



Red Hat

intel[®]



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."



vllny/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...



Settings



Model

Prompt

Knowledge Off

MCP

Model

vllny/RedHatAI/Llama-3.1-8B-Instruct

Temperature: 0 - 2



0.1

Streaming

Add audio transcription model

Agentic AI Demo

- MLOps Operations Persona
- Key Video Screenshots /
Slides



- 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	⋮



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



Hardware profiles

Manage hardware profiles for your organization. Administrators can use hardware profiles to determine resource allocation strategies for specific workloads or to explicitly define hardware configurations for users.

Name

Filter by name

Create hardware profile

		Name	Workload type	Enabled	Last modified	
>	⌵	default-profile This profile offers 2 CPUs and 4 GiB of memory. You can customize these...	All features	☒	--	⋮
>	⌵	intel gaudi This profile offers 2 CPUs and 4 GiB of memory. You can customize these...	All features	☒	2 months ago	⋮
>	⌵	Intel Xeon 6960P - vLLM large (TP=2, 2x24 cores) Intel Xeon 6960P (Granite Rapids-AP, AMX-BF16/INT8), SNC=3 so 24 physic...	All features	☒	--	⋮
▼	⌵	Intel Xeon 6960P - vLLM small (TP=2, 2x24 cores) Intel Xeon 6960P (Granite Rapids-AP, AMX-BF16/INT8), SNC=3 so 24 physic...	All features	☒	--	⋮

Node resources

Resource name	Resource identifier	Resource type	Default	Minimum allowed	Maximum allowed
Intel Xeon 69...	cpu	CPU	48 Cores	48 Cores	48 Cores
Memory (DD...	memory	Memory	192 GiB	16 GiB	512 GiB

- 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

Catalog Registry Deployments

Discover models that are available for your organization to register, deploy, and customize.

Validated models by Red Hat AI

Validated models by Red Hat AI offer confidence, predictability, and flexibility when deploying third-party generative AI models across the Red Hat AI platform.

[Explore Red Hat AI - validated models](#)



[LUL](#)

Model performance
view

Enable performance filters, display
model benchmark data, and exclude
unvalidated models.

Task

Q Search

- ☐ audio-text-to-text
- ☐ Audio-to-text
- ☐ automatic-speech-recognition
- ☐ automatic-speech-translation
- ☐ code-generation

Show more

Q gemma-4-26B

All models

Red Hat AI - validated models

Red Hat AI models

Other models



Validated

[gemma-4-26B-A4B-it-FP8-
dynamic](#)

FP8-dynamic variant of
gemma-4-26B-A4B-it.

[View 38 benchmarks](#)

Tool calling Text-to-text

6 more



Validated

[gemma-4-26B-A4B-it-
NVFP4](#)

NVFP4 variant of
gemma-4-26B-A4B-it.

[View 48 benchmarks](#)

Tool calling Text-to-text

7 more



Validated

[gemma-4-26B-A4B-it](#)

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

[View 60 benchmarks](#)

Tool calling Text-to-text

7 more

☰

Red Hat
OpenShift AI

🏠 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

💡🔔🗄️❓⚙️🌙

kube:admin

🔗

Models

CatalogRegistryDeployments

Model Catalog > gemma-4-26B-A4B-it

 **gemma-4-26B-A4B-it** Validated

Provided by Google

Deploy model

Register model

OverviewPerformance insightsSafety 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](#)

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

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 **redhatalgemma-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

Memory (DDR5-6400, NUMA-local) requests

192 GiB

Min = 16 GiB, Max = 512 GiB

[Learn more about requests and limits](#)

Intel Xeon 6960P cores (physical) limits

48 Cores

Min = 48 Cores, Max = 48 Cores

Memory (DDR5-6400, NUMA-local) limits

192 GiB

Min = 16 GiB, Max = 512 GiB

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



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

[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

[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

[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

>

>>

Agentic AI Demo:

- AI App Developer Persona
- Key Video Screenshots /
Slides



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

	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

- Home
- Projects
- AI hub
- Models
- Gen AI studio
- Playground Tech Preview
- AI asset endpoints Tech Preview
- Prompts
- Develop & train
- Learning resources
- Applications
- Enabled
- Explore
- 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 rhointel-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-tp2/RedHatAI/gemma-4-26B-A4B-it (AI) 9/16/2026, 12:26:27 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

Prompt

You are a loan advisor assistant for Northcrest Bank. You are warm, concise and human.
Every lending decision comes from an auditable rules engine that you reach through a tool – never from your own judgement or knowledge.

>>> ABSOLUTE RULE <<<
You have NO knowledge of any loan product, interest rate, limit, risk band, DTI or monthly payment. Every number and every product name you say MUST come from a tool result in this conversation. If you are about to state a product, rate, limit,

Red Hat
OpenShift AI

Home

Projects

AI hub

Models

Gen AI studio

PlaygroundTech Preview

AI asset endpointsTech Preview

Prompts

Develop & train

Learning resources

Applications

Enabled

Explore

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

PlaygroundProjectrhintel-loan

Compare chatNew chatView codeSettings

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-tp2/RedHatAI/gemma-4-26B-A4B-It

AI

9/16/2026, 12:26:27 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

ModelPromptKnowledgeOffMCP1

MCP servers3 tools enabled

NameI

View tools

Auth

☒

loan-engine

3 active

Red Hat

intel

Red Hat
OpenShift AI

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

kube:admin

9/16/2026, 1:38:53 PM

Hi, I'd like to borrow some money.

vllm-tp2/RedHatAI/gemma-4-26B-A4B-It (AI)

9/16/2026, 1:38:53 PM

Hello! I can help you find a suitable loan – would you like a recommendation?

2.07 s

T: 2820

1367 T/s

TTFT: 712ms

79 B

kube:admin

9/16/2026, 1:39:01 PM

I'm a citizen. I'd like EUR 50,000 over 36 months.

vllm-tp2/RedHatAI/gemma-4-26B-A4B-It (AI)

9/16/2026, 1:39:01 PM

That sounds like a plan. To get started, could you tell me your gross monthly income and how much you currently pay toward existing monthly debts?

3.06 s

T: 2892

944 T/s

TTFT: 912ms

146 B

kube:admin

9/16/2026, 1:39:10 PM

I earn 8,000 per month and I pay 500 per month on a car loan.

vllm-tp2/RedHatAI/gemma-4-26B-A4B-It (AI)

9/16/2026, 1:39:10 PM

Thank you. Before I pull up the offers, could I take your full name and a phone number so one of our advisors can follow up?

2.54 s

T: 2958

1165 T/s

TTFT: 445ms

124 B

kube:admin

9/16/2026, 1:39:40 PM

Alice Brown, +353 87 100 1001

vllm-tp2/RedHatAI/gemma-4-26B-A4B-It (AI)

9/16/2026, 1:39:40 PM

Send a message...

This chatbot uses AI. Check for mistakes.

Settings

Model

Prompt

Knowledge Off

MCP On

MCP servers 3 tools enabled

Name	Tools	Auth
☒ loan-engine	3 active	

 Red Hat

 intel

24.28 s

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_lmt": "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": 1504.51
  },
  {
    "loan_name": "Personal Finance without Salary Transfer",
    "max_lmt": "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
}
}

```


This chatbot uses AI. Check for mistakes.

x

Model

Knowledge Off

MCP 1

MCP servers 3 tools enabled

Name _____

Te

Auth

☒ loan-engine

3 active

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



Connect

Thank you



linkedin.com/company/red-hat



facebook.com/redhatinc



youtube.com/user/RedHatVideos



twitter.com/RedHat