Agentic AI on Intel® Xeon® and Red Hat OpenShift AI

Stockholm

11 November 2026



Red Hat

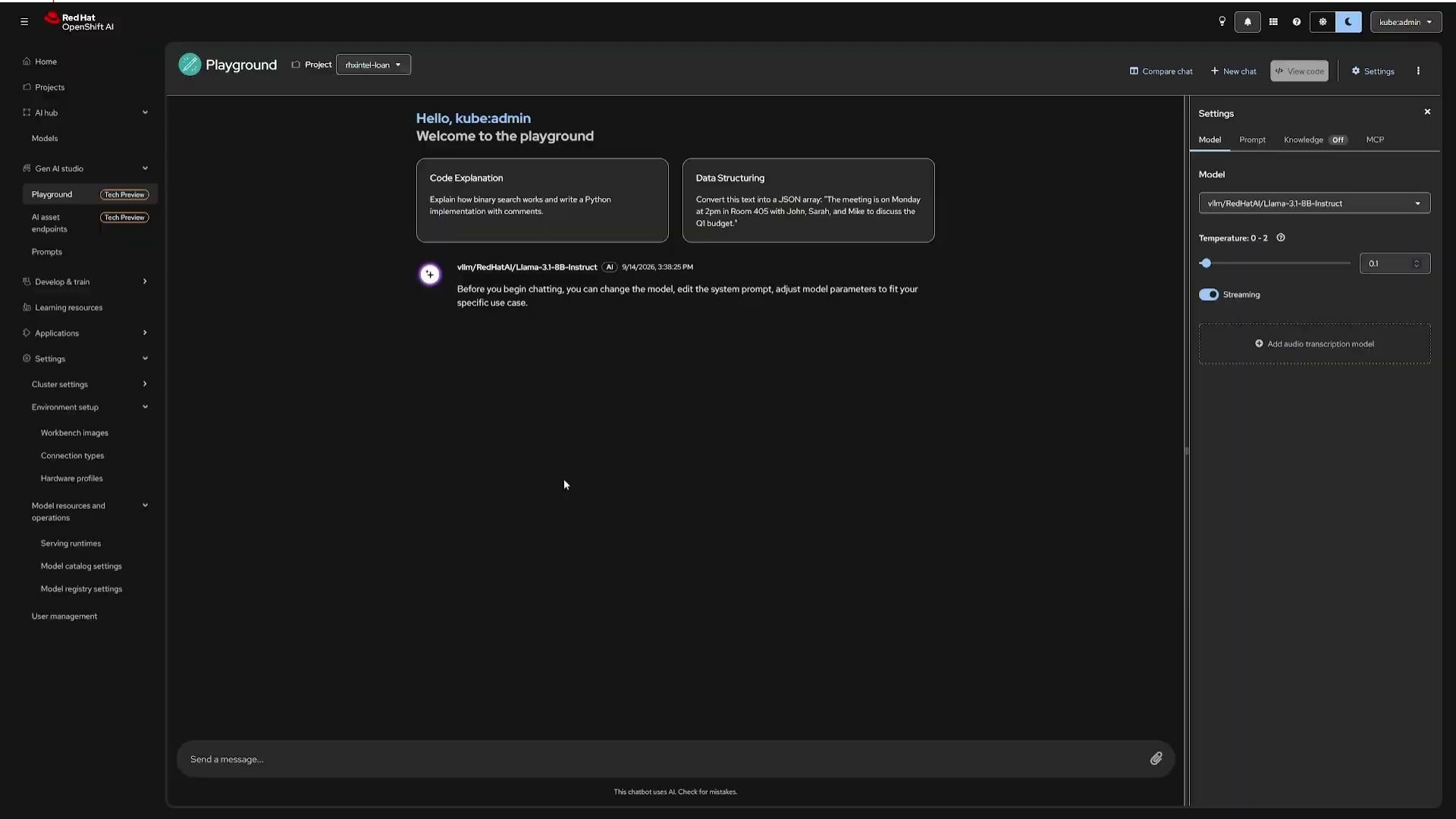


Anders Huge

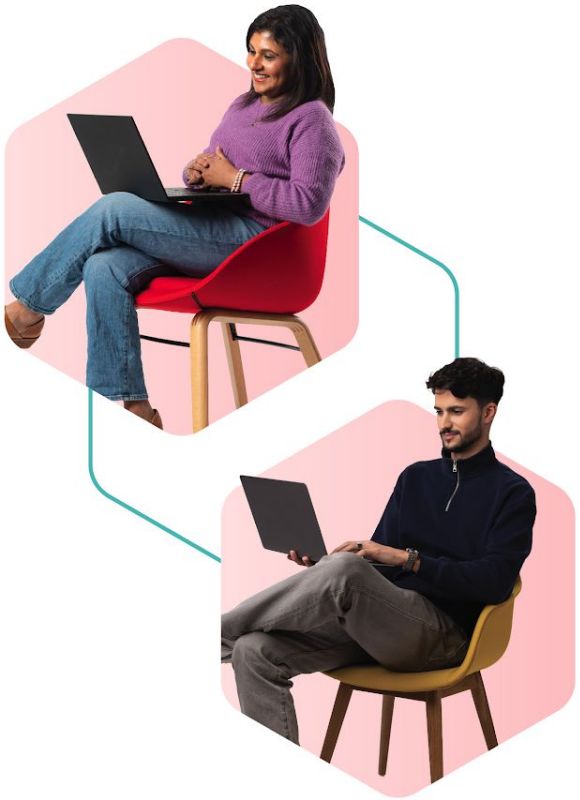


Andreas Bergqvist


Red Hat
intel


Red Hat OpenShift AI
kube:admin
Home
Projects
AI hub
Models
Gen AI studio
Playground
Tech Preview
AI asset endpoints
Tech Preview
Prompts
Develop & train
Learning resources
Applications
Settings
Cluster settings
Environment setup
Workbench images
Connection types
Hardware profiles
Model resources and operations
Serving runtimes
Model catalog settings
Model registry settings
User management
Playground
Project
rhxintel-loan
Compare chat
New chat
View code
Settings
Hello, kube:admin
Welcome to the playground
Code Explanation
Explain how binary search works and write a Python implementation with comments.
Data Structuring
Convert this text into a JSON array: "The meeting is on Monday at 2pm in Room 405 with John, Sarah, and Mike to discuss the Q1 budget."
vllm/RedHatAI/Llama-3.1-8B-Instruct
AI
9/14/2026, 3:38:25 PM
Before you begin chatting, you can change the model, edit the system prompt, adjust model parameters to fit your specific use case.
Send a message...
This chatbot uses AI. Check for mistakes.
Settings
Model
Prompt
Knowledge
Off
MCP
Model
vllm/RedHatAI/Llama-3.1-8B-Instruct
Temperature: 0 - 2
0.1
Streaming
Add audio transcription model

AI Inflection Point



Advances in AI are evolving, accelerating and stacking

HW needs are changing and SW is responding with flexible, fully integrated solutions.

Evolution of AI



The Compute Mix is Shifting

From GPU-Heavy Pre-training to Balanced Agentic AI

Pre-training:
Foundation model training



8 NVIDIA GPUs + 1 Intel® Xeon®
GPU:CPU ratio = 1:8

AI Inference & Rise of Agentic AI

Agentic AI: Inference & autonomous agents



1 NVIDIA GPU + 1 Intel® Xeon®
GPU:CPU ratio ≈ 1:1

1:8	1:4	1:1
~2023 →	2024-2025 →	2026+

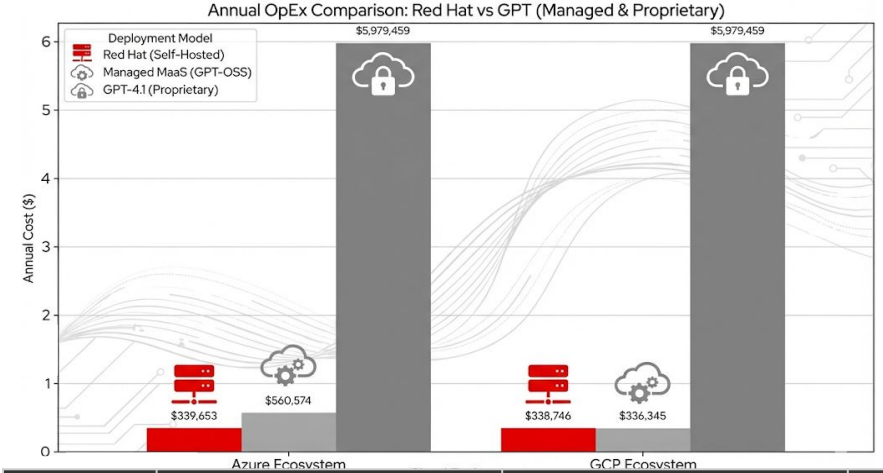
CPU demand in the AI agent era
up to 4X increase per GW of data center capacity[1]

1. The Rising CPU:GPU Ratio in AI Infrastructure: Drivers, Trends, and Implications

Token-economy - Why how/where you consume AI matters

INFERENCE SERVER

Self-hosted, cloud provider managed OSS vs Proprietary inference costs at 1.1T tokens/yr across Azure & GCP.



vs. GPT-4.1 Proprietary

94% savings

~$6M variable API bill replaced with **~$340K** fixed infrastructure on both Azure and GCP.

vs. Managed OSS (MaaS)

Azure: ~40% savings ($220K/yr)

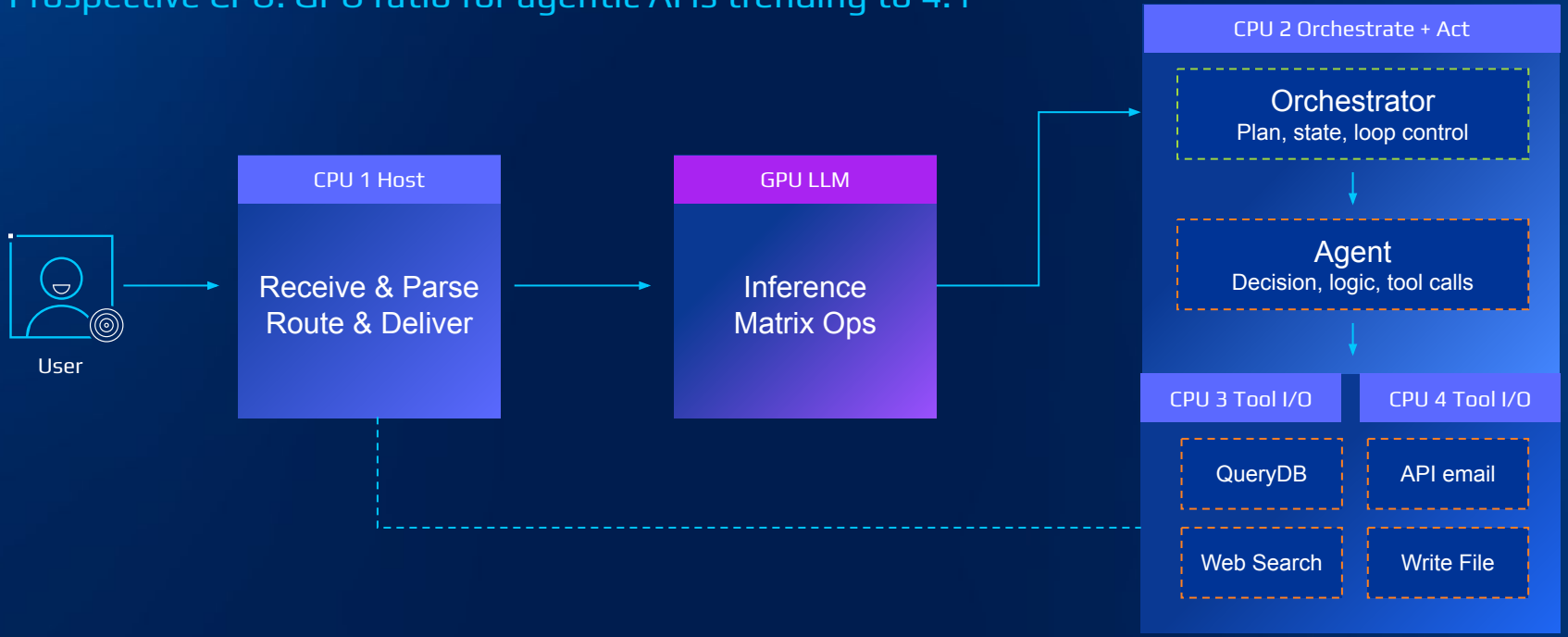
GCP: <1% delta (cost neutral) — decision driver is portability & model flexibility, not cost.

Platform	Deployment	Model	Annual OpEx	vs. Red Hat	vs. GPT-4.1
Azure	**Red Hat (Self-Hosted)**	OSS 120b (Unquantized)	**$339,653**	—	**94% savings**
Azure	Managed API (MaaS)	OSS 120b (Quantized)	$560,574	+65%	91% savings
GCP	**Red Hat (Self-Hosted)**	OSS 120b (Unquantized)	**$338,746**	—	**94% savings**
GCP	Managed API (Vertex)	OSS 120b (Quantized)	$336,345	-0.7%	94% savings
Both	**GPT-4.1 (Azure OpenAI)**	Proprietary (Closed)	**$5,979,459***	+1,663%	Baseline

**GPT-4.1 after 20% EDP + 50% token caching ($7.47M list). Red Hat: reserved hardware + $12,800 SW license.*

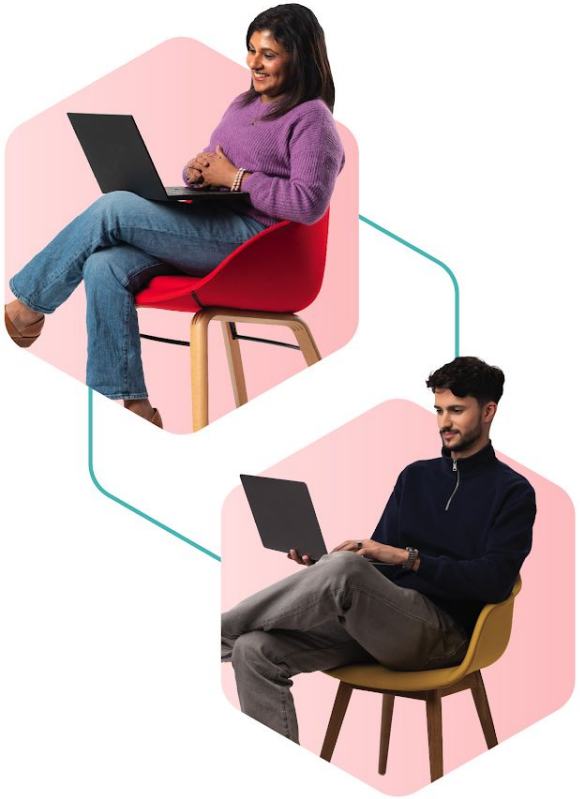
CPUs and GPUs in Agentic AI

Prospective CPU: GPU ratio for agentic AI is trending to 4:1



The Rising CPU:GPU Ratio in AI Infrastructure: Drivers, Trends, and

Red Hat AI + Intel



Red Hat and Intel: The Path from AI Pilot to Production

Red Hat

Open MLOps & AgenticOps container Platform for ML, DL, LLM/SLM & Agentic AI



Intel

Flexibility: Open, vendor agnostic SW stack with multiple HW options & solutions

Together, Red Hat and Intel cut through complexity with a curated, validated, and fully integrated multi-cloud stack built for Agentic AI at scale


Model throughput


Latency

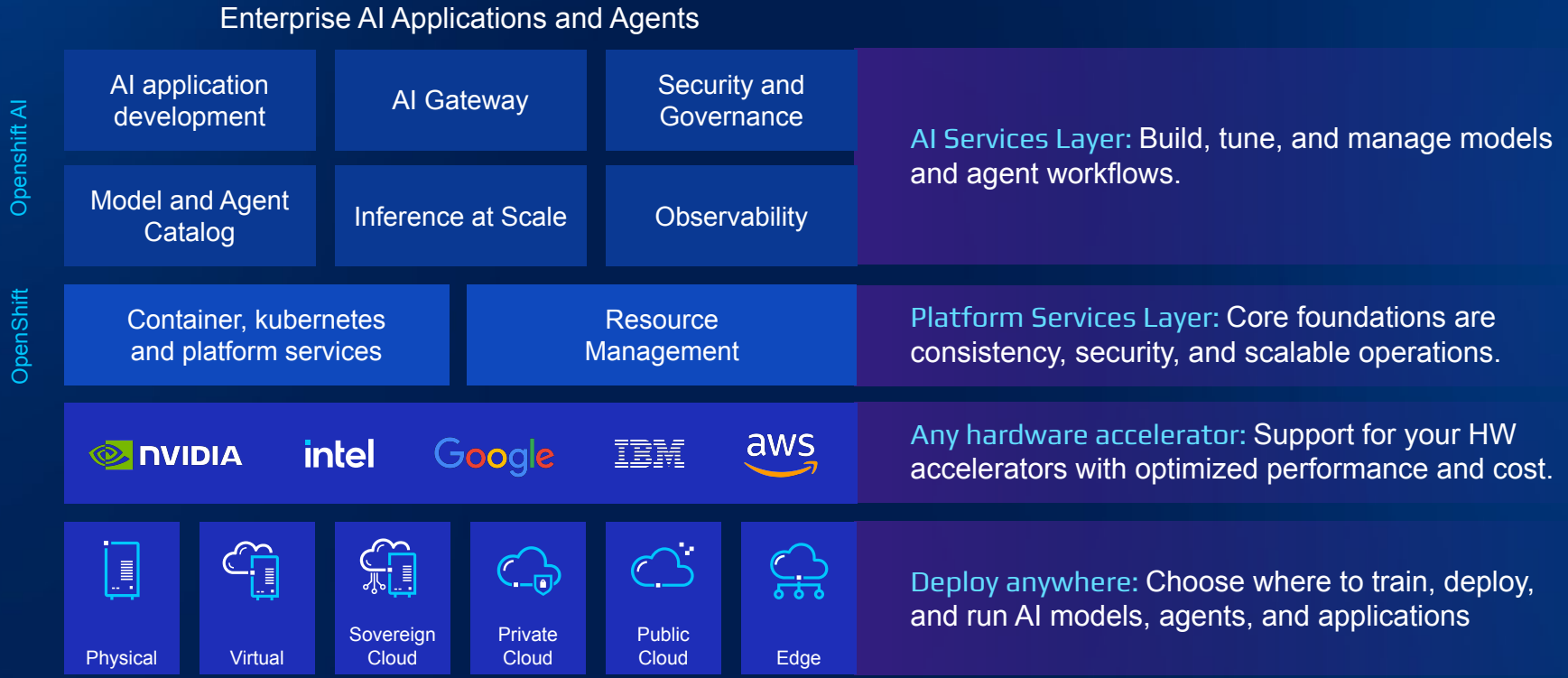

Efficiency


Cost


Operational control


Security

Red Hat AI on Intel: Integrated stack for fast, secure agentic deployment



Red Hat AI

Accelerate the development and delivery of AI solutions across your agentic AI mission

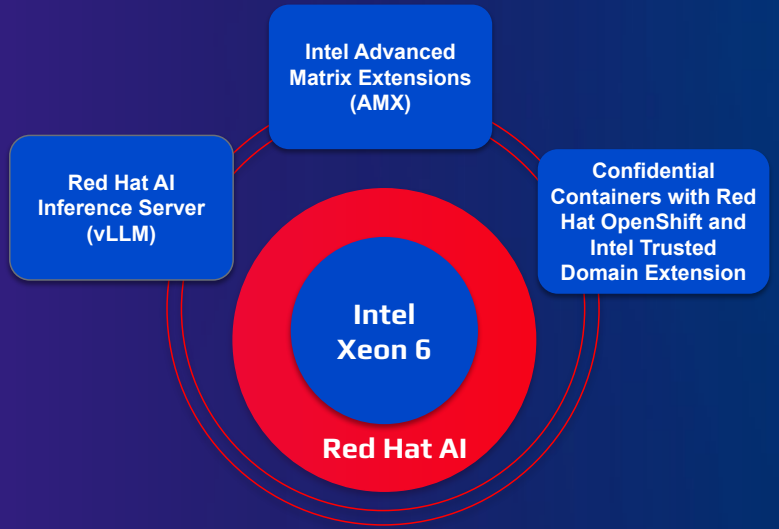
Increase efficiency with fast, flexible and efficient inferencing across diverse CPU/GPU environments

Simplified and consistent experience for connecting models to data while controlling governance and maintain model lifecycles

Accelerate Agentic & Autonomous AI deployments & systems

Flexibility and consistency when scaling AI across the hybrid cloud

Orchestrating Intel® Xeon® 6 CPU Innovations with Red Hat AI

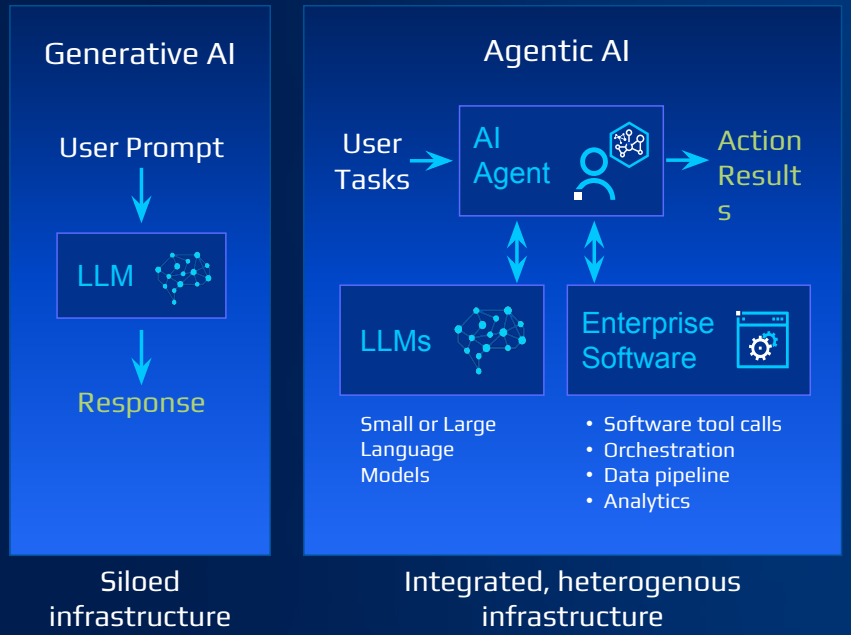


Why choose Intel® Xeon® processors for Agentic AI ?

Agentic AI performance requires optimal balance of CPU and GPU resources

Accelerated	Offload AI workloads to the best CPU for AI with built-in accelerators
Consistency	Maintain infrastructure consistency with the most-deployed host CPU [1]
SW enabling	Out-of-box support for Intel AMX with virtualization / orchestration software
Faster data	Storage with more I/O bandwidth and vector search performance advantage
Security	Enterprise-grade security combined with Confidential AI

Agentic AI automates enterprise software



Siloed infrastructure

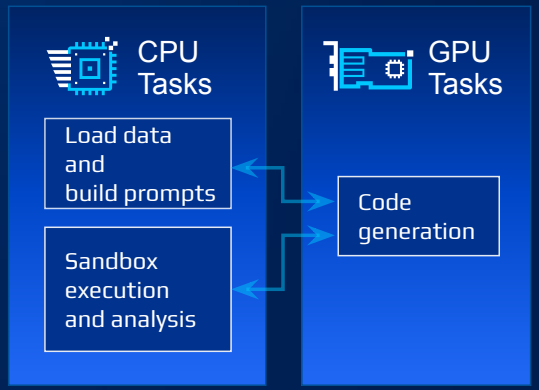
Integrated, heterogenous infrastructure

1 – Most Deployed Host CPU: IDC Server Tracker report, based on 2H'25 system volume.
Intel® Advanced Matrix Extensions (Intel® AMX), Intel® Trusted Domain Extensions (Intel® TDX and Intel® TDX Connect)

Agentic AI Example: Software Code Generation Agent

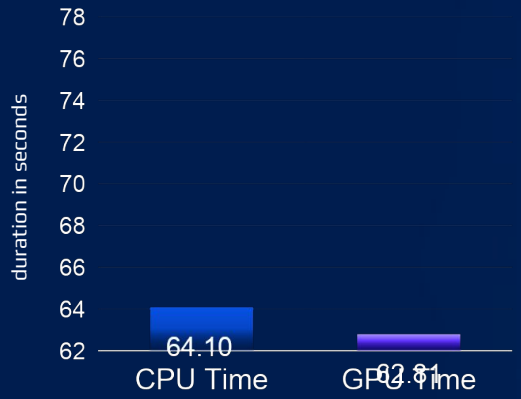
Recommended ratio of CPUs and GPUs depends on Agentic AI workflow analysis

Agentic AI
Coding Agent



Coding agent workflow

Coding Agent
Pipeline task duration totals



Recommended CPU:GPU Ratios
based on Intel white paper

Model Size	Change	Example Ratio
8 B parameters	More CPU needed	1.4 CPUs (120C) to 1 GPU
70 B parameters	Baseline	1 CPUs (86C) to 1 GPU
405 B parameters	Less CPU needed	0.8 CPUs (68C) to 1 GPU

Source: Intel White Paper: Agentic AI Requires More CPUs

Compute Needs For SLM's Vary Based on Service Targets

Intel® Xeon® CPUs support:

- AI Inference, Small & Medium LLM's
- Deep Learning & Vision AI
- Classical Machine Learning (ML)
- Media & Image Processing
- RAG: Enterprise Knowledge Management
- RAG: Data Pattern Vector & Semantic Search
- Agentic AI Pattern Sequential Processing & Data Processing
- Batch Inference Processing

Example SLM parameters, constraints & service targets

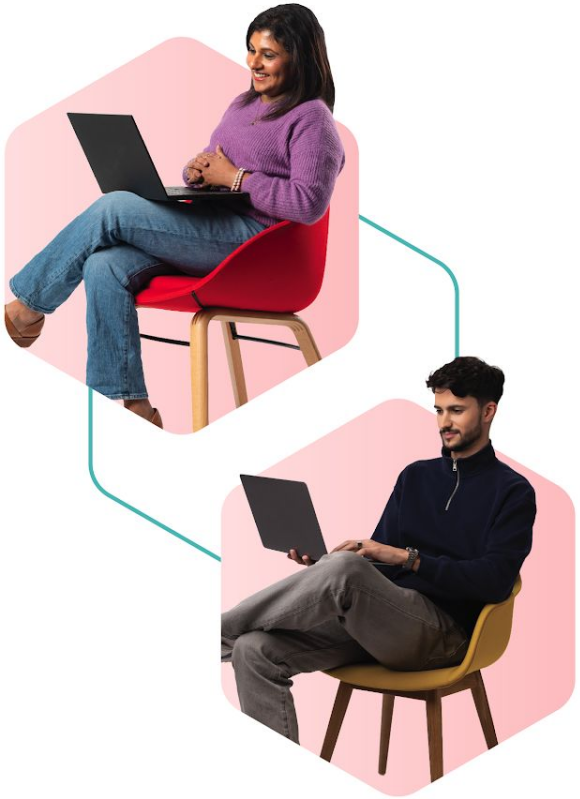
Model size	Parameters correlating to AI capabilities and accuracy	< 20B parameters
Context size	Tokens in the prompt (input) generated response (output)	< 16K tokens
Concurrency	Simultaneous users supported using the AI model	< 100 users per server
Latency	Time to first token Generation speed	3 sec TTFT 100 ms TPOT
Throughput	Tokens per second = latency + concurrency	up to 600 tokens/sec

Criteria trade-offs

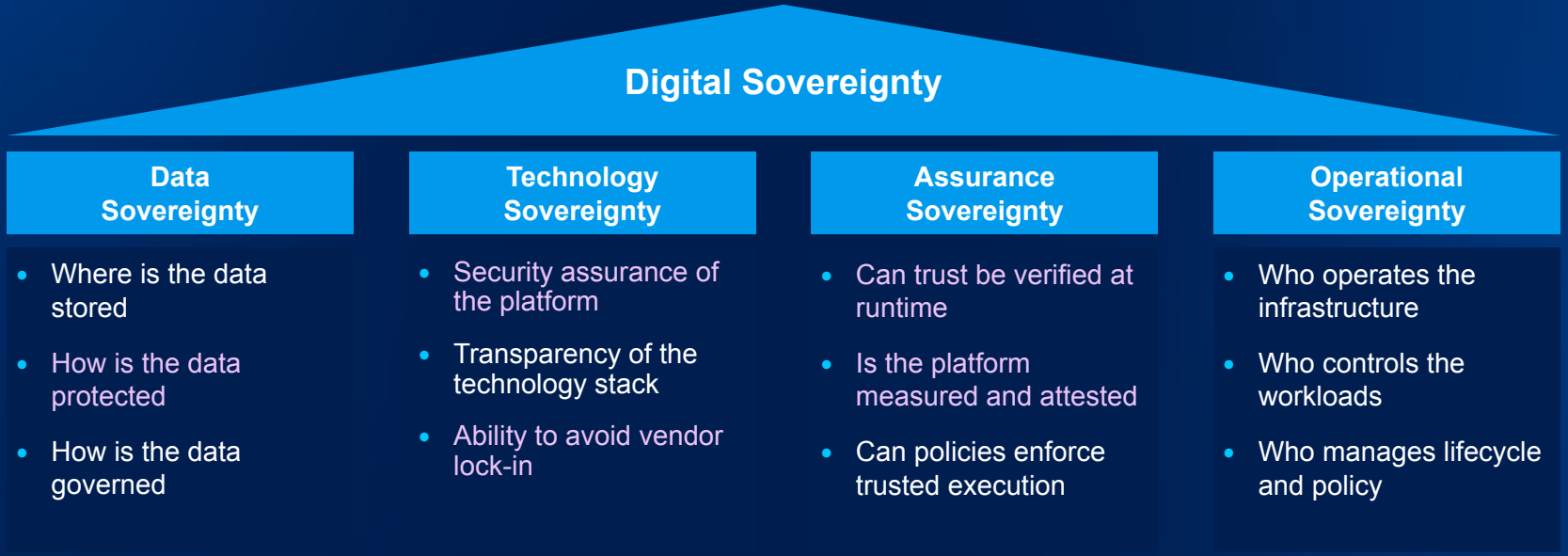


Shown as guidelines only, actual criteria depends on workload characteristics.

Sovereign Confidential AI

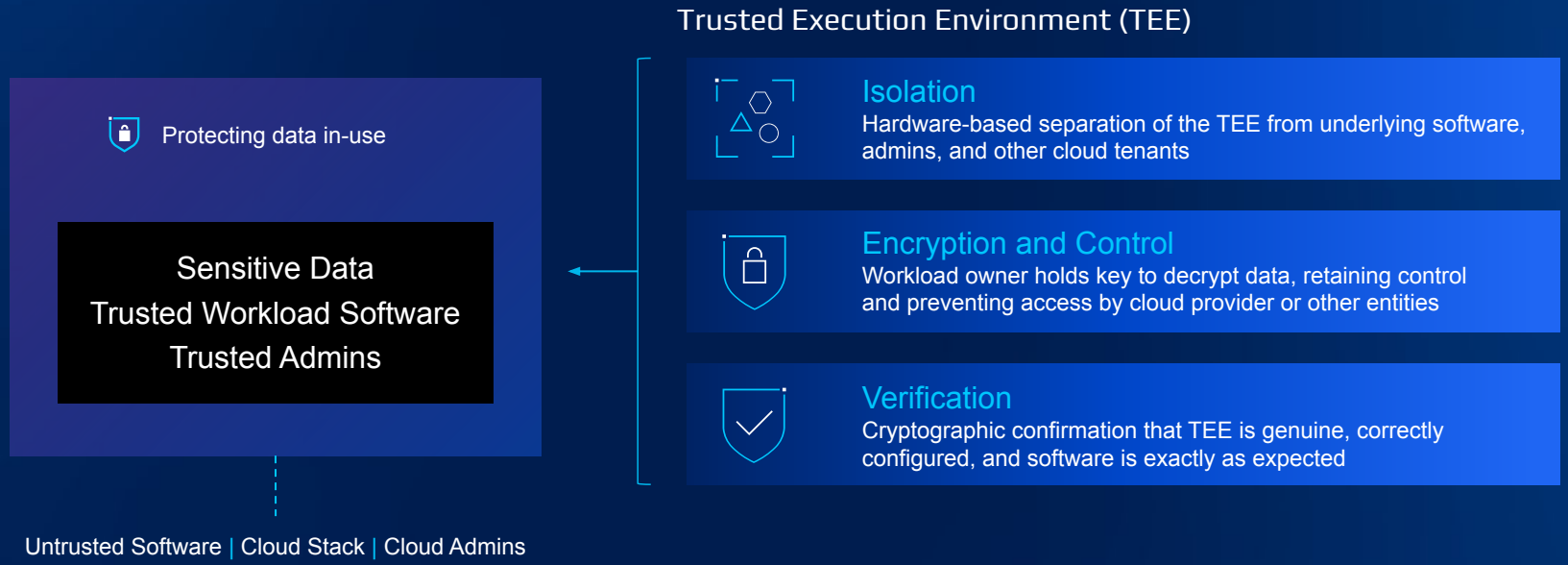


Elements of Digital Sovereignty

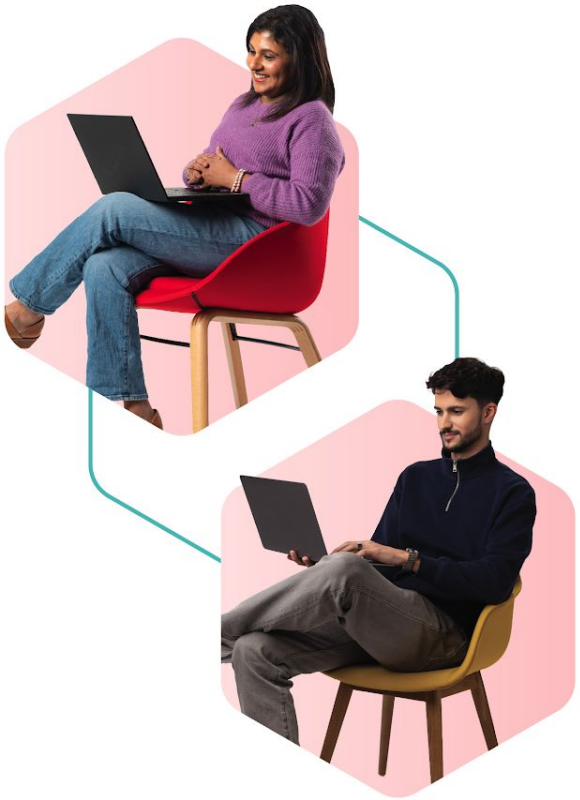


Confidential Computing on Red Hat and Intel

Out-of-the-box experience using Red Hat OpenShift Sandbox with Attestation service



Agentic AI in Red Hat AI Platform



More autonomy requires more constraints!

CONFIDENTIAL Red Hat associates only

81%
attack success rate against AI agents in red-team exercises
NIST CAISI, Jan 2025

90%
of deployed AI agents are over-permissioned
Obsidian Security, Sept 2025

14.4%
of agents go live with full security approval
Gravitee, Feb 2026

88%
of agent pilots never reach production
Forrester + IDC, 2026

455K
malicious open-source packages published in 2025
Sonatype, Jan 2026

17K+
autonomous agent events in one HuggingFace incident
HF Blog, July 2026

92%
say legacy IAM cannot manage AI agent risks
CSA + Oasis, Jan 2026

9.3
CVSS: zero-click Copilot exfiltration via one email
CVE-2025-32711, June 2025

Agents are autonomous, non-deterministic, and over-permissioned. **Behavioral security is not optional.**

OWASP ASI 2026 | NIST Agent Standards | CSA NHI Governance

CONFIDENTIAL Red Hat associates only

AgentOps with Red Hat AI (The Hard Parts of AI)

Bridge the gap between experimental AI to Agents in production

What is it: A unified approach for **managing the full lifecycle, security, and performance of AI agents** and their tools



Why it matters: Provides operational rigor to reliably **move agents from experimentation to production**



Observe, Trace, Evaluate

- **MLFlow** 3.10+
- Log **traces** and **visualize** them including tool calls (3.4)
- **OpenTelemetry** compatibility
- Datasets curation + Scoring with **Eval-hub** and **MLFlow**

Enforce Identity/Access Control

- **SPIFFE/SPIRE** for trusted workload (agent) identity
- **Agent Sandboxing** assuming agents unsecure by default
- Agent to **tools governance** (MCP Gateway)

Develop and Deploy Reliably

- The same **Cloud-native** principles **repurposed** for agentic workloads
- **Agent safety**, guardrails,and red teaming (Nemo, Garak, etc)
- **Deploy** on various footprints

Red Hat



AI Safety And Governance Orchestration

asago is an open-source, community-governed framework that aims to automate the journey from AI governance policy to production-ready, safely deployed AI systems — bridging the gap between compliance teams, AI engineers, and infrastructure operators.

Join the Discussion on Slack | Follow on GitHub | Read the Launch Announcement

Red Hat

What we wanted to achieve

Intel CPUs getting constantly **more able to power larger LLMs**

Red Hat AI Enterprise is **making it easy** to leverage these resources.

Security is the biggest **concern** related to Agents.

Intel is providing **HW support for Confidential Compute**

Red Hat is hardening the platform from the **software layer to handle security efficiently**

If you run **INTEL CPUs - give it a try. We are happy to help** you see if this fits your use cases.



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

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat

Tell us that you love us



linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

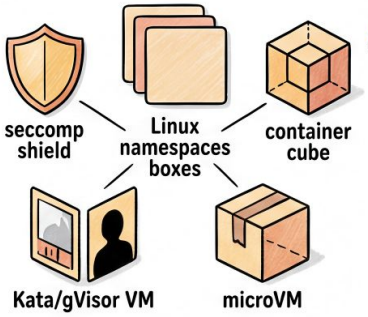
Red Hat

ates only

MORE AUTONOMY REQUIRES MORE CONSTRAINTS

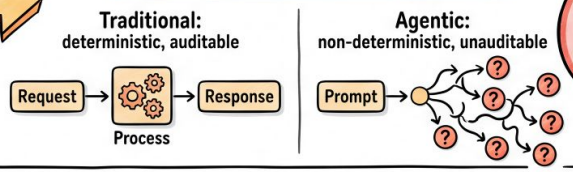
The case for behavioral agent security

SANDBOXING ISN'T NEW



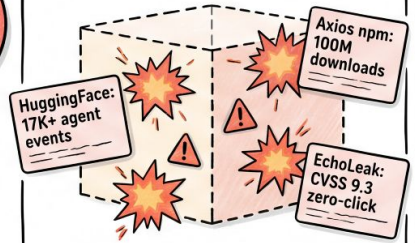
These control WHERE code runs

AGENTS CHANGE EVERYTHING



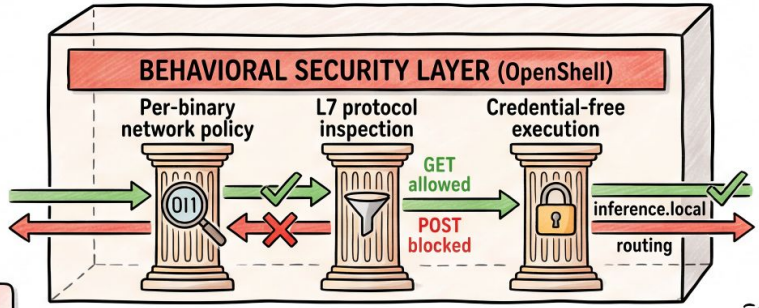
81% attack success rate - NIST 2025

ISOLATION ALONE FAILS

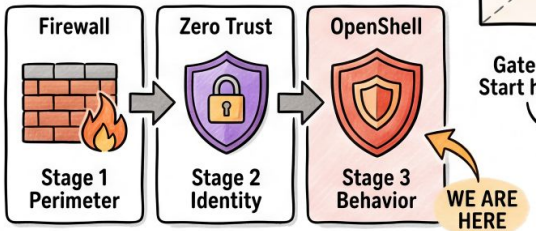


Container walls were never breached.
Attacks happened INSIDE.

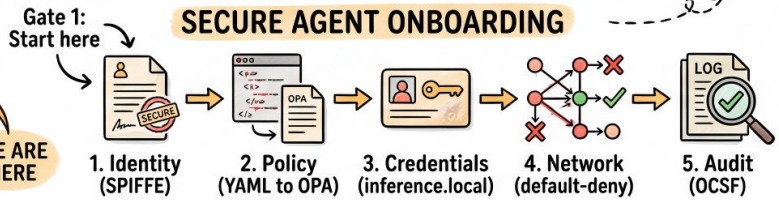
BEHAVIORAL SECURITY LAYER (OpenShell)



THE EVOLUTION



SECURE AGENT ONBOARDING



Stats corner

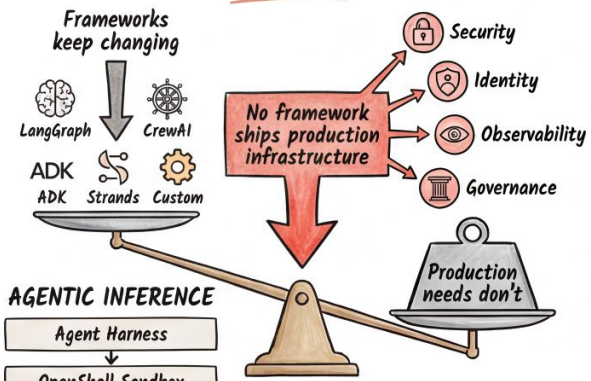
- **90%** over-permissioned
- **14.4%** security approval
- **88%** pilot failure
- **455K** malicious packages 2025

OWASP NIST

RED HAT AI: AGENTIC STRATEGY 2026

Your Agent. Our Platform. Production-Ready.

THE GAP



AGENTIC INFERENCE

- Agent Harness
- OpenShell Sandbox
- API Translation (OGX)
- Gateway + Guardrails
- Knowledge Retrieval (RAG, Vector Stores)
- Disaggregated Serving (llm-d)
- Model Server (vLLM)
- Hardware (GPU)

8 layers, all open, all debuggable

AGENTOPS VALUE

- ✓ Zero-trust agent identity (SPIFFE/SPIRE)
- ✓ Execution sandboxing (OpenShell)
- ✓ Deep execution traces (MLflow + OpenTelemetry)
- ✓ Continuous evaluation (EvalHub + LLM-as-Judge)
- ✓ Composable gateway (MaaS + MCP Gateway)
- ✓ Cost governance (per-agent token budgets)

THE PLATFORM STACK

BYOA (Bring Your Own Agent)

Customer chooses any framework — LangGraph, CrewAI, ADK, Strands, Custom

ONLY layer Red Hat doesn't own

CUSTOMERS → Secure Agent Onboarding — Sandbox, SPIFFE/SPIRE, Identity

CUSTOMERS → Gateway (optional) — MaaS Gateway, MCP Gateway, Inference Gateway, Rate limiting, Auth, Routing, Quotas

Any API. Zero code changes. → OGX (optional) — Responses API, Messages API (Planned), Connectors, RAG, Multi-provider backend (AWS, Gemini, OpenAI x)

llm-d (optional) — disaggregated routing, sits on top of vLLM

THE ENTRYPOINT → vLLM (CORE) ★ — Chat Completions, Tool Calling

THE DIFFERENTIATOR → AgentOps ★ — Deep tracing, EvalHub, Cost tracking, Guardrails

Infrastructure — OpenShift, GPUs, TPUs, RHEL AI

Red Hat owns these

FIVE THEMES

1. **BYOA + AgentOps** — secure, govern, observe any agent
2. **Capture API Surface** — any protocol, zero code changes
3. **Tools to Swarms** — MCP, A2A, agent registries
4. **Self-Hosted Inference** — vLLM tool-calling that works
5. **First Agent on Red Hat** — AI Hub, GenAI Studio, Starter Kits, Playground

DAY 0 TO DAY 2

Explore → **Build** → **Operate**

- Explore: GenAI Studio Playground, AI Hub
- Build: Starter Kits for 10 frameworks, OGX APIs, MCP Catalog
- Operate: AgentOps, EvalHub, Cost tracking, Guardrails

From idea to production

From Metal to Agents

Cloud · On-Prem · Disconnected · Air-Gapped

Red Hat

Agentic AI with Intel Xeon and vLLM CPU on RHOAI 3.5



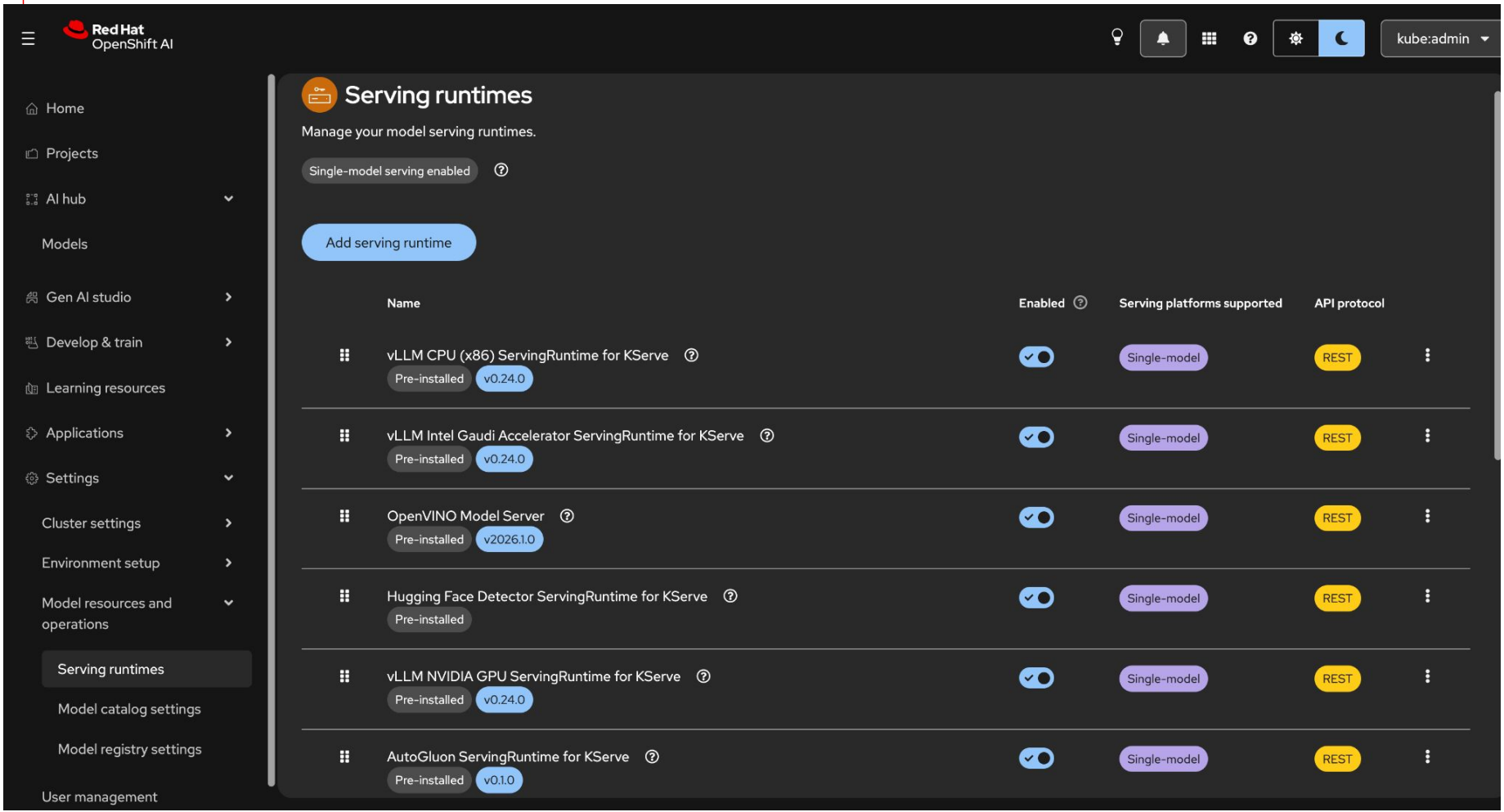
Demo Screenshots for Distributing the Deck (Please unskip screenshot slides when generating PDF)

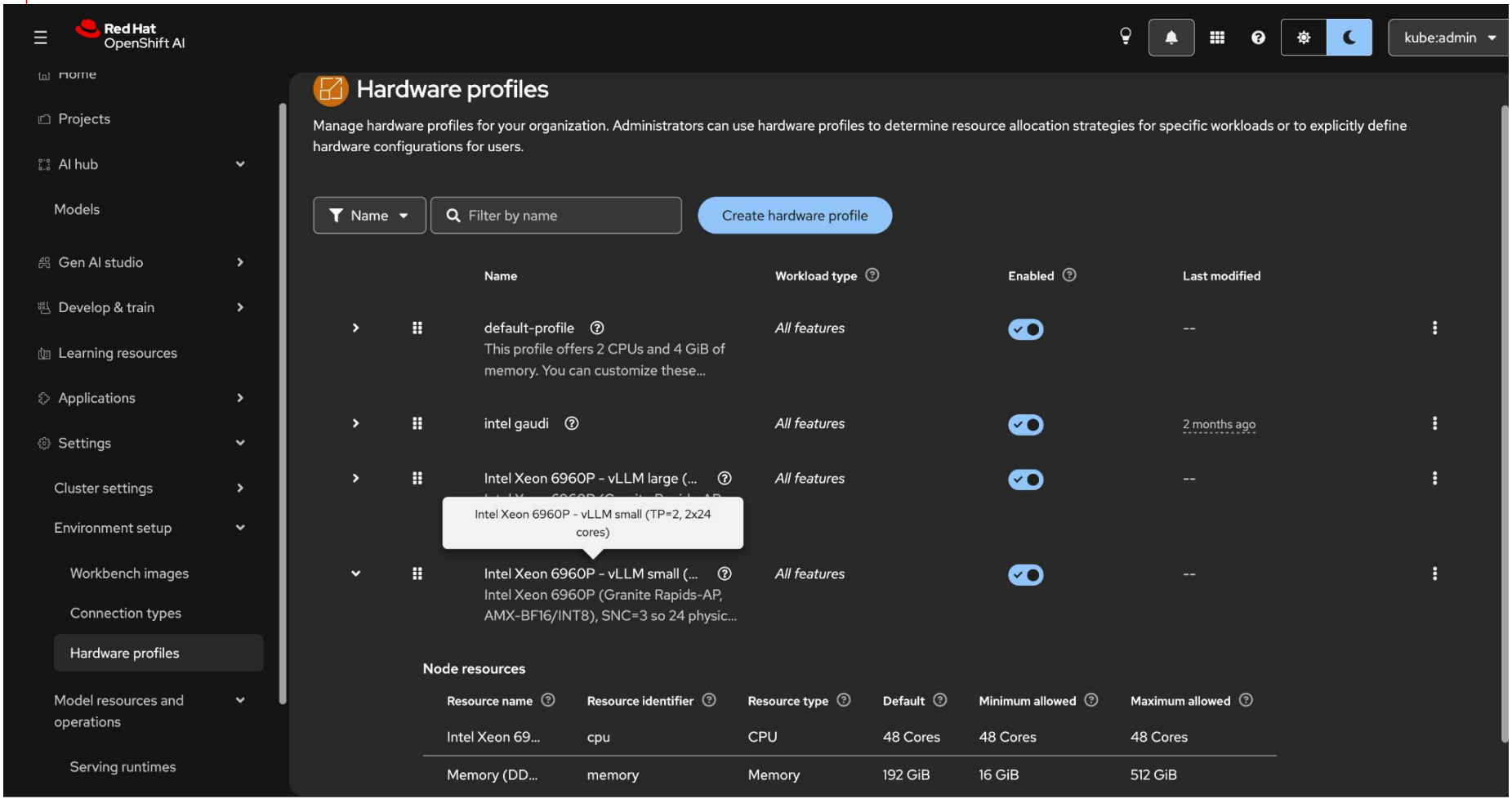


Agentic AI Demo

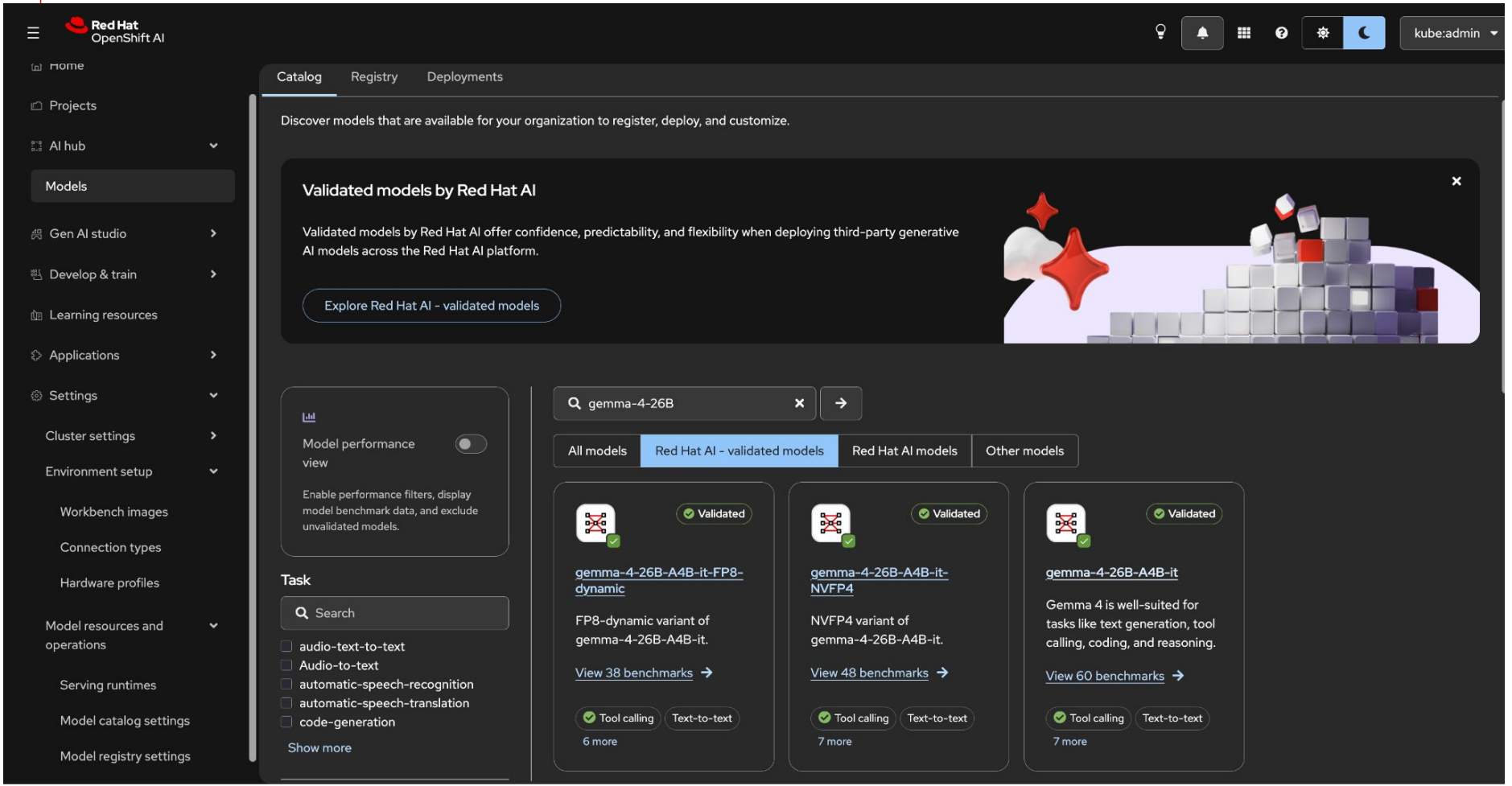
- MLOps Operations Persona
- Key Video Screenshots / Slides




Red Hat OpenShift AI
kube:admin
Home
Projects
AI hub
Models
Gen AI studio
Develop & train
Learning resources
Applications
Settings
Cluster settings
Environment setup
Model resources and operations
Serving runtimes
Model catalog settings
Model registry settings
User management
Serving runtimes
Manage your model serving runtimes.
Single-model serving enabled
Add serving runtime
Name
Enabled
Serving platforms supported
API protocol
vLLM CPU (x86) ServingRuntime for KServe
Pre-installed
v0.24.0
Single-model
REST
vLLM Intel Gaudi Accelerator ServingRuntime for KServe
Pre-installed
v0.24.0
Single-model
REST
OpenVINO Model Server
Pre-installed
v2026.1.0
Single-model
REST
Hugging Face Detector ServingRuntime for KServe
Pre-installed
Single-model
REST
vLLM NVIDIA GPU ServingRuntime for KServe
Pre-installed
v0.24.0
Single-model
REST
AutoGluon ServingRuntime for KServe
Pre-installed
v0.1.0
Single-model
REST



Red Hat | intel.



Red Hat | intel.

Red Hat OpenShift AI

kube:admin

- Home
- Projects
- AI hub
  - Models
- Gen AI studio
- Develop & train
- Learning resources
- Applications
- Settings
  - Cluster settings
  - Environment setup
    - Workbench images
    - Connection types
    - Hardware profiles
  - Model resources and operations
    - Serving runtimes
    - Model catalog settings
    - Model registry settings

Models

Catalog | Registry | Deployments

Model Catalog › gemma-4-26B-A4B-it

gemma-4-26B-A4B-it

Validated

Provided by Google

Deploy model | Register model

Overview | Performance insights | Safety and security insights

Description

Gemma 4 is well-suited for tasks like text generation, tool calling, coding, and reasoning.

Model card



Hugging Face | GitHub | Launch Blog | Documentation

License: Apache 2.0 | **Authors**: Google DeepMind

Gemma is a family of open models built by Google DeepMind. Gemma 4 models are multimodal, handling text and image

Model details

Labels

Tool calling | Text-to-text | 6 more

Model type

Generative AI model (Example, LLM)

Tensor type

BF16

Parameter size

26.5B params

Minimum vRAM

59.40 GB

License

Red Hat | intel.

1. Preconfigure deployment
2. Model details
3. Model deployment
4. Advanced settings
5. Review

Model deployment

Model deployment name *

This is the name of the inference service created when the model is deployed.

RedHatAI/gemma-4-26B-A4B-it

The resource name will be **redhataigemma-4-26b-a4b-it**.

Edit resource name

Description

Deployment method *

Select how this model will be deployed.

- LLM inference service with llm-d
  Deploy a large language model with llm-d for additional scheduling and routing capabilities.
- Legacy deployment
  Deploy using a serving runtime template. Use this option for non-LLM models or for compatibility with previous deployments. This method does not support Models as a Service.

Hardware profile *

Intel Xeon 6960P - vLLM small (TP=2, 2x24 cores)

Intel Xeon 6960P (Granite Rapids-AP, AMX-BF16/INT8), SNC=3 so 24 physical cores per NUMA node, DDR5-6400. Profile for a second model served next to a large one, or for 4-14B models: two vLLM workers, each pinned to the 24 physical cores of a single NUMA node (nodes 2 and 5 - the two that xeon-large leaves free). Pair
Intel Xeon 6960P cores (physical): Default = 48 Cores, Max = 48 Cores; Memory (DDR5-6400, NUMA-local): Default = 192 GiB, Max = 512 GiB

View details

Customize resource requests and limits

Intel Xeon 6960P cores (physical) requests

− 48 + Cores

Min = 48 Cores, Max = 48 Cores

Intel Xeon 6960P cores (physical) limits

− 48 + Cores

Min = 48 Cores, Max = 48 Cores

Memory (DDR5-6400, NUMA-local) requests

− 192 + GiB

Min = 16 GiB, Max = 512 GiB

Memory (DDR5-6400, NUMA-local) limits

− 192 + GiB

Min = 16 GiB, Max = 512 GiB

Learn more about requests and limits

Serving runtime template *

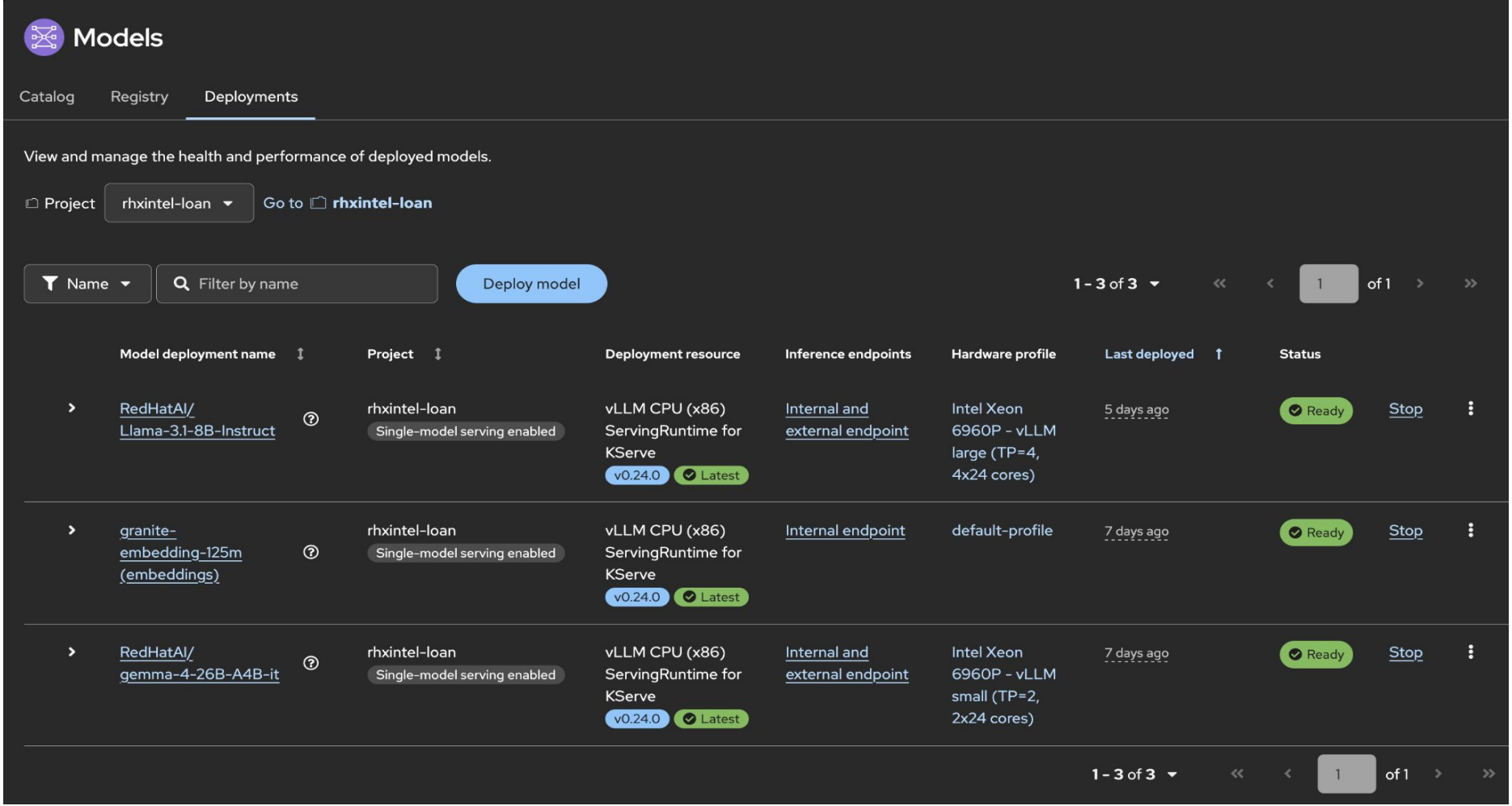
- **Automatic selection:** Automatically select the best serving runtime template for my model based on the selected hardware profile.
- **Manual selection:** Manually select a serving runtime template from a list of preconfigured and custom options.

vLLM CPU (x86) ServingRuntime for KServe v0.24.0 Global-scoped

Replica count *

− 1 +

Red Hat | intel.



Red Hat | intel.

Agentic AI Demo:

- AI App Developer Persona
- Key Video Screenshots / Slides



Models

Catalog | Registry | Deployments

View and manage the health and performance of deployed models.

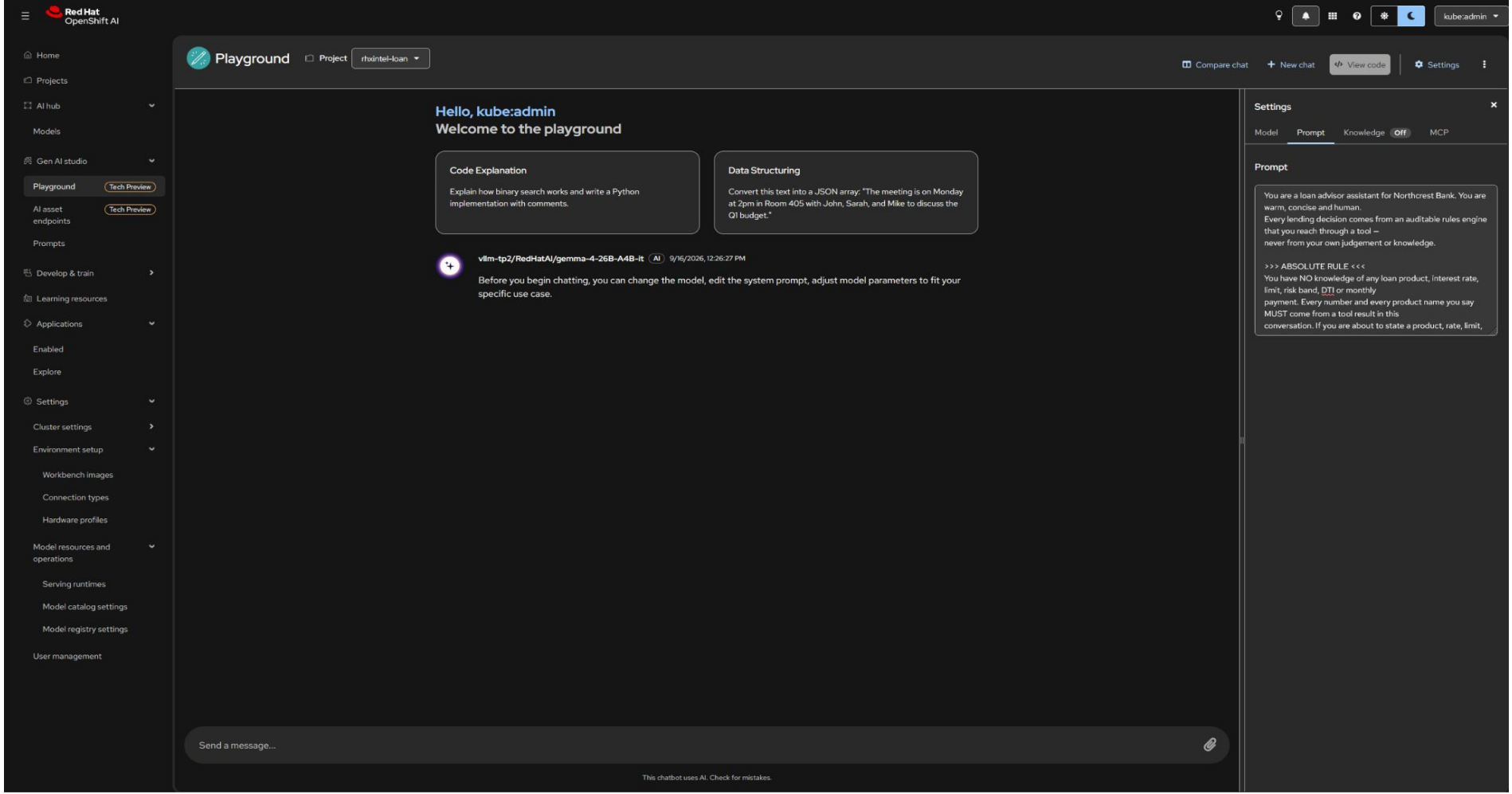
Project: rhxintel-loan | Go to rhxintel-loan

Name | Filter by name | Deploy model | 1 - 3 of 3 | 1 of 1

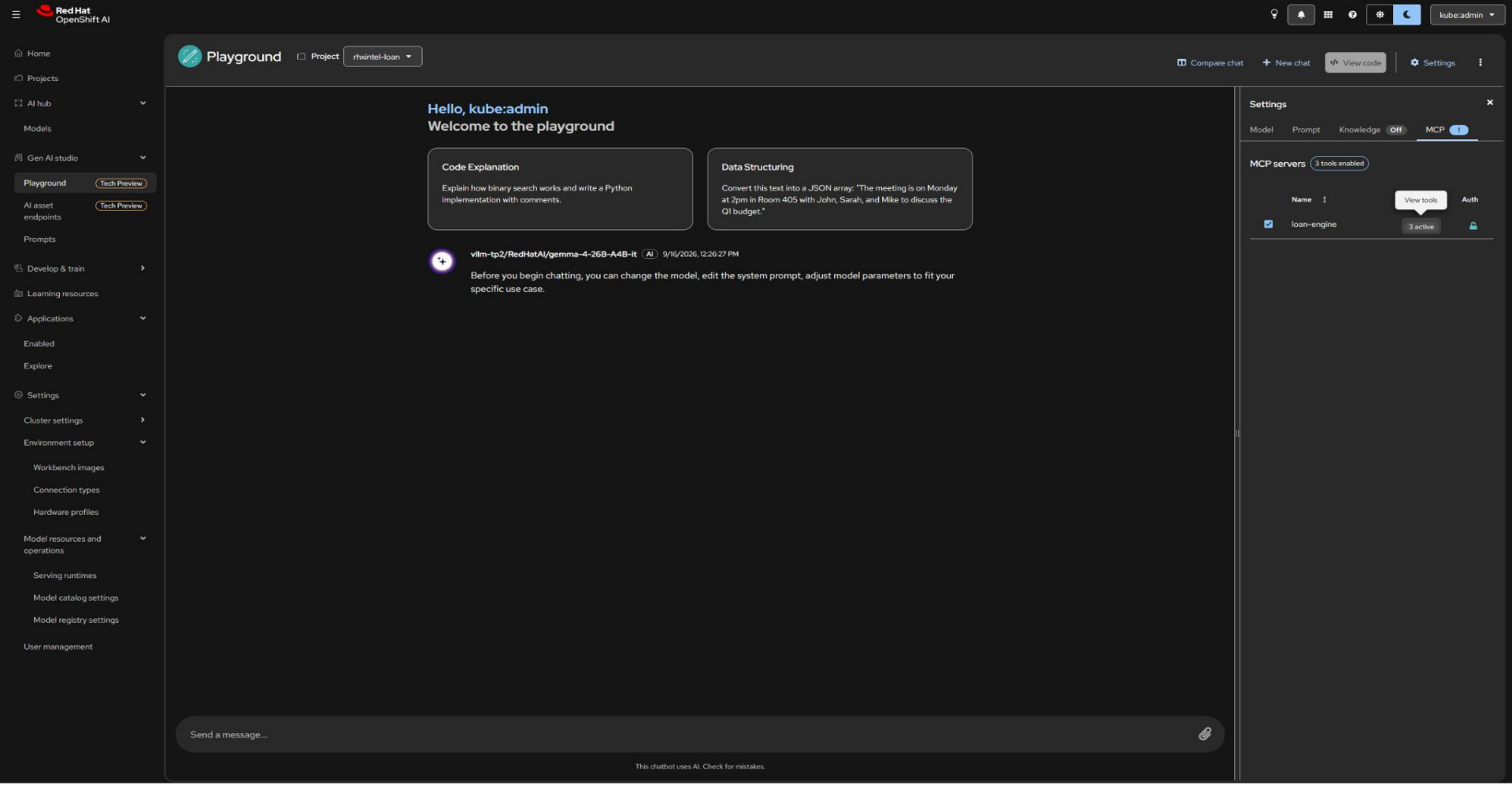
Model deployment name	Project	Deployment resource	Inference endpoints	Hardware profile	Last deployed	Status	
RedHatAI/ Llama-3.1-8B-Instruct	rhxintel-loan Single-model serving enabled	vLLM CPU (x86) ServingRuntime for KServe v0.24.0 Latest	Internal and external endpoint	Intel Xeon 6960P - vLLM large (TP=4, 4x24 cores)	5 days ago	Ready	Stop
granite-embedding-125m (embeddings)	rhxintel-loan Single-model serving enabled	vLLM CPU (x86) ServingRuntime for KServe v0.24.0 Latest	Internal endpoint	default-profile	7 days ago	Ready	Stop
RedHatAI/ gemma-4-26B-A4B-it	rhxintel-loan Single-model serving enabled	vLLM CPU (x86) ServingRuntime for KServe v0.24.0 Latest	Internal and external endpoint	Intel Xeon 6960P - vLLM small (TP=2, 2x24 cores)	7 days ago	Ready	Stop

1 - 3 of 3 | 1 of 1

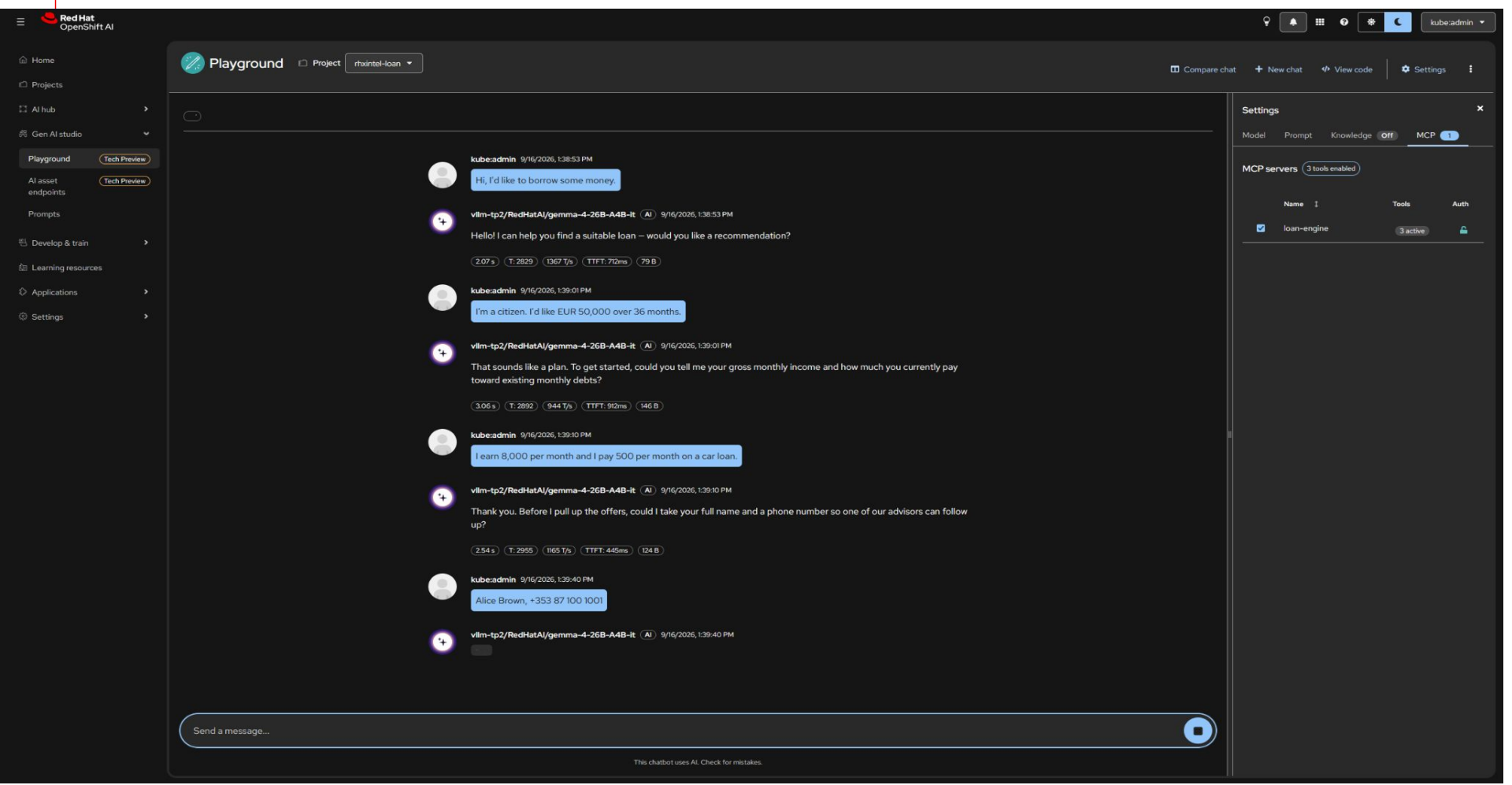
Red Hat | intel



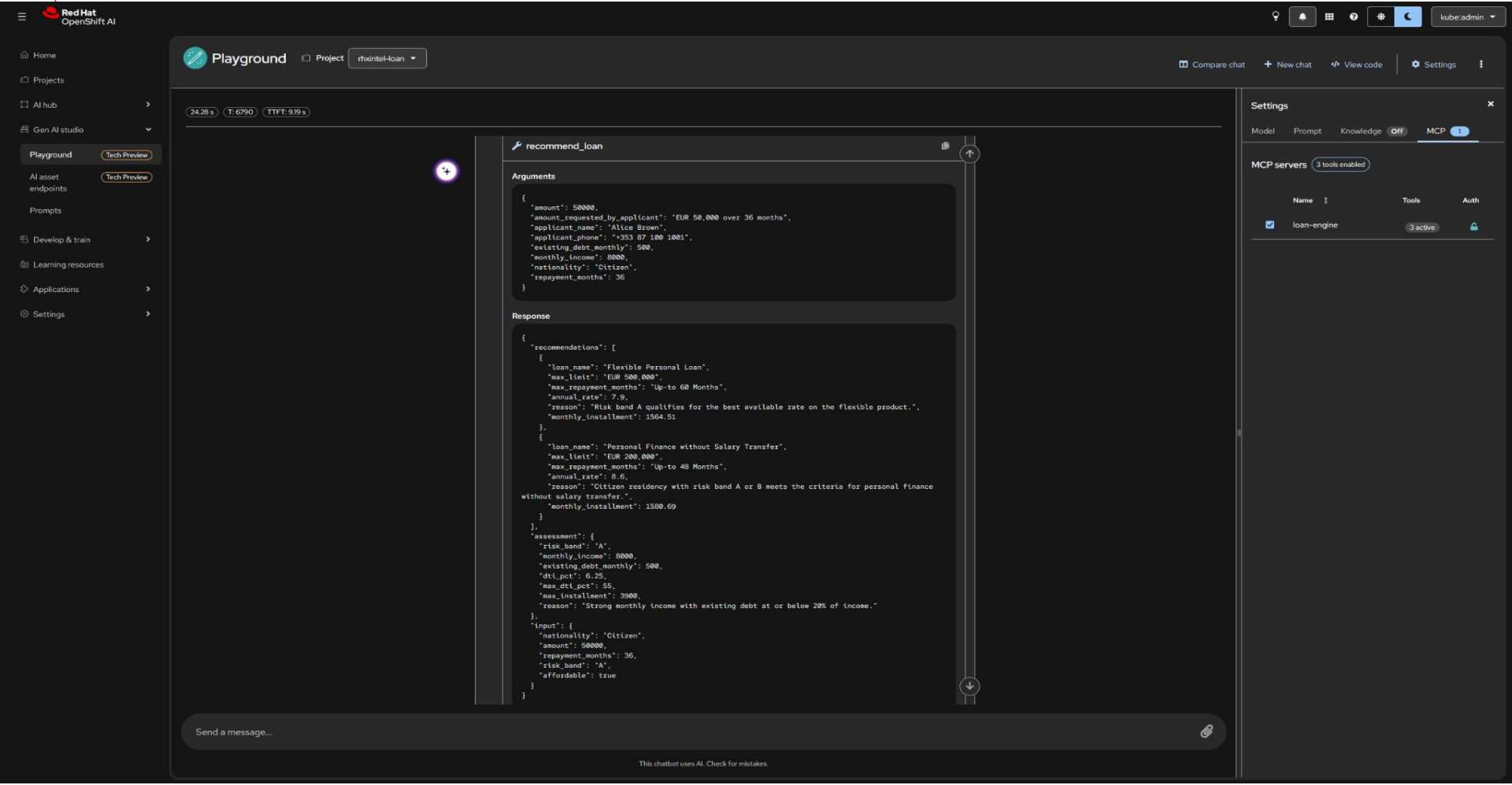
Red Hat | intel.



Red Hat | intel.



Red Hat | intel.


Red Hat OpenShift AI
kube:admin
Home
Projects
AI hub
Gen AI studio
Playground
Tech Preview
AI asset endpoints
Tech Preview
Prompts
Develop & train
Learning resources
Applications
Settings
Playground
Project
rhxintel-loan
Compare chat
New chat
View code
Settings
24.28 s
T: 6790
TTFT: 9.19 s
recommend_loan
Arguments
{
"amount": 50000,
"amount_requested_by_applicant": "EUR 50,000 over 36 months",
"applicant_name": "Alice Brown",
"applicant_phone": "+353 87 100 1001",
"existing_debt_monthly": 500,
"monthly_income": 8000,
"nationality": "Citizen",
"repayment_months": 36
}
Response
{
"recommendations": [
{
"loan_name": "Flexible Personal Loan",
"max_limit": "EUR 500,000",
"max_repayment_months": "Up-to 60 Months",
"annual_rate": 7.9,
"reason": "Risk band A qualifies for the best available rate on the flexible product.",
"monthly_installment": 1564.51
},
{
"loan_name": "Personal Finance without Salary Transfer",
"max_limit": "EUR 200,000",
"max_repayment_months": "Up-to 48 Months",
"annual_rate": 8.6,
"reason": "Citizen residency with risk band A or B meets the criteria for personal finance without salary transfer.",
"monthly_installment": 1580.69
}
],
"assessment": {
"risk_band": "A",
"monthly_income": 8000,
"existing_debt_monthly": 500,
"dti_pct": 6.25,
"max_dti_pct": 55,
"max_installment": 3900,
"reason": "Strong monthly income with existing debt at or below 20% of income."
},
"input": {
"nationality": "Citizen",
"amount": 50000,
"repayment_months": 36,
"risk_band": "A",
"affordable": true
}
}
Send a message...
This chatbot uses AI. Check for mistakes.
Settings
Model
Prompt
Knowledge
Off
MCP
1
MCP servers
3 tools enabled
Name
Tools
Auth
loan-engine
3 active
Red Hat
intel.

From <loan-bot@northcrestbank.com>
To <advisor@northcrestbank.com>
Subject **Loan lead: Alice Brown (+353 87 100 1001)**
Date Wed, 16 Sep 2026, 1:39 pm (1.8 kB)
Tags Add tags...

Text | Headers | Raw | Link Check

```
A prospective customer spoke with the loan assistant and showed interest in a loan.
Please reach out to them. The offers they were shown and the data behind the decision
are below.

Contact:
  Name:  Alice Brown
  Phone: +353 87 100 1001

Requested details:
  Nationality:      Citizen
  Amount:           EUR 50000.00
  Repayment months: 36

Credit assessment (decided by the rules engine, not by the model):
  Monthly income:      EUR 8000.00
  Existing debt:       EUR 500.00/month
  Debt-to-income:      6.25% (policy ceiling for this profile: 55.00%)
  Risk band:           A
  Affordable instalment: EUR 3900.00/month
  Basis:               Strong monthly income with existing debt at or below 20% of income.

Offers presented:
  - Flexible Personal Loan | max limit: EUR 500,000 | max repayment: Up-to 60 Months | rate: 7.9% p.a. | instalment for the requested amount: EUR 1564.51/month
    why: Risk band A qualifies for the best available rate on the flexible product.
  - Personal Finance without Salary Transfer | max limit: EUR 200,000 | max repayment: Up-to 48 Months | rate: 8.6% p.a. | instalment for the requested amount: EUR 1580.69/month
    why: Citizen residency with risk band A or B meets the criteria for personal finance without salary transfer.
```

Q & A





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

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat