



# Red Hat AI Inference Server 3.2

## Validated models

Red Hat AI Inference Server Validated models



## Red Hat AI Inference Server 3.2 Validated models

---

Red Hat AI Inference Server Validated models

## Legal Notice

Copyright © 2025 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux<sup>®</sup> is the registered trademark of Linus Torvalds in the United States and other countries.

Java<sup>®</sup> is a registered trademark of Oracle and/or its affiliates.

XFS<sup>®</sup> is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL<sup>®</sup> is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js<sup>®</sup> is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack<sup>®</sup> Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

## Abstract

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

Table of Contents

PREFACE ..... 3

CHAPTER 1. RED HAT AI VALIDATED MODELS ..... 4



## 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	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct-quantized-w4a16:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/llama-4-scout-17b-16e-instruct-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct-quantized-w4a16:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-llama-4-scout-17b-16e-instruct-fp8-dynamic:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-4-Maverick-17B-128E-Instruct	FP8	<ul style="list-style-type: none"> <li>Baseline</li> <li>FP8</li> </ul>	<ul style="list-style-type: none"> <li>Baseline: <b>registry.redhat.io/rhelai1/llama-4-maverick-17b-128e-instruct:1.5</b></li> <li>FP8: <b>registry.redhat.io/rhelai1/llama-4-maverick-17b-128e-instruct-fp8:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>Baseline: <b>registry.redhat.io/rhelai1/modelcar-llama-4-maverick-17b-128e-instruct:1.5</b></li> <li>FP8: <b>registry.redhat.io/rhelai1/modelcar-llama-4-maverick-17b-128e-instruct-fp8:1.5</b></li> </ul>
Mistral-Small-3.1-24B-Instruct-2503	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>Baseline</li> <li>INT4</li> <li>INT8</li> <li>FP8</li> </ul>	<ul style="list-style-type: none"> <li>Baseline: <b>registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503:1.5</b></li> <li>INT4: <b>registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503-quantized-w4a16:1.5</b></li> <li>INT8: <b>registry.redhat.io/rhelai</b></li> </ul>	<ul style="list-style-type: none"> <li>Baseline: <b>registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503:1.5</b></li> <li>INT4: <b>registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-quantized-w4a16:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
			<div>1/mistral-small-3-1-24b-instruct-2503-quantized-w8a8:1.5</div> <div><ul style="list-style-type: none"><li>FP8: registry.redhat.io/rhelai1/mistral-small-3-1-24b-instruct-2503-fp8-dynamic:1.5</li></ul></div>	<div>● INT8: registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-quantized-w8a8:1.5</div> <div><ul style="list-style-type: none"><li>FP8: registry.redhat.io/rhelai1/modelcar-mistral-small-3-1-24b-instruct-2503-fp8-dynamic:1.5</li></ul></div>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.3-70B-Instruct	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">INT8</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/llama-3-3-70b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/llama-3-3-70b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/llama-3-3-70b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/llama-3-3-70b-instruct-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-llama-3-3-70b-instruct-fp8-dynamic:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.1-8B-Instruct	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">INT8</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/llama-3-1-8b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/llama-3-1-8b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/llama-3-1-8b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/llama-3-1-8b-instruct-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-8b-instruct-fp8-dynamic:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
granite-3.1-8b-instruct	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">INT8</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/granite-3-1-8b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/granite-3-1-8b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/granite-3-1-8b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/granite-3-1-8b-instruct-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-instruct-fp8-dynamic:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
phi-4	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">INT8</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/phi-4:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/phi-4-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/phi-4-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/phi-4-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-phi-4:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-phi-4-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/modelcar-phi-4-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-phi-4-fp8-dynamic:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Qwen2.5-7B-Instruct	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">INT4</a></li> <li>• <a href="#">INT8</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/qwen2-5-7b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/qwen2-5-7b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/qwen2-5-7b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/qwen2-5-7b-instruct-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct:1.5</b></li> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-quantized-w4a16:1.5</b></li> <li>• INT8: <b>registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-quantized-w8a8:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-qwen2-5-7b-instruct-fp8-dynamic:1.5</b></li> </ul>
	INT4, INT8, FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline:</li> </ul>	<ul style="list-style-type: none"> <li>• Baseline:</li> </ul>



Mistral-Small-24b-Instruct-2501 Model	Quantized variants	<ul style="list-style-type: none"><li>INT4</li></ul> Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
		<ul style="list-style-type: none"><li>FP8</li></ul>	<ul style="list-style-type: none"><li>registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501:1.5</li><li>INT4: registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-quantized-w4a16:1.5</li><li>INT8: registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-quantized-w8a8:1.5</li><li>FP8: registry.redhat.io/rhelai1/mistral-small-24b-instruct-2501-fp8-dynamic:1.5</li></ul>	<ul style="list-style-type: none"><li>registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501:1.5</li><li>INT4: registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-quantized-w4a16:1.5</li><li>INT8: registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-quantized-w8a8:1.5</li><li>FP8: registry.redhat.io/rhelai1/modelcar-mistral-small-24b-instruct-2501-fp8-dynamic:1.5</li></ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Mixtral-8x7B-Instruct-v0.1	None	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/mixtral-8x7b-instruct-v0-1:1.4</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-mixtral-8x7b-instruct-v0-1:1.4</b></li> </ul>
granite-3.1-8b-base	INT4 (baseline currently unavailable)	<ul style="list-style-type: none"> <li>• <a href="#">INT4</a></li> </ul>	<ul style="list-style-type: none"> <li>• INT4: <b>registry.redhat.io/rhelai1/granite-3-1-8b-base-quantized-w4a16:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• INT4: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-base-quantized-w4a16:1.5</b></li> </ul>
granite-3.1-8b-starter-v2	None	<ul style="list-style-type: none"> <li>• Unavailable on Hugging Face</li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/granite-3-1-8b-starter-v2:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-granite-3-1-8b-starter-v2:1.5</b></li> </ul>

Model	Quantized variants	Hugging Face model cards [1]	OCI artifact images [2]	ModelCar images [3]
Llama-3.1-Nemotron-70B-Instruct-HF	FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/llama-3-1-nemotron-70b-instruct-hf:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/llama-3-1-nemotron-70b-instruct-hf-fp8-dynamic:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-nemotron-70b-instruct-hf:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-llama-3-1-nemotron-70b-instruct-hf-fp8-dynamic:1.5</b></li> </ul>
gemma-2-9b-it	FP8	<ul style="list-style-type: none"> <li>• <a href="#">Baseline</a></li> <li>• <a href="#">FP8</a></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/gemma-2-9b-it:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/gemma-2-9b-it-fp8:1.5</b></li> </ul>	<ul style="list-style-type: none"> <li>• Baseline: <b>registry.redhat.io/rhelai1/modelcar-gemma-2-9b-it:1.5</b></li> <li>• FP8: <b>registry.redhat.io/rhelai1/modelcar-gemma-2-9b-it-fp8:1.5</b></li> </ul>

1. For use with standalone Red Hat AI Inference Server

2. For use with RHEL AI

3. For use with Red Hat OpenShift AI