

Cloud Studio (云端 IDE)







【版权声明】

©2013-2025 腾讯云版权所有

本文档(含所有文字、数据、图片等内容)完整的著作权归腾讯云计算(北京)有限责任公司单独所有,未经腾讯云 事先明确书面许可,任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成 对腾讯云著作权的侵犯,腾讯云将依法采取措施追究法律责任。

【商标声明】

🔗 腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算(北京)有限责任公司及其关联公司所有。本文档涉及的第三方主体的 商标,依法由权利人所有。未经腾讯云及有关权利人书面许可,任何主体不得以任何方式对前述商标进行使用、复 制、修改、传播、抄录等行为,否则将构成对腾讯云及有关权利人商标权的侵犯,腾讯云将依法采取措施追究法律责 任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况,部分产品、服务的内容可能不时有所调整。 您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定,除非双方另有约定,否则, 腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务,及相应的技术售后服务,任何问题请联系 4009100100或 95716。





实践教程

DeepSeek 使用指南



实践教程 DeepSeek 使用指南

最近更新时间: 2025-02-10 15:10:33

Cloud Studio 已内置 DeepSeek-R11.5B、7B、14B、32B 模型,并支持一键部署。

一、快速创建

1. 进入 Cloud Studio,在空间模板下单击 DeepSeek-R1 模板。



2. 根据需求选择不同规格(DeepSeek体验、免费基础型、HAI进阶型)来创建。





规格对比如下:

| 名称 | DeepSeek 体验规格 | 基础型 | HAI进阶型 | |
|------|--|---|--|--|
| 预置环境 | 预装 Ollama。不同 规格分别预置 DeepSeek 1.5B、 7B、8B、14B、 32B模型。 Chatbot-Ollama | 预装 Ollama、DeepSeek-R1 7B、14B、32B 植型。 支持 Open WebUI。 | | |
| 规格 | CPU: 8~64核不等 内存: 16~128GB不等 | 显存: 16GB + 算力: 8 + TFlops SP CPU: 8 核 内存: 32GB | 显存: 32GB + 算力: 15 + TFlops SP CPU: 8~10 核 内存: 32GB 选择后仍有两卡基础型、两卡 进阶型可供选择,具体参见 购 买页 提示。 | |
| 适用场景 | 适用于体验 DeepSeek 蒸馏模型推理。打开即可 访问可视化对话界面,上 手成本低。 | 16G显存,可用于体验 DeepSeek 7B,最大可 到32B但推理速度较慢。 | 32G显存,可用于体验 Deepseek 14B、32B。支 持模型部署。 | |



| 使用费用 | 每月可免费兑换使用时长 从12500分钟到1562分 钟不等。具体费用可参考 <mark>购买指南</mark> 。 | 默认使用每月10000分钟 高性能工作空间免费时 长,与通用工作空间免费 时长分别计算互不影响。 暂不支持付费增加时长。 | 该规格无法使用每月10000分 钟高性能工作空间免费时长, 需跳转到 HAI 购买,具体费 用可参见 <mark>计费说明</mark> 。 |
|------|---|--|---|
|------|---|--|---|

() 说明:

基础型资源有限且为动态库存,如出现库存不足提示,可稍后重试或选用其他规格。

二、使用轻量型进行对话与临时部署

轻量型已为您预装并启动了以下服务:

- Ollama 服务: 支持通过 API 调用 DeepSeek 模型。
- Chatbot-Ollama 前端服务:提供交互式聊天界面。
- 预装模型:
 - DeepSeek-R1-Distill-Qwen-1.5B: 适合轻量级任务,如简单文本生成、基础问答和低资源设备上的 应用。
 - DeepSeek-R1-Distill-Qwen-7B: 适合中等复杂度任务,如文本摘要、翻译、对话生成和内容创作。
 - DeepSeek-R1-Distill-Llama-8B: 与7B类似,性能略有提升,适合更复杂的文本生成和理解任务。
 - DeepSeek-R1-Distill-Qwen-14B:适合高复杂度任务,如长文本生成、深度问答、知识推理和高级 对话系统。
 - DeepSeek-R1-Distill-Qwen-32B: 适合高精度任务,如复杂推理、大规模知识库问答、专业领域内 容生成和研究级应用。

使用方法1:通过 Chat UI 快速体验

我们提供了一个开箱即用的聊天界面,您可以直接通过浏览器与模型交互。等待页面加载完成即可查看到下图中的界 面。单击红框区域,即可新页面打开聊天页面。





操作步骤

1. 访问 Chat UI

当前环境已配置为默认启动该服务,可直接访问。如果希望修改启动命令,可以前往 .vscode/preview.yml 修改。

2. 选择模型

在界面右上角的下拉菜单中,选择 deepseek-r1:1.5b 或 deepseek-r1:7b 。

▲ 注意:

加载7B模型需要更多内存,请先切换到旗舰型后,再选择7B模型。

3. 开始对话

输入问题或指令,单击发送即可获取模型回复。

4. 在其他设备打开

您可以复制链接地址后在其他设备打开。也可 hover 下方红框处并扫码,即可在移动端打开。





示例: 生成代码

输入:



预期输出:



如果遇到网页展示有问题,可返回编辑器终端,查看报错信息并询问 AI 代码助手。





使用方法2:通过命令行调用模型

您可以使用 curl 直接与 Ollama 服务交互,适用于自动化任务或脚本调用。

基础请求示例

调用 deepseek-r1:1.5b



调用 deepseek-r1:7b







} '

流式输出(实时逐句返回)

将 stream 参数设为 true:



使用方法3: Ollama 原生命令进阶使用

除了 HTTP API, 您还可以直接通过 ollama 命令行工具与模型交互。

1. 查看已安装模型

输入如下:

ollama list

输出示例:

NAME SIZE deepseek-r1:1.5b 1.5B deepseek-r1:7b 7B

2. 启动交互式对话

启动输入如下:

ollama run deepseek-r1:1.5b

输入 /bye 或者 Ctrl + D 退出对话。 示例对话流程:

>>> 推荐一个适合初学者的机器学习项目





3. 模型管理

| 命令 | 说明 |
|----------------------------|---------------|
| ollama pull deepseek-r1:7b | 重新拉取模型(更新时使用) |
| ollama rm deepseek-r1:7b | 删除模型 |

4. 切换模型

如需切换到7B模型,可在退出当前进程后再次输入以下代码即可:

ollama run deepseek-r1:7b

<u>小 注意:</u>

当前机器为 CPU,7B模型推理速度较慢,建议使用1.5B模型。更大参数模型建议使用 HAI进阶型。

常见问题

如何查看预览?

如需了解更多使用内置 web 预览的介绍,可参考 在线预览调试 文档。

如何释放端口,关闭某个进程?

如果需要释放端口,关闭某个进程,可以通过下述命令查看端口占用情况:

lsof -i -P -n

再使用下述命令关闭进程:

kill -9 **<进程**ID>

空间不足怎么处理?



如果右下角提示空间不足,可单击右下角的"2核4GB",选择"旗舰型",即可升级空间。



相关文档与资源

- Deepseek 官方介绍文档: GitHub deepseek-ai/DeepSeek-R1
- Ollama 官方介绍文档: GitHub ollama/ollama

更多帮助和支持

欢迎加入 Cloud Studio 用户反馈群。当您遇到问题需要处理时,您可以直接通过扫码进入 Cloud Studio 用户群 进行提问。

- 腾讯云工程师实时群内答疑。
- 扫码入群可先享受产品上新功能。
- 更多精彩活动群内优享。



三、使用基础型进行对话

高性能工作空间开机自启动 Ollama 服务,可单击新建终端。



| S | 文件 编辑 选择 查看 转到 运行 | 终端 帮助 \leftarrow \rightarrow | , C workspace □ □ □ □ □ □ |
|---------------|---------------------|--|---|
| ſЪ | 资源管理器 | 新建终端 へん、 | |
| LC- | V WORKSPACE | 拆分终端 郑\ | |
| \mathcal{Q} | | 运行任务 | |
| وم | | 运行生成任务 企 器B | |
| | | 运行活动文件 | |
| æ | | | |
| ₽₽ | | | |
| | | | |
| ١Ļ | | 配置任务 | |
| | | 配置默认生成任务 | |
| | | ○(base) root@VM-0-80-ubu >>> 推理是什么意思 | ntu:/workspace# ollama run deepseek-r1:1.5b |
| | | <tn1nk> 好的,我现在需要解释"推</tn1nk> | |
| | | 首先,推理通常指的是基 法、类比推理和演绎推理。 | F已有知识或信息做出某种推断或结论的过程。它不仅仅是简单的逻辑推理,还可能涉及到多种思维模式。我记得之前提到过不同的推理方式,比如归纳 这些方法在科学、哲学和日常生活中都应用广泛。 |
| | | 让我先回忆一下不同的人 题 于解决问题或生成新观点。 | & 如何理解 "推理"的。有人可能会认为推理是基于已有知识进行的逻辑推导,这与数学中的证明相似。而另一些人则认为它是一种创造性的思维活动,用 例如,在科学实验中,当有新的证据出现时,科学家会进行推断以验证之前的假设。 |
| | | 我还需要考虑推理在不同∜ 杂性和多样性。 | 顽域的具体应用。比如,法律推理通常用于论证案件的成立性,而哲学推理则涉及对存在、知识和真理的理解。这种灵活性可能意味着推理在实际中的复 |
| | | 接下来,我可以思考一些扑 推理则是通过已知命题和5 | 韭理的具体例子。比如,在日常生活中,我们可能会通过观察、实验或讨论来推断出结论。数学中的代数推理就是通过方程的变形得出新的结论,而逻辑 则则进行逐步推导。 |
| | | 另外,推理在艺术中也很重 | il要。艺术家们可能通过观察情感或材料的特征进行创作,这属于形象推理。同时,艺术家也会运用抽象逻辑来表达情感或思想,形成独特的作品。 |
| (8) | | 我还应该考虑推理与思维相 可能带来更全面、更准确的 | 模式的关系。不同的推理方式意味着有多种思维路径可供选择。比如,有些人喜欢用归纳法,而另一些人则更擅长类比推理和演绎推理。这些方法的结合 9结果。 |
| £633 | > 大纲 > 时间线 | 我还需要思考为什么推理如 们评估各种可能性,并做出 | 山此重要。在许多情况下,推理是解决问题的关键。例如,在工程设计中,测试不同的材料或结构方案时需要进行推理。在决策过程中,推理可以帮助我 出明智的选择。 |
| က Clo | ud Studio 🛛 🛞 0 🖄 0 | | 布局: U.S. 😒 🗘 |

输入以下命令查看已下载的模型并启动:

```
ollama list # 查看当前已经下载的模型,当前工作空间已预装 DeepSeek-R1 1.5B、7B、
8B、14B、32B 五种模型。
ollama run [NAME] # 运行想要启动的模型,NAME是模型的名称,如 ollama run
deepseek-r1:32b。
```

常见问题

其他常见问题可参考 常见问题。

四、使用进阶型进行对话与部署

1. 购买算力

选择 HAI进阶型后单击下一步。如遇到授权提示,根据提示信息完成授权即可。

△ 注意:

当前不支持协作者账号授权访问,请使用主账号或子账号。



| 反 AI 模板 前端模板 语言模板 | | | | | | |
|---|--|--|--|---|--|--|
| | | 选择空间规格 | | | | |
| · · · · · · · · · · · · · · · · · · · | FTML ಸುಪನ್ ≁ HTM | 轻量型 适用于体验 DeepSeek 兩個總型,就 认使用應用工作空间免费时长 CPU:8 核 | 基础理 | HAI 进行型 ① 高性能,适用于推进、训练活景,购 实后使用 显存:3268 + | | Jupyter Notebook Jupyter Notebook是実態は教徒 |
| | | 内存: 1668 | 第力: 8 + TFlops SP CPU: 8核 内存: 32GB | 第力: 15 + TFlops SP CPU: 8~10核 内存: 32G8 | | |
| | | | 线至HAI产品中心进行购买,购买先成后即可方 h、Tensorflow、ChatGLM3、Llama3等,后线 | 开始使用。 就被逐步支持更多模版,数道期待。 | | |
| | Tenso Tensorf Tensorf Tensorf CUDA | 下一步 取消 | | OF CLAUDING 60 CUDA Pytorch Jupyteria | 开設的 LLM 現型。支持工具範囲(Function Carl)、代語時… b | Liama3 88 Instruct Uama3 88 Instruct 小 Liama3 88 Instruct 第言部型 Liama3 |
| <u>東日田</u> | 混元 Di えた.ortst 少 混元:ortst | it (高佐施) 国際和目研め高性総領社度中文理解多分構成す 飲行。 知 | insformer機型,提供双演生成能力,J <mark>选</mark> 力 | | | Llama3.2 - RAG SFT练》 基于Llama-3.2-18-Instruct,内 少 llama32-ag-sft |
| unda Uma | | · 由于教室 我元 | | AI应用 语言语型 DeepSe | | AI应用。语言规型 Liama |
| 是一个用于构建高效。可扩展的 Nodels 服务器制应用的框架。它使用新进式 InviSc a (Net Lio | NuxtJS NuxtJs 是 ∳ nuxtis | 一个基于Vue,is 的轻量级应用框架,可用来创建服务i demo (Nuat/S) | ••• 靖道染(SSR) 应用,也可充当静态站… | Vue-TypeScript 描記 TypeScript 使用 Vue,可以 小 Vue-TypeScript (Vue-TypeScript | 在論評封書记錄去分析检測出很多常用错問。这成少了生产环境中… iof | VueJS Vuejs 是一个用于构建 JavaScri 小 VueJS (VueJS) |

根据个人意向配置选购方案,填写实例名称,阅读并勾选协议,立即购买后,自动跳转到 Cloud Studio 创建工作 空间。如账户余额为负请及时续费。

| 计费模式 | 按量计费 运用需求重有大幅波动的场景 | 包年包月 运用需求量长期稳定 | 約业务 | | | |
|--------|----------------------------------|-------------------------------------|---|-------------------------------------|---------------|------|
| 地域 | 北京 硅谷 | 南京 更多 | | | | |
| | 此地域支持学术加速能力,开启后可提高部分 | 分学术资源平台访问、下载加速 | | | | |
| | 不同地域的实例之间内网互不相通;选择靠近您客户 | 的地域,可降低网络时延、提高您客户的访问速度。 | | | | |
| 算力方案 | GPU基础型 | GPU进阶型 | 两卡GPU基础型 | 两卡GPU进阶型 | | |
| | 元/小时 高性价比,适用于推理场景 | <mark>元/小时</mark> 高性能,适用于推理、训练场景 | 元/小时 高性价比,适用于推理场景 | 元/小时 高性价比,适用于推理、训练场景 | | |
| | ✓ 显存: 16GB+ ✓ 算力: 8+TFlops SP | ✓ 显存: 32GB+ ✓ 算力: 15+TFlops SP | ✓ 显存: 2*16GB+ ✓ 算力: 16+TFlops SP | ✓ 显存: 2*32GB+ ✓ 算力: 30+TFlops SP | | |
| | ✓ CPU: 8核 ✓ 内存: 32GB | ✔ CPU: 8~10 核 ✔ 内存: 40GB | ✓ CPU: 16 核 ✓ 内存: 64GB | ✓ CPU: 18~20 核 ✓ 内存: 80GB | | |
| 实例名称 | 请输入实例名称 | | | | | |
| 云硬盘 | 200 GB 300 GB | | - 200 + GB | | | |
| | 免费提供80GB,包含操作系统(30GB)及应用环境占 | 用空间,可根据需求调整。详见 计费概述 [2 | | | | |
| 网络 🕄 | 每台实例免费提供500GB流量包,默认10Mbp | os带宽,每月刷新 | | | | |
| 协议 | 我已阅读并同意《腾讯云服务协议》、《腾 | 讯云禁止虚拟货币相关活动声明》 | | | | |
| | | | | | | ♥ |
| 数量 - 1 | + | | | | 费用总计 🚯 🚺 元/小时 | 立即购买 |



HAI 实例开机需要大约2-5分钟,状态修改为运行中,即可点击进入。

| = | Cloud Studio | | | \square | |
|----------|--------------------|----|------------------------------------|--|----|
| _ | • | | | | 2 |
| 88 | 通用工作空间 | 50 | | | |
| Ø | 高性能工作空间 🔼 | | DeepSeek 来了!不用等待模型下载,创建即可使用。 | | |
| ው | 空间模板 | | 内置 Ollama、DeepSeek-R1 1.5B 及 7B 模型 | | |
| Ŷ | 应用推荐 | | 立即创建 | | |
| | | | 高性能工作空间 | 新建 | |
| | | | 提供GPU算力,能够即开即用使用AI框架、AI模型、AI应用 | 本月贈送时长(分钟) ⑦ 🛛 🗕 🖛 🛛 🗠 🖬 2291 分钟 / 总计 10000 分額 | ŧ |
| 最近的 D | eepseek 高性制 运行中 | | Deepseek [暂无描述] ∲DeepSeek-R1 | 創建封詞 ● 运行中 HAI GPU通路型 ◎ ↓ 2025-02-08 ● 运行中 | |
| | | | | | Ρ. |

2. 在终端对话

与基础型类似,打开后新建终端并输入命令即可完成对话。



3. 使用 Chatbot UI / Openweb UI 对话



目前进阶型空间也具备 Chatbot UI 对话框功能,其体验路径为:

1. 在高性能工作空间列表中,单击已创建的进阶型空间最右侧图标,进入 HAI资源控制台。





2. 单击 ChatbotUI 图标或 OpenWebUI 图标。

| 高性能应用服务 | ●精选特惠 | 云服务器3年机 | /5年机 限时抢购,低至 3.5折! 查看详情 > | | | | | | × |
|----------|----------|------------------------|---|-------|------|-----------|--------------------------|------------|---|
| 器 算力管理 | ← hai-47 | ll7y5h (ha | 1) | | | | | | |
| ○ 应用管理 | | | | | | | | | |
| ☑ 资源社区 № | C) | hai 运 _{开机} | 行中 关机 重启 重置密码 应用存档 | 云硬盘扩 | 容 销毁 | 连接算力 | | | |
| | | 实例 ID 算力类型 | hai- GPU进阶型 - 32GB+ 15+TFlops SP CPU - 8~10 核 内存 - 40GB 云硬盘 - 200GB 网络 - 500GB (峰值带宽: 10Mbps) | | | ChatbotUI | CloudStudio | jupyterLab | |
| | | 应用名称 | DeepSeek-R1 | 公网 IP | | | | | |
| | | 创建时间 | 2025-02-08 01:19:14 | 内网 IP | | | | | |
| | | | | 地域 | 广州 | OpenWebUI | > 终端连接(SSH) | | |
| | 监控 | 端口配置 | | | | spennoor | | | 9 |
| | (?) ±i∓ | 1小时 | 台 ⑤ 时间粒度: = = moni | 10秒 | ¥ | | | | 4 |

3. 选择模型,进行对话。

| 43.136.6 | 0.55:6889 | | | | | | \$ | ·) (0 [2 |
|----------|-----------|-----------------|--|--|---------------------------------|----------------|----|-----------|
| | ← | | | Chatbot Ollama | | | → | + New J |
| | | | Model | | | | | |
| | | | deepseek-r1:32b | | ~ | | | |
| | | | Size: 19.85 GB Modified: 9 hours ag | go | | | | |
| | | | System Prompt | | | | | |
| | | | Enter a prompt or type | e "/" to select a prompt | | | | |
| | | | Temperature | | | | | |
| | | | Higher values like 0.8 wi 0.2 will make it more for | ill make the output more random, w cused and deterministic. | hile lower values like | | | |
| | | | | 1.0 | | | | |
| | | | Precise | Neutral | Creative | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | Chatbot Ollama. | Chatbot Ollama is an advance | d chatbot kit for Ollama models aiming to | mimic ChatGPT's interface and 1 | functionality. | | |

更多进阶型操作指南

HAI 实践文档:包含个人支持库搭建指引等。



OpenwebUI 文档:包含联网搜索指引等。