Red Hat Enterprise Linux AI 1.1

Release notes

Red Hat Enterprise Linux AI release notes

Last Updated: 2024-10-23

Red Hat Enterprise Linux AI 1.1 Release notes

Red Hat Enterprise Linux AI release notes

Legal Notice

Abstract

This document provides the release notes for Red Hat Enterprise Linux AI version 1.1

Table of Contents

CHAPTER 1. RED HAT ENTERPRISE LINUX AI 1.1 RELEASE NOTES

Red Hat Enterprise Linux AI version 1.1 is the general availability release of the product. RHEL AI provides organizations with a process to develop enterprise applications on open source Large Language Models (LLMs).

1.1. ABOUT THIS RELEASE

Red Hat Enterprise Linux AI version 1.1 includes various features for Large Language Model (LLM) fine-tuning on the Red Hat and IBM produced Granite model. A customized model using the RHEL AI workflow consisted of the following:

- Install and launch a RHEL 9.4 instance with the InstructLab tooling.
- Host information in s Git repository and interact with a Git-based taxonomy of the knowledge you want a model to learn.
- Run the end-to-end workflow of synthetic data generation, training, and evaluation.
- Serve and chat with the newly fine-tuned LLM.

1.2. FEATURES AND ENHANCEMENTS

Red Hat Enterprise Linux AI version 1.1 includes various features for Large Language Model (LLM) fine-tuning. This includes the end-to-end workflow of synthetic data generation (SDG), training and evaluation on various supported cloud platforms.

1.2.1. Installing

Red Hat Enterprise Linux AI is installable as a bootable image. This image contains various tooling for interacting with RHEL AI. The image includes: Red Hat Enterprise Linux 9.4, Python version 3.11 and InstructLab tools for model fine-tuning. For more information about installing Red Hat Enterprise Linux AI, see Installation overview.

1.2.1.1. Installing on bare metal

You can install Red Hat Enterprise Linux AI on various NVIDIA bare-metal hardware in RHEL AI version 1.1. For the documentation on installing Red Hat Enterprise Linux AI on bare metal, see Installing RHEL AI on bare metal.

RHEL AI currently supports 2x-8x NVIDIA A100, 2x-8x NVIDIA H100, and 4x-8x NVIDIA L40S GPU accelerators for the full end-to-end workflow. You can also inference serve LLMs provided by Red Hat on NVIDIA bare-metal hardware. For more details on the RHEL AI hardware requirements for NVIDIA bare-metal machines, see Red Hat Enterprise Linux AI hardware requirements .

1.2.1.2. Installing on Amazon Web Services (AWS)

You can install and deploy Red Hat Enterprise Linux AI on Amazon Web Services (AWS) in RHEL AI version 1.1. You can deploy an Red Hat Enterprise Linux AI AWS instance by converting a RHEL AI raw disk image into an Amazon Machine Image (AMI) and launching an instance with the AMI. For the documentation on converting the RHEL AI image to an AMI and deploying the instance, see Installing RHEL AI on AWS.

RHEL AI currently supports 8xA100 and 8xH100 accelerators on AWS instances for the full end-to-end workflow. You can also inference serve LLMs provided by Red Hat on AWS instances. For more details on the RHEL AI hardware requirements for AWS, see Red Hat Enterprise Linux AI hardware requirements.

1.2.1.3. Installing on IBM Cloud

You can install and deploy Red Hat Enterprise Linux AI on IBM Cloud in RHEL AI version 1.1. You can deploy an Red Hat Enterprise Linux AI IBM Cloud instance by converting the RHEL AI image into an QCOW image, then into a IBM Cloud image and launching an instance. For the documentation on converting the RHEL AI image to an IBM Cloud and deploying the instance, see Installing RHEL AI on IBM Cloud.

RHEL AI currently only supports inference serving Large Language Models (LLMs) on version 1.1. For more details on the RHEL AI hardware requirements for IBM Cloud, see Red Hat Enterprise Linux AI hardware requirements

1.2.2. Building your RHEL AI environment

After installing Red Hat Enterprise Linux AI, you can set up your RHEL AI environment in various ways with the InstructLab tools.

1.2.2.1. Initializing InstructLab

You can initialize and set up your RHEL AI environment by running the **ilab config init** command. This command creates the necessary configurations for interacting with RHEL AI and fine-tuning models. It also creates proper directories for your data files. Red Hat Enterprise Linux AI version 1.1 includes the ability to select a training profile that matches your hardware configurations through the InstructLab CLI. The training profiles in the CLI allows you to easily configure your machine for training when setting up your RHEL AI environment. For more information about initializing InstructLab, see the Initialize InstructLab documentation.

1.2.2.2. Downloading Large Language Models

You can download various Large Language Models (LLMs) provided by Red Hat to your RHEL AI machine or instance. You can download these models from a Red Hat registry after creating and logging in to your Red Hat registry account. RHEL AI allows you to fine-tune the **granite-7b-starter** base model with your own knowledge data. There are various models necessary for the end-to-end workflow as well that are downloadable on RHEL AI, including the **mixtral-8x7B-instruct-v0-1** teacher model for SDG and **prometheus-8x7b-v2.0** judge model for training and evaluation. For more information about downloading models, see the Downloading models documentation.

The **granite-8b-code-instruct** and **granite-8b-code-base** code models are currently in Technology Preview on RHEL AI version 1.1. For more information about the support scope of Red Hat Technology Preview features, see Technology Preview Features Support Scope .

1.2.2.3. Serving and chatting with models

Red Hat Enterprise Linux AI version 1.1 allows you to run an vLLM inference server on various LLMs. The vLLM tool is a memory-efficient inference and serving engine library for LLMs that is included in the RHEL AI image. After you serve a specified model, you can chat with the model via the InstructLab CLI. For more information about serving and chatting with models, see Serving and chatting with the models documentation.

1.2.2.3.1. Allowing chat access to a model from a secure endpoint

Red Hat Enterprise Linux AI general availability includes the ability to serve an inference endpoint and allow others to interact with models provided with Red Hat Enterprise Linux AI. This process is only currently supported on bare metal platforms. For more information on chatting from a secure endpoint, see the Optional: Allowing chat access to a model from a secure endpoint documentation.

1.2.2.3.2. Running **ilab model serve** as a service

Red Hat Enterprise Linux AI general availability includes the ability to set up a **systemd** service so that the ilab model serve command runs as a running service. For more information on running the **ilab model serve** as a service, see the Optional: Running ilab model serve as a service documentation.

1.2.3. Customizing a Large Language Model (LLM) on RHEL AI

Red Hat Enterprise Linux AI allows you to customize and fine-tune the **granite-7b-starter** base model with the RHEL AI end-to-end workflow.

1.2.3.1. Adding knowledge data to a Granite LLM.

On Red Hat Enterprise Linux AI, you can customize your taxonomy tree so a model can learn domain-specific information. You host your knowledge data in a Git repository and fine-tune a model with that data. Red Hat Enterprise Linux AI version 1.1 currently only supports knowledge data in markdown format. In the RHEL AI workflow, you create a **qna.yaml** file that includes questions and answers for the model to learn. This file gets run through the synthetic data generation (SDG) process, training, and evaluation, to then create a new LLM that contains the data from the Git repository and **qna.yaml** file. For detailed documentation on how to create a knowledge markdown and YAML file, see Adding knowledge to your taxonomy tree.

1.2.3.2. Synthetic Data Generation (SDG)

Red Hat Enterprise Linux AI includes the LAB enhanced method of synthetic data generation (SDG). You can use the **qna.yaml** files with your own knowledge data to create hundreds of artifical datasets in the SDG process. Red Hat Enterprise Linux AI includes the **mixtral-8x7B-instruct-v0-1** LLM, available for downloading, as the teacher and critic model for SDG. For more information on running the SDG process, see Generating a new dataset with Synthetic data generation (SDG) .

1.2.3.3. Multi-phase training

Red Hat Enterprise Linux AI includes the LAB enhanced method of multi-phase training: A fine-tuning strategy where datasets are trained and evaluated in multiple phases to create the best possible model. You can use the synthetic data set generated during SDG to train, evaluate and create a new model with your data. Red Hat Enterprise Linux AI includes the **prometheus-8x7b-v2.0** LLM, available for downloading, as the judge model for multi-phase training and evaluation. For more details on multi-phase training, see Training your data on the model .

1.2.3.4. Benchmark evaluation

Red Hat Enterprise Linux AI includes the ability to run benchmark evaluations on the newly trained models. On your trained model, you can evaluate how well the model knows the model you added with the **MMLU_BRANCH** benchmark. Red Hat Enterprise Linux AI includes the **prometheus-8x7b-v2.0** LLM, available for downloading, as the judge model for benchmark evaluation. For more details on benchmark evaluation, see Evaluating your new model .

1.2.4. Updating models

Red Hat Enterprise Linux AI version 1.1 allows you to update your local models to the most recent version of the model. For more information on upgrading the models, see Updating a model.

1.3. TECHNOLOGY PREVIEW FEATURE STATUS

1.3.1. Installation technology preview feature tracker

Table 1.1. Installation features

Feature	1.1
Installing on bare metal	Generally available
Installing on AWS	Generally available
Installing on IBM Cloud	Generally available

1.3.2. Large Language Models (LLMs) technology preview status

Table 1.2. LLM features

Feature	1.1
granite-7b-starter	Generally available
granite-7b-redhat-lab	Generally available
granite-8b-code-instruct	Technology preview
granite-8b-code-base	Technology preview
mixtral-8x7B-instruct-v0-1	Generally available
prometheus-8x7b-v2.0	Generally available

1.4. KNOWN ISSUES

1.4.1. Kdump over nfs

Red Hat Enterprise Linux AI version 1.1 does not support kdump over nfs out of the box. To use this feature, run the following commands:

```
mkdir -p /var/lib/kdump/dracut.conf.d
echo "dracutmodules="" > /var/lib/kdump/dracut.conf.d/99-kdump.conf
```

```
echo "omit_dracutmodules="" >> /var/lib/kdump/dracut.conf.d/99-kdump.conf
echo "dracut_args --confdir /var/lib/kdump/dracut.conf.d --install /usr/lib/passwd --install
/usr/lib/group" >> /etc/kdump.conf
systemctl restart kdump
```