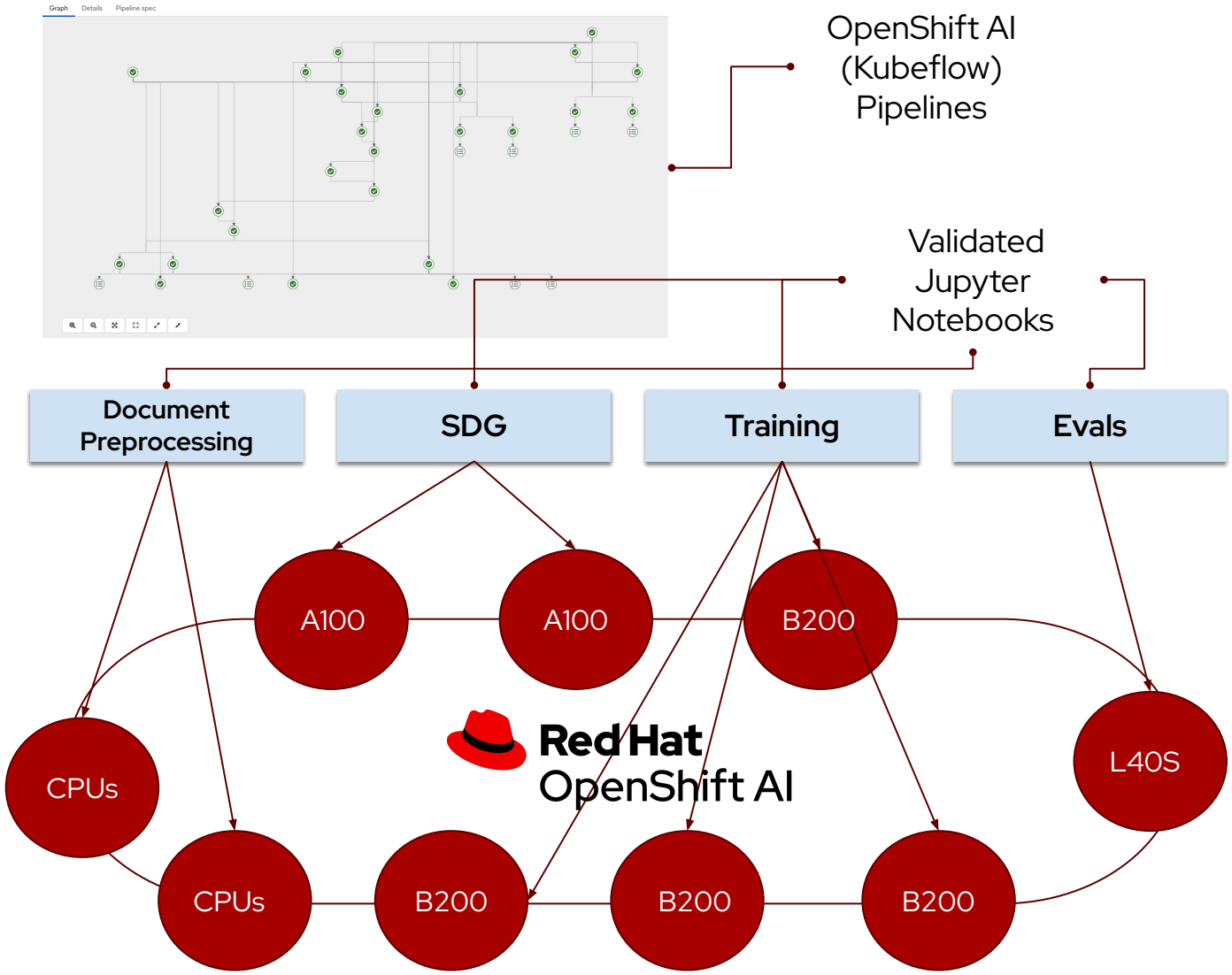
Distributed InstructLab

Overview

Modules for InstructLab model customization can run in a distributed manner for scale and resilience

Why does it matter?

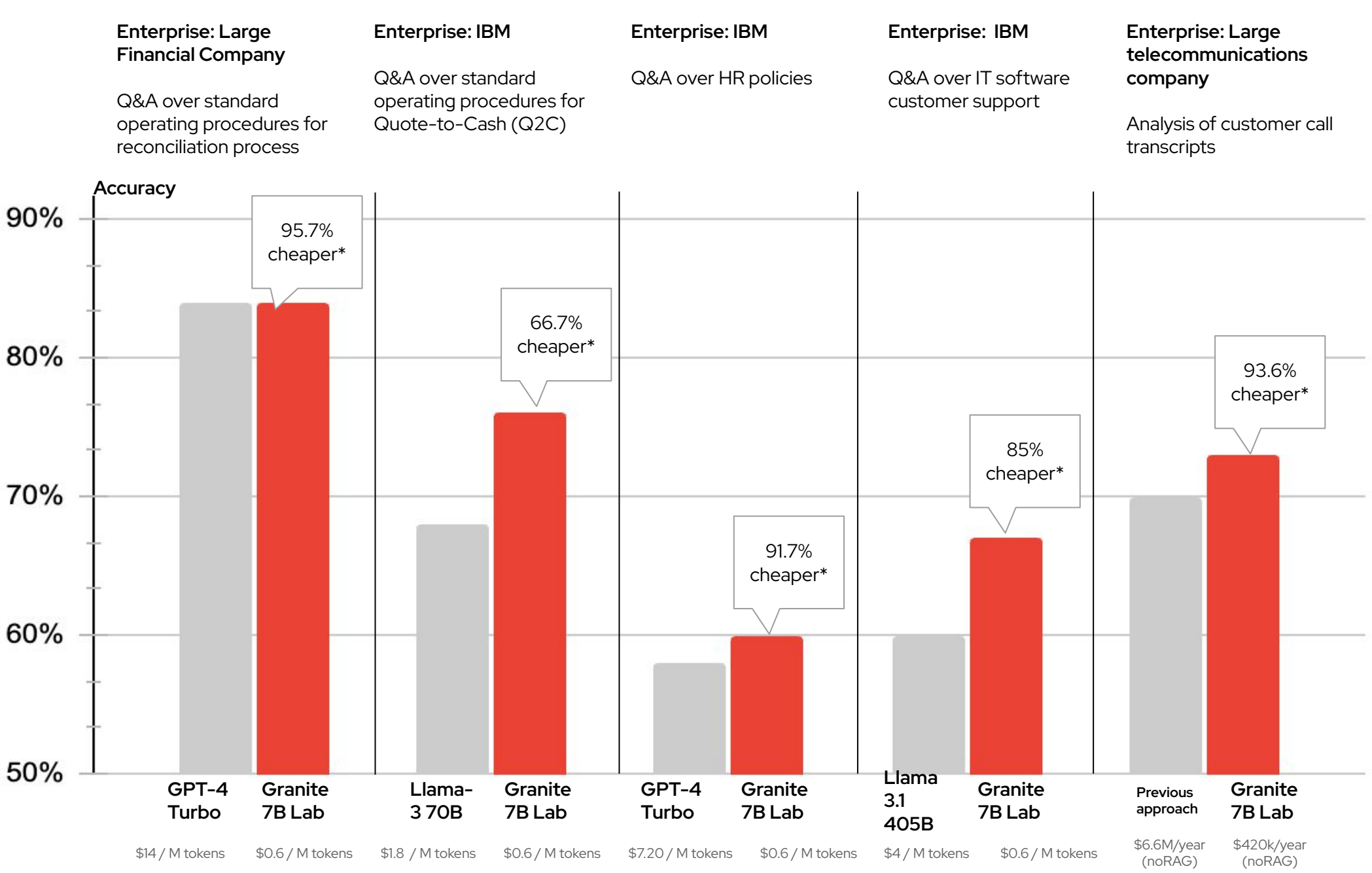
To get enterprise data into models in a secure, efficient, collaborative way. New approaches are needed all of the tools of the iLab toolkit are showcased in a distributed environment like RHOAI



Distributed InstructLab is Tech Preview RHOAI; These modules (Python SDKs) will also available on RHEL AI in the future.

Red Hat

The value of enterprise data can be seen in tuning smaller, targeted, optimized models to deliver state-of-the-art performance at considerably lower cost.



*SaaS cost per million tokes (assuming blend of 80% input, 20% output), https://artificialanalysis.ai/models/prompt-options/multiple/medium#pricing

Innovation's from AI researchers at Red Hat and the AI community

Product-driven. Open-source AI.

Inference-time scaling

From throughput to accuracy — redefining inference for real-world tasks.

Dr. Sow, Particle Filtering, SQuat

Customization of Reasoning Models

Turning enterprise knowledge into reasoning power.

Reasoning Blog, Async-GRPO

Customization of Instruct Models

Adapting instruction-tuned models without forgetting what matters.

Sculpting Subspaces

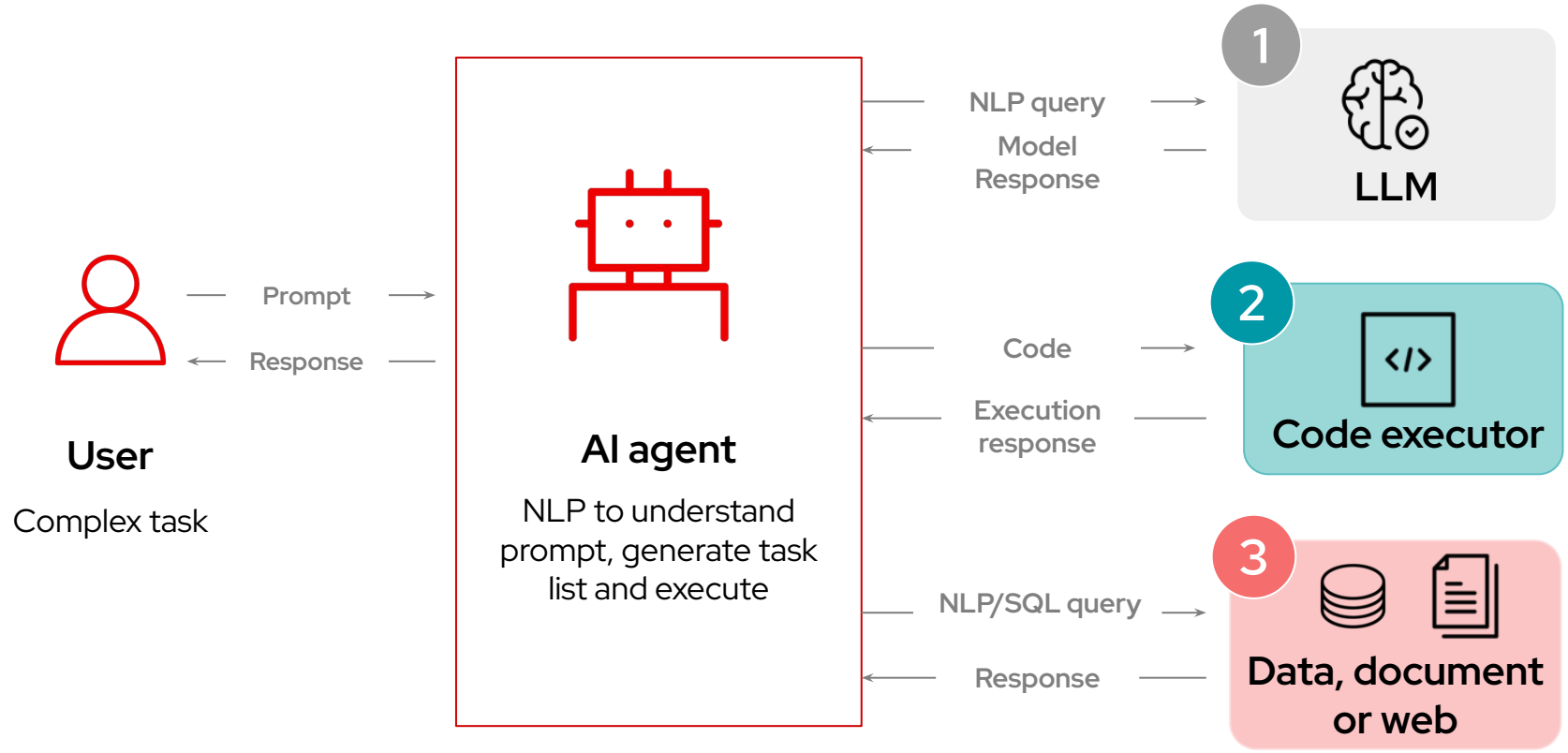
Red Hat

AI Agents

Red Hat

AI agents integrate models, functions & tools

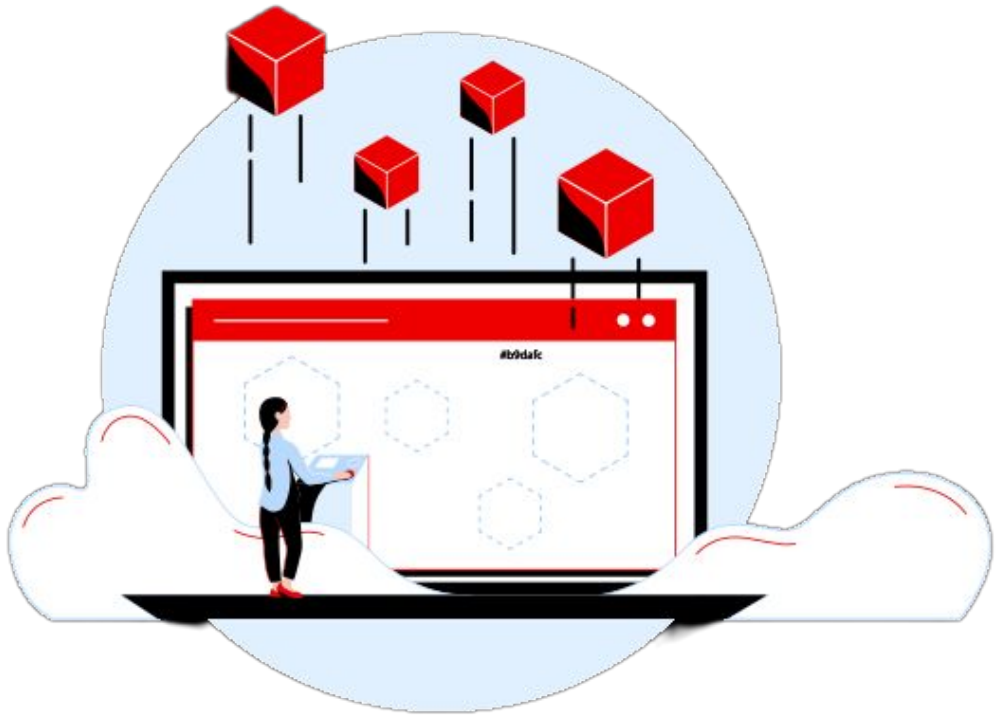
Gen AI Models, Predictive AI Models, Code Functions, Search & more



Red Hat

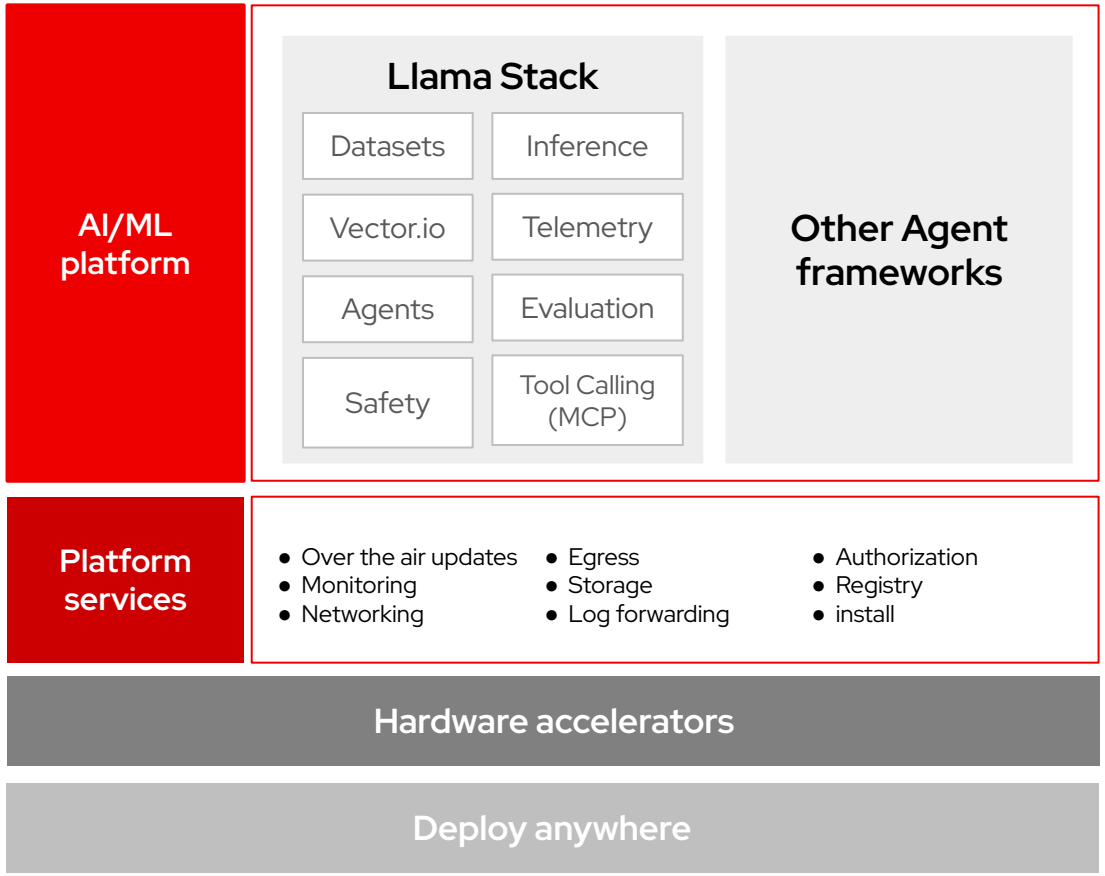
Red Hat AI provides an agile, stable foundation to accelerate the development and deployment of AI agentic workflows.

- Allows running and managing agents as microservices.
- Simplifies production deployment by managing LLM serving and scaling.
- Offers native capabilities to build and manage agents with Llama Stack, and standardized communication protocols (MCP).
- Provides the flexibility to integrate preferred tools like LangChain and Crew AI.



Red Hat

A modular approach to building AI agents



Red Hat AI allows to:

- Build agents using **Llama Stack's native capabilities and implementations**.
- **Bring compatible Llama Stack implementations** to OpenShift AI.
- **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- **Build with Core Primitives** and manage your own agent framework as a standard workloads.

Red Hat

AI Platforms

Red Hat

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

Red Hat OpenShift AI

Model Development & Tuning	Model Serving	Data & model pipelines	Model Monitoring
• Workbenches with Jupyter Notebooks as a service • Distributed training & tuning - Kueue & Ray • ISV images | Custom images • GPU support	• Serving runtimes : vLLM, other runtimes (Custom, OVMS) • Serving engines: Kserve	• Automated data science pipeline based on Kubeflow Pipelines • Visual editor • Model Registry	• Performance metrics • Operations metrics • Quality metrics

Red Hat Enterprise Linux AI

Red Hat AI Inference Server

Red Hat OpenShift

InstructLab

Data Ingestion | Synthetic Data Generation | Phase Training | Serving | Evaluation | RAG | Agents

Granite models | 3rd party models

vLLM

Red Hat Enterprise Linux — Pytorch| FSDP | NVIDIA CUDA | AMD ROCm | Intel Gaudi | Google TPU

Physical — Virtual — Private cloud — Public cloud — Edge

Red Hat

Red Hat's AI Partner Ecosystem

Integrated ISVs

- Anaconda
- elastic
- IBM
- intel
- NVIDIA
- Pachyderm
- run:ai
- stability ai
- Starburst

AI and general ISVs

- C3.ai
- Cloudera
- Couchbase
- crunchy data
- EDB
- Guise AI
- H2O.ai
- Hugging Face
- Lightbend
- Palantir
- redis
- SAS

Delivery partners

- accenture
- BCG
- Capgemini
- Deloitte.
- HCLTech
- Infosys
- kyndryl
- World Wide Technology

Cloud partners

- aws
- Azure
- Google Cloud
- IBM Cloud

Hardware

- AMD
- DELL
- cisco
- HPE
- intel
- Lenovo
- NVIDIA
- Qualcomm
- Supermicro

Conceptual machine learning architecture

Red Hat OpenShift AI

Gather and prepare data	Develop model	Deploy models in an application	Model monitoring and management
Data exploration Data preparation	ML notebooks ML libraries	Model lifecycle CI/CD	Monitor / alerts Model visualization Model drift

Red Hat OpenShift

Hybrid, multi cloud platform with self service capabilities

NVIDIA intel

Compute acceleration

aws Azure vmware

Infrastructure

Physical · Virtual · Private cloud · Public cloud · Edge

Red Hat

U.S. Department of Veterans Affairs

Suicide has no single cause, and no single strategy can end this complex problem. That's why Mission Daybreak is fostering solutions across a broad spectrum of focus areas.

A diversity of solutions will only be possible if a diversity of solvers answer the call to collaborate and share their expertise.

Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce Veteran suicides

Challenge

Develop new data-driven means of identifying Veterans at risk for suicide.

Solution

Red Hat teamed with global consulting services provider Guidehouse and Philip Held, Ph.D. of Rush University Medical Center, to **develop a new data-driven means of identifying Veterans at risk for suicide running on Red Hat OpenShift**, leveraging Red Hat OpenShift API Management **and Red Hat OpenShift AI**.

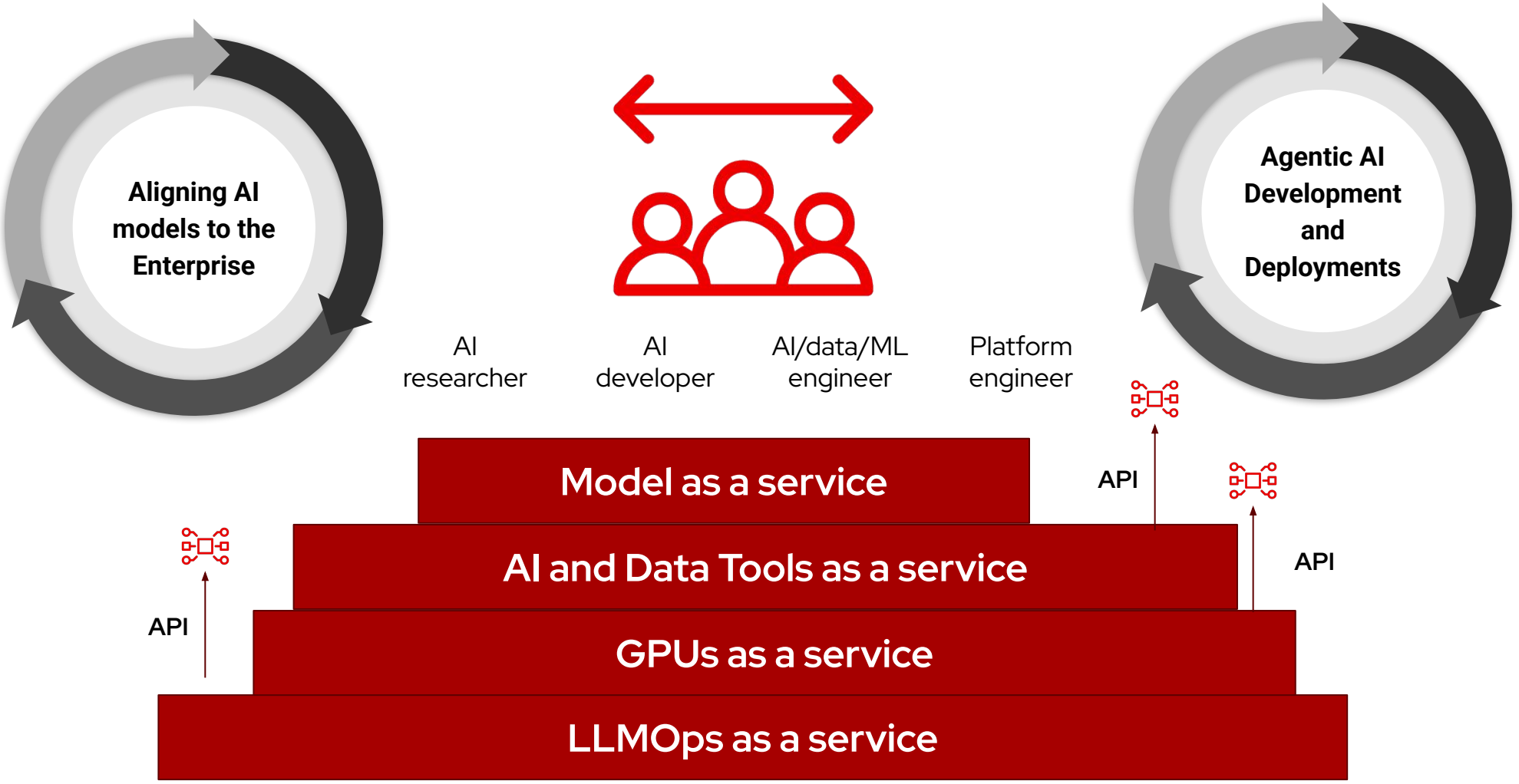
Results

- Named a winner in the Mission Daybreak challenge, Phase 2, of the U.S. Department of Veterans Affairs' (VA) Mission Daybreak Grand Challenge in support of cutting-edge suicide prevention solutions
- Moved forward with a solution for the VA's efforts to reduce Veteran suicides
- Showcased the repeatability and scalability of open source-enabled solutions

Red Hat

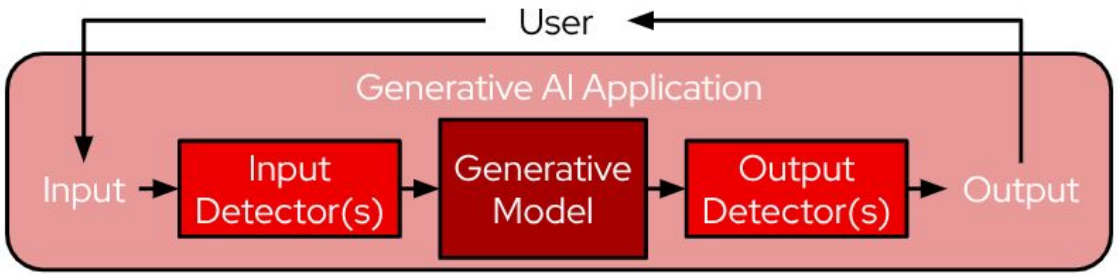
Source: Red Hat blog. Red Hat, Team Guidehouse named winner in Mission Daybreak challenge to reduce Veteran suicides, Mar. 2023.

The Red Hat AI Platform



Guardrails for Generative AI in Red Hat AI

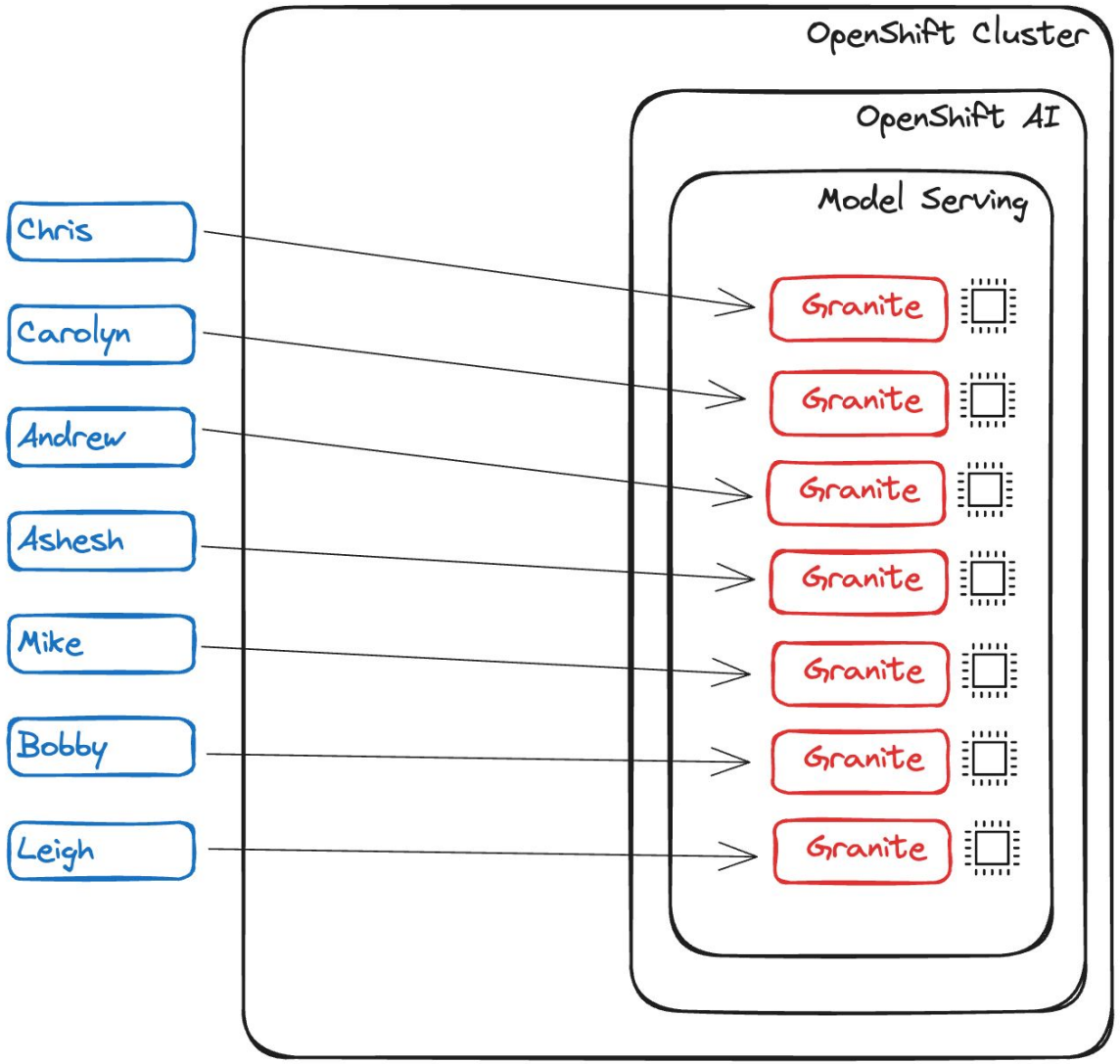
- Ensure **secure, compliant, and efficient** AI operations with These key features:
  - **Customizable Input and Output Validators**: Tailor the AI's behavior to meet your business needs
  - **Request-Time Configuration**: Dynamically apply guardrails on a per-request basis
  - **Role-Specific Detection**: Design targeted validation pathways for different user groups



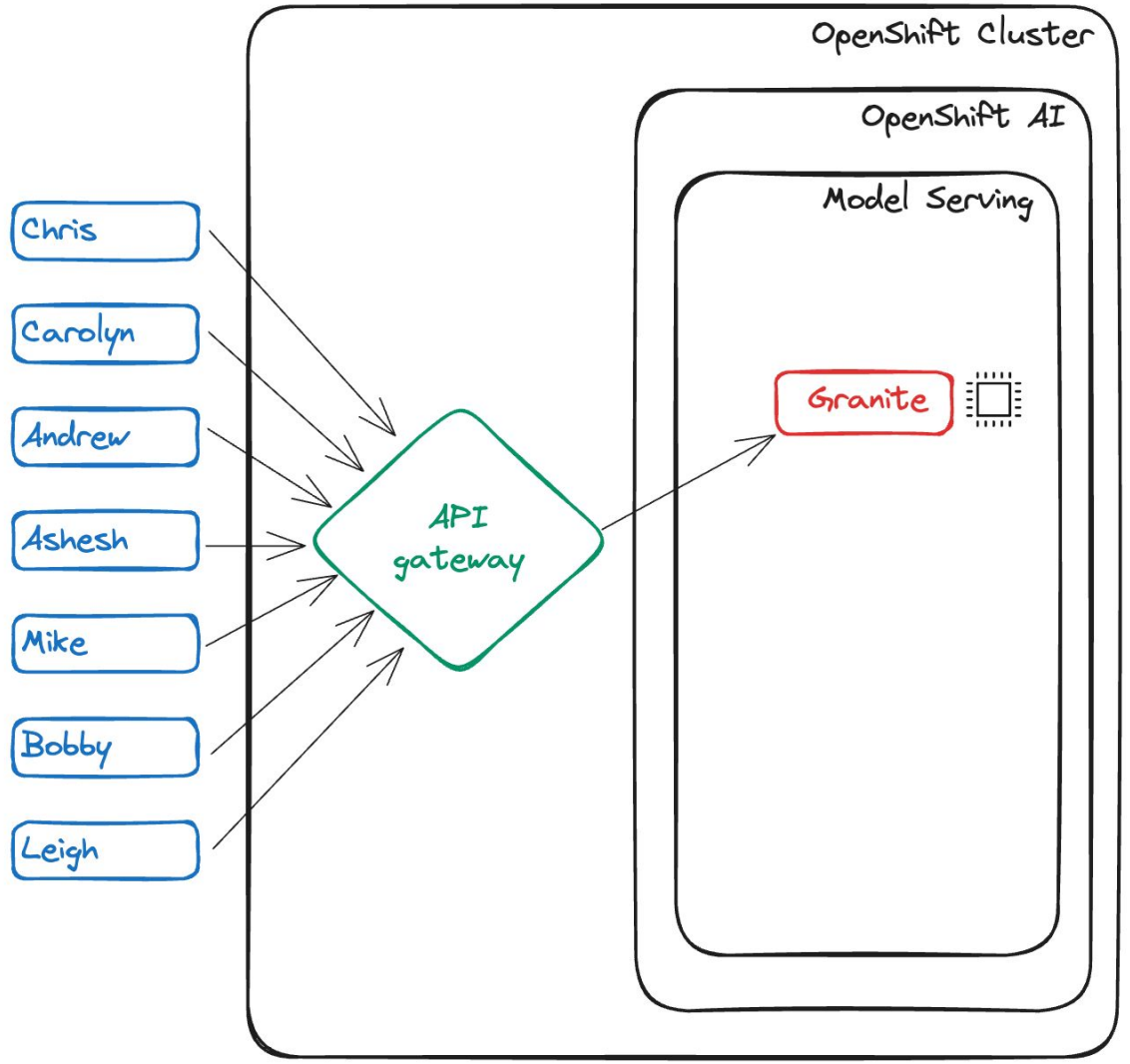
- **Protect customer's Brand**: Prevent mentions of competitors to maintain focus on your products.
- **Minimize Risk**: Restrict contract creation and negotiation to human oversight.
- **Enhance Customer Experience**: Provide role-specific, meaningful interactions for every user group.

PS: signing and securing model artifacts is part of the Model Registry's OCI-compliance storage provided by OpenShift AI

Before MaaS



After MaaS

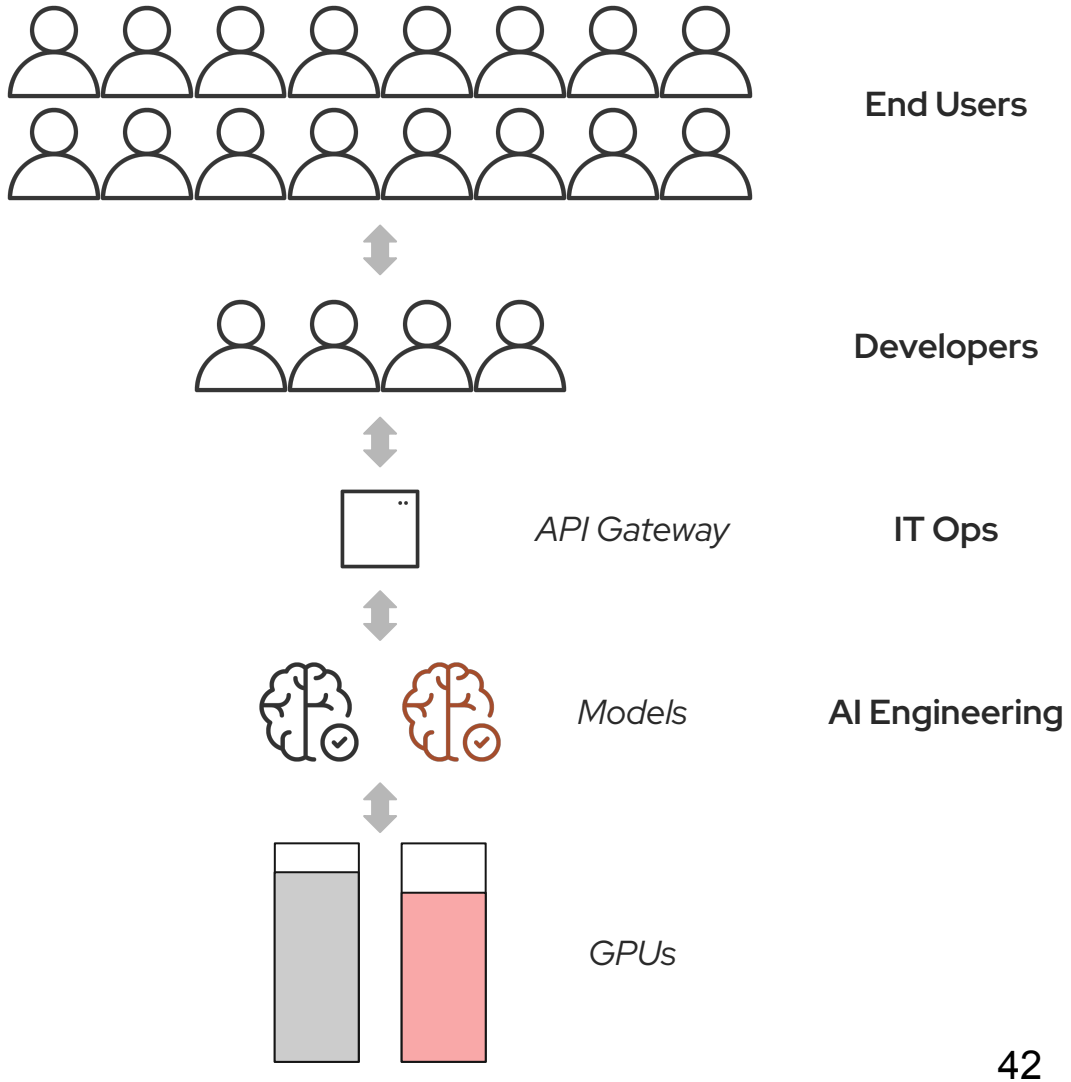


Red Hat

Models as a Service (MaaS)

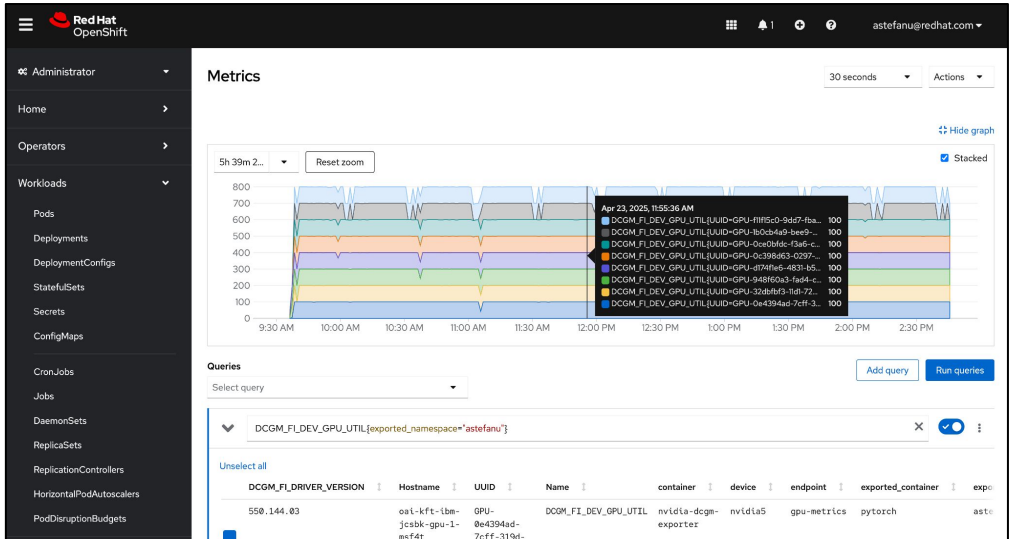
Offering AI **models as *the* service** to a larger audience

- 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



GPUaaS

- Enables efficient management and allocation of GPU resources for a variety of AI workloads: workbenches, training/tuning jobs, model serving
- Supports both whole and fractional GPU allocation
- Includes observability tools for resource optimization and to facilitate chargeback scenarios



Why does it matter?

- *Improving Resource Utilization*: Reclaiming idle GPUs and optimizing allocation to reduce waste
- *Supporting the complete AI Lifecycle*: Handling workloads from notebooks to model serving
- *Providing Visibility*: Offering metrics for both data scientists and administrators

Auto Scaling GPUs for Inferencing: https://www.youtube.com/watch?v=vsxg17uMFIo

Red Hat

Red Hat Summit

Thank you


linkedin.com/company/red-hat


facebook.com/redhatinc


youtube.com/user/RedHatVideos

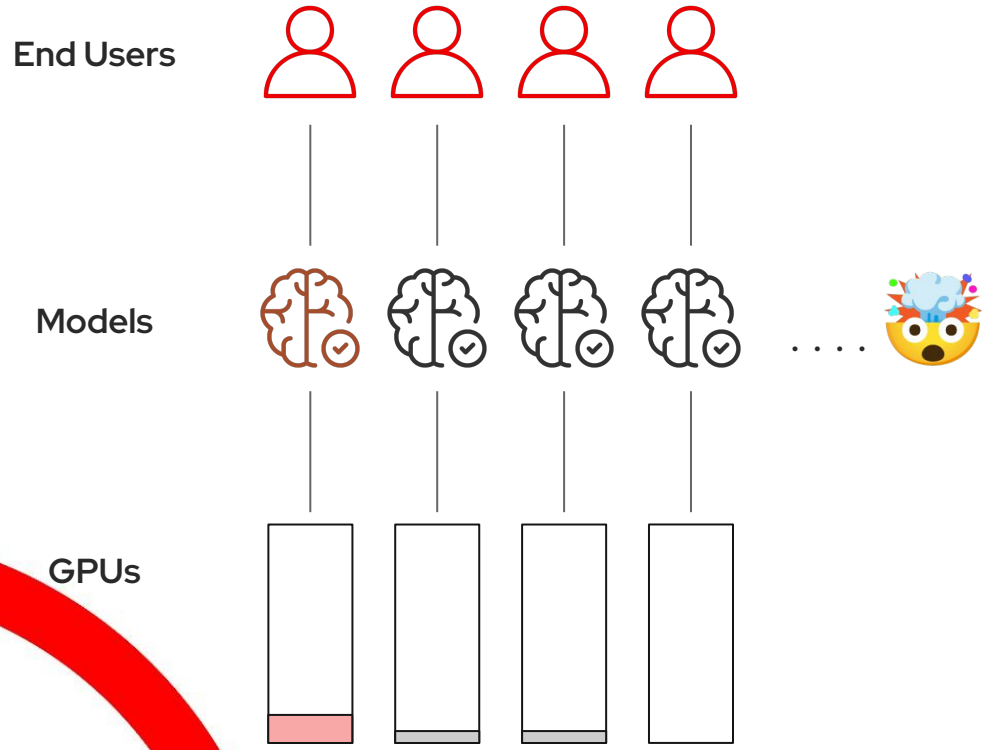

twitter.com/RedHat

Red Hat

Model as a Service

CPUs & **especially GPUs**

Infrastructure as a Service can be costly



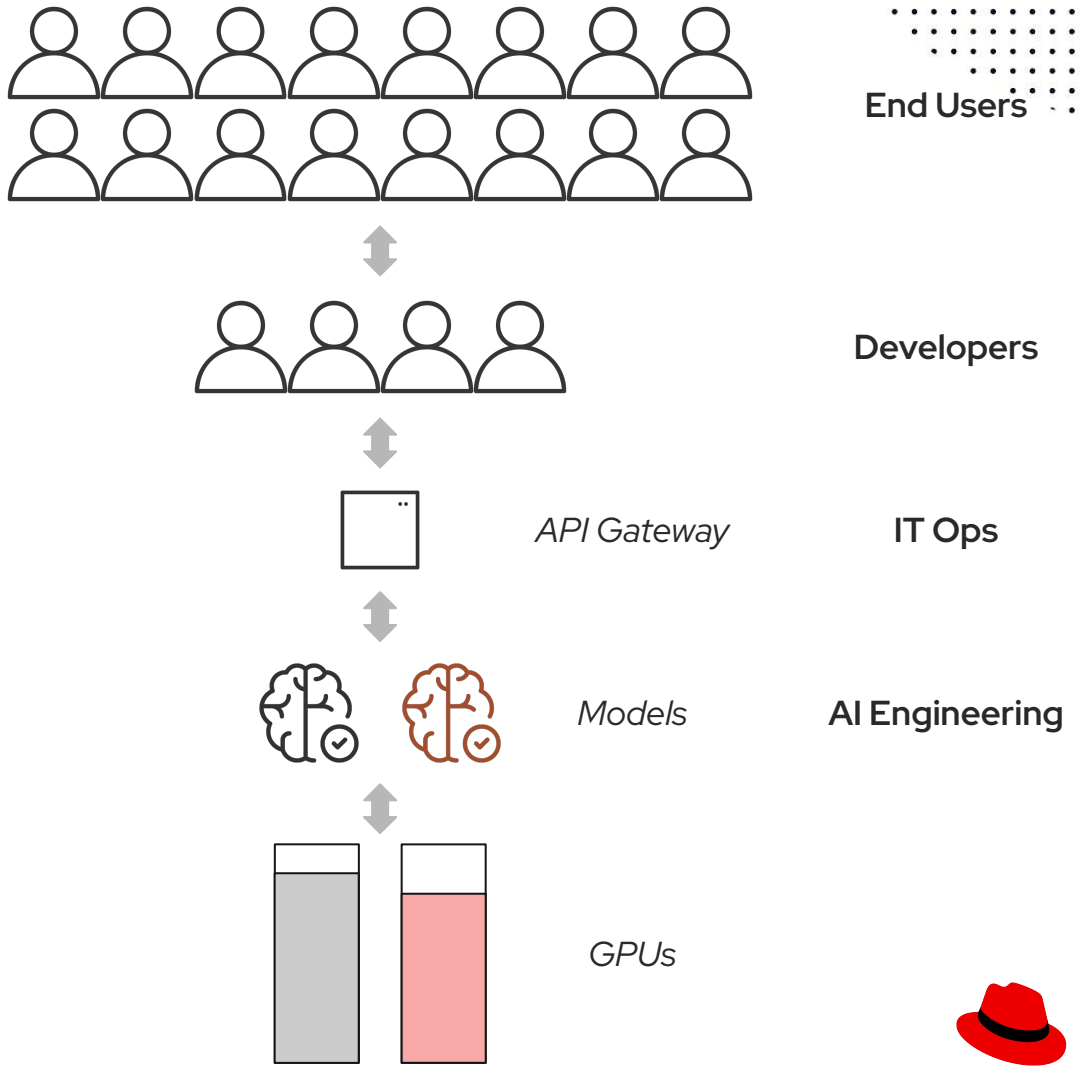
Self-Service is good for plentiful resources & small teams

- 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 (MaaS)

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

Risks & Challenges:

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

Become the Private AI Provider

LLaMA by Meta

IBM Granite

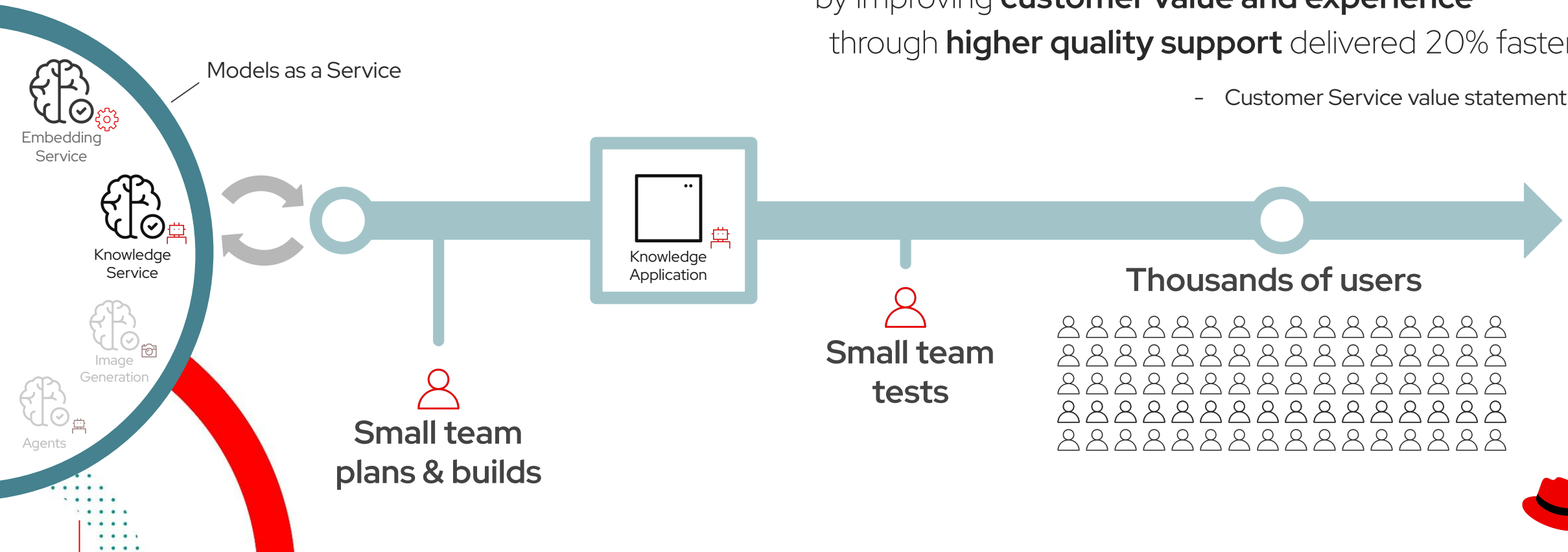
MISTRAL AI_

MaaS Benefits:

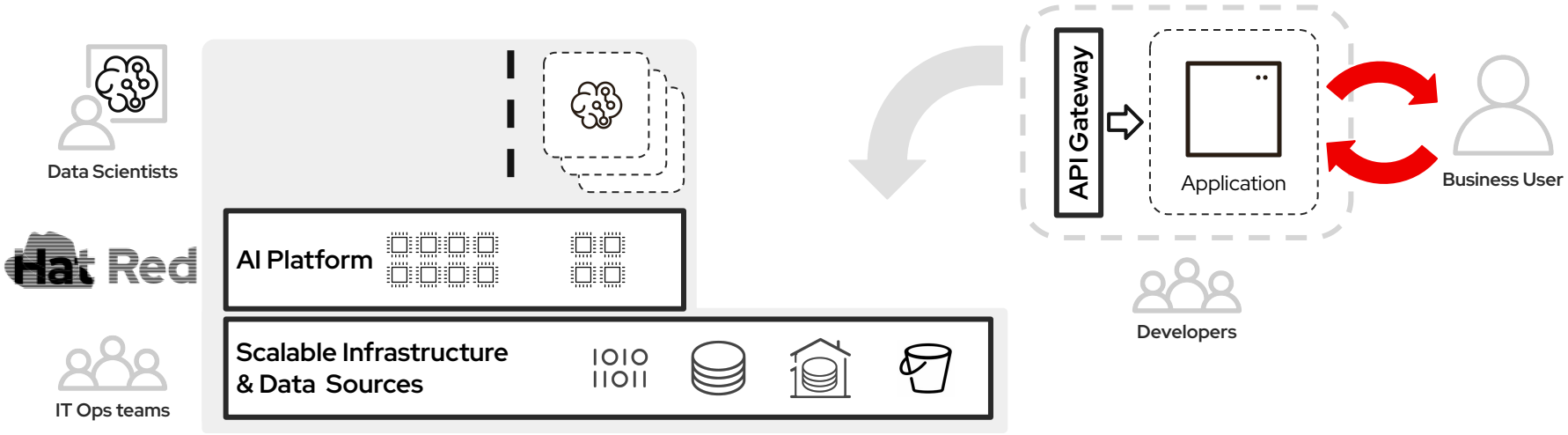
- Cost effective & optimize performance
- Easy to use
- Consistent with data & security requirements

Example: Private Customer Service **knowledge application**

> Goal to **increase transaction frequency** by improving **customer value and experience** through **higher quality support** delivered 20% faster.
>
> - Customer Service value statement



Today's **infrastructure** + tomorrow's **strategy**



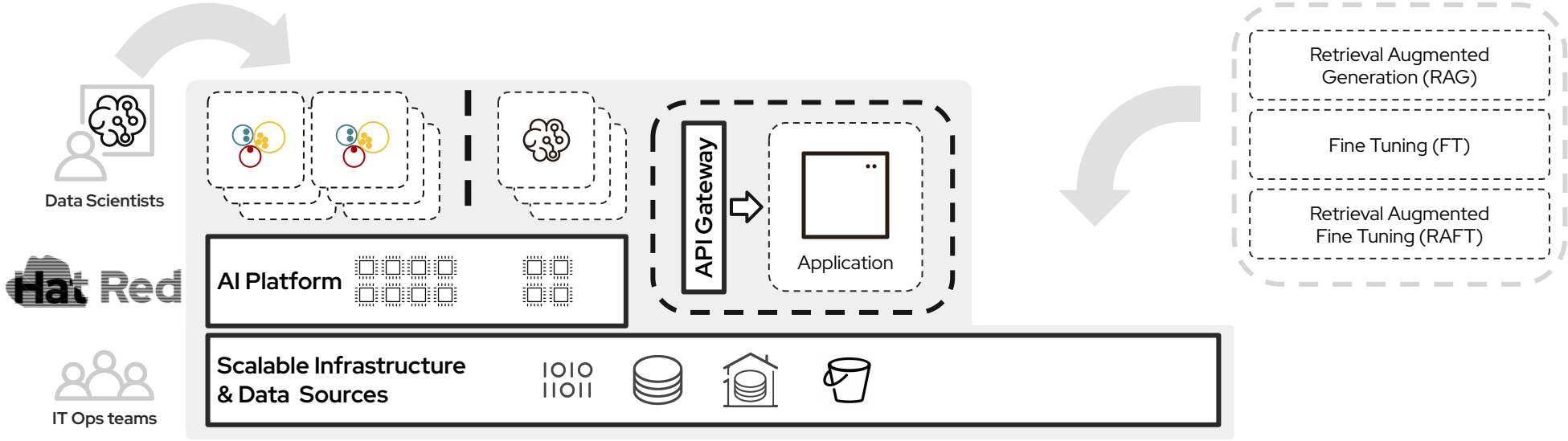
What's not new:

- IT manages AI infrastructure
- Platform, hardware & access centrally located
- Data Scientists use GPUs to customize models
- IT & Data Scientists monitor & evaluate performance

What's new:

- Models served to a wider audience
- IT adds API Gateway for production serving
- Developers build using standardized endpoints
- Associates consume Private AI services

Stable foundation for **expanding use cases** and techniques



Success generates success:

- Add models for new use cases and applications
- Data Scientists build specialized models
- Use RAG, FT, or RAFT for improved model results
- Expand proven scalable environment

RAG, FT, & RAFT in simple terms:

- RAG - supplement model's basic information with details
- FT - train model a bit more with detailed information
- RAFT - combines the two, retrieve details and train

* Most organizations progress in this order

Become the **Private AI Provider** for your organization

What is Models as a Service

- Strategy delivering central AI services privately
- Model service consumed by large audience
- Accessible to Developers and Associates
- GPUs invisible to user, critical for cost optimization

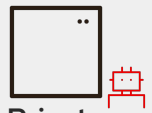
Why IT should become the Private AI Provider

- Compliant with existing security, data & privacy policies
- Predictable costs & increased utilization
- Reduce time to market with AI applications
- Unified & impactful service delivery

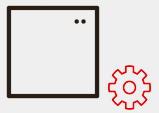
How value is created

- AI managed like any other workload
- Innovation across entire organization
- Plug into existing cost models
- Reduced costs, risk, and overhead

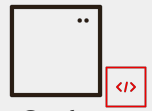
Use cases



Private Assistant (RAG)



Embedding Application



Code Assistant

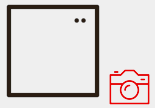
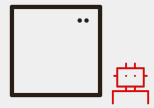


Image Generation



Agentic

Your next app

Red Hat can help

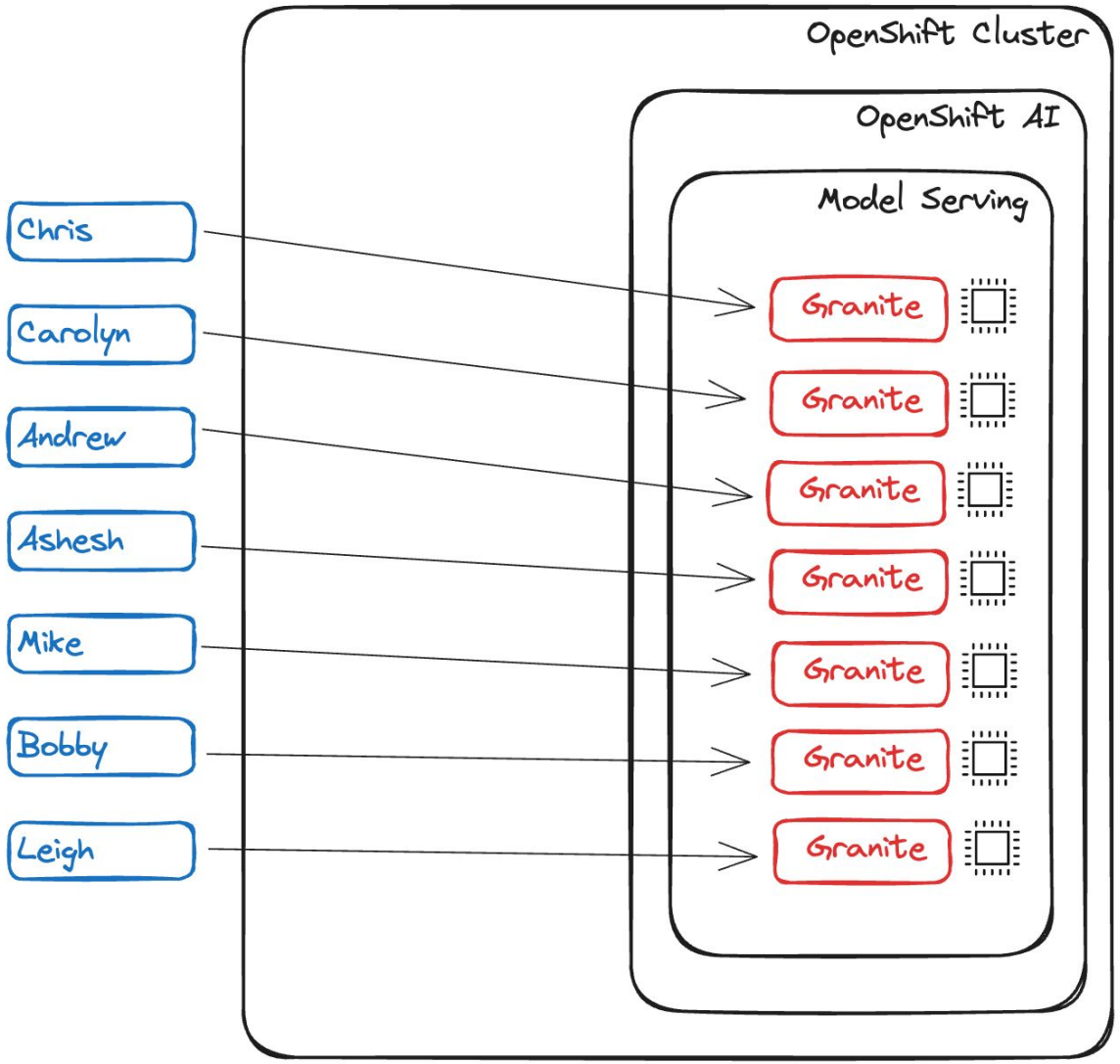
Red Hat 3scale API Management

Red Hat OpenShift AI

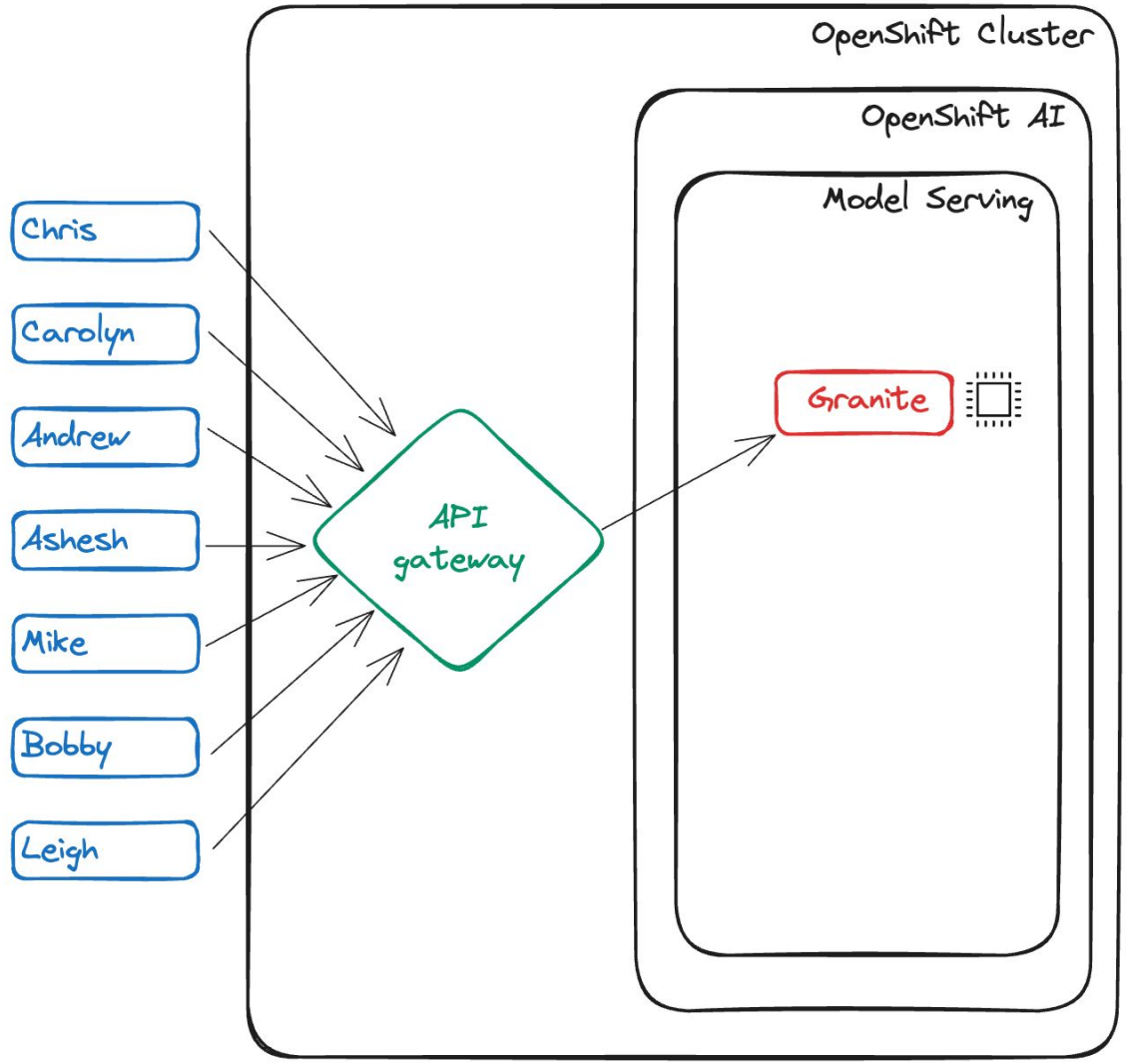
Red Hat OpenShift

Red Hat Consulting

Before MaaS



After MaaS



Red Hat

Before MaaS

- Self-Service access to GPUs can lead to great inefficiencies:
  - Duplication of models
  - Duplication of efforts
  - Lack of accountability
  - Low GPU utilization
  - Unnecessarily high costs

Red Hat

After MaaS

- STOP providing access to GPUs.
- Instead:
- Offer self-service access to Models
- Leads to great benefits:
  - High traceability
  - Most people would rather use the models (than the GPUs)
  - Lower TCO / chargeback, etc.
  - Increase utilization

Red Hat

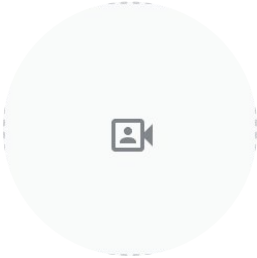
MaaS Principles

- Become the provider of Private AI:
  - Don't just "throw GPUs at the problem"
  - Team of experts serve each model only once
  - Provide self-service access to the models
  - API gateway to track model consumption

MaaS Principles

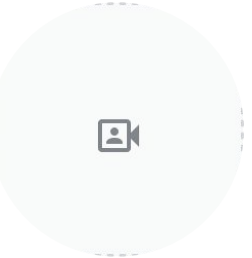
- Internally replicate the business structure of public AI providers (like Gemini, OpenAI, etc...)
  - None of them give you access to their GPUs

Readily available models for your Developers



- Getting access to Private AI models becomes as trivial as using OpenAI/Gemini/Etc..

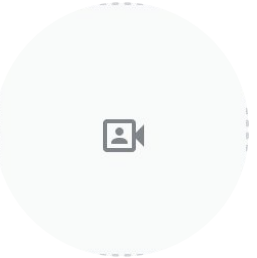
MaaS is your engine for "AI for all"



- Your devs can now easily build AI-Powered applications
- And now, your employees and customers can easily leverage AI-powered applications

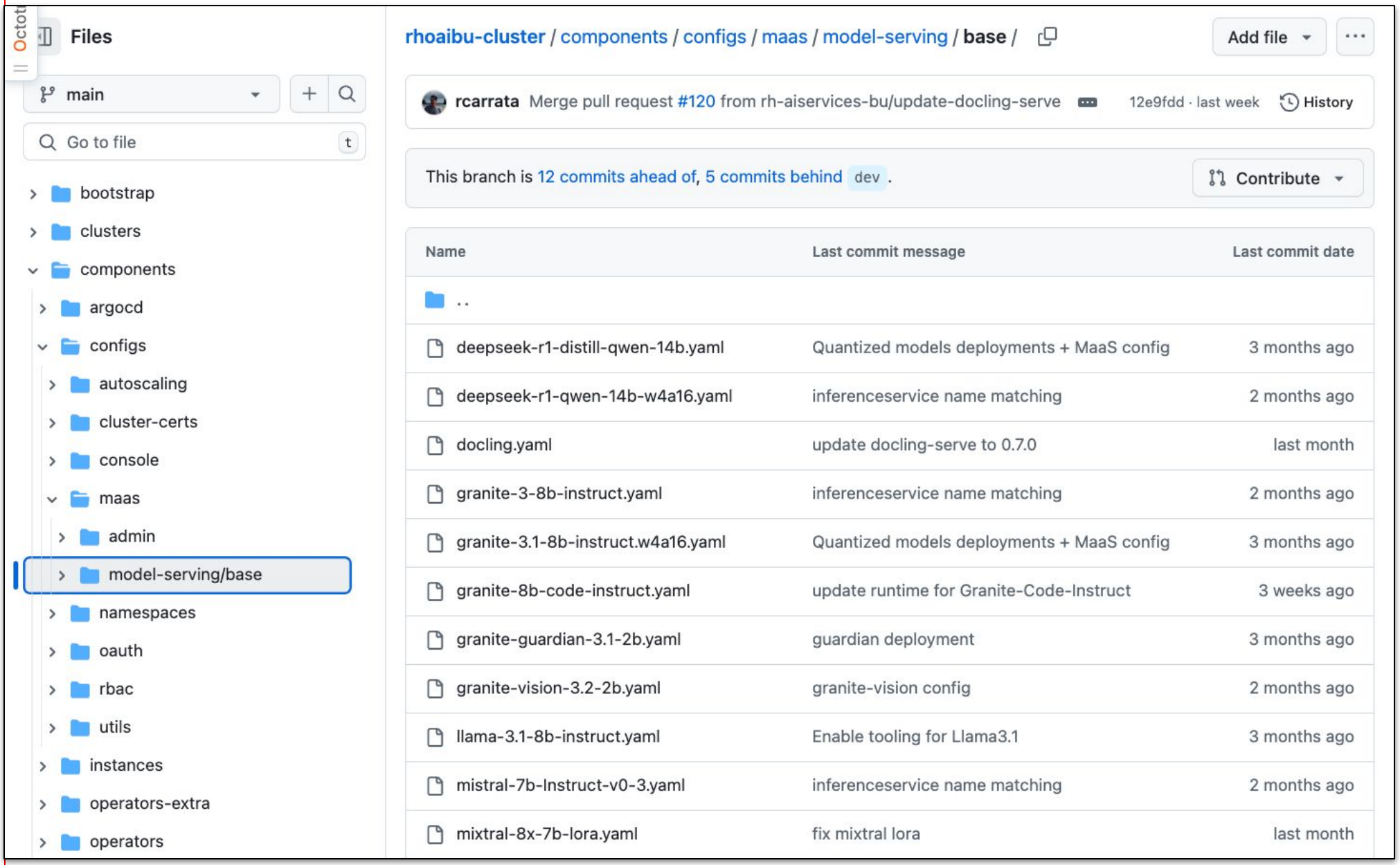
Red Hat

But who's keeping track?



- No such thing as a free AI lunch
  - But MaaS will minimize the costs
  - then further reduce costs, by using quantized models
- Manage your models with GitOps
- API gateway keeps tracks of hits and tokens
- Chargeback proportional to usage/tokens

Centrally and reliably manage your models via GitOps



Red Hat

Centrally and reliably manage your models via GitOps

rhoaibu-cluster / components / configs / maas / model-serving / base / mixtral-8x-7b.yaml

mixtral deployment

Code | Blame | 85 lines (85 loc) · 2.3 KB | Code 55% faster with GitHub Copilot

```yaml
apiVersion: serving.kserve.io/v1alpha1
kind: ServingRuntime
metadata:
  annotations:
    opendatahub.io/accelerator-name: ''
    opendatahub.io/apiProtocol: REST
    opendatahub.io/recommended-accelerators: '["nvidia.com/gpu"]'
    opendatahub.io/template-display-name: 'vLLM-RHOAI 2.17-Mixtral-8x7B'
    opendatahub.io/template-name: vllm-mixtral-8x7b-instruct-v0-1
    openshift.io/display-name: vllm-mixtral-8x7b-instruct-v0-1
  name: vllm-mixtral-8x7b-instruct-v0-1
  labels:
    opendatahub.io/dashboard: 'true'
spec:
  annotations:
    prometheus.io/path: /metrics
    prometheus.io/port: '8080'
  containers:
    - args:
        - '--port=8080'
        - '--model=/mnt/models'
        - '--served-model-name=mistralai/Mixtral-8x7B-Instruct-v0.1'
        - '--distributed-executor-backend=mp'
        - '--dtype=float16'
        - "--tensor-parallel-size=4"
```

sync

Project: llm-hosting

ServingRuntimes › ServingRuntime details

SR vllm-mixtral-8x7b-instruct-v0-1-lora

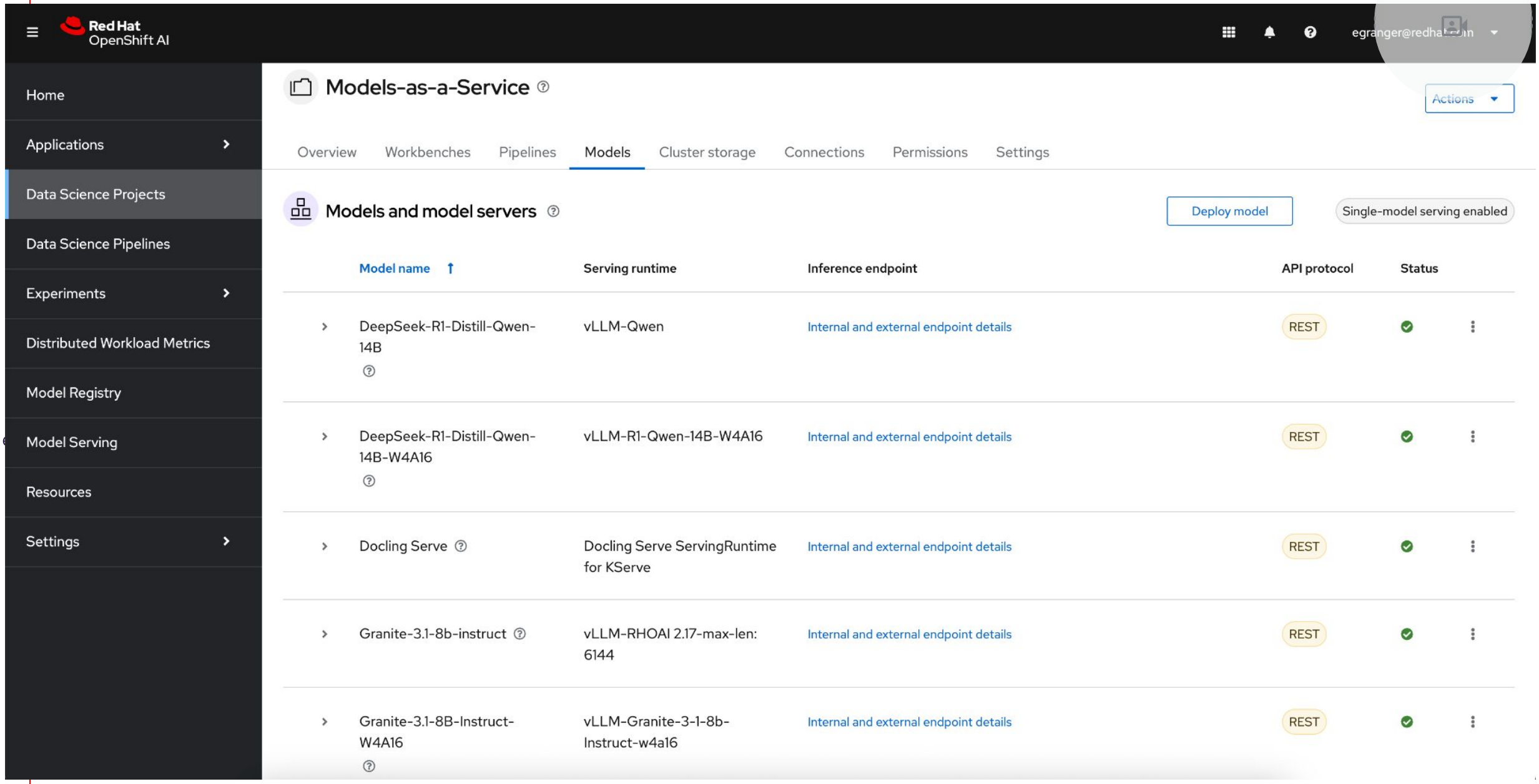
Details | YAML

```yaml
apiVersion: serving.kserve.io/v1alpha1
kind: ServingRuntime
metadata:
  annotations:
    opendatahub.io/accelerator-name: ''
    opendatahub.io/apiProtocol: REST
    opendatahub.io/recommended-accelerators: '["nvidia.com/gpu"]'
    opendatahub.io/template-display-name: vLLM-RHOAI 2.17-Mixtral-8x7B
    opendatahub.io/template-name: vllm-mixtral-8x7b-instruct-v0-1-lora
    openshift.io/display-name: vllm-mixtral-8x7b-instruct-v0-1-lora
  resourceVersion: '1231526846'
  name: vllm-mixtral-8x7b-instruct-v0-1-lora
  uid: 5428e15f-231f-42a1-93dd-1fb4e31ddbda
  creationTimestamp: '2025-03-26T20:02:02Z'
  generation: 3
  managedFields:
    - apiVersion: serving.kserve.io/v1alpha1
      fieldsType: FieldsV1
      fieldsV1:
        'f:metadata':
          'f:annotations':
```

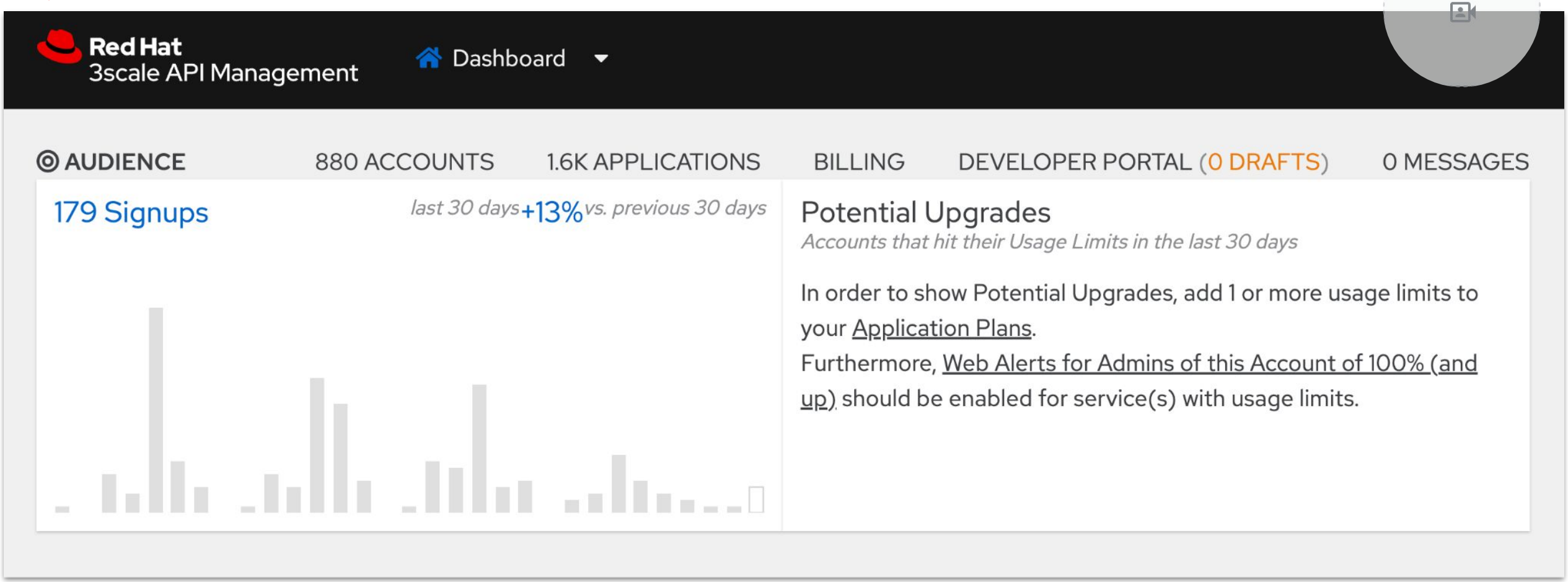
Save | Reload | Cancel

Red Hat

Centrally and reliably manage your models via GitOps

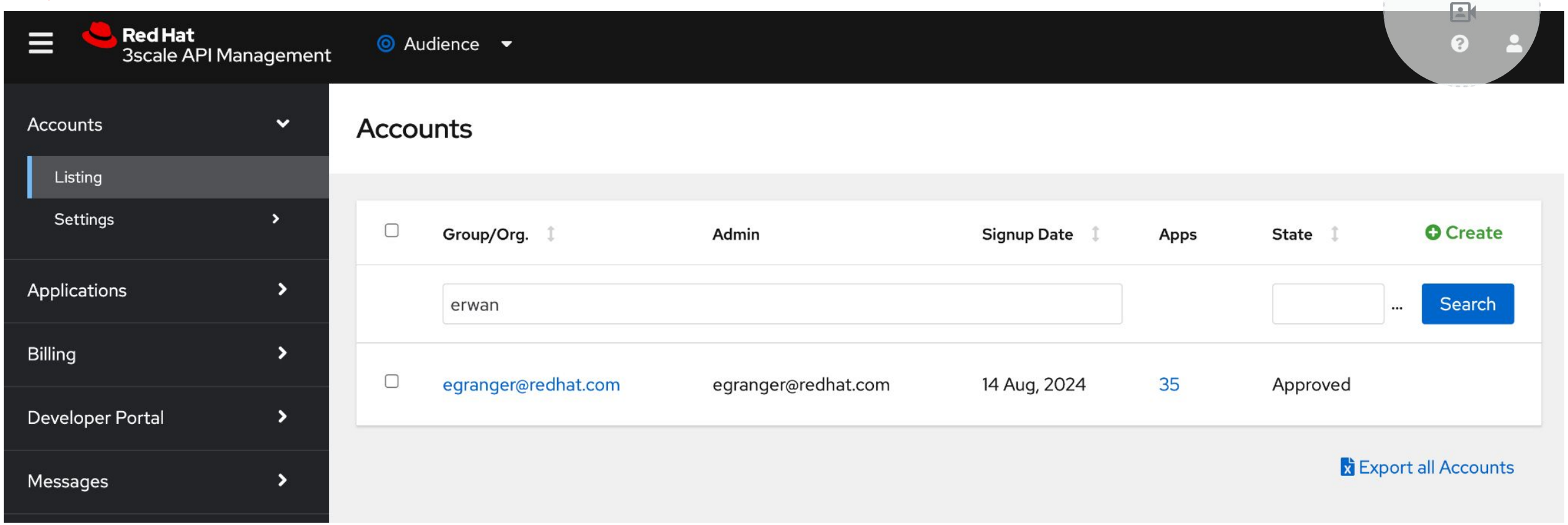


Your API Gateway tracks things

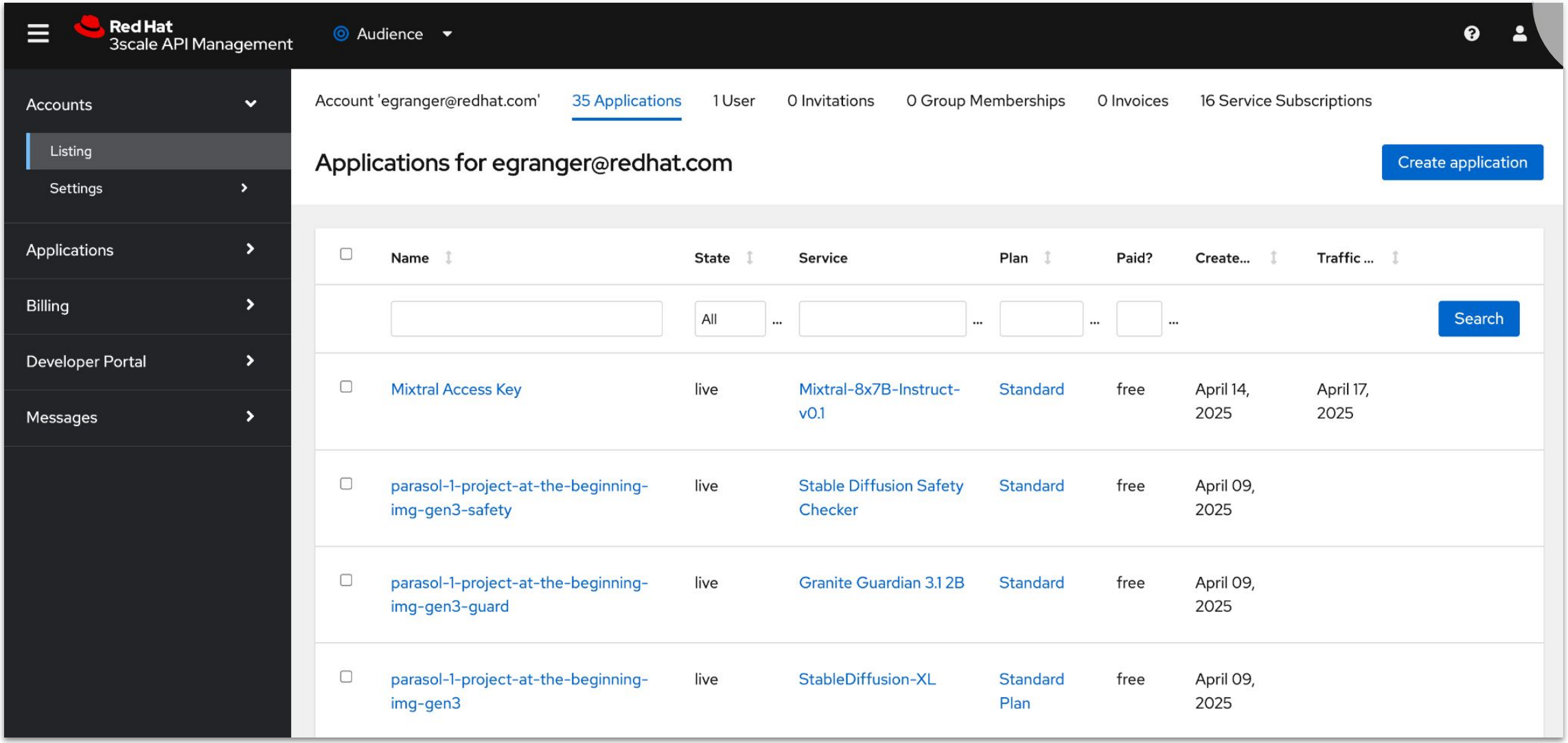


Red Hat

You can see how many "Apps" your developers have created

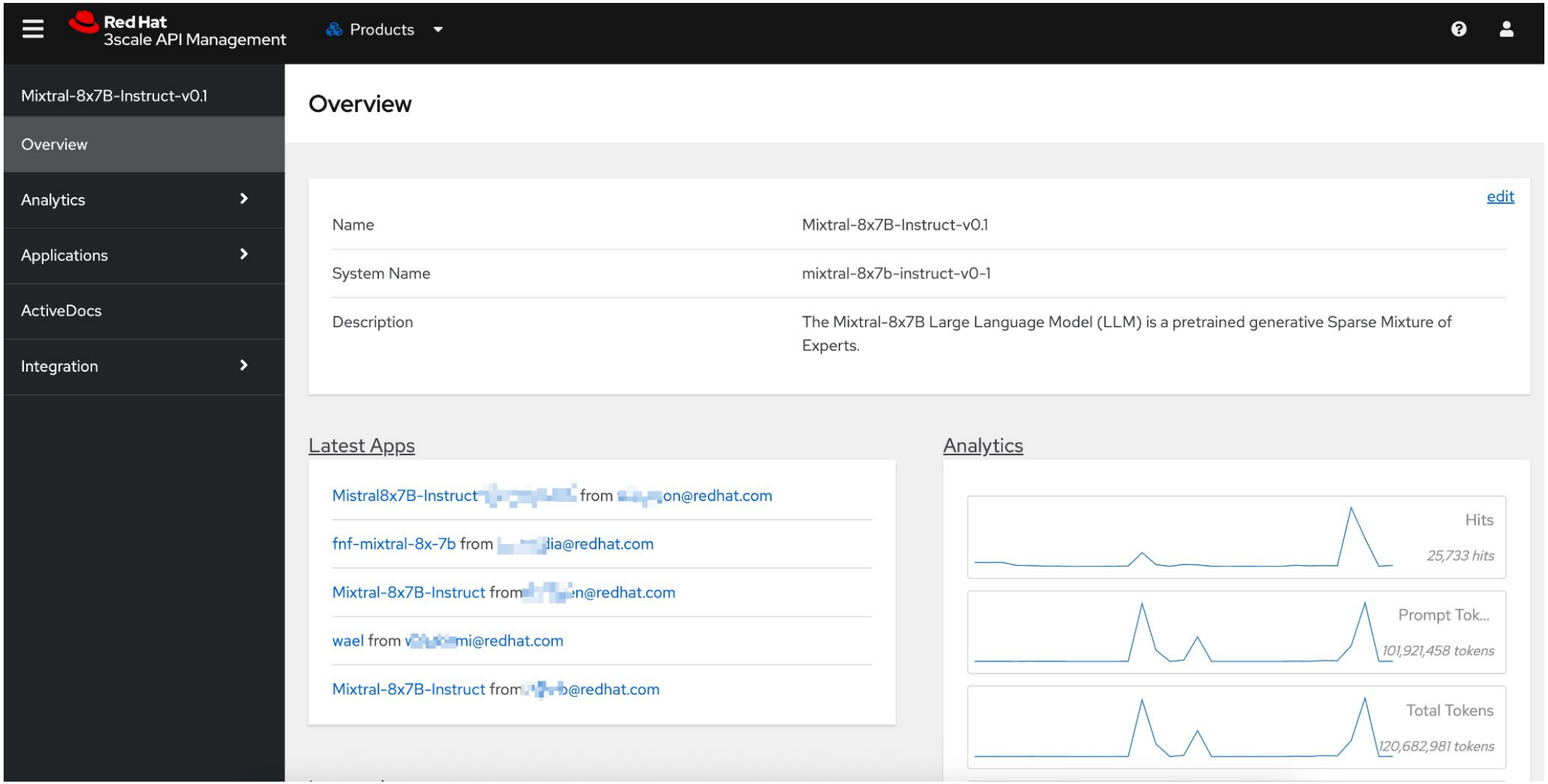


And the details for one user

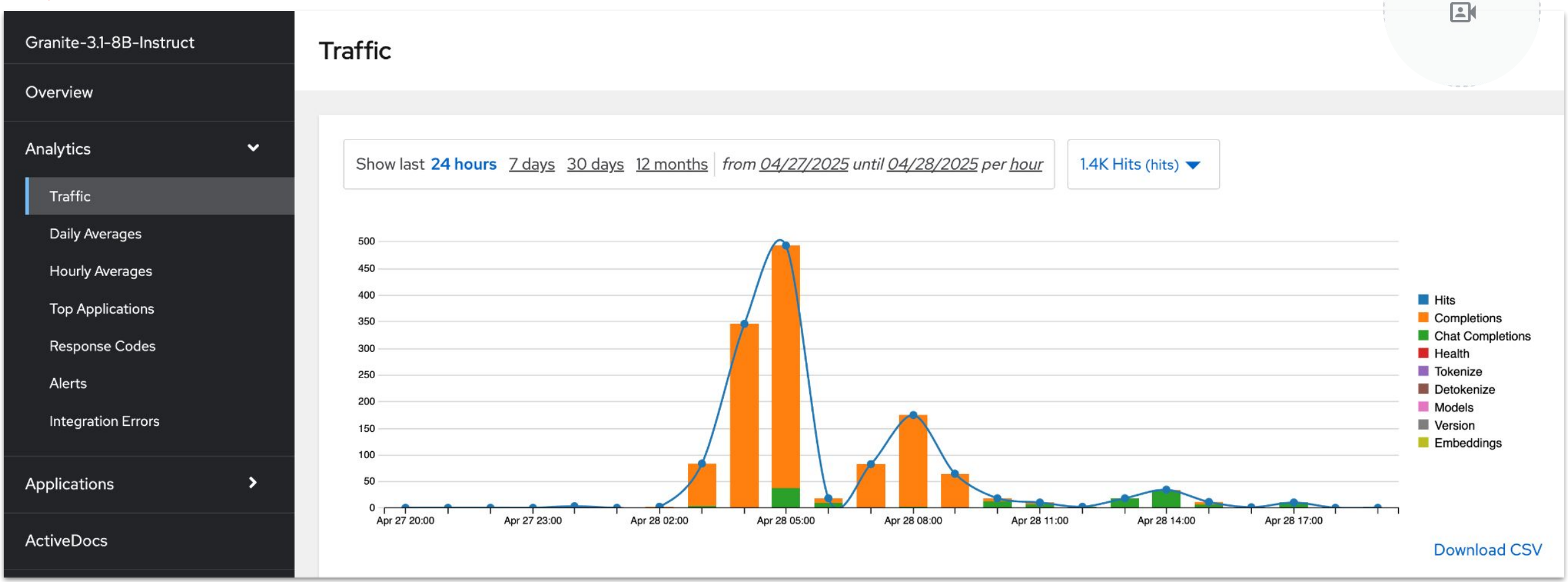


Red Hat

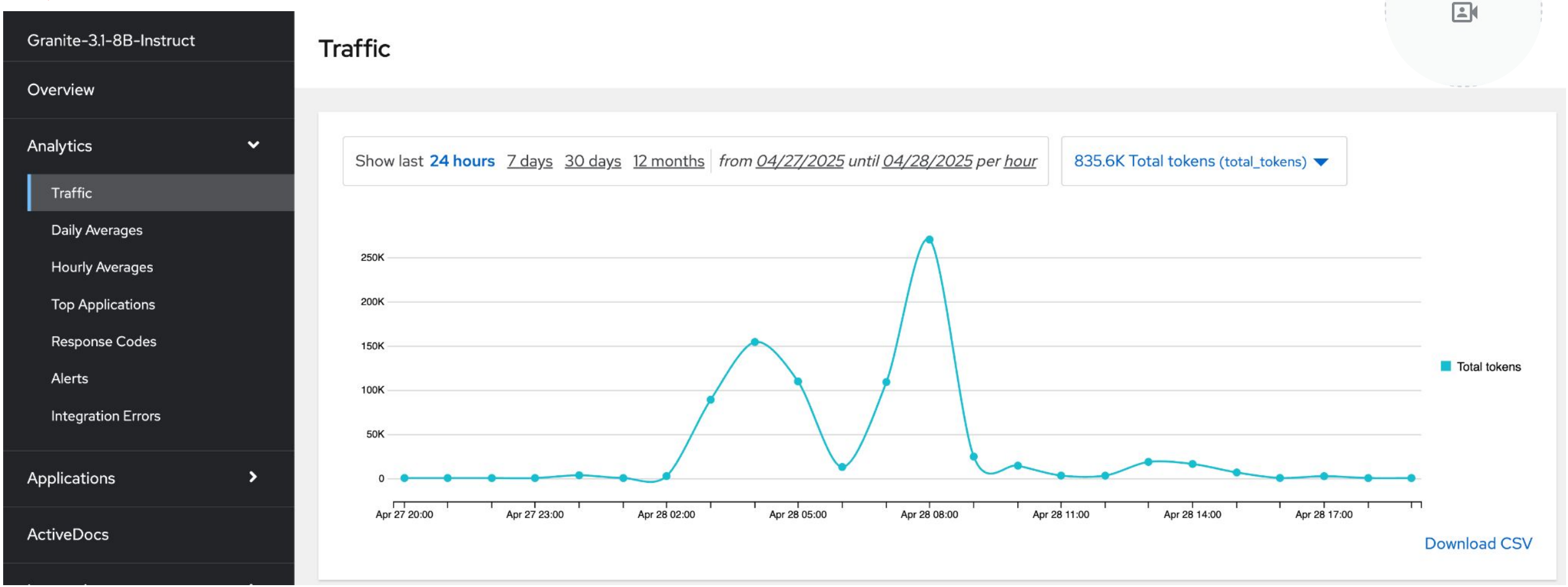
Track the usage of one model over time



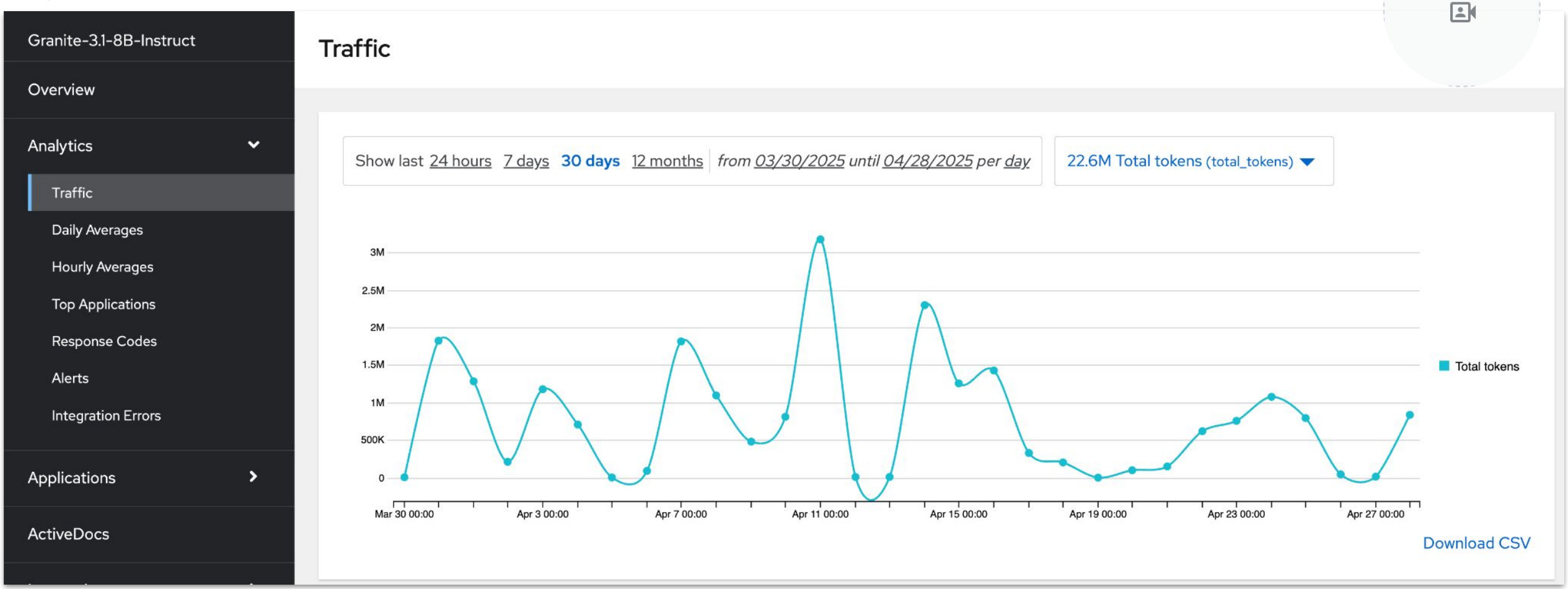
Number of Hits over last 24H



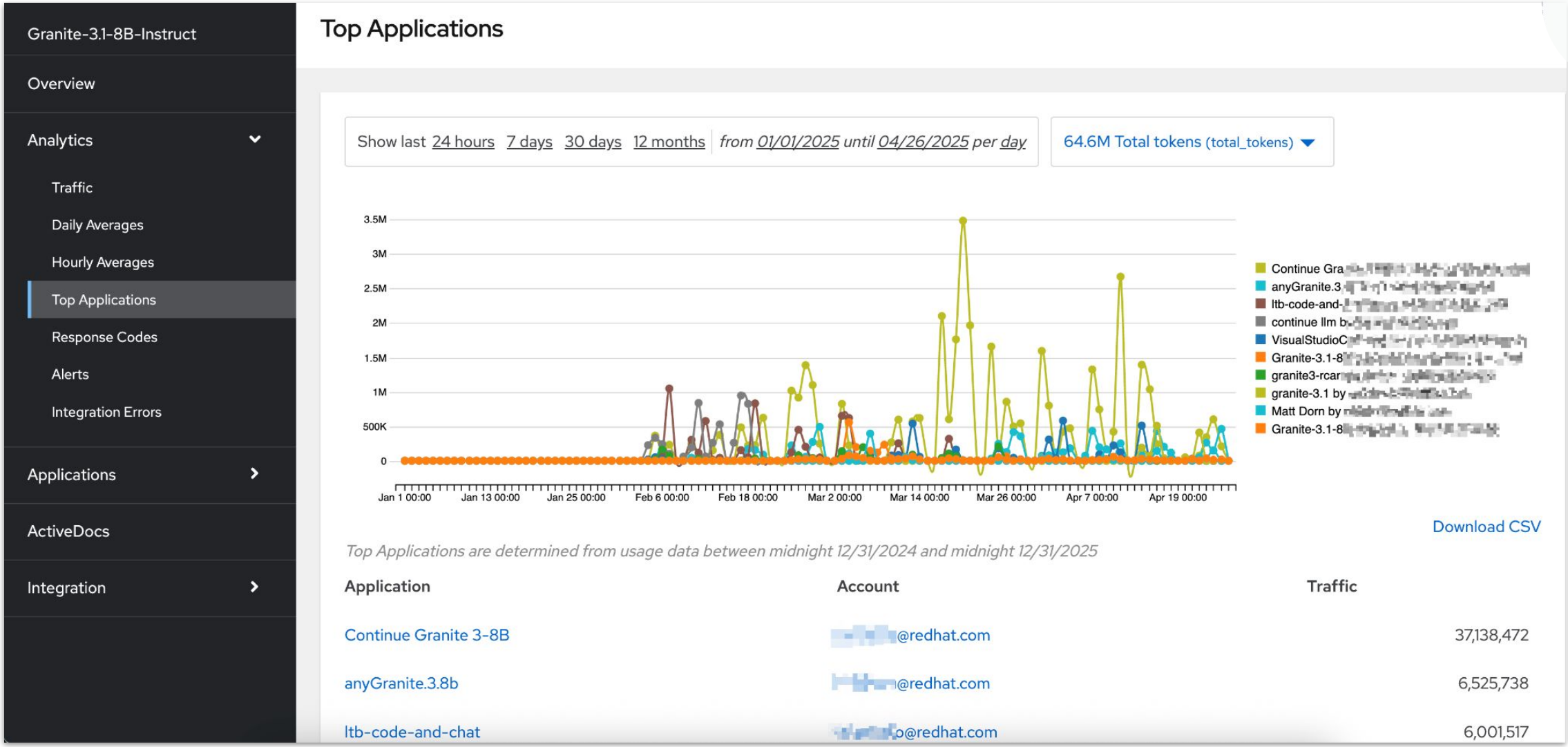
Number of Tokens over 24H



And over 30 days



But who's using this model the most?



Red Hat OpenShift AI

Integrated AI platform

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

Available as

- Fully managed cloud service
- Traditional software product on-site or in the cloud!

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

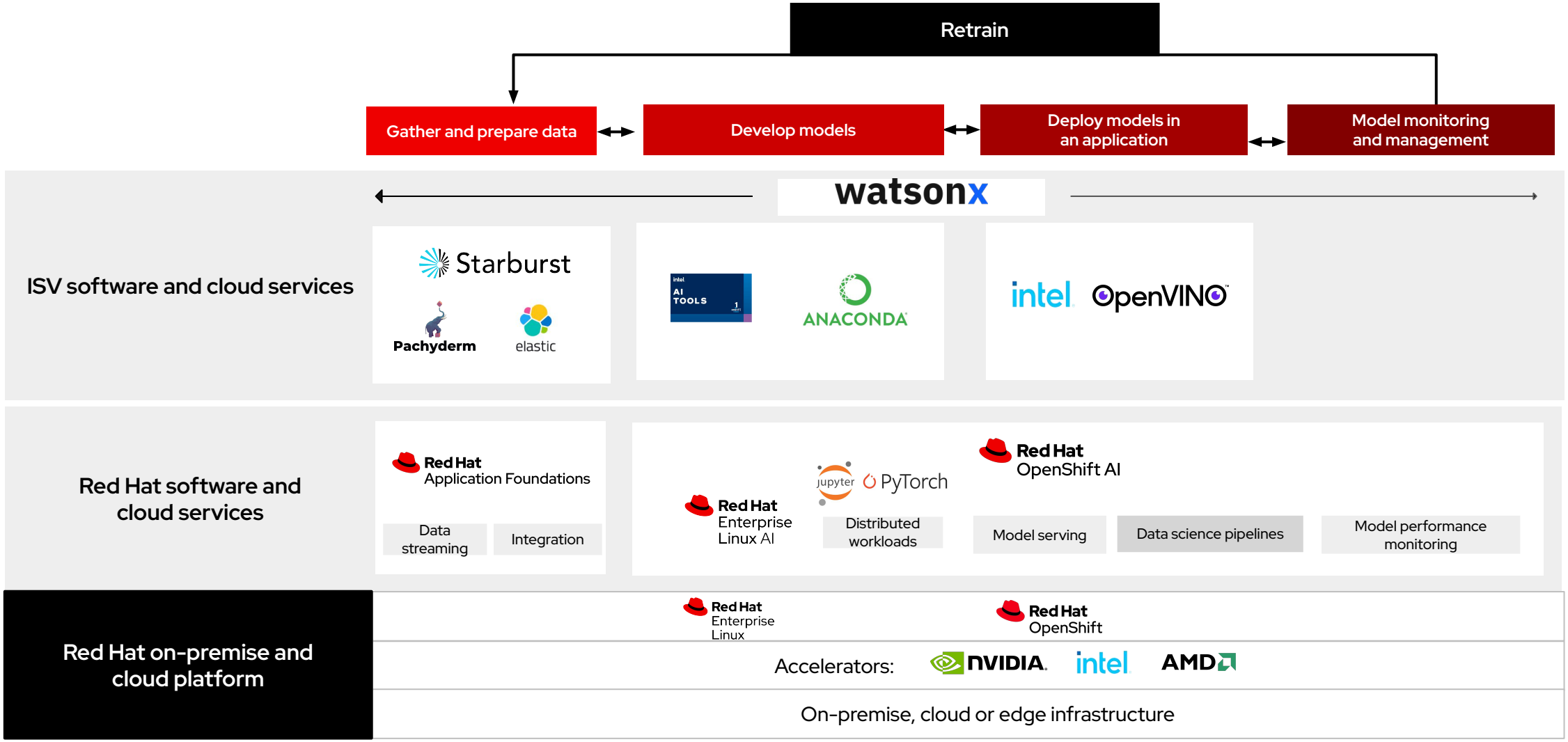
Expand **DevOps practices to MLOps** to manage the entire **AI/ML lifecycle**.

Resource optimization and management

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

Red Hat

OpenShift AI Components



Red Hat

Hybrid cloud deployment for AI

Across different hardware accelerators, on-prem OEM servers, and cloud environments

Hardware Accelerators

- NVIDIA GPU
- AMD Instinct
- intel Gaudi
- Google TPU
- aws Neuron
- IBM AIU

OEM Servers

- DELL
- Lenovo
- CISCO
- Hewlett Packard Enterprise

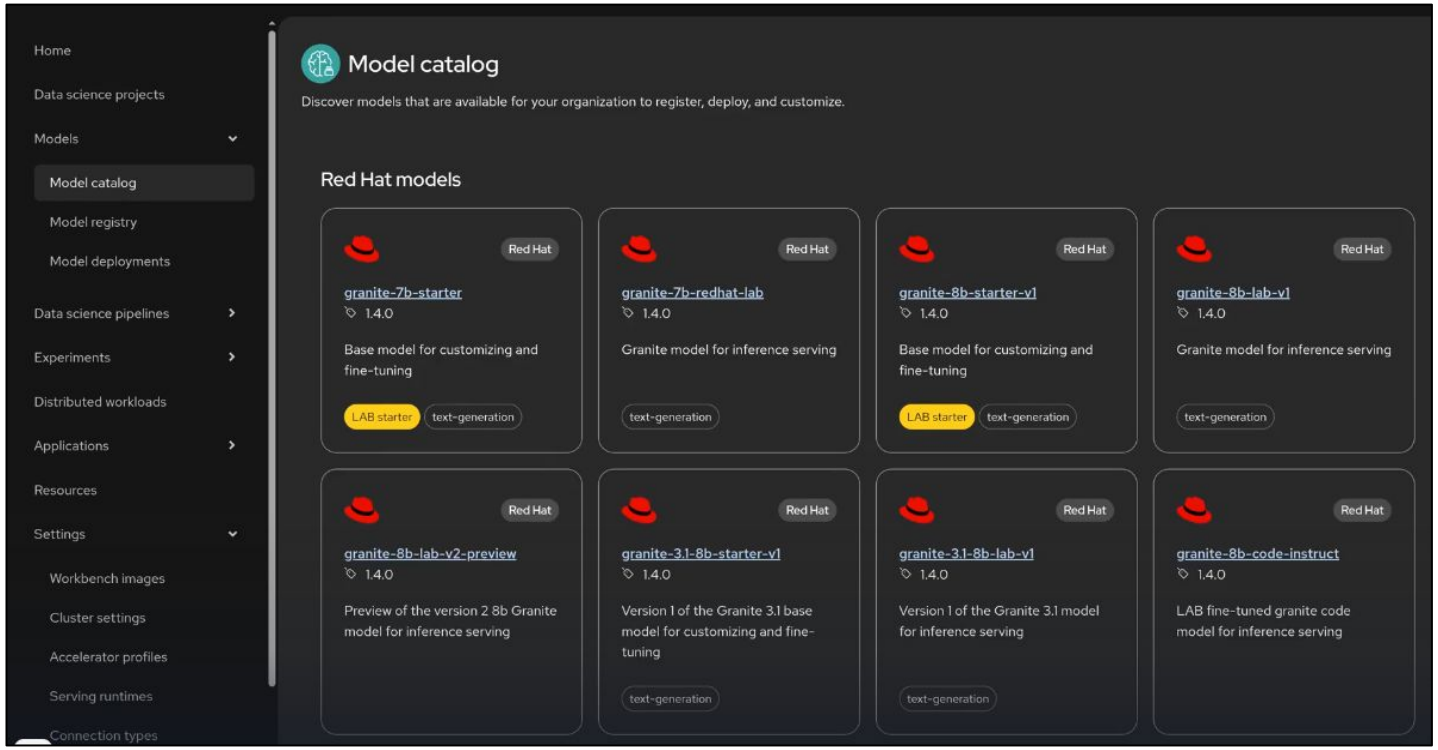
Cloud Environments

- IBM Cloud
- aws
- Azure
- Google Cloud

* NVIDIA GPUs fully supported in Red Hat AI. AMD Instruct $ Intel Gaudi in Preview, AWS Inferentia/Neuron, Google TPU, IBM AIU are on our roadmap

Model Catalog

OpenShift AI users now have the ability to access Red Hat and 3rd party models and easily deploy them from within the application.



Red Hat