Red Hat AI

Skill Up on Red Hat AI

Sales and Tech Sales Training

Eoin Crosbie

Global Senior Learning Architect - AI

Red Hat

Q: Where is your company at in it's AI journey?

1. New to AI
2. Have been experimenting internally
3. Currently involved AI projects
4. Have been involved in multiple projects

Q: Do you have Data Science / AI Dev capabilities in-house?

1. No - no plans to
2. No - but plan to grow in that area
3. No - we leverage our ecosystem
4. Yes, we are quite new in this area
5. Yes, we are mature in this area

Unlocking AI with Red Hat for Partners – Enablement

1. **Master Your First Meeting** - The Red Hat AI Opportunity
2. **Skill Up on Red Hat AI** - Sales and Tech Sales Training
3. **Openshift AI Architecture Workshop** - incl. Hands on lab
4. **AI PoC Enablement** - How to approach AI POCs
5. ***Openshift AI WaaS*** - *Workshop as a Service*
6. ***Customer POC***

What we'll discuss today

- AI Landscape
- Red Hat and AI
- Demoing Red Hat AI
- Customer Successes
- POC Opportunity

AI Landscape

AI is becoming a part of our everyday lives

Chat GPT

watsonx Code Assistant

IBM Granite Models

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

Generative AI is a strategic enabler across industries

Using AI to drive productivity and efficiency

Revenue Generation

- Chatbots
- Campaign and Content Marketing
- Developer assistants
- Guided selling
- Drive product innovation

Cost Optimization

- Automated AI support
- Knowledgebase Search & Summarization
- Doc summarization
- AI-optimized logistics
- Augmented Product R&D

Risk Management

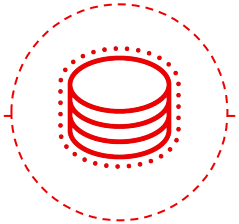
- Sentiment analysis
- Predict employee attrition
- Contract risk assessment
- Fraud detection
- AI-assisted Security Operations

Why is NOW a good time to invest in AI?

Investing in AI can allow organizations to gain competitive advantage

AI technologies are becoming **more accessible and affordable** for businesses of all sizes

Companies can **realize the value** from AI-enabled applications and AI-support



Organizations are better prepared to manage, transform and **use their ever-increasing data**

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 (as of 2022) 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

What are the challenges impeding progress?

Understanding the nuances

What type of AI should you invest in?

Building the right infrastructure

How do you choose and build the right environments for development and production?

Realizing value

How do you safely realize the business benefits of AI/ML innovation?

87% of data science projects never make it into production[1]

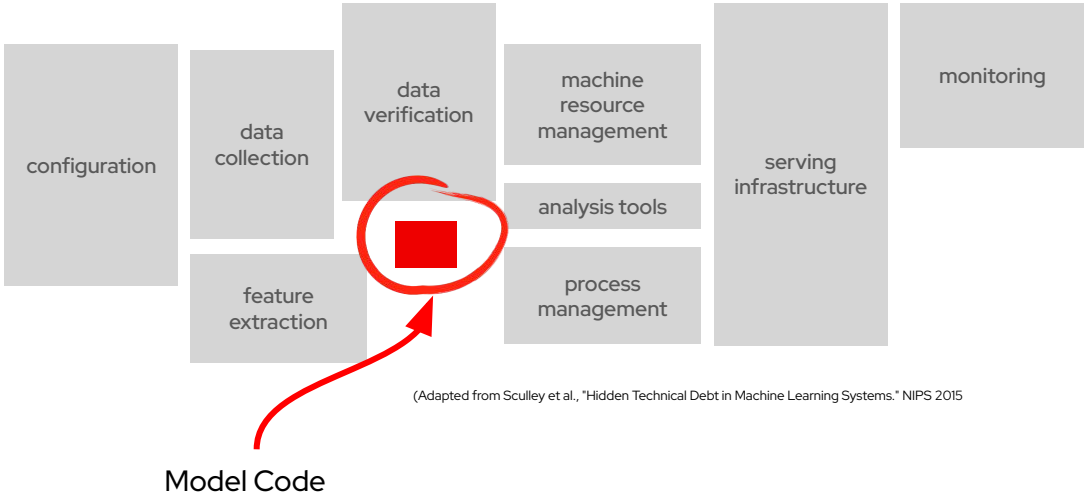
Source:
IDC - Market Analysis Perspective: Worldwide Artificial Intelligence Software, 2021

Operationalizing AI is one of the biggest challenges

Average timeline from idea to operationalizing the model

7-12 months to operationalize a model 50%

+1 year to operationalize a model 26%



(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015

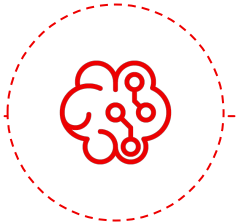
Source:
Gartner Peer Insights, Open Source AI for Enterprise survey, 2023

Personas

Operationalizing AI projects is a team sport

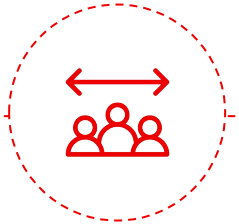
Line of Business

- Time to business value
- Reliability of predictions
- Responsive to evolving business



Data Scientist, Data Engineer, App Developer

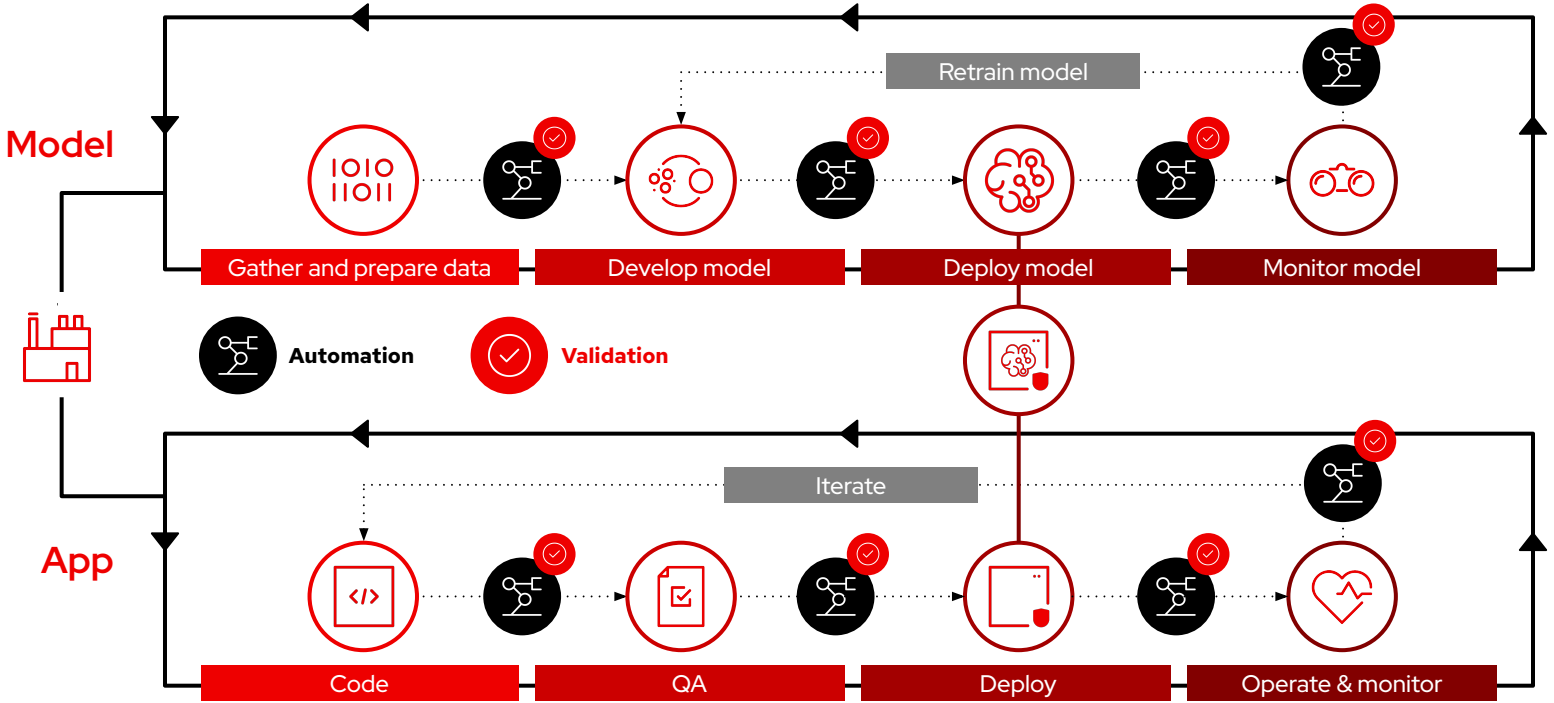
- Self-service access to tools, data & infrastructure
- Collaborative environments for fast AI-enabled apps creation



IT Operations

- High availability
- Security
- Ease of management
- Scalability
- Enhance existing vs “rip and replace”

Lifecycle for operationalizing models



The business value of Inference

AI Inference Market expected to grow from $106B (2025) to $255B (2030)

AI Inference Market worth $254.98 billion by 2030, Feb 2025, PR Newswire

Red Hat

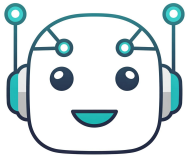
What is Inference?

User using AI app

User submits request

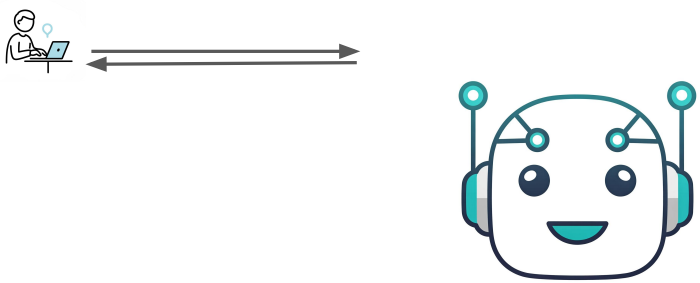
AI returns a response

AI Model processing answer with an **inference server**

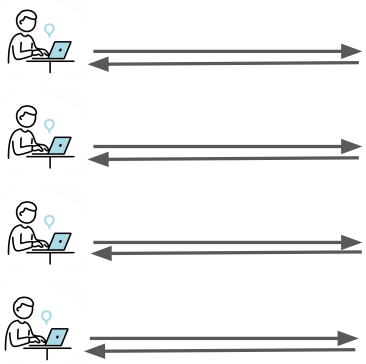


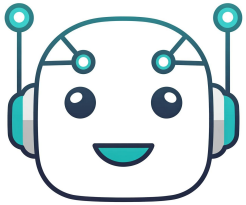
Goal: Meet user expectations. Return the most accurate response as fast as possible

Inference at scale



Inference at scale





Red Hat and AI

What are the challenges impeding progress?

1. *Operationalizing AI is hard*
2. *AI Inference at scale is Hard*

How do I maximise the efficiency of scarce resources?

Red Hat AI

Red Hat AI Inference Server

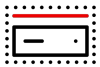
Red Hat Enterprise Linux AI

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

Red Hat AI Inference Server

Gen AI model inference

- Packaging: Linux container
- Red Hat vLLM inference server
- Validated & optimized model repository
- Granite family models (with indemnification)
- 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 Enterprise Linux AI

Gen AI model inference

- Packaging: Linux server appliance
- Optimized RHEL image with integrated accelerators
- **Includes entitlement to Red Hat AI Inference Server**

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

Red Hat OpenShift AI

Gen AI model inference, training & AIOps

Predictive AI inference, training and MLOps

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

For customers who need a complete distributed Gen AI/Predictive AI platform for inference, training and AIOps

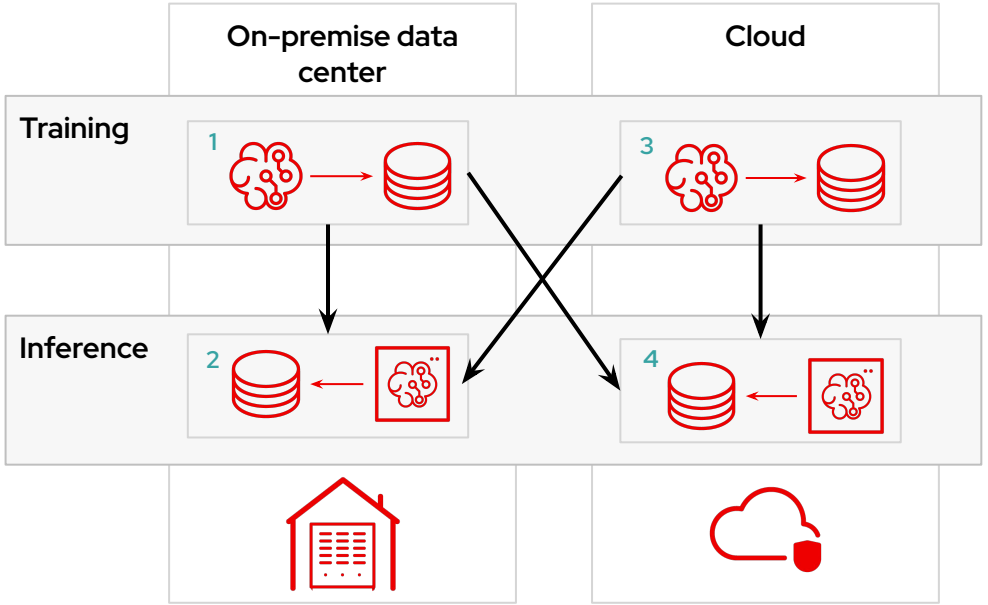
Red Hat

Gain hybrid cloud flexibility

Train and deploy models and AI-enabled apps on-premises, cloud or edge

What you do should not dictate
where you do it

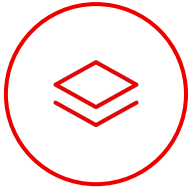
1. Data on-prem = Train on-prem
2. Data on-prem = Inference on-prem
3. Data in the cloud = Train on cloud
4. Data in the cloud = Inference on cloud



Red Hat

Red Hat OpenShift AI

Develop, train, serve, monitor, and manage the life cycle of AI/ML models and applications, from experiments to production.



Built on top of OpenShift®

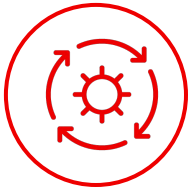
Deliver consistency, **cloud-to-edge** production deployment and monitoring capabilities

Designed for Gen and predictive AI

Scale to meet the workload demands of foundation models and traditional ML.

Empowered collaboration

Provide a **unified platform** for **data scientists** and **intelligent application developers**

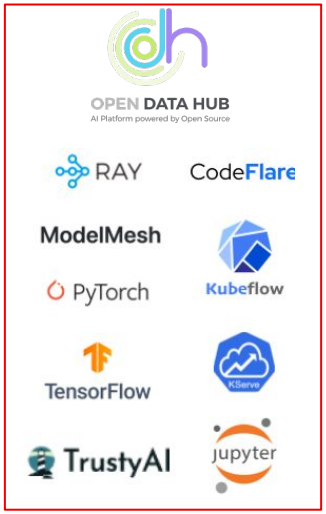


DevOps applied to ML

Set up rigorous **pipelines** and **workflows** to take you from **development** to **production**.

Simplify AI adoption

Designed to increase AI adoption and enhance trust in AI initiatives



Red Hat OpenShift AI

Flexible

A composable platform for rapid dev and delivery of AI-enabled apps

Expand

A certified AI partner ecosystem for delivering an E2E AI/ML experience

Red Hat

What are the challenges impeding progress?

1. *Operationalizing AI is hard*

   OpenShift's DevOps and MLOps capabilities allows all stakeholders to effectively contribute to the lifecycle of an application

2. *AI Inference at scale is Hard*

How do I maximise the efficiency of scarce resources?

Fast, flexible and scalable Inference

The requirements of an enterprise AI production systems

Identifying the tradeoffs of inference

Need to be fast and **accurate** in its responses

Manage processing times and token output to control **cost**

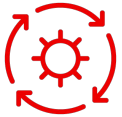
Deliver high throughput and lower latency for best **performance**

Red Hat AI helps address the trade-off challenges of inference

Gain consistent, fast and cost-effective inference at scale with vLLM

Select an optimal LLM

Red Hat AI third party validated and compressed models ready-to-use



An efficient inference runtime

vLLM

Broad support for hardware

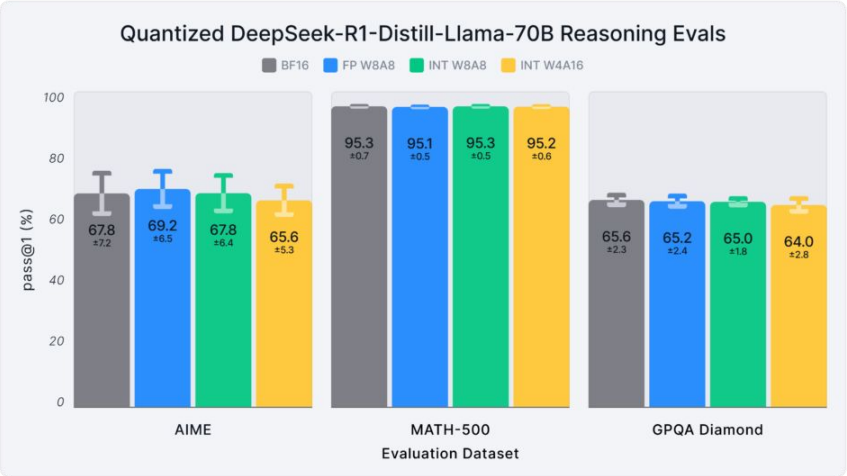
AMD NVIDIA IBM Google intel aws

Red Hat

Ex. 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)



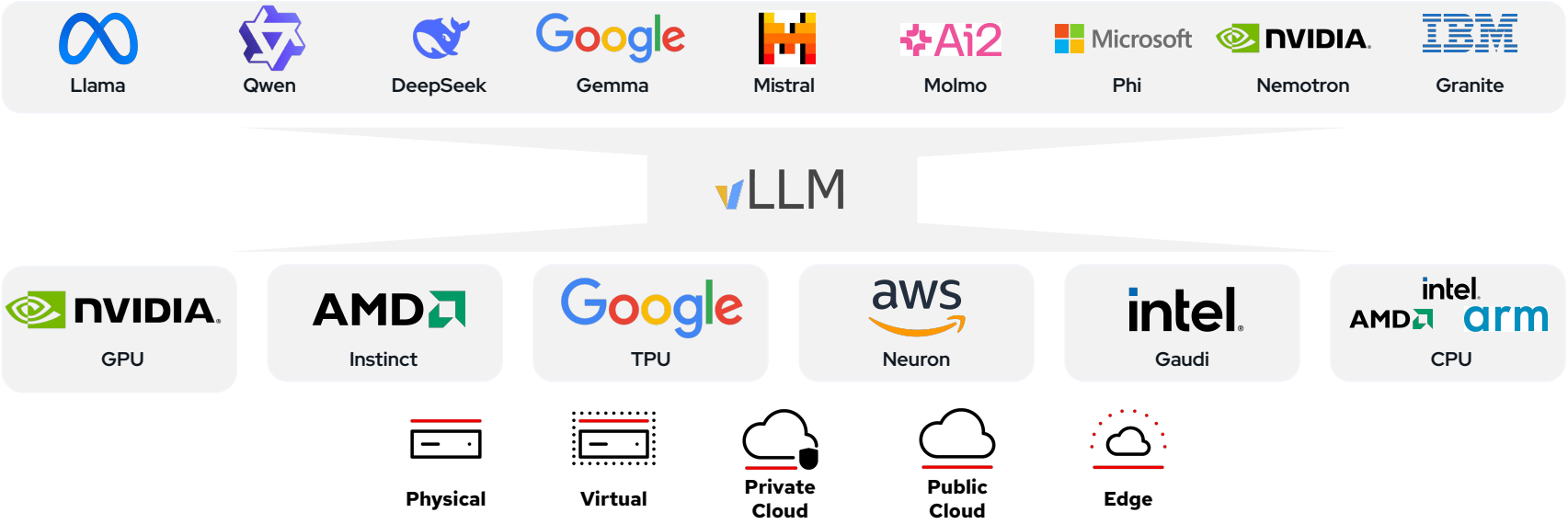
https://developers.redhat.com/articles/2025/03/03/deployment-ready-reasoning-quantized-deepseek-r1-models#

Interactive demo

Red Hat

vLLM inference server in Red Hat AI

vLLM supports the key models on the key hardware accelerators



Single platform to run your models across Accelerators and OEMs

Red Hat

Summary: Why Red Hat for *you*?

- We solve important problems for your customers
- MLOps and Model-as-a-Service can become the long-term platforms for customer AI projects
  - Become a strategic partner for your customers
  - Opportunity for significant product revenue
  - Opportunity for significant consulting revenue

Demoing Red Hat AI

Demo Platforms for Partners

Link	Access	Lifetime	Description	Use Case
demo.redhat.com	RH Account	1-2 days*	Structured demos on real infrastructure	Multi-user workshop, Technical Demos, Labs, Solution experimentation, PoC
partner.demo.redhat.com	RH Partner	1-3 days	Structured demos on real infrastructure	Partner Demo, Labs, Workshops
https://www.redhat.com/en/interactive-labs (Instruqt)	public	1-2 hours	Structured scenarios on locked down infrastructure	Feature preview, specific scenario
developers.redhat.com	RH Account	30 days	Small Sandbox environment	Platform exploration, Solution experimentation
Interactive Experience (Arcade)	public	infinite	Static wireframe click-through	Product Storytelling, shareable resource

- How do you Demo?
- Who do you demo to?
- What story are you telling?

Scenario

Administrator:

Setting: **informal**

Sentiment: **relaxed**

Outcome: **inspire**

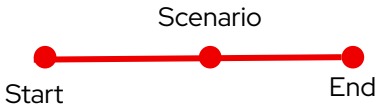
CIO:

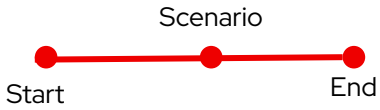
Engagement: **25 minutes to demonstrate the technology**

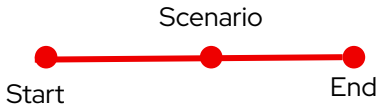
Have you ever seen such a Demo?

Red Hat

How you COULD Demo!



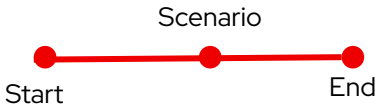




How you COULD Demo

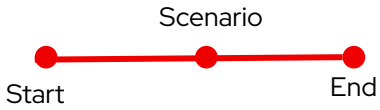
Episode 1

"It's easy to build, serve, and infer from an AI model with OpenShift AI"



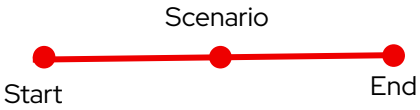
Episode 2

"Take advantage of enhanced AI features, like Data Science Pipelines"



Episode 3

"A single platform for data scientist, operations, and developers to collaborate - accelerating time to value"



How you COULD Demo

Episode 1

Administrator:

Scenario

Start — End

Episode 2

Administrator:

Scenario

Start — End

Episode 3

CIO:

Scenario

Start — End

Red Hat

How you COULD Demo

Turn your features into benefits

Which value should be demonstrated?

What demand is satisfied in showing something?

Integration details (how it works)

Does it work for everyone?

The set up

Various places of demo assets

Know your/the limitations

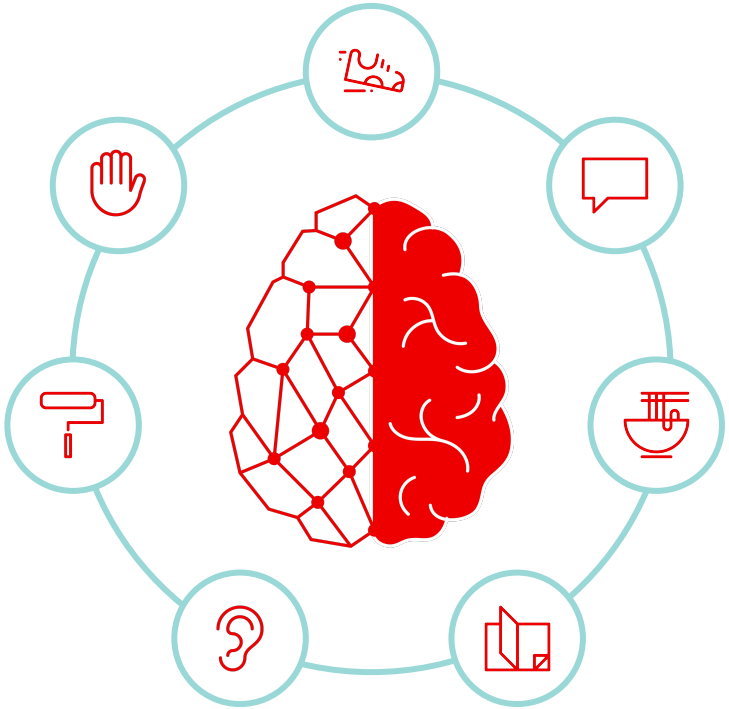
Make the demo available to run and analyze

Storytelling

Your prospects are wired for storytelling

Storytelling is an art form, but science explains why the human brain loves stories.

Infact, storytelling activates seven regions of the brain (including sensory areas such as visual, auditory, and haptic).[1]



1 Bruder, Emma. "The neuroscience behind storytelling in sales," Hubspot. Accessed Aug 2022.

Storytelling

Storytelling makes sales memorable

Storytelling - an opportunity to bring personality to what you are selling.

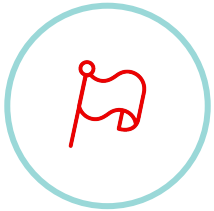
Think about the kind of sales pitch you'd want to hear:

Features or real-life scenarios?

Get your storytelling basics down

Who is the main character?

What main challenge does the character face?



How will the character overcome the challenge?

Determine the takeaway

Envision the endgame:
It determines if your story is informative, inspiring or ends as a drama.

What's the key takeaway you want the listener to get after you finish your story?

Use your experience
(such as projects; case studies) to tailor the message to your prospect.

Preparation is key

Now Your Turn

Customer

Persona

Demo

Story

Now You Try

30 minutes

Pick a Customer - Pick Persona(s) - Pick Demo Asset(s)

Demo Assets

https://www.redhat.com/interact

>100 Red Hat AI *demos*

1. Red Hat helps with AI Adoption
2. MaaS
3. InstructLab: Training an LLM
4. Serving an LLM locally with InstructLab
5. Red Hat AI at the Edge
6. RHELAI and InstructLab*

*multi-part demo

Now You Try

30 minutes

Pick a Customer - Pick Persona(s) - Pick Demo Asset(s)

Demo Assets

https://www.redhat.com/interact

>100 Red Hat AI *demos*

1. **A real estate company**
2. **A banking company**
3. **A city government**
4. **A large global telecommunications company**

Now You Try

30 minutes

Pick a Customer - Pick Persona(s) - Pick Demo Asset(s)

Demo Assets

https://www.redhat.com/interact

>100 Red Hat AI *demos*

1. **A real estate company** wants to streamline how their teams address real estate-related inquiries. (want an LLM chatbot)
    a. The model must understand how the layers of documents from different time periods interact to answer queries correctly (some land deals occurred in the 1800s, for example, and laws/leases/agreements have changed over time).
    b. InstructLab: Train a model on real estate data to achieve above needs.
2. **A banking company** wants to help customer service representatives handle credit card disputes more easily. (want LLM chatbot)
    a. InstructLab: train model on bank-specific data.
3. **A city government** wants to improve and streamline how it serves its constituents.
    a. They want to:
        i. Generate summaries of legislation
        ii. Create knowledge agents to assist with eligibility questions for public health services, assist junior engineers in the department of transportation, and support public transportation programs
    b. InstructLab: fine-tune a model on city data to serve above use cases.
4. **A large global telecommunications company** receives over 300k customer calls daily. Call transcripts need to be quickly analyzed for business & client impact to roll back into client-facing processes to improve automation and customer satisfaction.
    a. Desire for quality improvement, lower cost, quick results, data privacy built-in.
    b. InstructLab: align their own custom LLMs with their private corporate data.
5.

Red Hat

Any Volunteers?

Customer

Persona

Demo

Story

Customer Successes

Public sector entity built a chatbot for confidential documents

Infrastructure agency

Customer is responsible for road and rail infrastructure of a country.

30,000+ employees

Challenge

The documents that employees reference for their daily duties are needed but are also preventing the organization from being effective.

Approach

Use an AI framework & LLMs for different parts of the organization to support the thousands of associates that need guidance from the internal documentation.

Technical requirements

- Air gapped functionality.
- Enable product teams to own the service.
- Support a variety of accelerators for current and future needs.
- Ability to optimize and manage the compute resources for training and inferencing.

Implementation approach

Define business & technical requirements.

Build a proof-of-concept of a RAG solution for a subset of the documents covering a specific domain.

Prove the functionality for the organization and confirm value to get buy in to build a MVP.

International airline starts applying AI across their organization

International Airline

One of the top ten largest airlines in the world – based on number of passengers.

It operates more than 300 destinations across several continents.

Challenge

Enable rapid AI innovation efforts across the organization – while also optimizing compute resources.

Approach

Help business and IT teams adopt MLOps practices with an analytical sandbox and scalable MLOps environment. .

Technical requirements

- Consistent self-service approach for data scientists
- Unified platform that allow for model experimentation, development, and deployment.
- Ability to optimize and manage the compute resources for training and inferencing.

Implementation approach

Define business & technical requirements.

Build an analytical sandbox and a scalable MLOps environment to adopt MLOps practices

Prove the value then scale to production

POC Opportunity

GenAI POCs

1000 / 200 / 50

Basic Requirements

All qualified PoCs must contain the following:

- An LLM served using Red Hat Inference Server (vLLM) on customer infrastructure

What's in it for Partners

- Find the low hanging fruit (**OCP | AI strategy**)
- Initiate discussions with your Customers (**MYFM**)
- Connect with GAP partner to speed up sales process
  - Cisco
  - Dell
  - HPE
  - IBM
  - ....
- Co-create a solution with Red Hat
- Position and sell your services

What's Next?

- Plan your *Episodes*
- Connect with the Red Hat Ecosystem team
  - Create a joint initiative in your Partner Success Plan (RHPSP)
  - Connect with Red Hat Distributor
- Complete the AI267 Certification course in PTP
- *Attend the RHOAI Architecture Workshop*
- Attend one of our Meet the Expert sessions to discuss real-world situations and findings

Q: What aspects of our Product/features/messaging Are still unclear for you?

Let us know in the chat…

Q: What other next steps might you need to excel?

Let us know in the chat…

Q: What do you expect from Red Hat?

Let us know in the chat…

Unlocking AI with Red Hat for Partners – Enablement

1. **Master Your First Meeting** - The Red Hat AI Opportunity
2. **Skill Up on Red Hat AI** - Sales and Tech Sales Training
3. **Openshift AI Architecture Workshop** - incl. Hands on lab
4. **AI PoC Enablement** - How to approach AI POCs
5. ***Openshift AI WaaS*** - *Workshop as a Service*
6. ***Customer POC***

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make
Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

x.com/RedHat

Red Hat

Appendix

Generative AI is a strategic enabler across industries

Using AI to drive productivity and efficiency

Revenue Generation

- Chatbots
- Campaign and Content Marketing
- Developer assistants
- Guided selling
- Drive product innovation

Cost Optimization

- Automated AI support
- Knowledgebase Search & Summarization
- Doc summarization
- AI-optimized logistics
- Augmented Product R&D

Risk Management

- Sentiment analysis
- Predict employee attrition
- Contract risk assessment
- Fraud detection
- AI-assisted Security Operations

Generative AI customer adoption challenges

Cost

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

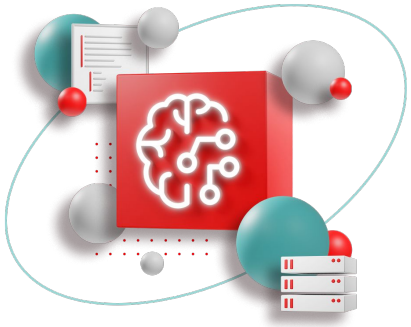
Complexity

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.

Red Hat AI accelerates the development and delivery of AI solutions across hybrid-cloud environments.

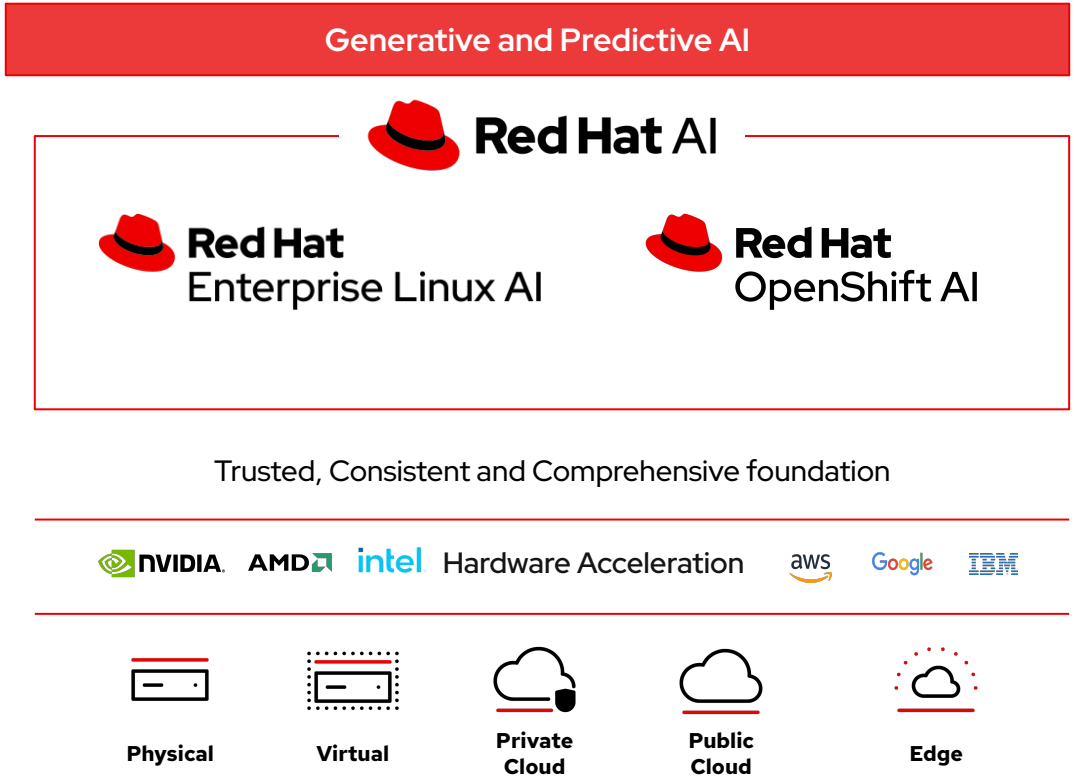


Tune small, purpose-built models efficiently with enterprise-relevant data

Deploy with flexibility and consistency wherever your data resides

Manage and monitor the lifecycle of predictive and generative AI models at scale

Maximize your technology investments with operational guidance & advisory services



* 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

Red Hat AI can help

Portfolio of products to help address the high costs associated with delivering AI solutions

Challenge	Red Hat AI can help...
Cost of building and inference models	Provides access to Granite Models, and bring your own LLM or build your machine learning model
Complexity of building and tuning models	Model development and tuning tools -InstructLab- to support customization with enterprise private or synthetic data
Constraints of deploying models anywhere	Hybrid-cloud flexibility for individual server environments and scaled-out distributed deployments

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

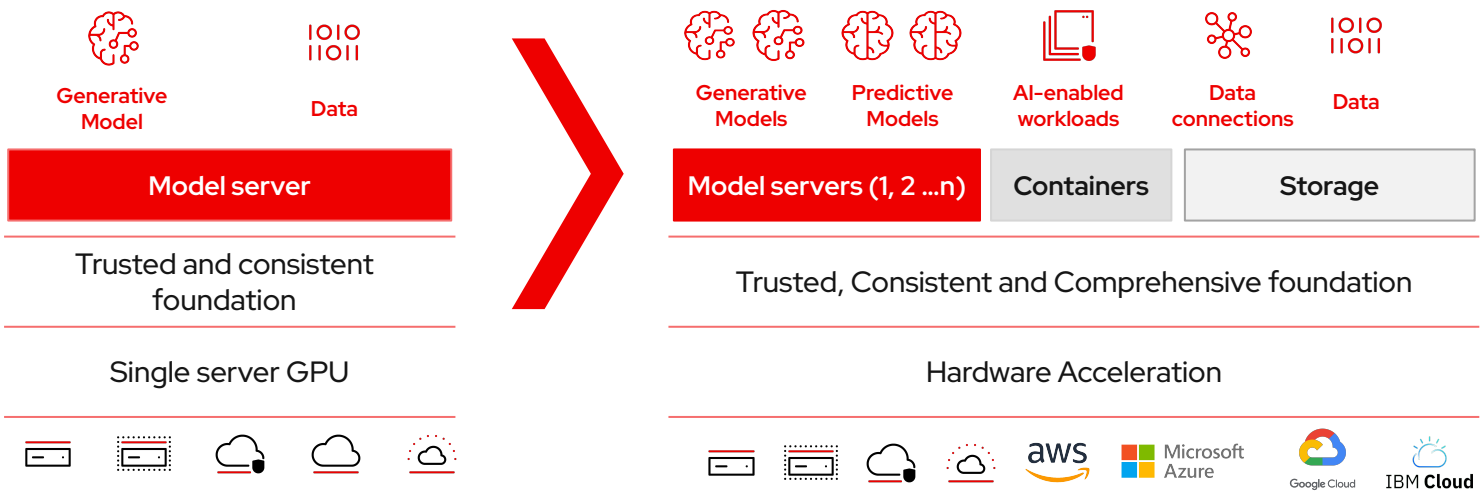
Cloud Environments

- 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

Red Hat AI supports each stage of the AI adoption journey

From single server deployments to highly scaled-out platform architectures



Increasing flexibility and choice with an open source approach

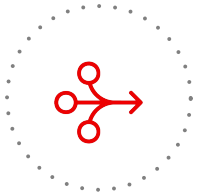
Red Hat prioritizes investments on open source AI and building a certified AI partner ecosystem

Flexibility

Access to cutting-edge open source innovations to keep up with a fast moving market.

Choice

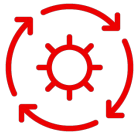
Access to an open ecosystem of communities, technology providers, ISVs and customers.



Abstract the complexity

Reduce the complexity of switching and adapting to new technologies.

The value of Red Hat AI



Increased efficiency

Access to enterprise-grade, **open source Granite models** fully supported and indemnified by Red Hat

Simplified experience

InstructLab model alignment tooling to support developers and domain experts to tailor models more efficiently

Flexibility to deploy anywhere

Mitigate risks and generate value while controlling costs by **owning the decision of where to train and deploy AI models**

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

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.

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 technologies.

Results

- Allows providers to more **easily identify and help specific veterans in need,** using artificial intelligence and machine learning to sift through vast volumes of data.
- Offers an API-first approach that streamlines integration into existing systems, **providing ready access** to medical histories that are key to identify veterans at risk in support of timely interventions.
- Uses a managed cloud service for data scientists and developers to **rapidly develop, train and test machine learning models** in the public cloud before deploying to production

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

Red Hat Summit

"Red Hat's work with AGESIC exemplifies our dedication to improving the user experience for both our and their customers."

Steven Huels
Vice President and General Manager – AI Business Unit, Red Hat

Presentation abstract

AGESIC, Uruguay's Agency for Electronic Government and Information and Knowledge Society, is responsible for e-government strategy and implementation. With Red Hat®, it led Uruguay's AI strategy and provided a more consistent, hybrid AI/ML platform to build and host models while delivering innovative applications.

Presentation summary

- With the proliferation of AI, AGESIC knew that infusing it into its operations would be key to meeting Uruguay's evolving needs.
- AGESIC optimized its AI infrastructure with Red Hat OpenShift®, which brought a containerized approach to workload management and automation of key processes while also bringing development, operations, and systems security functions together on a centralized platform.
- AGESIC evolved its offerings to include Platform as a Service (PaaS), enabling other government agencies to develop, run, and manage applications without the build and maintenance of complex infrastructure.
- AGESIC has begun automating the creation and development process of its AI models and managing model lifecycles, which has enabled standardization of AI usage across all Uruguayan governmental agencies

Products and services

Red Hat OpenShift

Red Hat OpenShift AI

Red Hat

DenizBank transforms AI operations and empowers innovation

"As an invaluable AI-driven solution, Red Hat OpenShift AI provides a streamlined environment that enables our data scientists to build and deploy more robust and secure models."

Okan Çetinkaya
CDO – CAO
DenizBank

Challenge

Intertech - IT subsidiary of DenizBank - wanted to build a comprehensive, standardized, holistic solution for data scientists that would improve time to market while delivering AI/ML process cost efficiencies across multiple business lines, including risk management, marketing and customer relations.

Solution

Red Hat Consulting helped the team design and architect the Red Hat OpenShift AI solution - on premise - providing self-service capabilities and capacity to scale model serving while improving operational efficiency.

Results

- Provided more than 120 data scientists, from different lines of business, with greater autonomy and more consistent standards
- Accelerated time-to-market while ensuring more robust and secure models
- Optimized GPU usage with slicing

Source: Red Hat case study DenizBank empowers its data scientists. Jan, 2025.

Red Hat AI

Next best steps you can take

Learn more and get hands-on experience

TRY RED HAT ENTERPRISE LINUX AI

A single, 60-day, self-supported subscription to Red Hat® Enterprise Linux® AI

TRY RED HAT OPENSHIFT AI

A single, 60-day, self-supported subscription to Red Hat® OpenShift® AI (Self-Managed)

PROOF OF CONCEPT (POC) DESIGNED FOR YOU

Let us bring your vision to life: Request your personalized POC today!

Red Hat saves $5 million with AI-powered innovation in IT support

"The troubleshooting optimization project alone has avoided an estimated $1.5 million in support costs in just 10 months, by reducing the number of users who open a case after clicking on our content."

Mike Clark
Senior Manager
Software Engineering
Red Hat

Challenge

The Experience Engineering (XE) team at Red Hat wanted to increase the efficiency and scalability of customer and technical support services to its growing customer base with AI solutions.

Solution

Utilized the Red Hat® AI® portfolio, comprising Red Hat OpenShift® AI and Red Hat Enterprise Linux® AI, to develop, test, and deploy a range of solutions powered by artificial intelligence (AI), all with the aim of simplifying IT support.

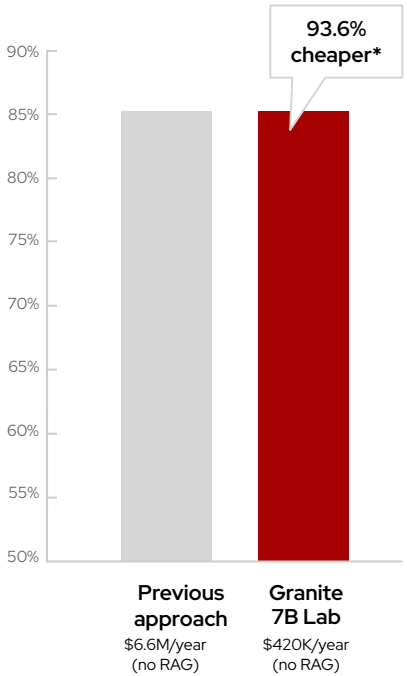
Results

- Delivered over $5m in cost avoidance, with estimated $1.5m in just 10 months
- Increased availability of knowledge content and minimized repetitive tasks
- Provided faster responses to customers with AI, enhancing overall experience

Source: Red Hat case study Red Hat saves $5 million in IT support costs with AI augmentation. Dec, 2024.

RHEL AI InstructLab Telco Customer POC

Enterprise: Large telecommunications company
Analysis of customer call transcripts



USE CASE:
A large global telecommunications company receives over **300K customer calls daily.** Call transcripts need to be quickly analyzed for business & client impact, to roll back into client-facing processes to improve automation and customer satisfaction.

The evaluation of **General Large Language Models** (LLMs) for this challenge found them to be:

- Expensive to operate
- Limited in domain knowledge critical to answering client issues and reducing call volumes

Through POC testing of **InstructLab to align their own custom LLMs** with their private corporate data, this company achieved:

- An **88% quality improvement** over existing solutions
- An **80% lower cost** over a generic commercial LLM
- Ability to process the full call volume **daily** for up-to-the-minute results
- Secure and auditable **data privacy built-in**, with efficient GPU utilization as needed and matched to the workloads

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

Red Hat OpenShift AI use cases:

- A joint venture with Harvard University Medical School. Red Hat OpenShift AI is helping them process their large quantity of structured data using an NVIDIA GPU.
- Using LLM to pinpoint patients for preventive medication or closer inspections. We need to look at doctors' notes to find these people; for that, we use LLMs such as LLaMa.
- Another use case allows us to develop image-based machine learning processes.

Overview

The patient lies at the center of everything Clalit Health Services (Clalit) does. Its 14 hospitals include 8 general hospitals, 2 mental health hospitals, 2 geriatric hospitals, and a children's hospital. It also operates community clinics, dental clinics, imaging facilities, and a lifestyle program. Clalit recently established an advanced AI platform based on Red Hat OpenShift AI and Red Hat® OpenShift®. We recently interviewed Eyal Dviri, Innovation Team Leader in the Data Department at Clalit:

What led you to Red Hat OpenShift AI?

"Our central IT department implemented Red Hat OpenShift to provide a modern solution for a wide range of use cases across the caregiving part of our organization. When they made it available to us, we adopted it immediately for our research needs. We didn't think twice. To address our ML/AI use cases, because we had OpenShift already, we decided to go with Red Hat OpenShift AI."

What's next?

"We have two challenges ahead. The first thing is the increase in demand when people hear about our OpenShift AI platform. We need to figure out how to use our computing resources—our GPUs—optimally. The second thing is the pipeline. We need to understand best practices for deploying AI algorithms from training to production."

Source: Red Hat Q&A: Clalit accelerates research and innovation with AI and GPUs. Interview with Eyal Dviri, Innovation Team Leader in the Data Department at Clalit. May 2024.

Large global airline

Challenge

Needed a unified platform for model development and inference

Goals

- **Operationalize AI** - The organization aims at reducing the time it takes to provision development environments, build a library of machine learning models, and deploy models to production environments.
- **Predictive AI models** - Produce multiples models targeting different use cases across the organization: asset management, revenue management, marketing, customer management, operations and support functions. A few examples on operations are: Crew planning, automated customer onboarding, fuel optimization, inflight catering prediction, cargo no-show and baggage handling optimization.

Results

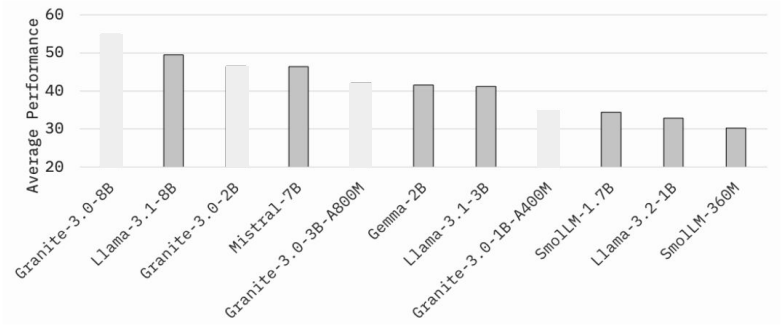
Provide a sandbox environment that accelerates innovation allowing business and IT teams to develop a variety of predictive models for optimizing business operations across the organization.

- **Enable ITOps to optimize and manage the compute resources** during model training and inference.
- **Adopt an MLOps approach** for managing the lifecycle of the model and the application in a singular platform.

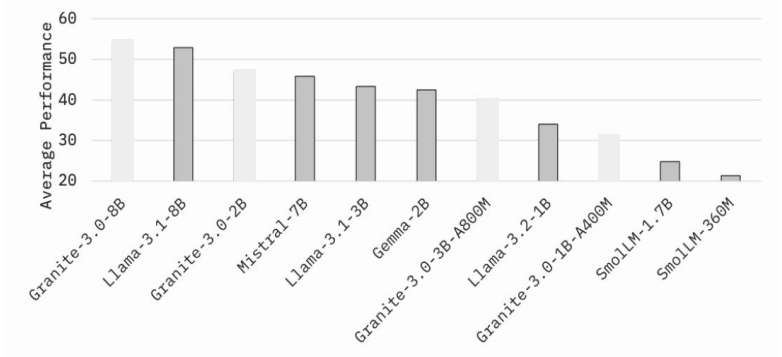
Red Hat

IBM Granite 3.0

01 **Base Models:** Average performance across 19 tasks / 6 domains[1]



02 **Instruct Models:** Average performance across 23 tasks / 8 domains[1]



- State-of-the-art training[1] and open source data recipes.[2]
- 12T+ tokens training data in Granite 8B + 2B.
- Designed for enterprise tasks:
  - **Language** (RAG, summarization, entity extraction, classification, etc.)
  - **Code** (generation, translation, bug fixing)
  - **Agents** (tool use, advanced reasoning)
  - **Multilingual support** (en, de, es, fr, ja, pt, ar, cs, it, ko, nl, zh)
- Additional models including MoE, Guardian, and more.
- Trained on the Blue Vela cluster, which runs on 100% renewable energy to minimize the environmental impact.

Sources:
1. "Granite 3.0 Models," Granite Team, IBM. https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf
2. Open source data recipes available in the IBM Data Prep Kit: https://github.com/ibm/data-prep-kit

Red Hat OpenShift AI

Learn more ▸ redhat.com

Contact us ▸

Red Hat