Red Hat Enterprise Linux AI 1.5

Hardware Requirements

Hardware requirements for RHEL AI

Last Updated: 2025-06-27

Red Hat Enterprise Linux AI 1.5 Hardware Requirements

Hardware requirements for RHEL AI

Legal Notice

Abstract

This document provides details for RHEL AI hardware requirements

Table of Contents

CHAPTER 1. RED HAT ENTERPRISE LINUX AI HARDWARE REQUIREMENTS

Various hardware accelerators require different requirements for serving and inferencing as well as installing, generating and training the Granite starter model on Red Hat Enterprise Linux AI.

1.1. HARDWARE REQUIREMENTS FOR END-TO-END WORKFLOW OF GRANITE MODELS

The following charts show the hardware requirements for running the full InstructLab end-to-end workflow to customize the Granite student model. This includes: synthetic data generation (SDG), multi-phase training, and evaluating a custom Granite model.

1.1.1. Bare metal

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU memory	Recommended additional disk storage
NVIDIA	2xA100 4xA100 8xA100	160 GB 320 GB 640 GB	3 TB
NVIDIA	2xH100 4xH100 8xH100	160 GB 320 GB 640 GB	3 TB
NVIDIA	2xH200 4xH200 8xH200	282 GB 564 GB 1128 GB	3 TB
NVIDIA	4xL40S 8xL40S	192 GB 384 GB	3 TB
AMD	2xMI300X 4xMI300X 8xMI300X	384 GB 768 GB 1536 GB	3 TB

1.1.2. IBM Cloud

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU Memory	IBM Cloud Instances	Recommended additional disk storage
NVIDIA	2xA100	160 GB	gx3d-48x240x2a100p	3 TB
NVIDIA	8xH100	640 GB	gx3d-160x1792x8h100	3 TB
NVIDIA	8xH200	1128 GB	gx3d-160x1792x8h200	3 TB
AMD	8xMI300X	1536 GB	gx3d-208x1792x8mi300x	3 TB

1.1.3. Amazon Web Services (AWS)

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU Memory	AWS Instances	Recommended additional disk storage
NVIDIA	8xA100	320 GB	p4d.24xlarge	3 TB
NVIDIA	8xA100	640 GB	p4de.24xlarge	3 TB
NVIDIA	8xH100	640 GB	p5.48xlarge	3 TB
NVIDIA	8xL40S	384 GB	g6e.48xlarge	3 TB

1.1.4. Azure

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU Memory	Azure Instances	Recommended additional disk storage
NVIDIA	8xA100	640 GB	Standard_ND96amsr_A100_v4	3 TB
NVIDIA	4xA100	320 GB	Standard_ND96asr_A100_v4	3 TB
NVIDIA	8xH100	640 GB	Standard_ND96isr_H100_v5	3 TB

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU Memory	Azure Instances	Recommended additional disk storage
AMD	8xMI300X	1535 GB	Standard_ND96is_MI300X_v5	3 TB

1.1.5. Google Cloud Platform (GCP)

Hardware vendor	Supported accelerators (GPUs)	Aggregate GPU Memory	GCP Instances	Recommended additional disk storage
NVIDIA	8xA100	640 GB	a2-highgpu-8g	3 TB
NVIDIA	8xH100	640 GB	a3-highgpu-8g a3-megagpu-8g	3 TB

1.2. HARDWARE REQUIREMENTS FOR INFERENCE SERVING GRANITE MODELS

The following charts display the minimum hardware requirements for inference serving a model on Red Hat Enterprise Linux AI.

1.2.1. Bare metal

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU memory	Recommended additional disk storage
NVIDIA	A100	80 GB	1 TB
NVIDIA	H100	80 GB	1 TB
NVIDIA	H200	141 GB	1 TB
NVIDIA	GH200 (Technology Preview)	192 GB	1 TP
NVIDIA	L40S	48 GB	1 TB
NVIDIA	L4	24 GB	1 TB

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU memory	Recommended additional disk storage
AMD	MI300X	192 GB	1 TB
Intel	Gaudi 3 (Technology Preview)	128 GB	1 TB

1.2.2. Amazon Web Services (AWS)

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU Memory	AWS Instance family	Recommended additional disk storage
NVIDIA	A100	40 GB	P4d series	1 TB
NVIDIA	H100	80 GB	P5 series	1 TB
NVIDIA	L40S	48 GB	G6e series	1 TB
NVIDIA	L4	24 GB	G6 series	1 TB

1.2.3. IBM cloud

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU Memory	IBM Cloud Instance family	Recommended additional disk storage
NVIDIA	L4	24 GB	gx3 series	1 TB
NVIDIA	L40S	48 GB	gx3 series	1 TB
NVIDIA	A100	80 GB	gx3 series	1 TB
NVIDIA	H100	80 GB	gx3 series	1 TB
NVIDIA	H200	141 GB	gx3 series	1 TB
AMD	MI300X	192 GB	gx3 series	1 TB
Intel	Gaudi 3 (Technology Preview)	128 GB	gx3 series	1 TB

1.2.4. Azure

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU Memory	Azure Instance family	Recommended additional disk storage
NVIDIA	A100	80 GB	ND series	1 TB
NVIDIA	H100	80 GB	ND sereis	1 TB
AMD	MI300X	192 GB	ND series	1 TB

1.2.5. Google Cloud Platform (GCP)

Hardware vendor	Supported accelerators (GPUs)	Minimum Aggregate GPU Memory	GCP Instance family	Recommended additional disk storage
NVIDIA	A100	40 GB	A2 series	1 TB
NVIDIA	H100	80 GB	A3 series	1 TB
NVIDIA	4xL4	96 GB	G2 series	1 TB