媲美科技一体机用户手册-V1版本

媲美科技 DeepSeek 一体机

用户手册

文档版本 1.0

发布日期 2025-02-25

概述

媲美科技推出基于 DeepSeek 深度思考大模型的 RAG 一体机解决方案,整合算力服务器与 RAG 服务,提供全面的后台管理、用户端、API 接口,为企业解决算力部署复杂、技术门槛高 和安全可控等三大核心痛点提供创新路径。

本手册旨在指导用户快速掌握媲美一体机的使用方法。核心功能模块:

- 后台管理系统使用指南
- 用户操作指南
- API 接口操作说明
- 系统部署与配置
- 日常维护与技术支持

特别说明:当前版本为体验版,主要用于客户体验和培训使用。正式商用版将后续迭代更新。

媲美科技基于 DeepSeek 深度思考大模型的 RAG 一体机解决方案

し のです。 日研客戸端	API	(次) (次) (次) (定) (不) (不) (不) (不) (不) (不) (不) (不) (不) (不	し 工作流	() 数字员工			
	后台管理	(RAG、客服和	大模型)				
	RAG软件与服务						
OpenSeek DeepSeek 深度思考大模型 深度思考大模型							
8卡4090算力服务器							

业务操作流程

管理后台		用户端
登录后台 🗁 创建知识库	新建AI会话	使用AI会话 全录用户端

请对照以上流程,参考后续操作指引,完成相关操作。

一、 后台管理系统使用指南

1. 功能概述

为企业提供一体机相关软件的管理功能,包括但不限于知识库管理、AI会话管理、登录及个 人信息管理等功能。后期会逐步完善企业/微信客服,以及 API 管理功能等。

- 访问地址:
 - O 外网试用地址: DeepSeek 后台管理端外网试用地址
 - O 测试账号: admintest/123456
- 主界面包含以下功能模块:
 - O 登录管理
 - O 知识库管理
 - O 成员管理
 - O AI 会话管理
 - O 工作流【敬请期待】
 - O 微信客服 【敬请期待】
 - 数字员工 【敬请期待】

2. 登录管理

- 输入分配好的账号和密码,点击登录,即可登录到媲美一体机管理后台。
- 前后端使用统一的登录页面。注意:当下为了避免账号登录冲突的情况,请使用两个浏览
 器分别登录管理端和用户端账号进行测试。或者单独开一个无痕浏览器操作。



3. 知识库管理

知识库管理包括以下功能:

- 新建知识库。
- 知识库的编辑和删除。知识库文档内容更新。

	È业DS—体机		
AI会话	[] test1	[]]小学语文 …	test
知识库	无描述	小学语文	无描述
用户管理	文件:《四大门》_贾行家解读.pdf 状态:知识库可用	文件: 教大家如何使用Deepseek Al进行超级降维知 状态: 知识库可用	文件: DeepSeek从入门到精通-清华.pdf 状态: 知识库可用

◇ 创建知识库

● 新建知识库。企业可以根据不同的需求场景,创建不同的知识库,满足各种需求场景。



● 知识库基本信息。可配置知识库名称,以及知识库描述。

新建知识库	×
①	
*知识库名称	
媲美科技机器人走播操作教学	
知识库描述	
媲美专属机器人走播,展示用户如何使用机器人走 操作流程,以及销售人员操作完整手册。	播,以及走播机器人的部署安装和配置
	取消 确定

- 上传知识库文件。上传和该知识库关联的文件,支持上传多个文件,后期也可以再编辑。
 - O 支持格式: txt/docx/pptx/xlsx/pdf 文件格式。



 创建并初始化知识库。信息确认无误后,点击初始化,开始创建知识库。初始化需要等 1-2 分钟,初始化完成后即可创建会话进行会话验证。

♦ 知识库列表管理

● 编辑知识库。可修改知识库名称、描述,以及增减知识库文件,更新知识库。

修改知道	识库			×
*知识库谷	3称			
2024年	度AI十大趋势			
知识库描	述			
2024年	度AI十大趋势			
知识库文	件			
	上传知	识库文 档		
	支持 txt/doc	x/pptx/pdf 文件		
PDF	最新图谱榜单.pdf	5.2 MB	已上传	匝
PDF	2024年度AI十大趋势报告_线上	52.8 MB	已上传	匝
		取消	保存	保存并初始化

 删除知识库。删除知识库二次确认,知识库删除后,关联知识库的会话将无法获取知识库 相关数据。



4. 成员管理

◇ 概述

管理企业成员,包括管理后台的管理员,以及用户端的使用用户,都可以通过成员管理进行统一的维护。指定用户可以登录系统使用相关功能,外部用户无法随意使用,实现系统的权

限管控。

◇ 成员列表

//	企业DS—体机			
L _				
AI会话	用户名	用户类型	创建时间	更新时间
@	test4	管理员	2025-02-20 17:34:37	2025-02-
知识库	test	普通用户	2025-02-21 15:21:05	2025-02-
用户管理	test5	管理员	2025-02-21 15:08:39	2025-02-
个人中心	admin	管理员	2025-02-20 17:34:17	2025-02-
	shan123	普通用户	2025-02-24 14:49:11	2025-02-
	test6	普通用户	2025-02-21 16:02:31	2025-02-
	test1	普通用户	2025-02-21 18:53:40	2025-02-
	dylan	普通用户	2025-02-24 11:31:53	2025-02-
_				

● 添加成员

后台管理员暂不支持直接添加,由数据库配置。

新增用户					×	
用户名						
密码						
			提交	取消		
	2025	02 20 47.2	4.47			

● 编辑成员

修改用户	×	
用户名	test4	
旧密码		
新密码		
	提交 取消	

● 删除成员

后台管理员暂不支持删除。普通用户可以删除。

警告		×
! 您确定要删除用户[test]吗?		
	取消	确定

5. 会话管理

◇ 概述

企业可以针对不同的应用场景,创建多个会话场景需求,比如企业业务咨询、企业员工培训等,满足不同用户人群和场景的需求。

◇ 创建会话

○ 创建会话入口。在 AI 会话菜单的会话列表中,点击创建会话,即可进入创建流 程。

		Û
。 媲美科技机器人走播操作教学	 AI发展趋势	
会话ID: 7efb630881c73e00cf379e99b0a6f1e0 自色名称: 直播培训讲师 印识库: 媲美科技机器人走播操作教学 干场白: 您好,请问有什么问题可以帮到您吗?	会话ID: b997e4db7fb499192e973abd90dd8e7e 角色名称: AI专家 知识库: AI发展趋势 开场白: 您好,请问有什么问题可以帮到您吗?	

heqingshan 🔻

会话

O 创建会话



■ 输入会话名称。

- 选择知识库。可以从配置的知识库中,选择一个知识库,作为当前会话的 检索对象。如需更新知识库内容,请前往知识库更新文件内容。
- 设定角色名称。指定角色名称,让会话更加真实亲切。【暂时未涉及,后续

为智能体做准备】

■ 引导词-开场白配置。为每次对话设置开场白,引导用户开启对话。

♦ 会话管理

企业os—体机				
▲ 认知破局		上 DeepSeek学习专区		▲ 媲美科技机器人走播操作教学
 知明年 会话ID: 429739f6c 角色名称: AI张琦 知识车: 认知破局 开场白: 您好, 请问 	1893c08016d9252ed3c9b48b 可有什么问题可以帮到您吗?	会话ID: 72bc2a738a10984dd9f464a57681 角色名称: DeepSeek专家 知识库: DeepSeek学习专区 开场白: 您好,请问有什么问题可以帮到您	18886 3吗?	会话ID: 7efb630881c73e00cf379e99b 角色名称: 直播培训讲师 知识车: 媲美科技机器人走播操作教学 开场白: 您好,请问有什么问题可以帮

- O 编辑会话。可编辑会话的标题、知识库,以及开场白和角色配置。
- 删除会话。会话删除后,后台列表以及前台列表都将同步移出。删除时,需要 二次确认。

提示		×
🤨 确定删除会话: DeepSeek从入门到料	青通?	
	取消	确定

- O 会话预览。可对于创建的会话进行会话效果测试。操作包含:
 - 文字消息发送。

- 欢迎语。
- 深度思考。当前固定深度思考模式。
- 文案复制、编辑,以及回答结果的重新回答。



二、 用户操作指南

1、 功能概述

企业在管理后台完成知识库和会话配置,将整套 DeepSeek 大模型服务部署后,我们提供了一整套的前端会话交互入口,满足用户和大模型交互的产品功能设计和能力。

- 访问地址:
 - O 外网试用地址: DeepSeek 用户端外网试用地址
 - O 测试账号: usertest/123456
- 主要功能模块包括:
 - O PC端 AI 会话
 - O 移动端 AI 会话

2、 PC端AI会话

- 功能清单
 - 输入文字,发送给服务器,在通过路由调度,RAG的问题检索,以及大模型的推理, 最终完成结果的输出,返回给用户。
 - O 大模型回答分两部分:深度推理+答案。
 - O AI会话切换。
 - O 主题色切换功能。

以上功能同时适合 PC 端和移动端。更多功能,包括自定义对话等功能将陆续完善。

◇ 新建会话



♦ AI 对话



♦ 主题色切换

主题切换入口



浅色主题



深色主题

3、 移动端 AI 会话

浅色主题

2:40 💴 …	HD] ⁵⁶ atl ⁵⁶ atl <u>56</u> p	2:40 🛤 …	HD 56 11 56
× DeepSeek会话 ds.lycheeai.com.cn	>	K DeepSeek会	话
	操作教学 2:74)	企业DS一体机	× 教学
(2025-02-25 11.4	2.30)	+ 新建会话	0)
□ 告诉我如何开始	自我的直播工作	AI发展趋势 (2025-02-25 02:	29:38) 直播工作
● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○		媲美科技机器人走播操作教学 (2025-02-25 02:27:54)	
根据提供的上下文信息,以下是关 工作的指导: 设置互动动作和播报内容 设置专门的互动动作和播报内容,	于如何开始直播 让直播间更加生	DeepSeek学习专区 (2025-02 02:27:30)	7开始直播 2-25 释间更加生
动自然,增强观众的参与感。 完成荔枝App开播 按照荔枝App的开播流程完成准备	工作。	媲美科技机器人走播操作教学 (2025-02-25 11:42:36)	
切换到抖音开播流程 打开抖音App,进入直播界面。 点击直播按钮,开启直播。 设置好欢迎话术,吸引观众的注意	力。		
点击开炉视频直播 确认一切准备就绪后,点击开始视 播状态。 一、一一	频直播,进入直		₹, 进入直
开始直播 进入直播后,按照之前设置的内容 保持直播的活跃度。	,与观众互动,		₿众互动,
希望以上步骤能帮助您顺利开始直	播工作!		E!
19 TO 19 TO 19	0	浅色	深色

深色主题

三、 API 接口操作说明

1、 接口参数说明

♦ 接口请求参数示例

● 参数解释:

类型	是否必填	默认值	描述
Array of	是	-	由目前为止的对话组成的消息列表

 MessageParam			
Dict	否	-	如果需要使用 RAG 构建的知识库,需要在"knowledgebase_id" 成的知识库 ID; "init_character"是拓展参数,可以通过该参数自定义自己对话
Boolean	否	false	响应内容是否流式返回 false:模型生成完所有内容一次性返回结果 true:逐块返回模型生成内容
•	MessageParam	·	

类型	是否必填	默认值	描述
String	是	-	发出该消息的对话参与者角色,可选值包括: system: System Message 系统消息 user: User Message 用户消息 assistant: Assistant Message 对话助手消息
String	是	-	消息内容, 文本生成模型仅支持 String 类型。

◇ 响应参数示例

● 非流式回复

```
{
   "id": "chat-d008b0b713e4492db67bd69093f80a7c",
   "choices": [
      {
          "finish reason": "stop",
          "index": 0,
          "logprobs": null,
          "message": {
             "content": "***",
             "refusal": null,
             "role": "assistant",
             "function call": null,
             "tool calls": []
          },
          "stop reason": null
      }
   ],
   "created": 1733195265,
   "model": "***",
   "object": "chat.completion",
   "service_tier": null,
   "system_fingerprint": null,
   "usage": {
      "completion_tokens": 28,
      "prompt tokens": 263,
      "total_tokens": 291
   },
   "prompt logprobs": null
}
```

● 流式回复

"我 ","function_call":null,"refusal":null,"role":null,"tool_calls":nu ll},"finish_reason":null,"index":0,"logprobs":null}],"created":17 39501982,"model":"model name","object":"chat completion chunk"."s

89bfd45f6159422b96e8acb96612e27b", "choices": [{"delta": {"content":

```
","function_call":null,"refusal":null,"role":null,"tool_calls":nu
ll},"finish_reason":null,"index":0,"logprobs":null}],"created":17
39501982,"model":"model_name","object":"chat.completion.chunk","s
ervice_tier":null,"system_fingerprint":null,"usage":null}
```

```
data: {"id":"chatcmpl-
89bfd45f6159422b96e8acb96612e27b","choices":[{"delta":{"content":
```

",

data: {"id":"chatcmpl-

"嗯 ","function_call":null,"refusal":null,"role":null,"tool_calls":nu ll},"finish_reason":null,"index":0,"logprobs":null}],"created":17 39501982,"model":"model_name","object":"chat.completion.chunk","s ervice_tier":null,"system_fingerprint":null,"usage":null}

```
data: {"id":"chatcmpl-
89bfd45f6159422b96e8acb96612e27b","choices":[{"delta":{"content":
```

data: {"id":"chatcmpl89bfd45f6159422b96e8acb96612e27b","choices":[{"delta":{"content":
"\n","function_call":null,"refusal":null,"role":null,"tool_calls"
:null},"finish_reason":null,"index":0,"logprobs":null}],"created"
:1739501982,"model":"model_name","object":"chat.completion.chunk"
,"service tier":null,"system fingerprint":null,"usage":null}

data: {"id":"chatcmpl89bfd45f6159422b96e8acb96612e27b","choices":[{"delta":{"content":
"<think>","function_call":null,"refusal":null,"role":null,"tool_c
alls":null},"finish_reason":null,"index":0,"logprobs":null}],"cre
ated":1739501982,"model":"model_name","object":"chat.completion.c
hunk","service_tier":null,"system_fingerprint":null,"usage":null}

data: {"id":"chatcmpl89bfd45f6159422b96e8acb96612e27b","choices":[{"delta":{"content":
"","function_call":null,"refusal":null,"role":"assistant","tool_c
alls":null},"finish_reason":null,"index":0,"logprobs":null}],"cre
ated":1739501982,"model":"model_name","object":"chat.completion.c
hunk","service_tier":null,"system_fingerprint":null,"usage":null}

四、系统部署与配置

以下是本项目推荐服务器配置以及部署流程。

1、 硬件

● 服务器: 推荐如下:

高度	5U
CPU	2 颗 AMD EPYC 7542 32C/64H 2.9G
GPU 卡	8 颗 Geforce RTX4090 24G
内存	16 条 DDR4 3200MHz RDIMM
	8个 PCIE4.0x16 扩展用于 GPU 卡
	3个 PCIE4.0x16 扩展
前置 IO	4个 USB 接口, 1个 VGA 接口, 1个 RJ45 管理口, 2个 10GbE 接口
	2 块 2.5 寸 SAS/SATA 480G 硬盘
~~ ¹ 4 ¹ 1	1 块 3.84T NVME U.2
操作系统	Ubuntu 22.04
电源	4 个 2400W RPSU
机箱尺寸	宽 445mm,高 222mm,深 950mm

● 存储: 见上图 (本地存储), 推荐 U.2 存储选择 英特尔(Intel) P5500 系列硬盘

2、 软件

● 操作系统:推荐 Ubuntu 或者 Debian 系衍生发行版均可。本项目操作系统为

"Ubuntu 22.04.5 LTS"。

● 依赖环境:本项目唯一需要配置的环境为容器环境,建议选择最新 Docker Engin。

L	:~\$ docker version
Client: Docker Engi	ne - Community
Version:	27.5.1
API version:	1.47
Go version:	go1.22.11
Git commit:	9f9e405
Built:	Wed Jan 22 13:41:31 2025
0S/Arch:	linux/amd64
Context:	default
Server: Docker Engi	ne - Community
Engine:	
Version:	27.5.1
API version:	1.47 (minimum version 1.24)
Go version:	go1.22.11
Git commit:	4c9b3b0
Built:	Wed Jan 22 13:41:31 2025
0S/Arch:	linux/amd64
Experimental:	false
containerd:	
Version:	1.7.25
GitCommit:	bcc810d6b9066471b0b6fa75f557a15a1cbf31bb
runc:	
Version:	1.2.4
GitCommit:	v1.2.4-0-g6c52b3f
docker-init:	
Version:	0.19.0
GitCommit:	de40ad0
Parre	ar:∼\$ docker compose version
Docker Compose vers	ion $\sqrt{2.32.4}$

3、 部署流程

3.1 安装 nvidia 驱动

0.查询型号

lspci | grep NVIDIA

(base) root@stone:/data/ops# lspci | grep NVIDIA 01:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) 01:00.1 Audio device: NVIDIA Corporation Device 22ba (rev a1) 26:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) 26:00.1 Audio device: NVIDIA Corporation Device 22ba (rev a1) 41:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) 41:00.1 Audio device: NVIDIA Corporation Device 22ba (rev a1) 61:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) 61:00.1 Audio device: MVIDIA Corporation Device 22ba (rev a1) 81:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) 81:00.1 Audio device: MVIDIA Corporation Device 22ba (rev a1) a1:00.0 VGA compatible controller: **NVIDIA** Corporation Device 2684 (rev a1) a1:00.1 Audio device: NVIDIA Corporation Device 22ba (rev a1) c1:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) c1:00.1 Audio device: MVIDIA Corporation Device 22ba (rev a1) e1:00.0 VGA compatible controller: NVIDIA Corporation Device 2684 (rev a1) e1:00.1 Audio device: MVIDIA Corporation Device 22ba (rev a1)

apt install -y gcc g++ make

需要对驱动包授权

chmod 755 NVIDIA-Linux-x86_64-550.144.03.run

(base) root	(base) root@stone:/data/ops# 11							
total 30009	96							
drwxr-xr-x	5	root	root	4096	Feb	27	02:44	•/
drwxr-xr-x	19	root	root	4096	Feb	26	08:13	**/
drwxr-xr-x	3	root	root	4096	Feb	26	08:24	DS/
drwxrwxr-x	2	root	root	4096	Feb	26	08:35	MW /
- WXR-XR-X)1	root	root	307272669	Feb	26	08:55	NVIDIA-Linux-x86_64-550.144.03.run*
drwxrwxr-x	3	root	root	4096	Feb	26	08:40	RAG/
- rw-rr	1	root	root	395	Feb	26	08:11	running.sh

2. 安装驱动

./NVIDIA-Linux-x86_64-550.144.03.run

3. 查看驱动

cat /etc/os-release

(base) root@stone:/data/ops# cat /etc/os-release PRETTY_NAME="Ubuntu 22.04.3 LTS" NAME="Ubuntu" VERSION_ID="22.04" VERSION="22.04.3 LTS (Jammy Jellyfish)" VERSION_CODENAME=jammy ID=ubuntu ID_LIKE=debian HOME_URL="https://www.ubuntu.com/" SUPPORT_URL="https://help.ubuntu.com/" BUG_REPORT_URL="https://bugs.launchpad.net/ubuntu/" PRIVACY_POLICY_URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy" UBUNTU_CODENAME=jammy

从这里我们可以知道系统是 Debian 系下 ubuntu 22.04 版本 那么我们选择执行中一个章节的命令

Table of Contents

NVIDIA Container Toolkit

Overview

Installing the Toolkit Running a Sample Workload Platform support Troubleshooting Release Notes

Advanced Configuration

Architecture Overview Container Device Interface Specialized Configurations with Docker NVIDIA Docs Hub > Cloud Native Technologies > NVIDIA Container Toolkit > Installing the NVIDIA Container Toolkit

Installing the NVIDIA Container Toolkit

Installation

Prerequisites

1 Note

- 1. Read this section about platform support.
- 2. Install the NVIDIA GPU driver for your Linux distribution. NVIDIA recommends installing the driver by using the package manager for your distribution. For information about installing the driver with a package manager, refer to the <u>NVIDIA Oriver Installation Quickstart Guide</u>. Alternatively, you can install the driver by <u>downloading</u> a <u>run</u> installer.

With apt : Ubuntu, Debian #

These instructions should work for any Debian-derived distribution.

1. Configure the production repository:

\$ curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey | sudo gpg --&& curl -s -L https://nvidia.github.io/libnvidia-container/stable/deb/nvidia sed 's#deb https://#deb [signed-by=/usr/share/keyrings/nvidia-container-to sudo tee /etc/apt/sources.list.d/nvidia-container-toolkit.list

Optionally, configure the repository to use experimental packages:

\$ sed -i -e '/experimental/ s/*#//g' /etc/apt/sources.list.d/nvidia-container-

\$ sudo apt-get install -y nvidia-container-toolkit

我们验证 CUDA Toolkit 的完整性

2. Update the packages list from the repository:
\$ sudo apt-get update
3. Install the NVIDIA Container Toolkit packages:

sudo docker run --rm --runtime=nvidia --gpus all ubuntu nvidiasmi

(base) root@10-0-0-10:~# docker run --rm --runtime=nvidia --gpus all ubuntu nvidia-smi Fri Mar 28 08:13:28 2025 NVIDIA-SMI 555.42.06 Driver Version: 555.42.06 CUDA Version: 12.5 Persistence-M | Bus-IdDisp.A | Volatile Uncorr. ECCPwr:Usage/Cap |Memory-Usage | GPU-Util Compute M. GPU Name Fan Temp Perf Pwr:Usage/Cap MIG M. 0 NVIDIA GEForce RTX 4090 0ff | 00000000:1B:00.0 0ff | 31% 28C P8 18W / 450W | 6651MiB / 24564MiB | 0ff Default 31% 28C P8 N/A . _ _ _ + _ 1 NVIDIA GeForce RTX 4090 0ff | 00000000:1E:00.0 0ff | 31% 29C P8 21W / 450W | 4MiB / 24564MiB | 0% Default +-----
 2
 NVIDIA GeForce RTX 4090
 0ff
 00000000:22:00.0 0ff
 30%

 30%
 31C
 P8
 20W / 450W
 4MiB / 24564MiB
 44000
 0ff Default N/A
 3
 NVIDIA GeForce RTX 4090
 0ff
 00000000:23:00.0 0ff
 30%
 30C
 P8
 12W / 450W
 1183MiB / 24564MiB
 24564MiB
 1
 0ff 0% Default 4 NVIDIA GEForce RTX 4090 0ff | 0% 29C P8 18W / 450W | 00000000:4F:00.0 Off | 0ff 1149MiB / 24564MiB 30% Default N/A

特别提醒:

Cuda compilation tools, release 12.4, V12.4.131: 这部分告诉你当前使用的 CUDA 工具包的版本是 12.4 版本,内部版本号为 12.4.131。这表示你的开发环境支持 CUDA 12.4 所有特性以及改进。

至此 cuda/toolkit 安装完成 GPU 服务器具备计算环境。 下面我们将进入 DS 一体机的部署环节

/OPS	所有部署	文件均在
running.sh —	键部署	这里不做介绍
./MW	Middlew	ware 中间件 这里需要执行两条命令
./DS	DS 大模型	型部署
./RAG	Rag 服务	算法-后端-前端

以上是我们所有的部署脚本和文件。

3.3 设置环境

执行~/ops/MW/start.sh

说明: 主要是安装 docker--建立容器网络--设置持久化存储

使用命令

```
bash start.sh
```

这里我们需要分条执行一下命令 验证上面执行结果

```
docker --version
docker compose version
docker network ls
ll /data/
```

```
(base) root@stone:/data/ops/MW# docker --version
                                                     docker 版本
Docker version 28.0.0, build f9ced58
(base) root@stone:/data/ops/MW# docker compose version
Docker Compose version v2.33.0
                                                   docker compose 版本
(base) root@stone:/data/ops/MW# docker network ls
NETWORK ID
              NAME
                           DRIVER
                                     SCOPE
bf2eb036f0be
              bridge
                           bridge
                                     local
976f40f37573
              ds_online
                          bridge
                                     local 自定义网
5d67566d8874
                                     local
              host
                           host
397a65657ab3
              none
                           null
                                     local
(base) root@stone:/data/ops/MW# 11 /data/
total 96
drwxr-xr-x 19 root root 4096 Feb 26 08:13 ./
drwxr-xr-x 20 root root 4096 Feb 21 19:40 .../
drwxr-xr-x 5 root root 4096 Feb 19 14:16 aio/
drwxr-xr-x 4 root root 4096 Feb 21 06:15 damo/
drwxr-xr-x 3 root root 12288 Feb 13 21:35 DeepSeek-R1/
drwxr-xr-x 4 root root 4096 Feb 7 10:01 DeepSeek-R1-Distill-Llama-70B/
drwxr-xr-x 2 root root 4096 Feb 26 03:42 🚮
drwxr-xr-x 2 root root 4096 Feb 22 15:31 432/
drwx----- 2 root root 16384 Feb 12 13:46 lost+found/
drwxr-xr-x 47 root root 4096 Feb 27 02:58 lycheeairag/
drwxr-xr-x 4 root root 4096 Feb 21 09:55
drwxrwxrwt 8 root root 4096 Feb 27 03:18 mm
drwxr-xr-x 7 root root 4096 Feb 12 16:3 ncel-tests/
drwxr-xr-x 5 root root 4096 Feb 27 03:18 cms/
drwxr-xr-x 2 root root 4096 Feb 27 02:44 rag-chat/
drwxr-xr-x 2 root root 4096 Feb 21 08:4 rag ui/
drwxr-xr-x 3 root root 4096 Feb 27 03:18 redis/
                                                            自定义持久化存储文
drwxr-xr-x 5 root root 4096 Feb 21 06:18
                                               temp/
drwxr-xr-x 2 root root 4096 Feb 26 08:45 yy/
(base) root@stone:/data/ops/MW#
```

3.4 安装中间件

文件位置 ~/ops/MW/docker-compose.yml 说明:安装桶文件存储 minio\缓存服务 redis\数据库 mysql

do	cker compose up -d			
这里我们需要	P.验证上面执行结果			
docker	compose ps			
(base) root@stone	e:/data/ops/MW# docker compose ps			
NAME	IMAGE	COMMAND	SERVICE	CREATED
ds-minio-1	ccr.ccs.tencentyun.com/pmtech/minio	"/usr/bin/docker-ent"	(minio)	5 seconds ago
000-9001->9000-90	901/tcp			
ds-minio-init-1	ccr.ccs.tencentyun.com/pmtech/minmc	"/bin/sh -c 'minio-i…"	minio-init	5 seconds ago
ds-mysql-1	ccr.ccs.tencentyun.com/pmtech/mysql8	"docker-entrypoint.s"	mysq1	5 seconds ago
:]:13306->3306/to	ср		\leq	
ds-redis-1	ccr.ccs.tencentyun.com/pmtech/redis	"docker-entrypoint.s"	redis	5 seconds ago
tcp				

3.5 安装 DS

文件位置~/ops/DS/docker-compose.yml 说明: 部署 DS 大模型服务

文件简介

```
(base) root@stone:~/ops/DS# 11
total 20
drwxr-xr-x 3 stone 1001 4096 Feb 21 06:58 ./
drwxr-xr-x 5 stone 1001 4096 Feb 21 06:58 ../
drwxr-xr-x 2 root root 4096 Feb 21 07:00 deepseek-ai/
      #权重文件集
-rw-rw-r-- 1 stone 1001 642 Feb 21 05:45 run_rag.sh
      #大模型启动文件
-rw-rw-r-- 1 stone 1001 695 Feb 21 05:47 docker-compose.yml
   #DS 大模型服务配置文件
(base) root@stone:/data/ops/DS# 11 deepseek-ai/
total 16
drwxr-xr-x 4 root root 4096 Feb 26 03:20 ./
drwxr-xr-x 3 root root 4096 Feb 26 08:24 ../
drwxr-xr-x 4 root root 4096 Feb 26 03:29 DeepSeek-R1-Distill-
Llama-70B/
                #70B
drwxr-xr-x 3 root root 4096 Feb 6 15:00 DeepSeek-R1-Distill-
Qwen-32B/ #32B 此次选用
```

执行命令 用以启动部署大模型

我们执行两条命令验证大模型的部署

(base) root@stone:/dat NAME IMAG ds-ds1-1 ccr. ds-ds2-1 ccr. ds-minio-1 ccr. ds-minio-init-1 ccr. ds-mysql-1 ccr. , [::]:13306->3306/tcp ds-redis-1 ccr. 379/tcp ; 建议大家使用 nvitop	a/ops/DS# docker c E ccs.tencentyun.com ccs.tencentyun.com 1/tcp ccs.tencentyun.com ccs.tencentyun.com ccs.tencentyun.com <u>ccs.tencentyun.com</u> <u>ccs.tencentyun.com</u>	ompose ps /pmtech/ds:lates /pmtech/ds:lates /pmtech/minio /pmtech/minmc /pmtech/mysql8 /pmtech/redis を行进程	COMMAND st "/opt/nvidia/nvidia" st "/opt/nvidia/nvidia" "/usr/bin/docker-ent" "/bin/sh -c 'minio-i" "docker-entrypoint.s" "docker-entrypoint.s"	SERVICE ds1 ds2 minio minio-init mysq1 redis
NVITOP 1.4.2 Driver Ver	sion: 550.144.03 CUD	A Driver Version: 12.4		
GPU Fan Temp Perf Pwr:Usg/Ca	p Memory-Usage	GPU-Util Compute M.		
0 30% 50C P2 147W / 450	√ 19317MiB / 23.99GiB	0% Default	: MEM:	78.6% UTL:
1 30% 49C P2 147W / 450	W 19315MiB / 23.99GiB	0% Default	T MEM:	78.6% UTL:
2 30% 51C P2 156W / 450	W 19315MiB / 23.99GiB	0% Default	: MEM:	78.6% UTL:
3 30% 50C P2 148W / 450	W 19315MiB / 23.99GiB	0% Default	MEM:	78.6% UTL:
4 30% 42C P2 152W / 450	W 19315MiB / 23.99GiB	0% Default	MEM:	78.6% UTL:
5 30% 43C P2 154W / 450	W 19313MiB / 23.99GiB	0% Default	MEM:	78.6% UTL:
6 30% 42C P2 152W / 450	₩ 19313MiB / 23.99GiB	1% Default	MEM:	78.6% UTL:
7 30% 44C P2 135W / 450	W 19313MiB / 23.99GiB	0% Default	MEM:	78.6% UTL:
[CPU: CPU: 6.1% [MEM: 5.6% CMUTION: SUPERUSER LOGGED-IN				UPTIME: USED:
Processes: GPU PID USER GPU-M	EM %SM %CPU %MEM TIME	COMMAND		
0 1665429 C root 16328% 1 1666119 C root 16326% 2 1666120 C root 16326% 3 1666183 C root 16326% 4 1665426 C root 19328% 5 1666189 C root 19368% 6 1666190 C root 19268% 7 1666191 C root 19208%	IB 36 11.02 0.3 1:02 IB 37 109.0 0.3 0:37 IB 27 109.8 0.3 0:37 IB 27 110.1 0.3 0:36 IB 65 112.2 0.3 1:03 IB 64 113.0 0.3 0:36 IB 54 114.8 0.3 0:36 IB 56 113.6 0.3 0:36	/opt/conda/envs/deeps /opt/conda/envs/deeps /opt/conda/envs/deeps /opt/conda/envs/deeps /opt/conda/envs/deeps /opt/conda/envs/deeps /opt/conda/envs/deeps	seek/bin/python -c "from multiproces seek/bin/python -c "from multiproces	sing.spawn import spawr sing.spawn import spawr

3.6 安装 RAG

文件位置在~/ops/RAG/docker-compose.yml 说明: 部署 RAG 服务算法服务、后端服务、前端服务

```
(base) root@stone:~/ops/RAG# 11
total 16
drwxrwxr-x 3 stone 1001 4096 Feb 21 07:15 ./
drwxr-xr-x 5 stone 1001 4096 Feb 21 06:58 ../
-rw-rw-r-- 1 stone 1001 1339 Feb 21 07:15 docker-compose.yml
    #RAG 服务配置文件
drwxrwxr-x 3 stone 1001 4096 Feb 21 05:41 web_ui/
    #网站入口配置文件
```

执行命令用以启动 Rag 服务

docker compose up -d

这里我们需要 验证上面执行结果

(base) root@st	tone:/data/ops/RAG# docker ps -a		
CONTAINER ID	IMAGE	COMMAND	CREATED
	NAMES		
8ef95ba06988	ccr.ccs.tencentyun.com/pmtech/rag:1.1.2-prod-rag	"/opt/nvidia/nvidia"	32 seconds ago
14/tcp	ds-rag-1		
84042db56305	ccr.ccs.tencentyun.com/pmtech/rag-ui	"/docker-entrypoint"	32 seconds ago
	ds-rag-ui-1		
a6ab13b02f56	ccr.ccstencentyun.com/pmtech/rag-chat	"sh -c 'java -jar ra"	32 seconds ago
/tcp	ds-rag-chat-1		
cbf3091c2d5a	ccr.ccs.tence ntyun.co m/pmtech/ds:latest	"/opt/nvidia/nvidia"	2 hours ago
	ds-ds1-1		
cab2be1592b4	ccr.ccs.tencentyun.com/pmtech/ds:latest	"/opt/nvidia/nvidia"	2 hours ago
	ds-ds2-1		
6eb46558a145	ccr.ccs.tencentyun.com/pmtech/minmc	"/bin/sh -c 'minio-i…"	2 hours ago
	ds-minio-init-1		
5dff43d95da8	ccr.ccs.tencentyun.com/pmtech/minio	"/usr/bin/docker-ent"	2 hours ago
00-9001->9000	-9001/tcp ds-minio-1		
7894b86a08db	ccr.ccs.tencentyun.com/pmtech/redis	"docker-entrypoint.s"	2 hours ago
ср	ds-redis-1		
98032c637fa7	ccr.ccs.tencentyun.com/pmtech/mysql8	"docker-entrypoint.s"	2 hours ago
12206 32206/1			

整个服务验证 可以使用下面命令

企业DS一体机	
▲ 请输入账号 ▲ 请输入密码	

五、日常维护与技术支持

提供7*12h的日常维护支持。技术支持:

- 邮箱: info@lycheeai.com.cn
- 电话: 18701446128 (同微信号)
- 联系人: 何先生

版权所有 © 2025 媲美科技