Red Hat Enterprise Linux AI 1.5

发行注记

Red Hat Enterprise Linux AI 发行注记

Last Updated: 2025-09-24

Red Hat Enterprise Linux AI 1.5 发行注记

Red Hat Enterprise Linux AI 发行注记

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 ® is the registered trademark of Linus Torvalds in the United States and other countries.

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

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

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

Node.js ® 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 ® 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

本文档提供了 Red Hat Enterprise Linux AI 版本 1.5 发行注记

Table of Contents

第 1 章 RED HAT ENTERPRISE LINUX AI 1.5 发行注记

RHEL AI 为组织提供了在开源大型语言模型(LLM)上开发企业应用程序的过程。

1.1. 关于此版本

Red Hat Enterprise Linux AI 版本 1.5 包括 Large Language Model (LLM)对红帽和 IBM 生成的 Granite 模型进行微调的各种功能。使用 RHEL AI 工作流的自定义模型包括：

- 使用 InstructLab 工具安装和启动 RHEL 9.4 实例。
- Git 存储库中的主机信息，并与希望了解模型的基于 Git 的税务知识进行交互。
- 运行复合数据生成(SDG)、多阶段培训以及基准测试评估的端到端工作流。
- 通过新调优的 LLM 提供和聊天。

1.2. 功能和增强功能

Red Hat Enterprise Linux AI 版本 1.5 包括大型语言模型(LLM)微调的各种功能。

1.2.1. Suppored 加速器

1.2.1.1. NVIDIA H200 加速器

现在，您可以在 RHEL AI 版本 1.5 上使用 NVIDIA H200 加速器来推测服务并运行完整的端到端工作流。在初始化 RHEL AI 环境时，请选择与机器中加速器数匹配的 H200 配置集。有关 RHEL AI 支持的硬件的更多信息，请参阅 Red Hat Enterprise Linux AI 硬件要求 。

1.2.1.2. NVIDIA Grace Hopper GH200 加速器（技术预览）

现在，您可以在 RHEL AI 版本 1.5 上使用 NVIDIA H200 加速器作为技术预览。RHEL AI 默认不包含 Grace Hopper 加速器的系统配置文件。要使用 GH200 加速器，请使用 **h200_x1** 配置集初始化 RHEL AI 环境，并将 **max_startup_attempts: 1200** 参数添加到 **config.yaml** 文件中。

```
$ ilab config edit
```

```
serve:
vllm:
    gpus: 1
    llm_family: ''
    max_startup_attempts: 1200
    vllm_args: ["--tensor-parallel-size", "1"]
```

1.2.1.3. AMD MI300X 加速器

AMD MI300X 加速器现在可用于推测服务并运行完整的端到端工作流。有关 RHEL AI 支持的硬件的更多信息，请参阅 Red Hat Enterprise Linux AI 硬件要求 。

1.2.2. 安装

Red Hat Enterprise Linux AI 可作为可引导镜像安装。此镜像包含各种与 RHEL AI 交互的工具。该镜像包括：用于建模微调的 Red Hat Enterprise Linux 9.4、Python 版本 3.11 和 InstructLab 工具。有关安装 Red Hat Enterprise Linux AI 的更多信息，请参阅安装概述 和"安装功能跟踪器"

1.2.3. 构建 RHEL AI 环境

安装 Red Hat Enterprise Linux AI 后，您可以使用 InstructLab 工具设置 RHEL AI 环境。

1.2.3.1. 初始化 InstructLab

您可以通过运行 **ilab config init** 命令初始化和设置 RHEL AI 环境。这个命令创建与 RHEL AI 和微调模型交互所需的配置。它还为您的数据文件创建正确的目录。有关初始化 InstructLab 的更多信息，请参阅 Initialize InstructLab 文档。

1.2.3.2. 下载大型语言模型

您可以将红帽提供的各种大型语言模型(LLM)下载到 RHEL AI 机器或实例。您可以在创建并登录到红帽 registry 帐户后，从红帽注册中心下载这些模型。有关支持的 RHEL AI LLMs 的更多信息，请参阅 下载模型文档和 "大语言模型(LLMs)技术预览状态"。

1.2.3.2.1. 3.1 Granite 模型的版本 2

RHEL AI 版本 1.5 现在支持 **granite-3.1-8b-starter-v2** student 模型和 **granite-3.1-8b-lab-v2** inference 模型。有关模型的更多信息，请参阅 下载大型模型 文档。

1.2.3.3. 使用模型提供和聊天

Red Hat Enterprise Linux AI 版本 1.5 允许您在各种 LLM 上运行 vLLM inference 服务器。vLLM 工具是 RHEL AI 镜像中包含的 LLM 的内存效率推测和服务引擎库。有关使用模型提供服务和聊天的更多信息，请参阅 使用 模型文档 的 Serving 和聊天。

1.2.4. 创建技能和知识 YAML 文件

在 Red Hat Enterprise Linux AI 上，您可以使用自定义 YAML 文件自定义税务树，以便模型可以了解特定域的信息。您在 Git 存储库中托管您的知识数据，并使用这些数据微调模型。有关如何创建知识标记和 YAML 文件的详细信息，请参阅自定义税务树。

1.2.5. 使用 RHEL AI 生成自定义 LLM

您可以使用 Red Hat Enterprise Linux AI 来根据您的域特定知识和知识自定义大量入门 LLM。RHEL AI 包括 Synthetic Data Generation (SDG)和多阶段培训的增强方法。

1.2.5.1. 合成数据生成(SDG)

Red Hat Enterprise Linux AI 包括同分析数据生成(SDG)的 LAB 增强方法。您可以将 **qna.yaml** 文件与自己的知识数据一起使用，在 SDG 进程中创建数百个人工数据集。有关运行 SDG 进程的更多信息，请参阅使用 Synthetic data generation (SDG)生成新数据集。

1.2.5.1.1. 使用 **llama-3.3-70B-Instruct** 模型作为手模型运行 SDG （技术预览）

现在，当运行 Synthetic Data Generation (SDG)作为技术预览时，RHEL AI 版本 1.5 支持将 **llama-3.3-70b-Instruct** 用作老板模型。如需更多信息，请参阅使用 llama-3.3-70B-Instruct 模型作为手模型（技术预览） 文档。

1.2.5.2. 使用您的数据培训模型

Red Hat Enterprise Linux AI 包括多阶段培训的 LAB 增强方法：一个微调策略，其中在多个阶段被接受并评估，以创建最佳模型。有关多阶段培训的详情，请参阅 培训模型上的数据。

1.2.5.3. 基准评估

Red Hat Enterprise Linux AI 包括对新受培训模型运行基准评估的能力。在受培训的模式中，您可以评估模型知道通过 **MMLU_BRANCH** 或 **MT_BENCH_BRANCH** 基准添加的知识或技能的程度。有关基准评估的详情，请参阅 评估您的新模型。

1.2.6. 红帽跨产品功能

1.2.6.1. 使用 Ansible Automation Platform 自动化 RHEL AI

现在，您可以使用 Ansible Automation Platform hub 在 playbook 中运行 RHEL AI 工作负载。这包括两个 Ansible 集合：

infra.ai

可以在各种云供应商基础架构上置备 RHEL AI 环境的内容集合，包括 AWS、GCP 和 Azure。这个集合简化了不同云供应商中的 AI 工作负载部署。

redhat.ai

用于在 RHEL AI 中管理工作负载的内容集合。您可以使用 Ansible playbook 选项快速创建 RHEL AI 中的部署，这可以更有效地进行模型培训和认证。

如果您是现有的 Ansible Automation Platform 客户，则这些集合会包含在您当前的订阅中。

1.3. RED HAT ENTERPRISE LINUX AI 功能跟踪器

这个版本中的一些功能当前还处于技术预览状态。它们并不适用于在生产环境中使用。有关红帽技术预览功能支持范围的更多信息，请参阅技术预览功能支持范围。

在以下表格中，功能被标记为以下状态：

- 不可用
- 技术预览
- 公开发行
- Deprecated
- 删除

1.3.1. 安装功能跟踪器

表 1.1. 安装功能

功能	1.1	1.2	1.3	1.4	1.5
在裸机上安装	正式发布	正式发布	正式发布	正式发布	正式发布

功能	1.1	1.2	1.3	1.4	1.5
在 AWS 上安装	正式发布	正式发布	正式发布	正式发布	正式发布
在 IBM Cloud 上安装	正式发布	正式发布	正式发布	正式发布	正式发布
在 GCP 上安装	不可用	技术预览	正式发布	正式发布	正式发布
在 Azure 上安装	不可用	正式发布	正式发布	正式发布	正式发布

1.3.2. 平台支持功能跟踪器

表 1.2. 端到端 InstructLab 工作流

功能	1.1	1.2	1.3	1.4	1.5
裸机	正式发布	正式发布	正式发布	正式发布	正式发布
AWS	正式发布	正式发布	正式发布	正式发布	正式发布
IBM Cloud	不可用	正式发布	正式发布	正式发布	正式发布
Google Cloud Platform	不可用	技术预览	正式发布	正式发布	正式发布
Azure	不可用	正式发布	正式发布	正式发布	正式发布

表 1.3. inference service LLMs

功能	1.1	1.2	1.3	1.4	1.5
裸机	正式发布	正式发布	正式发布	正式发布	正式发布
AWS	正式发布	正式发布	正式发布	正式发布	正式发布
IBM Cloud	正式发布	正式发布	正式发布	正式发布	正式发布
Google Cloud Platform (GCP)	不可用	技术预览	正式发布	正式发布	正式发布
Azure	不可用	正式发布	正式发布	正式发布	正式发布

表 1.4. Cloud Marketplace 支持

功能	1.1	1.2	1.3	1.4	1.5
AWS	不可用	不可用	正式发布	正式发布	正式发布

功能	1.1	1.2	1.3	1.4	1.5
Azure	不可用	不可用	正式发布	正式发布	正式发布

1.4. 大型语言模型跟踪器

1.4.1. RHEL AI 版本 1.5 硬件供应商 LLM 支持

表 1.5. 对硬件供应商的支持

功能	NVIDIA	AMD	Intel
granite-3.1-8b-starter-v2.1	正式发布	正式发布	不可用
granite-3.1-8b-lab-v2.1	正式发布	正式发布	不可用
granite-3.1-8b-starter-v2	不可用	不可用	技术预览
granite-3.1-8b-lab-v2	不可用	不可用	技术预览
granite-8b-code-instruct	技术预览	技术预览	技术预览
granite-8b-code-base	技术预览	技术预览	技术预览
mixtral-8x7B-instruct-v0-1	正式发布	正式发布	技术预览
llama-3.3-70b-Instruct	技术预览	技术预览	不可用
prometheus-8x7b-v2.0	正式发布	正式发布	不可用

1.5. 已知问题

运行 MMLU 评估

在 RHEL AI 版本 1.5 中，在运行 MMLU 时，您需要使用 the **-skip-server** 标志。

某些 NVIDIA A100 系统上的自动检测不正确

RHEL AI 有时使用 A100 加速器自动检测机器上不正确的系统配置文件。

您可以通过重新初始化并传递正确的系统配置文件来选择正确的配置集。

```
$ ilab config init --profile <path-to-system-profile>
```

fabric Manager 并不总是以 NVIDIA 加速器开头

在 NVIDIA 系统上安装 Red Hat Enterprise Linux AI 后，您可能会在提供模型或培训模型时看到以下错误。

```
INFO 2024-11-26 22:18:04,244 instructlab.model.serve_backend:56: Using model '/var/home/cloud-user/.local/share/instructlab/phased/phase2/checkpoints/hf_format/samples_29117' with -1 gpu-layers and 4096 max context size.
INFO 2024-11-26 22:18:04,244 instructlab.model.serve_backend:88: '--gpus' flag used alongside '--tensor-parallel-size' in the vllm_args section of the config file. Using value of the --gpus

File "/opt/app-root/lib64/python3.11/site-packages/vllm/entrypoints/openai/api_server.py", line 105, in build_async_engine_client
    async with build_async_engine_client_from_engine_args(
  File "/usr/lib64/python3.11/contextlib.py", line 210, in __aenter__
    return await anext(self.gen)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/opt/app-root/lib64/python3.11/site-packages/vllm/entrypoints/openai/api_server.py", line 192, in build_async_engine_client_from_engine_args
    raise RuntimeError(
RuntimeError: Engine process failed to start
```

要解决这个问题，您需要运行以下命令：

```
$ sudo systemctl stop nvidia-persistenced.service
$ sudo systemctl start nvidia-fabricmanager.service
$ sudo systemctl start nvidia-persistenced.service
```

UI AMD 技术预览安装

Red Hat Enterprise Linux AI 版本 1.5 目前不支持使用技术预览 AMD ISO 基于图形安装。确保为非互动安装配置了 **kickstart** 文件中的 **text** 参数。您还可以在互动安装过程中在 shell 中传递 **inst.text**，以避免安装时间崩溃。

SDG 可以在 4xL40s 上失败

要使 SDG 在 4xL40s 上运行，您需要使用 **--num-cpus 标志运行** SDG，并设置为 **4** 的值。

```
$ ilab data generate --num-cpus 4
```

granite-8b-starter-v1 模型上的 MMLU 和 MMLU_BRANCH

当评估从 **granite-8b-starter-v1** LLM 构建的模型时，在运行 MMLU 和 MMLU_BRANCH 基准时，vLLM 可能无法启动。

如果 vLLM 没有启动，请在 **config.yaml** 文件的 **serve** 部分中添加以下参数：

```
serve:
  vllm:
    vllm_args: [--dtype bfloat16]
```

kdump over nfs

Red Hat Enterprise Linux AI 版本 1.5 不支持在没有配置的情况下通过 nfs 进行 kdump。要使用这个功能，请运行以下命令：

```
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
```

1.6. 异步 Z-STREAM 更新

RHEL AI 1.5 的安全更新、程序漏洞修正、功能增强更新将会作为异步 z-stream 更新发布。

本节的内容将会持续更新，以提供以后发行的 RHEL AI 1.5 的 z-stream 版本改进和程序错误修复。异步子版本（例如，RHEL AI 1.5.z）的具体信息会包括在相应的子章节中。

1.6.1. Red Hat Enterprise Linux AI 1.5.1 特性和程序错误修复

发布日期： 2025 年 6 月 11 日

Red Hat Enterprise Linux AI 版本 1.5.1 现已正式发布。此发行版本包括程序错误修正和产品改进。

1.6.1.1. 功能

RHEL AI 1.5.1 及后续的 1.5.z 版本支持 Intel Gaudi 3 加速器用于推测服务模型。您可以在 Download Red Hat Enterprise Linux AI 页面下载 Red Hat Enterprise Linux AI 镜像，并在带有 Gaudi3 加速器的机器上部署 RHEL AI。

1.6.1.2. Upgrade（升级）

要将 RHEL AI 系统更新至最新的 z-stream 版本，您必须登录到 Red Hat registry 并运行以下命令：

```
$ sudo bootc upgrade --apply
```

1.6.2. Red Hat Enterprise Linux AI 1.5.2 功能和程序错误修复

发布日期： 2025 年 6 月 24 日

Red Hat Enterprise Linux AI 版本 1.5.2 现已正式发布。此发行版本包括程序错误修正和产品改进。

1.6.2.1. Upgrade（升级）

要将 RHEL AI 系统更新至最新的 z-stream 版本，您必须登录到 Red Hat registry 并运行以下命令：

```
$ sudo bootc upgrade --apply
```