Azure AI Foundry 上部署 DeepSeek R1

原创 木字头 RockyLinux 2025年02月04日 12:11 广东

注意：Azure AI Foundry 目前提供的 DeepSeek R1 为 671B 满血版，活动参数 37B，并且免费。但由于资源池限制，所以响应速度较慢。

DeepSeek 简介

深度求索人工智能基础技术研究有限公司(简称“深度求索”或“DeepSeek”)，成立于2023年，是一家专注于实现 AGI 的中国公司。

GitHub 仓库地址：DeepSeek · GitHub
DeepSeek 官网：DeepSeek | 深度求索

维基百科简介：深度求索 – 维基百科，自由的百科全书

> 深度求索（英语：DeepSeek），全称杭州深度求索人工智能基础技术研究有限公司，是中华人民共和国的一家人工智能与大型语言模型公司。该公司的总部位于中国大陆浙江省杭州市，由中资对冲基金幻方量化创立，创始人和首席执行官为梁文锋。

Azure AI Foundry

在 Azure AI Foundry 部署 DeepSeek 之前，先简单了解一下 Azure AI Foundry。Azure AI Foundry 是一个综合性的平台，用于构建、训练和部署机器学习和人工智能模型。在这个平台中 Hub、Project 和 Models 有着紧密的关联关系，这些组件共同构成了一个完整的工作流程。

1. **Hub（中心）**：
   - Hub 是 Azure AI Foundry 的核心组件，是整个系统的中央管理点。它负责管理所有的资源和配置，包括：数据集、模型、项目、部署环境等。
   - 在 Hub 中，您可以查看和管理所有的 Project 和 Models，并且可以进行权限管理和资源分配。
2. **Project（项目）**：
   - Project 是在 Hub 中为了实现特定业务需求而创建的具体工作单元。每个 Project 包含了为实现该项目目标所需的全部内容，如：数据集、模型训练脚本、配置文件和其他相关资源。

- Project 是实际进行数据处理、模型开发和实验管理的地方。它提供一个框架，确保这些过程的有序和可重复性。

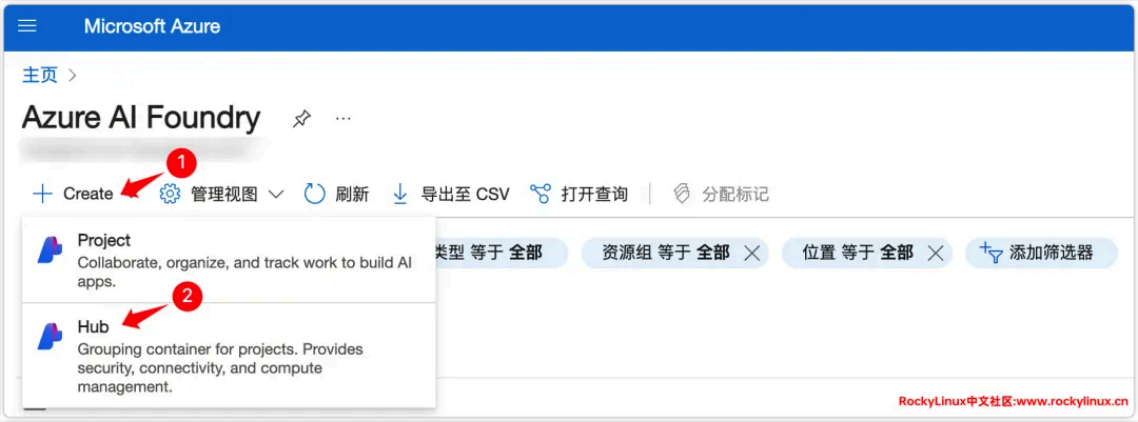
3. Models（模型）：

- 模型是根据 Project 中的数据和特定的机器学习算法训练出来的，可用于进行预测或分类等任务。
- 在模型开发完成后，它会部署到指定的环境中进行测试和使用。Hub 管理这些模型的版本和部署状态，以确保模型能够被正确调用和监控。

因此，在 Azure AI Foundry 中，您需要先在 Hub 里创建和管理 Project，在每个 Project 中进行模型的开发和训练，最后通过 Hub 来进行模型部署和管理，以确保整个流程的顺畅和高效。

创建 Hub

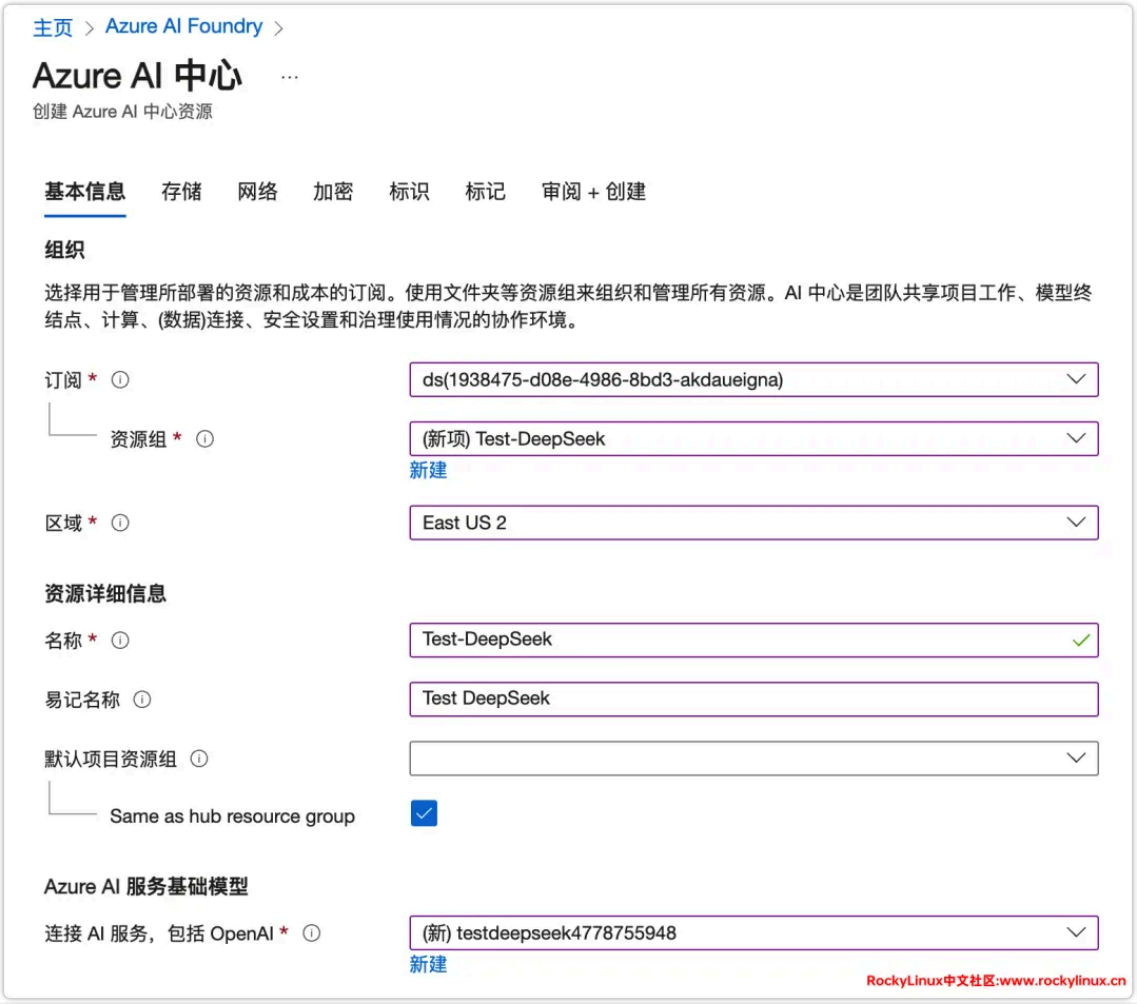
打开 Azure AI Foundry 链接，点击“+ Create” -- “Hub”。



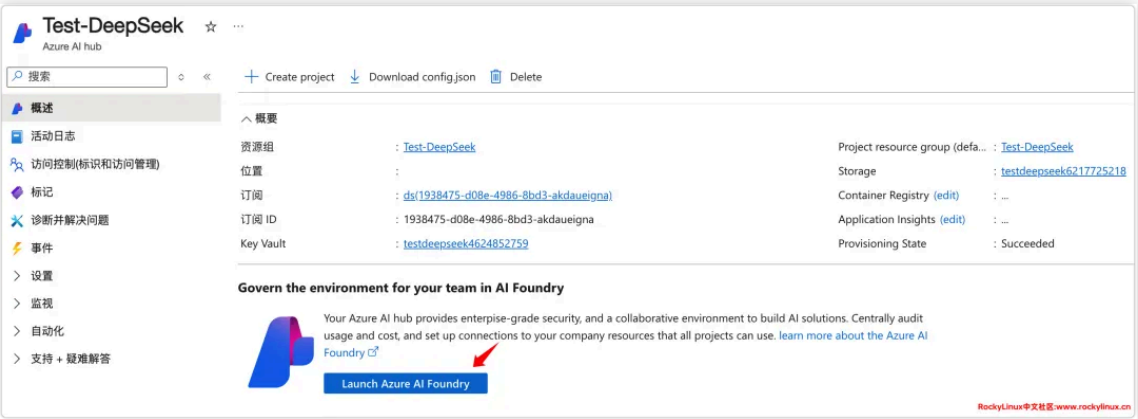
注意：区域的选择，确保对应区域存在资源，可用区域参考：Region availability for models in Serverless API endpoints – Azure AI Foundry | Microsoft Learn

目前 DeepSeek 只在以下几个区域提供：

- East US
- East US 2
- North Central US
- South Central US
- West US
- West US 3

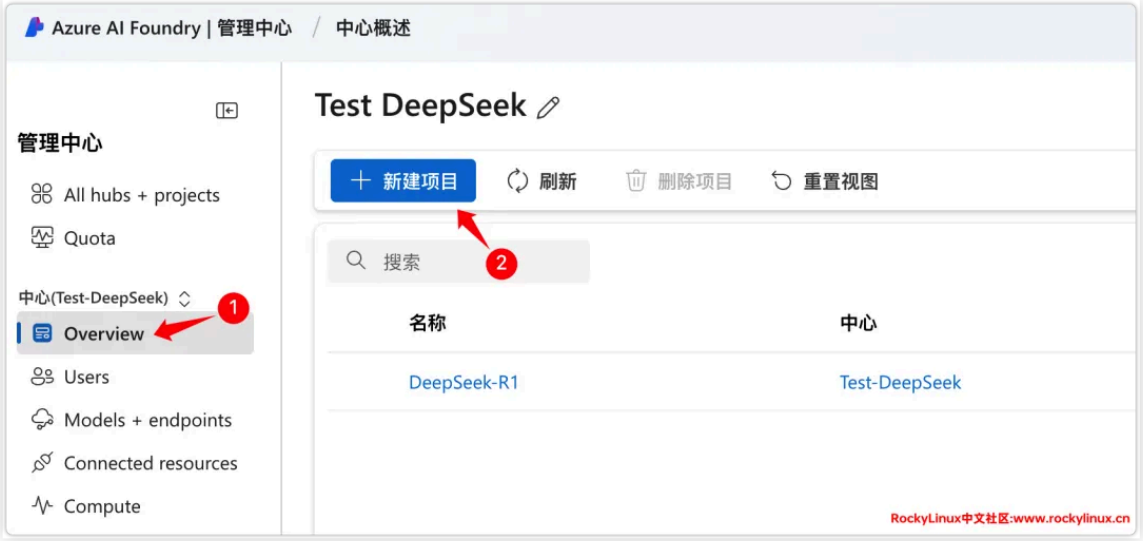


部署完 Azure AI Hub 后，点击“Launch Azure AI Foundry”。



创建 Project

点击“Overview”--“新建项目”。

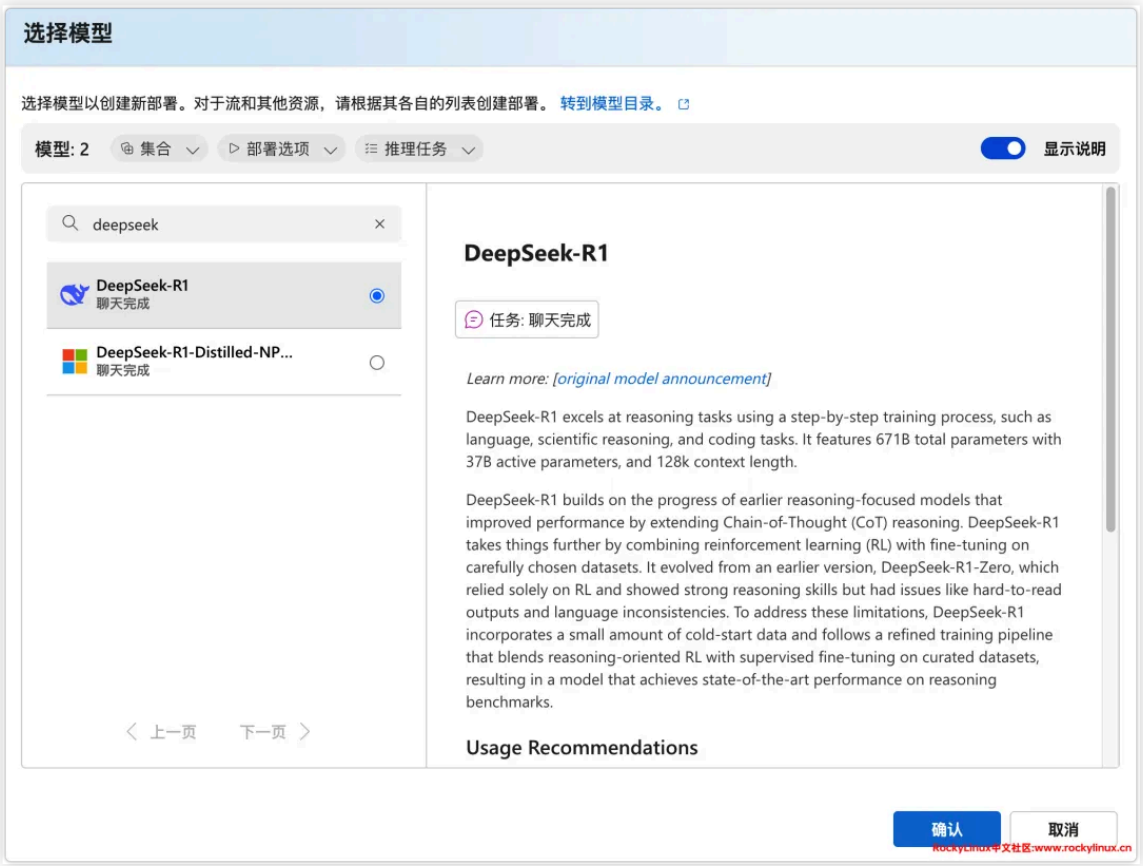


输入项目名称，点击“创建”。

创建 Models

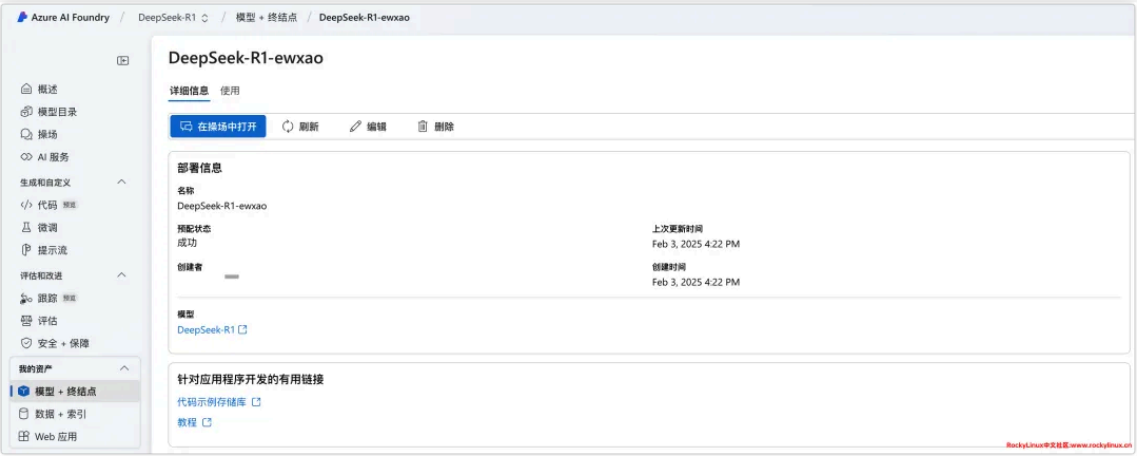
点击“模型+终结点”--“部署模型”--“部署基本模型”。

搜索“deepseek”，点击“确认”。



选择项目：“deepseek-r1”，部署名称默认自动生成，可以根据需要进行修改，然后点击“部署”。

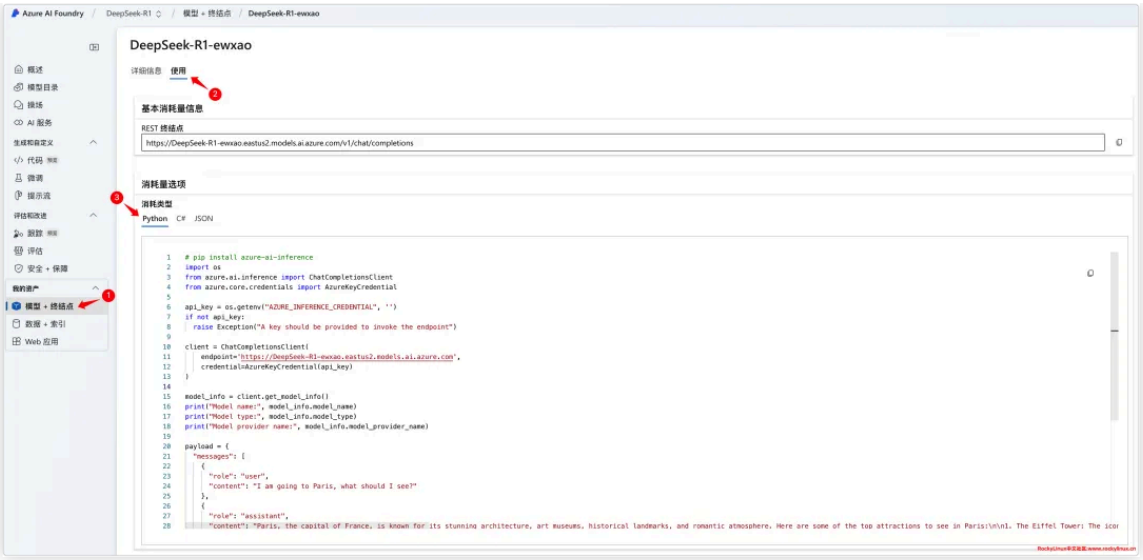
部署完成以后，点击“在操场中打开”，开启聊天会话。



开始愉快的聊天。

API 调用

API 调用可以使用如下图所示演示代码，也可以使用木子下方提供的测试代码。



chat.py 文件：

```
import os
from azure.ai.inference import ChatCompletionsClient
from azure.core.credentials import AzureKeyCredential
from dotenv import load_dotenv
from azure.ai.inference.models import SystemMessage, UserMessage

load_dotenv()

AZURE_ENDPOINT = os.getenv("AZURE_ENDPOINT")
AZURE_KEY = os.getenv("AZURE_KEY")

client = ChatCompletionsClient(
    endpoint=os.environ["AZURE_ENDPOINT"],
    credential=AzureKeyCredential(os.environ["AZURE_KEY"]),
)

response = client.complete(
    messages=[
        SystemMessage(content="你是一名运维工程师"),
        UserMessage(content="Linux 有哪些发行版？"),
    ],
)

print("Response:", response.choices[0].message.content)
```

.env 文件：

```
# 根据实际情况进行设置
AZURE_ENDPOINT=https://xxx.eastus2.models.ai.azure.com
```

```
AZURE_KEY=xxxx
```

requirements.txt 文件：

```
azure-core
azure-ai-inference
python-dotenv
```

测试结果：

```
❯ python ./chat.py
Response: <think>

</think>

Linux 有很多不同的发行版（Distribution），以下是常见的主要发行版及其分类：

---

### **基于 Debian 的发行版**
1. **Ubuntu**
   - 最流行的桌面发行版之一，适合新手。
   - 衍生版本：Kubuntu（KDE 桌面）、Lubuntu（轻量级）、Xubuntu（XFCE 桌面）、Ubuntu Server 等。
2. **Linux Mint**
   - 基于 Ubuntu，界面友好，适合从 Windows 转来的用户。
3. **Debian**
   - 以稳定著称，是 Ubuntu 的基础。
4. **Pop!_OS**
   - 由 System76 开发，专注于开发者和硬件兼容性。

---

### **基于 Red Hat 的发行版**
1. **Fedora**
   - 由社区支持，注重新技术，稳定性较好。
2. **CentOS**
   - 曾经是企业级的免费稳定版（基于 RHEL），现转向 CentOS Stream（滚动更新）。
3. **Red Hat Enterprise Linux (RHEL)**
   - 商业付费版本，主打企业级支持。
4. **Rocky Linux / AlmaLinux**
   - CentOS 替代品，旨在延续免费的企业级稳定系统。

---

### **基于 Arch Linux 的发行版**
```

1. **Arch Linux**
   - 滚动更新，高度可定制，适合高手。
2. **Manjaro**
   - 基于 Arch，但更用户友好，适合新手。
3. **EndeavourOS**
   - 简化 Arch 安装过程，保留灵活性。

独立/其他派系发行版

1. **openSUSE**
   - 分为 Leap（稳定版）和 Tumbleweed（滚动更新），企业级工具丰富。
2. **Gentoo**
   - 需从源码编译软件，高度优化，适合高级用户。
3. **Slackware**
   - 最古老的发行版之一，以简洁稳定著称。
4. **NixOS**
   - 强调配置化和可复现性，适合开发环境。

轻量级/专用发行版

1. **Alpine Linux**
   - 轻量级，适合容器和嵌入式系统。
2. **Raspberry Pi OS**
   - 专为树莓派设计的系统。
3. **Kali Linux**
   - 专注于渗透测试和安全审计。
4. **Tails**
   - 强调隐私和匿名上网的发行版。

其他特色发行版

- **Elementary OS**：模仿 macOS 设计。
- **Zorin OS**：针对 Windows 用户优化的界面。
- **Deepin**：国产发行版，界面美观。

选择发行版时，需考虑用途（桌面/服务器/嵌入式）、易用性、社区支持等因素。新手推荐 Ubuntu 或 Linux Mint，

总结

在 Azure AI Foundry 部署 DeepSeek 相对简单，但需要注意的是，这只是一个部署示范。根据 DeepSeek 的跑分数据显示，DeepSeek 在执行复杂推理任务时表现优异，其性能与 OpenAI-o1 模型相当，尤其在数学、编码和解决问题等领域表现突出。然而，对于

DeepSeek 在实际生产环境中的表现，还需要在具体的业务场景中进一步验证其效果和实用性。

注: 因为小编不太懂技术，所以后台留言，不一定会回复。如果对于文章有问题、疑问等，可以去网站对应文章后发表评论，会有更多热心朋友解答。

我们强烈建议您去 https://rockylinux.cn 阅读相关文档，更有利于系统性知识迭代，原因有三:

1. 因为微信公众号对于markdown扩展语法不支持，所以有一些文本会出现乱码等情况。
2. 微信公众号没办法及时修正错误和持续文档更新迭代。
3. 社区会有一些小伙伴原创投稿，我们可能不会在 RockyLinux 微信公众号发布。

目前 RockyLinux 微信公众号已经对接 Rocky Linux 中文社区官网，您可以在微信公众号对话框中输入任何您想搜索的内容关健字，比如: Redis，它将返回官网相关文章、话题、说说、页面中的关键字文档。

Rocky Linux 官网：https://rockylinux.org
Rocky Linux 中文社区官网：https://rockylinux.cn

微信公众号：RockyLinux
QQ群：626649599
微信群：加微信号 rockylinuxcn 备注: 入群
邮箱地址：muzi@vip.rockylinux.cn

AI 27 LLM 21

AI · 目录

上一篇 · DeepSeek 全面解析：全球各大公有云厂商价格一览表

阅读原文