Red Hat AI Inference Server 3.2

Validated models

Red Hat AI Inference Server Validated models

Last Updated: 2025-07-24

Red Hat AI Inference Server 3.2 Validated models

Red Hat AI Inference Server Validated models

Legal Notice

Abstract

Learn about the validated models that you can run with Red Hat AI Inference Server.

Table of Contents

PREFACE

Red Hat provides validated third-party models that you can serve with AI Inference Server. These models are designed for efficient deployment on the Red Hat AI platform. Models are validated using open source tools. Red Hat uses GuideLLM for performance benchmarking and Language Model Evaluation Harness for accuracy evaluations.

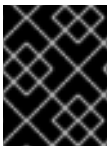
NOTE

You can explore the validated models, complete with model details and deployment instructions, in the Red Hat AI validated models - v1.0 collection on Hugging Face.

CHAPTER 1. RED HAT AI VALIDATED MODELS

The following table lists the Red Hat AI validated models for use with Red Hat AI Inference Server 3.2.

- If you are using AI Inference Server as standalone product, use the Hugging Face images.
- If you are using AI Inference Server as part of a RHEL AI deployment, use the model OCI artifact image.
- If you are using AI Inference Server as part of a OpenShift AI deployment, use the model ModelCar image.



IMPORTANT

AMD GPUs support FP8 (W8A8) and GGUF quantization variant models only. For more information, see Supported hardware.

Table 1.1. Red Hat AI validated models

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-4-Scout-17B-16E-Instruct	INT4, FP8	• Baseline • INT4 • FP8	• Baseline: **registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct-quantized-w4a16:1.5** • FP8: **registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct-quantized-w4a16:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-4-Maverick-17B-128E-Instruct	FP8	• Baseline • FP8	• Baseline: **registry.redhat.io/rhelai1/llama-4-maverick-17b-128e-instruct:1.5** • FP8: **registry.redhat.io/rhelai1/llama-4-maverick-17b-128e-instruct-fp8:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-llama-4-maverick-17b-128e-instruct:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-llama-4-maverick-17b-128e-instruct-fp8:1.5**
Mistral-Small-3.1-24B-Instruct-2503	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503:1.5** • INT4: **registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai**	• Baseline: **registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-quantized-w4a16:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
			1/mistral-small-3-1-24b-instruct-2503-quantized-w8a8:1.5 • FP8: **registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503-fp8-dynamic:1.5**	• INT8: **registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.3-70B-Instruct	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/llama-3-3-70b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/llama-3-3-70b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/llama-3-3-70b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/llama-3-3-70b-instruct-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.1-8B-Instruct	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/llama-3-1-8b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/llama-3-1-8b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/llama-3-1-8b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/llama-3-1-8b-instruct-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
granite-3.1-8b-instruct	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/granite-3-1-8b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/granite-3-1-8b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/granite-3-1-8b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/granite-3-1-8b-instruct-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
phi-4	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/phi-4:1.5** • INT4: **registry.redhat.io/rhelai1/phi-4-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/phi-4-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/phi-4-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-phi-4:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-phi-4-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-phi-4-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-phi-4-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Qwen2.5-7B-Instruct	INT4, INT8, FP8	• Baseline • INT4 • INT8 • FP8	• Baseline: **registry.redhat.io/rhelai1/qwen2-5-7b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/qwen2-5-7b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/qwen2-5-7b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/qwen2-5-7b-instruct-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-fp8-dynamic:1.5**
	INT4, INT8, FP8	• Baseline	• Baseline:	• Baseline:

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Mistral-Small-24B-Instruct-2501		• INT4 • INT8 • FP8	**registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501:1.5** • INT4: **registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-fp8-dynamic:1.5**	**registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501:1.5** • INT4: **registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-quantized-w4a16:1.5** • INT8: **registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-quantized-w8a8:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-fp8-dynamic:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Mixtral-8x7B-Instruct-v0.1	None	• Baseline	• Baseline: **registry.redhat.io/rhelai1/mixtral-8x7b-instruct-v0-1:1.4**	• Baseline: **registry.redhat.io/rhelai1/modelcar-mixtral-8x7b-instruct-v0-1:1.4**
granite-3.1-8b-base	INT4 (baseline currently unavailable)	• INT4	• INT4: **registry.redhat.io/rhelai1/granite-3-1-8b-base-quantized-w4a16:1.5**	• INT4: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-base-quantized-w4a16:1.5**
granite-3.1-8b-starter-v2	None	• Unavailable on Hugging Face	• Baseline: **registry.redhat.io/rhelai1/granite-3.1-8b-starter-v2:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-starter-v2:1.5**

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.1-Nemotron-70B-Instruct-HF	FP8	• Baseline • FP8	• Baseline: **registry.redhat.io/rhelai1/llama-3-1-nemotron-70b-instruct-hf:1.5** • FP8: **registry.redhat.io/rhelai1/llama-3-1-nemotron-70b-instruct-hf-fp8-dynamic:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-llama-3-1-nemotron-70b-instruct-hf:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-llama-3-1-nemotron-70b-instruct-hf-fp8-dynamic:1.5**
gemma-2-9b-it	FP8	• Baseline • FP8	• Baseline: **registry.redhat.io/rhelai1/gemma-2-9b-it:1.5** • FP8: **registry.redhat.io/rhelai1/gemma-2-9b-it-fp8:1.5**	• Baseline: **registry.redhat.io/rhelai1/modelcar-gemma-2-9b-it:1.5** • FP8: **registry.redhat.io/rhelai1/modelcar-gemma-2-9b-it-fp8:1.5**

1. For use with standalone Red Hat AI Inference Server
2. For use with RHEL AI

3. For use with Red Hat OpenShift AI