



# Red Hat AI Inference Server 3.0

## Release notes

Highlights of what is new and what has changed with this Red Hat AI Inference Server release



## Red Hat AI Inference Server 3.0 Release notes

---

Highlights of what is new and what has changed with this Red Hat AI Inference Server release

## 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

The release notes for Red Hat AI Inference Server summarize all new features and enhancements, notable technical changes, major corrections from the previous version, and any known bugs upon general availability.

Table of Contents

PREFACE ..... 3

CHAPTER 1. ABOUT THIS RELEASE ..... 4

CHAPTER 2. NEW FEATURES AND ENHANCEMENTS ..... 5



# PREFACE

Red Hat AI Inference Server provides developers and IT organizations with a scalable inference platform for deploying and customizing AI models on secure, scalable resources with minimal configuration and resource usage.

## CHAPTER 1. ABOUT THIS RELEASE

Red Hat AI Inference Server is now available. This Red Hat AI Inference Server 3.0 release provides container images that optimizes inferencing with large language models (LLMs) for NVIDIA and ROCm accelerators. The container images are available from [registry.redhat.io](https://registry.redhat.io):

- **`registry.redhat.io/rhaiis/vllm-cuda-rhel9:3.0.0`**
- **`registry.redhat.io/rhaiis/vllm-rocm-rhel9:3.0.0`**

With Red Hat AI Inference Server, you can serve and inference models with higher performance, lower cost, and enterprise-grade stability and security. Red Hat AI Inference Server is built on the upstream, open source [vLLM](#) software project.



## CHAPTER 2. NEW FEATURES AND ENHANCEMENTS

For a complete list of new features and enhancements for vLLM, review the upstream [vLLM v0.8.4 release notes](#).