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

Last Updated: 2025-06-24

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

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

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:

- **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.