解决方案实践

快速搭建 DeepSeek 推理系统

文档版本1.1.0发布日期2025-02-19





版权所有 © 华为技术有限公司 2025。保留一切权利。

非经本公司书面许可,任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部,并不得以任何形式传播。

商标声明

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束,本文档中描述的全部或部分产品、服务或 特性可能不在您的购买或使用范围之内。除非合同另有约定,华为公司对本文档内容不做任何明示或暗示的声 明或保证。

由于产品版本升级或其他原因,本文档内容会不定期进行更新。除非另有约定,本文档仅作为使用指导,本文 档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以"漏洞处理流程"为准,该流程的详细内容请参见如下网址: https://www.huawei.com/cn/psirt/vul-response-process 如企业客户须获取漏洞信息,请参见如下网址: https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目录

1 方案概述	1
2 资源和成本规划	3
3 实施步骤	9
3.1 准备工作	9
3.2 快速部署	
3.3 开始使用	
3.4 快速卸载	
4 附录	
5 修订记录	26



应用场景

随着互联网信息量的爆炸式增长,企业和个人面临着海量数据管理和高效检索的需求。传统的搜索引擎虽然能够满足基本的搜索需求,但在面对复杂多样的数据类型和个性化需求时,往往显得力不从心。因此,DeepSeek应运而生,DeepSeek(深度求索)作为一款国产AI大模型,凭借其高性能、低成本和多模态融合能力,在人工智能领域崛起,并在多个行业中展现出广泛的应用潜力。

该解决方案帮助您在华为云Flexus云服务器X实例(弹性云服务器 ECS)上快速搭建 DeepSeek-R1轻量级推理系统。DeepSeek-R1是一个高性能的AI推理模型,专注于数 学、代码和自然语言推理任务,通过Ollama在云服务器中部署DeepSeek-R1蒸馏版轻 量模型,快速打造您的私人AI助手,主要适用如下场景:

- 自然语言处理:能够理解和生成自然语言文本,适用于对话、翻译、摘要等任务。
- 文本生成:能够生成连贯、逻辑清晰的文本,适用于内容创作、故事编写等。
- 问答系统:能够回答用户提出的问题,适用于客服、知识库查询等场景。
- 情感分析:能够分析文本中的情感倾向,适用于市场调研、舆情监控等。
- 文本分类:能够对文本进行分类,适用于垃圾邮件过滤、新闻分类等。
- 信息抽取:能够从文本中提取关键信息,适用于数据挖掘、知识图谱构建等。

方案架构

该解决方案帮助您在华为云Flexus云服务器X实例(弹性云服务器 ECS)上快速搭建 DeepSeek-R1蒸馏版模型。

图 1-1 方案架构图



该解决方案将会部署如下资源:

- 创建一个弹性公网IP EIP,用于提供访问公网和被公网访问能力。
- 创建一台Flexus云服务器X实例(弹性云服务器 ECS),用于搭建DeepSeek-R1蒸 馏版模型。
- 创建一个安全组,通过配置安全组规则,为云服务器提供安全防护。

方案优势

高性能

DeepSeek通过强化学习技术显著提升推理能力,支持多步骤逻辑推理,能够逐步 分解复杂问题并解决问题。

- 低成本
 提供高性价比的云服务器,用户可以根据实际需求自定义不同规格的云服务器。
- 一键部署
 一键轻松部署,即可快速完成云服务器和公网IP等资源的下发以及DeepSeek-R1
 蒸馏版模型的搭建。

约束与限制

 该解决方案部署前,需注册华为账号并开通华为云,完成实名认证,且账号不能 处于欠费或冻结状态。如果计费模式选择"包年包月",请确保账户余额充足以 便一键部署资源的时候可以自动支付;或者在一键部署的过程进入费用中心,找 到"待支付订单"并手动完成支付。



该解决方案主要部署如下资源,以下费用仅供参考,具体请参考华为云官网<mark>价格详</mark> 情,实际收费以账单为准。

华为云服务	资源名称	配置示例	数 量	每月预估花 费
虚拟私有云 VPC	building-a-deepseek- Inference-system-demo	 VPC网段: 172.16.0.0/16 区域:华北-北京 四 	1	0.00元
子网 Subnet	building-a-deepseek- Inference-system- demo-subnet	 子网网段: 172.16.1.0/24 区域:华北-北京 四 	1	0.00元
安全组 SecurityGro up	building-a-deepseek- Inference-system-demo	 允许ping: 0.0.0.0/0 开放端口22允许 Cloud Shell 登 录: 121.36.59.153/3 2 区域:华北-北京 四 	1	0.00元

表 2-1 资源和成本规划	(按需计费)	CPU	版
---------------	--------	-----	---

华为云服务	资源名称	配置示例	数量	每月预估花 费
Flexus云服 务器X实例	building-a-deepseek- Inference-system-demo	 按需计费: 0.38/ 小时 区域: 华北-北京 四 规格: Flexus云 服务器X实例 性 能模式(关闭) x1.4u.4g 4核 4GB 镜像: Ubuntu 22.04 server 64bit 系统盘: 通用型 SSD 40GB 	1	271.30元
弹性公网IP EIP	building-a-deepseek- Inference-system- demo-eip	 按需计费: 0.80 元/GB 区域: 华北-北京 四 线路: 动态BGP 公网带宽: 按流 量计费 带宽大小: 300Mbit/s 	1	0.80元/GB
合计		-		271.30元 + 弹性公网IP EIP费用

表 2-2 资源和成本规划(包年包月) CPU 版

华为云服务	资源名称	配置示例	数量	每月预估花 费
虚拟私有云 VPC	building-a-deepseek- Inference-system-demo	 VPC网段: 172.16.0.0/16 区域:华北-北京四 	1	0.00元
子网 Subnet	building-a-deepseek- Inference-system-demo- subnet	 子网网段: 172.16.1.0/24 区域:华北-北京 四 	1	0.00元

华为云服务	资源名称 配置示例 数 量		资源名称 配置示例		数量	每月预估花 费
安全组 SecurityGro up	building-a-deepseek- Inference-system-demo	 允许ping: 0.0.0/0 开放端口22允许 Cloud Shell 登 录: 121.36.59.153/3 2 区域:华北-北京 四 	1	0.00元		
Flexus云服 务器X实例	building-a-deepseek- Inference-system-demo	 包年包月 区域:华北-北京四 规格:Flexus云服务器X实例 性能模式(关闭) x1.4u.4g 4核 4GB 镜像:Ubuntu 22.04 server 64bit 系统盘:通用型SSD 40GB 	1	190.00元		
弹性公网IP EIP	building-a-deepseek- Inference-system-demo- eip	 按需计费: 0.80 元/GB 区域: 华北-北京 四 线路: 动态BGP 公网带宽: 按流 量计费 带宽大小: 300Mbit/s 	1	0.80元/GB		
合计		-		190.00元 + 弹性公网IP EIP费用		

表 2-3 资源和成本规划(按需计费) GPU 版

华为云服务	资源名称	配置示例	数量	每月预估花 费
虚拟私有云 VPC	building-a-deepseek- Inference-system-demo	 VPC网段: 172.16.0.0/16 区域:华东-上海 一 	1	0.00元
子网 Subnet	building-a-deepseek- Inference-system-demo- subnet	 子网网段: 172.16.1.0/24 区域: 华东-上海 一 	1	0.00元
安全组 SecurityGro up	building-a-deepseek- Inference-system-demo	 允许ping: 0.0.0.0/0 开放端口22允许 Cloud Shell 登 录: 121.36.59.153/3 2 区域: 华东-上海 一 	1	0.00元
弹性云服务 器 ECS	building-a-deepseek- Inference-system-demo	 按需计费: 6.37/ 小时 区域: 华东-上海 一 规格: GPU加速 型g6 g6.xlarge.4 16 核 32GB NVIDIA T4 16GB 镜像: Ubuntu 22.04 server 64bit 系统盘: 通用型 SSD 40GB 	1	4585.54元
弹性公网IP EIP	building-a-deepseek- Inference-system-demo- eip	 按需计费: 0.80 元/GB 区域: 华东-上海 一 线路: 动态BGP 公网带宽: 按流 量计费 带宽大小: 300Mbit/s 	1	0.80元/GB

华为云服务	资源名称	配置示例	数量	每月预估花 费
合计		-		4585.54元 + 弹性公网IP EIP费用

表 2-4 资源和成本规划(包年包月) GPU 版

华为云服务	资源名称	配置示例	数量	每月预估花 费
虚拟私有云 VPC	building-a-deepseek- Inference-system-demo	 VPC网段: 172.16.0.0/16 区域: 华东-上海 一 	1	0.00元
子网 Subnet	building-a-deepseek- Inference-system- demo-subnet	 子网网段: 172.16.1.0/24 区域: 华东-上海 一 	1	0.00元
安全组 SecurityGro up	building-a-deepseek- Inference-system-demo	 允许ping: 0.0.0.0/0 开放端口22允许 Cloud Shell 登 录: 121.36.59.153/3 2 区域: 华东-上海 一 	1	0.00元
弹性云服务 器 ECS	building-a-deepseek- Inference-system-demo	 包年包月 区域:华东-上海 规格:GPU加速 型g6 g6.xlarge.4 16 核 32GB NVIDIA T4 16GB 镜像:Ubuntu 22.04 server 64bit 系统盘:通用型 SSD 40GB 	1	3064.40元

华为云服务	资源名称	配置示例	数量	每月预估花 费
弹性公网IP EIP	building-a-deepseek- Inference-system- demo-eip	 按需计费: 0.80 元/GB 区域: 华东-上海 一 线路: 动态BGP 公网带宽: 按流 量计费 带宽大小: 300Mbit/s 	1	0.80元/GB
合计		-		3064.40元 + 弹性公网IP EIP费用



- 3.1 准备工作 3.2 快速部署
- 3.3 开始使用
- 3.4 快速卸载

3.1 准备工作

当您首次使用华为云时注册的账号,则无需执行该准备工作,如果您使用的是IAM用 户账户,请确认您是否在admin用户组中,如果您不在admin组中,则需要为您的账号 授予相关权限,并完成以下准备工作。

创建 rf_admin_trust 委托 (可选)

步骤1 进入华为云官网,打开<mark>控制台管理</mark>界面,鼠标移动至个人账号处,打开"统一身份认证"菜单。

图 3-1 控制台管理界面



图 3-2 统一身份认证菜单

徿	御中文	ELMENMENMEN SUSSES
	基本信息	已实名认证
	安全设置	
τ γ	我的凭证	
ŧ	统一身份认	人证
	切换角色	
2	标签管理	
	操作日志	
2		退出

图 3-3 委托列表

委托 ③							创建委托
制除 您还可以创建49个委托。					全部类型 マ	请输入委托名称进行搜索	Q
● 要托名称/ID 1Ξ	要托对象 1Ξ	要托时长 ↓Ξ	创建时间	描述 ↓=		操作	
rt_admin_trust	云級侮 RFS	永久	2023/05/31 11:07:56 GMT+08:00	-		授权 修改 翻除	

- 如果委托存在,则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托
- **步骤3** 单击步骤2界面中的"创建委托"按钮,在委托名称中输入"rf_admin_trust",委托 类型选择"云服务",选择"RFS",单击"下一步"。

步骤2 进入"委托"菜单,搜索"rf_admin_trust"委托。

图 3-4 创建委托

委托 /	创建委托		
	★ 委托名称	rf_admin_trust	
	★ 委托类型	 普通帐号 将帐号内资源的操作权限委托给其他华为云帐号。 云服务 将帐号内资源的操作权限委托给华为云服务。 	
	* 云服务	RFS •	
	★ 持续时间	永久 🔻	
	描述	请输入委托信息。	
•		0/255 下一步 取消	

步骤4 在搜索框中输入"Tenant Administrator"权限,并勾选搜索结果,单击"下一步"。

图 3-5 选择策略

() REFERENCE () REER-HERCER () PAR	RYDK
经HEY1_admin_tust2*将将有所活动顺路	1773acht #2
整要已经(1) 从其他区域项目假制权限	全部映型 ▼ 新有芸服時 ▼ Tenant Administrator × Q
☑ 約約	类型
Image: Constraint of the second se	系統衛告

步骤5选择"所有资源",并单击"下一步"完成配置。

图 3-6 设置授权范围

() 25484	周列日版
● 客風当都認得活動和法律化、其他自弊以下我の注意方案、要用于包裹小化表現、可能行告報、了解如何用面容的应用系統最后命全面的消化可能力案	×
也接续反应差方質	
 新聞表題 新公元, WARP-可以成面容現得用板型中將有更思, 包括立日成之間面景思潮, 	
批行時間が常	

步骤6 "委托"列表中出现"rf_admin_trust"委托则创建成功。

图 3-7 委托列表

0						£13
删除 想还可以创建49个委托。				全部类型	▼ 请输入委托名称进行搜索	
● 要托名称/ID 1目	要托对象 1Ξ	要托时长 ↓Ξ	创建时间 15	描述 ↓=	操作	
rt_admin_trust	 一 服务 RFS	永久	2023/05/31 11:07:56 GMT+08:00	-	提权 修改 删除	

----结束

3.2 快速部署

操作场景

本章节帮助用户高效地部署"快速搭建DeepSeek推理系统"解决方案。一键部署该解决方案时,参照本章节中的步骤和说明进行操作,即可完成快速部署。

操作步骤

步骤1 登录**华为云解决方案实践**,选择"快速搭建DeepSeek推理系统",支持区域下拉选择 部署的区域(以北京四CPU版为例),单击"一键部署(CPU版)",跳转至解决方 案创建资源栈界面。

图 3-8 解决方案实施库

方案架构



步骤2 在选择模板界面中,单击"下一步"。

图 3-9 选择模板

< 立即创建资源制		
	(2) #\$\$\$\$\$ (3) HIRINGE (4) REMN	
* 自因散方式		
* 欄板来源	134468 Unit 上标数据 每个书题时和描述是出了任何的意思的。 第四年4月20日月 (1) 第三5月21日 (2001)	
★ 欄板 URL	竹町5/10cumentation-samples.dis.ch-north-4.myfx 新聞の目的では同じたり、2017年間目について、2017年間目的の目的では、約1千年の中枢の目的な話、2015年中型時間の長期の55回の目的目的をあ	
		#

步骤3 在配置参数界面中,参考"**表1** 参数填写说明"完成自定义参数填写,部分参数会自 动默认填充参数值。如需修改请在参数配置页面删除文本框内的默认值后填写新的参 数值,所有参数填写完成后方可单击"下一步"。

图 3-10 配置参数

< 立即创建资	5源栈						
 这時機能 — 	2 参数配置 —						
* 资源线名称	building-a-deepseek-inferenc 资源线名称以中文或英文开始。	Hoystem 2月中文、大小写苑文、裴永、下刻派、中刻派,128个字符之内, <mark>波澜城名称不能重</mark> 名	٤,				
編述	快速证面OHp5ex他理题间						
		16255 //					
配置参数	输业数名称	Q. 🔮 時機販賣茶打節分資源加率 ①					
受数名称		di .	类型	描述			
* vpc_name		building-a-deepseek-inference-system-demo	string	查报机局云条条,该模板发完新编VPC,不允许重备,取量定属:1-54个字符,支持中文,英文字母,数字,(下划线),-(中划线),-(点),数以…			
* security_group	_name	building-a-deepseek-inference-system-demo	string	安全组合称。该煤版新建安全组,安全组织则清参考部署颁集进行展盟,取量活展:1-54个字符,支持数字、字母、中文、《下划线》、-(中划线)、-(
* ecs_name		building-a-deepseek-Inference-system-demo	string	云服务器实形实命、不变地重杂、取重范围:1-4个字符、支持中文、英文字母、数字、(下划线)、-(中划线)、-(信),数以:building-a-deepse…			
* distilled_mode	i.	DeepSeek-R1-DistII-Quen-1.58	string	DeepSeek-R1-Diss制模型,支持Quien-158, Quien-78, Liama-88, 默认Quien-1.58,			
* ecs_flavor		x1.4u.4g	string	云服芳糊实的现象,1.56维考虑用v1.4u.4g、706386维考使用v1.16u.16g、默认: x1.4u.4g、			
* ecs_password			string	云服芳馨或符,长度为8-28位,实阴至少必须包含大写字母、小写字母、数字和特殊字符(106%~_=*们5.77)中的三种,仅支持小写字母、数字、中划…			
				(<u>+</u> -#)			

参数名称	类型	是否可 选	参数解释	默认值
vpc_name	strin g	必填	虚拟私有云名称,该模板使用 新建VPC,不允许重名。取值 范围:1-54个字符,支持中 文、英文字母、数字、_(下划 线)、-(中划线)、.(点)。	building-a- deepseek- Inference- system-demo
security_grou p_name	strin g	必填	安全组名称,该模板新建安全 组,请参考 安全组规则修改 进 行配置。取值范围:1-64个字 符,支持字母、数字、中文、 下划线(_)、中划线(-)、 英文句号(.)。	building-a- deepseek- Inference- system-demo

表 3-1 参数填写说明

参数名称	类型	是否可 选	参数解释	默认值
ecs_name	strin g	必填	云服务器实例名称,不支持重 名。取值范围:1-64个字符, 支持中文、英文字母、数字、_ (下划线)、-(中划线)、. (点)。	building-a- deepseek- Inference- system-demo
distilled_mod el	strin g	必填	DeepSeek-R1-Distill模型。支 持Qwen-1.5B、Qwen-7B、 Llama-8B,默认Qwen-1.5B。 GPU版支持部署 Qwen-14B。	DeepSeek-R1- Distill- Qwen-1.5B
ecs_gpu_map ping	strin g	必填	(GPU版部署参数,涉及华北- 北京四、华南-广州区域)弹性 云服务器 ECS规格与GPU显卡 大小的映射关系。A-1对应显存 1*16、B-1对应显存1*24G、 B-2对应显存2*24G、B-4对应 显存4*24G、B-6对应显存 6*24G,推荐显存大小16GB及 以上(Qwen-32B推荐显存大 小24G及以上)。当参数 {ecs_flavor}有值时,此参数不 生效。默认A-1。	A-1
ecs_flavor	strin g	必填	云服务器实例规格,CPU版部 署1.5b推荐使用x1.4u.4g、7b 和8b推荐使用x1.16u.16g,默 认:x1.4u.4g;GPU版部署推荐 显存大小16GB及以上 (Qwen-32B推荐显存大小24G 及以上)。规格信息具体请参 考官网 <mark>弹性云服务器规格清</mark> 单。	x1.4u.4g
ecs_password	strin g	必填	云服务器密码,长度为8-26 位,密码至少必须包含大写字 母、小写字母、数字和特殊字 符(!@\$%^=+[{}]:,/?)中的 三种,仅支持小写字母、数 字、中划线(-)、英文句号 (.)。修改密码,请参考重置 云服务器密码登录ECS控制台修 改密码。管理员账户默认 root。	空
system_disk_si ze	num ber	必填	云服务器系统盘大小,磁盘类 型默认为通用型SSD,单位: GB,取值范围为40-1,024,不 支持缩盘。	40

参数名称	类型	是否可 选	参数解释	默认值
charging_mod e	strin g	必填	云服务器计费模式,默认自动 扣费,可选值为:postPaid (按需计费)、prePaid(包年 包月)。	postPaid
charging_unit	strin g	必填	云服务器订购周期类型,仅当 charging_mode为prePaid(包 年/包月)生效,此时该参数为 必填参数。取值范围:month (月),year(年)。	month
charge_period	num ber	必填	云服务器订购周期,仅当 charging_mode为prePaid(包 年/包月)生效,此时该参数为 必填参数。取值范围: charging_unit=month(周期类 型为月)时,取值为1-9; charging_unit=year(周期类型 为年)时,取值为1-3。	1

步骤4 (可选,如果使用华为主账号或admin用户组下的IAM子账户可不选委托)在资源设置 界面中,在权限委托下拉框中选择"rf_admin_trust"委托,单击"下一步"。

图 3-11 资源栈设置



步骤5 在配置确认界面中,单击"创建执行计划"。

图 3-12 配置确认

< 立即创建资源栈			
	***** · · · · · · · · · · · · · ·		
() 资源编排服务使用先费、但资	建线内包含的资源可能需要付费。目前暂不支持在当前页重词价格,您可遗过	创建执行计划 (免费) 后获取努性	島代 .
选择模板			
资源地名称	building-a-deepseek-inference-system		描述 快速乐谱DeepSeek的现在分词
配置参数 🖉			
参数名称	值	英型	描述
vpc_name	building-a-deepseek-inference-system-demo	string	虚拟私有云名称,读模板使用新聞VPC,不允许重名。取值范围:1-54个字符,支持中文,英文字母、数字、_ (下划线) 、・(中划线)、・(件),数以:building-a-deeps…
security_group_name	building-a-deepseek-inference-system-demo	string	安全很高你,读 我 版新建安全组,安全组现则请参考部署指笔进行配置,取道范围:1-64个字符,支持数字、字母、中文、(下划线),-(中划线)、-(点),默以:buildn
ecs_name	building-a-deepseek-inference-system-demo	string	云服务器实例名称,不支持重名,取撤范围:1-64个字符,支持中文,英文字母、数字、_ (下划线),- (中划线),. (点) , 默从: building-a-deepoeek-inference-syste
ecs_flavor	x1.4u.4g	string	云服务器实例规模,支持弹性云服务器 ECS及华为云Fierus 云服务器X实例,Fierus 云服务器X实例规模D参名规则为x1.7u.7g,他和2xCPUs4GB规模D为x1.4u.4g,具体…
ecs_password		string	云虽易器密码,长度为9-26位,密码至少必须检查大写字母、小写字母、数字和特殊字符((@6%/=+(①,/?) 中的三种,仅支持小写字母、数字、中划线(-) ,高文句号…
parameter_quantity	1.5b	string	Deepseek-R1模型争数量大小、支持1.5b、7b、8b、默认1.5b。
system_disk_size	40	number	云细质器系统曲大小、磁曲频型数以为通用型SSD、单位:08、取量压制为40-1,024、不变持缩曲。数认:40、
鼻用预估: 创建执行计划 (免费) 后言	可放取預估適用		(Issterski

步骤6 在弹出的创建执行计划框中,自定义填写执行计划名称,单击"确定"。

图 3-13 创建执行计划



步骤7 单击"部署",并且在弹出的执行计划确认框中单击"执行"。

图 3-14 执行计划

< building-a-deepseek-inference-syst	em				(制除) 更新機能成參数)	С
基本信息 资源 输出 事件 模板	执行计划					
17 <u>7</u>					诸能入关键字	a
执行计划名称ID	秋志	義用預估 ③	创建时间	描述	操作	
executionPlan_20250207_1147_n2tq	创建成功,特部署	兼新期用明细	2025/02/07 11:47:30 GMT+08:00	-	部署書約	

图 3-15执行计划确认

执行计划	\times
您确定要执行该计划吗?	
执行计划名称 状态 创建时间	
executionPlan_20250207_11 创建成功, 2025/02/07 11:47:30 GMT+0.	-
 确定执行后,资源栈会按照该计划更新,并且 会开通模板内的资源,根据资源付 费要求,可能会产生费用。 	
执行取消	\supset

- **步骤8** (可选)如果计费模式选择"包年包月",在余额不充足的情况下(所需总费用请参考表2-2)请及时登录费用中心,手动完成待支付订单的费用支付。
- **步骤9** 待"事件"中出现"Apply required resource success",堆栈部署成功,表示顺利完成资源的下发和部署。堆栈部署成功后,搭建DeepSeek脚本开始执行,耐心等待5-10分钟左右(受网络波动影响)。

图 3-16 部署完成

k building-a-deepseek-inference 基本信息 资源 输出 事件 相	ə-system1 朝板 执行计划			(新) (東新聞板或条数) (C)
				資源名称 > 資給入关键学 Q
antestie 🔶	血体菜面 Δ	甲件描述	资源名称/类型	关联资源ID
2025/02/07 11:51:09 GMT+08:00	日志	Apply required resource success.	-	-
2025/02/07 11:51:06 GMT+08:00	生成完成	husweicloud_compute_instance.compute_instance: Creation complete after 46s [id=5ce7c2dc-b66d.4ca3- 84d6-277541c46fa6]	compute_instance ECS	5ce7c2dc-b66d-4cs3-84d6-277541c46fa6
结束				

3.3 开始使用

该解决方案使用22端口CloudShell方式远程登录云服务器,默认已配置IP地址白名单, 若需远程登录云服务器,可直接使用CLoudShell远程登录。

该解决方案部署成功后,环境初始化预计5-10分钟,受网络、带宽影响,部署时间会 有波动,部署完成方可正常访问。

安全组规则修改(可选)

安全组实际是网络流量访问策略,包括网络流量入方向规则和出方向规则,通过这些 规则为安全组内具有相同保护需求并且相互信任的云服务器、云容器、云数据库等实 例提供安全保护。 如果您的实例关联的安全组策略无法满足使用需求,比如需要添加、修改、删除某个 TCP端口,请参考以下内容进行修改。

- 添加安全组规则:根据业务使用需求需要开放某个TCP端口,请参考添加安全组规则添加入方向规则,打开指定的TCP端口。
- 修改安全组规则:安全组规则设置不当会造成严重的安全隐患。您可以参考修改 安全组规则,来修改安全组中不合理的规则,保证云服务器等实例的网络安全。
- 删除安全组规则:当安全组规则入方向、出方向源地址/目的地址有变化时,或者 不需要开放某个端口时,您可以参考删除安全组规则进行安全组规则删除。

CPU 版

步骤1 登录Flexus云服务器X实例控制台,选择部署DeepSeek的服务器,单击远程登录。

图 3-17 道	5.择远程登录	
华为云Flexus云服务	Flexus 云服 经器 X 实例 ①	TexusX
总元 [2] Flexus应用服务器L实例 C ² NEW		×
Flexus云服务器X实例 NEW	我的xxx例: \$P\$(24) = \$P\$(24) = \$P\$(16) = 中国:雪陽(10) = 亚大·新加坡(4) = \$P\$(上海一(2) = 西歐·费阳一(2) = \$P\$(3兰第6一(1) = 亚大·勒加达(1)	Q
Flexus 云数图MIRDS 🕑	HI MI BI BES Pata (a) ABR: Image: Address of the second se)
	名称の0 盆腔 安全 状态(0) 可…(0) 規胞が開発(0) 温作系统…(0) IP地址(0) 計像…(0) 金虹類目(0) 構造(0) 温作	
	> butling = despoe ○ 6918035-bit5-lace-8c ○ 6918035-bit5-lace-8c ○ 6918035-bit5-lace-8c ○ 6918035-bit5-lace-8c ○ 5018035-bit5-lace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 5018035-bit5-1ace-8c ○ 501805-5c ○ 501805-5c.	•
	忠勝政:1 (10) く	1 >

步骤2 选择使用CloudShell登录,在连接配置页面输入云服务器的密码,单击"连接"。

图 3-18 选择 CloudShell 登录

登录Linux云服务器

Х

▲ 当前未开通21端口、22端口、80端口、443端口、3389端口,会影响登录与使用。请添加安全组规则 查看 端口功能
使用CloudShell登录New! 登录不上
请确保安全组已放通Cloud Shell连接实例使用的端口(默认使用22端口) 优势:操作更流畅,命令支持复制粘贴,支持浏览输出历史和多终端分区布局。了解更多 Cloud Shell登录
其他方式
1、使用控制台提供的VNC方式登录 立即登录

图 3-19 配置连接信息

🌒 CloudShell 蒼賓(V) 终端 帮助(H) 🛛 >_ 🗌	RemoleShell,支持升级为全型DE	的资源访问工具(点击下载)(
◎ 连接远程服务器 ×		
	- 区域: 华北-北京四	~ 刷新
	云服务器: building-a-deepseek-Inference-system-demo	
	● 1(公网) ● 172.16.1.65 (私网)	
	端口: 22	
	用户名: root	
	以证方式: 密码认证	
	• 宏码: [
	会话名称: root@'	
	2 打开远程主机文件树	
	注意: - 为确保连接的安全性,系统将对超过 四外嵴 没有活跃的会话进行自动顺开。 - 为确保连接的安全性,系统将对超过 四外嵴 没有活跃的会话进行自动顺开。 - 诸确也没全世中来源为了doud5helf代理的远程端口 (Set能)从境口为22) 已经分许。 - 当记程登录记程作书码时,通过查看一下和圆板(Setu)、存存带况,请定义云监应在主机异常时通过坦信等多种方式通知。 - 华为云CloudShell不会保存您的密码,请义善保管以防丢失。	

步骤3 在命令行输入"ollama run deepseek-r1:\$parameter_quantity", \$parameter_quantity支持1.5b、7b、8b,请使用3.2快速部署步骤3中参数 "parameter_quantity"的实际填写值,执行命令即可测试预览。

图 3-20 效果测试

- 0.14m3 Turn deegesenk-1:1-2-0- >> 介部一下計 名量王子則種
17.17.11%。 第一户让我介绍一下量子纠缠。首先,我得明确什么是量子纠缠,这是量子力学中的一个重要概念。
我记得量子纠缠是指当两个或更多的粒子同时存在时,它们的行为会受到彼此的强烈影响,即使它们分开得很远,这可能涉及到爱因斯坦相对论中的时间膨胀和空间收缩的问题。
然后,我需要解释它的核心原理,也就显所谓的"信息非可分离性"。就是说,信息在量子纠缠过程中无法单独存在,只能在关职在一起时共享。比如,如果两个粒子 entangle,当其中一个被测量到某个状态, 另一个也必须以将定的状态出现,这样信息就被按照下来。
接下来,我应该考虑量子纠缠的实际应用。它在量子通信中可能用得上,比如量子_key分布(QKO),可以通过纠缠未确很通信的安全性。另外,量子计算中的 entanglement也是关键因素,因为它允许某些运 算比经贵计算机更快完成。
我还应该提到量子纠缠与其他物理学概念的关系,比如爱因斯坦的相对论和量子力学之间的矛盾,以及量子纠缠如何解决了这些矛盾的问题。这可能涉及到解释信息显如何在宇宙中传播的。
最后,总统一下量子纠缠的重要性及其在未来的科技应用,让该者明白它的价值和潜力。
量子纠缠(Quantum EntangLement)是量子力学中的一个重要概念,描述的是当两个或多个粒子同时存在时,它们的行为会受到彼此的强烈影响,即使它们分开得很远。这个现象由爱因斯坦相对论中的时间膨胀与空间收缩问窗所引发,并且在量子力学中被广泛探讨。
44.公园理 量子列模型枚心 原理显 "信息非可分离性",即当两个粒子 entangle 时,它们的性质(如位置、动量等)无法单张确定,必须同时满足某种关系。例如,当一个粒子被测量为处于某个状态时,另一个 entangled 的赵子伯必须以将定的状态出现。这种现象端示了信息在量子系统中是不可分离的,只能通过关联来共享。
功能与应用 1. **置子通信**:量子纠缠在量子通信中被广泛应用。例如量子关键分布(Quantum Key Distribution, QKD)。通过利用 entanglement 未确保通信的安全性。因为即使在量子环境中,两个 Alice 和 Bob 的器号的无法极分篇。
2、**量子计算与算法**:量子纠缠是量子计算中计算某些问题的关键因素。例如、量子 Fourier 变换和 Shor 算法砍载于 entanglement 的存在,以实现指数级的时间发杂度。
3、**盘子检测与测试**:最子纠缝被用于验证最子力学的非可分离性,以及对经典物理模型进行挑战。

🛄 说明

在交互模式下,可以测试 DeepSeek-R1 的多种功能,例如:

- 智能客服: 输入客户常见问题, 如"如何安装软件?"。
- 内容创作: 输入"为一款智能手表撰写广告文案"。
- 编程辅助:输入"用 Python 实现快速排序"。
- 教育辅助: 输入"解释牛顿第二定律"。

----结束

GPU 版

步骤1 登录弹性云服务器 ECS控制台,如下图所示获取3.2快速部署步骤3中的部署的弹性云服务器的弹性公网 IP地址和私有 IP地址。

图 3-21 获取公网 IP

〇 名称: G5r × 添加筛选条件										× () ()
✓ 名称ID 令	监控	安全	状态 ⇔	可 🖯	規格/機像 ⇔	操作	3 (弹性公网IP) 300 Mbit/s 172.16.1.101 (私有IP)	模式 ⇔	标签 ⊖	企业项目 ⊖
building-a-deepseek-Inference-system-demog5 GPU 8080076c-	8	٠	😏 运行中	可用区1	16vCPUs 32GiB g5r.4 Ubuntu 22.04 server 64bi	Linux	1 2 (弹 C 按 172.16.1.101 (私有 C 20	蜀计费 25/02/07 16:30:51 G	-	default

步骤2 浏览器输入http://[弹性公网IP],访问Dify的开发平台。首次登录需注册管理员账号, 依次填写邮箱、账号、密码。

图 3-22 Dify 开发平台

Dify_	🌐 简体中文
设置管理员账户	
管理员拥有的最大权限,可用于创建应用和管理 LLM 供应商等。	
邮箱	
avartaya. a	
用户名	
admin	
密码	
•••••••• 密码必须包含字母和数字,且长度不小于8位	(
设置	
启动 Dify 社区版之前, 请阅读 GitHub 上的 开源协议	

步骤3 依次输入上一步骤中的"邮箱"、"密码"登录Dify平台。

图 3-23 登录 Dify 平台

🜓 Dify_			\$
	嗨,近来可好		
	🔇 欢迎来到 Dify, 登录以继续		
	邮箱		
	_ 163.com		
	密码	忘记密码?	
		6	
	登录		
	使用即代表您同意我们的 使用协议 & 關		
	如来感应;双方切后化成尸,南时住切后1	6次回 这里吕桂只知了	
© 2025 LangGenius, Inc. All rights reserved.			

步骤4 单击右侧"用户名称"下拉并单击"设置"。

图 3-24 设置

Dify_ O star 68,663	0.1928 • 10 20 10.00m ? IM		admin ~
88 全部 口 聊天助手 咒 Agent 63 工作曲		● 我創建的 S 全部标签 ~ A admin	
▲建立用□ 台線中白成用		工作经时 A admin's Worksp	ace >
同一人应用概形的職 2) 号入 DSL 文件		账户	
		用户反馈社区	
		帮助文档	
	未拢影应用	anaxiz 关于	0.15.3 🔹
		费出	

步骤5 单击左侧"模型供应商",在Ollama下单击"添加模型"。

设置	模型供应商		×
工作空间	LLM TEXT EMBEDDING RERANK	LLM	LLM TEXT EMBEDDING RERANK
◆ 模型供应商 ⊕ 成员 ● 数据来源	1(pstage <mark>筹</mark> Upstage 揚供的模型,例如 Solar-1- mini-chat.	aws & Bedrock	together.ai
C API 扩展	LLM TEXT EMBEDDING	LLM) (TEXT EMBEDDING) (RERANK)	LLM
■用 文 语言	≪OpenRouter	🛱 Ollama	WI MISTRAL AI_ MistralAl 提供的模型,例如 open- mistral-7b 和 mistral-large-latest,
		+ 添加模型	
	Groq Cloud GroqCloud 提供对 Groq Cloud API 的访问,其中托管了 LLama2 和 Mixtral 等模型。	lifeplicate	😕 Hugging Face
	LLM SPEECH2TEXT	LLM TEXT EMBEDDING	LLM TEXT EMBEDDING
	Korbits Inference		· 智谱·AI
	LLM TEXT EMBEDDING RERANK	LIM	LLM TEXT EMBEDDING

图 3-25 添加模型

步骤6 模型名称填写3.2快速部署中部署的模型,如"deepseek-r1:7b",基础URL填写**步骤1** 中获取的私网IP地址,端口号11434,单击右下角"保存"并关闭"设置"。

图 3-26 添加 Ollama

添加 Ollama	ြီး Ollama
模型类型*	
	 Text Embedding
模型名称 *	
deepseek-r1:7b	
基础 URL * 私	网IP地址:端口
http://172.16.1.101:11434	
模型类型*	
对话	×
模型上下文长度*	
4096	
最大 token 上限 *	
4096	
如何集成 Ollama 🖸	移除 取消 保存

步骤7选择"创建空白应用",单击"聊天助手"并填写"应用名称&图标",单击右下角 "创建"。

图 3-27 创建空白应用

Dify_ O Star 62,706	© 探索 ● I作室	□知识库 掌工具		A admin
88 全部 回 聊天助手 R Agent S 工作流			我的關節的 🛇 全部标签 🗸 Q 接索	
 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)				
	未找到应用			

图 3-28 创建应用

建空日应用 择应用类型			聊天郎 通过篇 Chatfl	手 单的配置快速描述一个基于 LLM 的对话机器 ww 编排。 了解更多	1人。支持切换为	
(運用)	1	_	•	Drchestrate	() an	to cur a v to Autor O
		T A A A A A A A A A A A A A A A A A A A	4		DEBUG & PREVIEW	Features x
简单配置即可构建基于 LLM 的对话 机器人	具备推理与自主工具调用的智能助 手	用于文本生成任务的 AI 助手	10 10 10		 Helio, Lam L. Loan andwer your questions related Here to make a brand sherd suff. Ber 	Conversation Opener ()
			0		Tige for analyzing competitors in marker © Conversation Openary (688)	Speech to Text
BETA	BETA					File Lipinad () United TALE TYPES West Instance Heads, Deers 3
hatflow 特记忆的复杂多轮对话工作流	工作流 面向单轮自动化任务的编排工作流			- VARABLES + Add		Container Monderation (*)
				O unnieled Jopie A promotion for spice. Science If		Selline-up
				- KNONLEDGE 1; Facult Settings + Add		a benar chut.
5称 & 图标		1-1		Orand Stategy ico recento		Conversation messages can be converted to speech.
mo				Market Research Castle extra control of Market Low	Talk to Difylliot	Citations and Abilitudians Draw source observant and antibulied section of the generated context.
(可远)			D	• ····· • ··· • ··· • ··· •	COC Alexans Coded	Anvailation Reply

步骤8 单击左侧"编排",在右下角"和机器人聊天"中输入内容即可调试预览。

emo ~ ~ · · · · · · · · · · · · · · · · ·	编排				deepseek-r1:7b CH] 幸 发布
	提示词 💿		* 118	调试与	1968.	
	在这里写你的想示词,输入""抓入支量、输入"/ 抓入提示内容块					
및 访问 API					请解释一下什么是重	子纠编
13. 日志与标注						
(9) 当別				۲	嗯,好的,我现在要试着理解一下什么是量子纠缠。从名字上看,"纠缠"可能就是说两个或多个事物 连续在一起吗? 都在量子物理里,这种现象是体是怎样的吗?	皮质密地
					1. 2010年1月1日1日1日1日1日1日1日1日1日1日1日1日1日1日1日1日1日1日1	中的一部 都会被影
	°				可是,我还不太确定这是怎么回事。是不是说如果两个粒子被某种方式"绑定"在一起,那么即使分开	良远,它
	(*) 安景 ①		+ 添加		11.857万位是相关的/1230,如果一个粒子间上编码,另一个数型向下编码,影着在是个将走的方向来?	1.506.00
	亚量能使用产输入表举引入继示调成开场台,你可以试试在继示调中输入 ((input))				然后,我记得有一个实验叫做贝尔实验。对吗? 这些实验是用来给验量子纠缠是否真实存在的。实验 有光子或素其他粒子使力超成两部分,然后分别预量了的对意思重性。如果成果将合量因新组、逐次 现的规思,能必须能则有 \$pooly setion at a distance 这种变象存在。	中通常会 和埃尔代
	上下文	1 召回设置	+ 添加		不过,我有点混乱,因为有时候人们会说"瞬间通讯"或者"超光速传递信息",但实际上这并不意味着	可以传递
	您可以导入现识库作为上下文				信息或影响事物的速度超过光速。那是因为在量子纠缠中,并没有实际的信息被传递,只是概率的状 起来了?	志被关联
					还有,在观量其中一个粒子后,另一个粒子的状态就会确定下来吗? 比如说,如果一个粒子的白腹间 另一个载一走显向下的,这并不是因为我测量了它,而是它们本来就存在这样的关系?	上,那么
					不过,其有点不确定。因为在经典物理中,两个物体的行为也是相互关联的,比如拉弓射箭,拉力相 运动方向和力量也会制成此。那量子持续是不是在微观尺度上的一 分类 似现象吧?或者说,这只是 关联,而并非确定性的国民关系?	同,新的 概率上的
					再想想,如果我有一个量子系统,比如说双缝实验中的粒子,它们通过玻罐时会表现出波动的行为,	但当它们
				和机器	8人聊天	
_				0	加肥开度	管理 →

图 3-29 调试与预览

----结束

🛄 说明

拓展应用请参考:

- 华为云ModelArts Studio,助力快速搭建专属大模型
- 探索Dify: 开启AI应用开发的新篇章

3.4 快速卸载

步骤1 登录资源编排 RFS资源栈,找到该解决方案创建的资源栈,单击资源栈名称右侧"删除"按钮。

Х

图 3-30 一键卸载

资源编排服务 📉	資源機 ①					19 使用描用	立即创建资源代
並近 資源 稿	资源线 (1)				building-a-deepseek-inference-	system1	X Q O C
资源线集	资源线名称/ID	秋窓 🖓	描述	创建时间 🔶	JEMANIA ↔	操作	
マスト マスト マス	building-a-deepseek-inference-system1	部署成功	快速描建DeepSeek推理系统	2025/02/07 11:49:56 GMT+08:00	2025/02/07 11:51:09 GMT+08:00	制除 更新	

步骤2 在弹出的删除资源栈确定框中,删除方式选择删除资源,输入Delete,单击"确定",即可卸载解决方案。

图 3-31 删除资源栈确认

删除资源栈

您确定要删除该资源栈及资源栈内资源吗? 删除后不能恢复,请谨慎操作 资源栈名称 状态 创建时间 building-a-deepseek-inference-sy... 部署成功 2025/02/07 11:49:56 GMT+08:00 资源列表 (7) building-a-deepseek-inference-system-d... 虚拟私有云 生成完成 虚拟私有云 生成完成 с 虚拟私有云 生成完成 building-a-deepseek-inference-system-d... 虚拟私有云 生成完成 building-a-deepseek-inference-system-d.. 弹性公网IP 生成完成 (building-a-deepseek-inference-system-d... 生成完成 虚拟私有云 🔵 保留资源 (仅删除资源栈) 删除方式 💿 删除资源 如您确认要删除资源栈或其资源,请输入Delete以确认删除 Delete

确定

取消



名词解释

- Flexus云服务器X实例:Flexus云服务器X实例是新一代面向中小企业和开发者打造的柔性算力云服务器。Flexus云服务器X实例功能接近ECS,同时还具备独有特点,例如Flexus云服务器X实例具有更灵活的vCPU内存配比、支持热变配不中断业务变更规格、支持性能模式等。
- 弹性云服务器 ECS: 是一种云上可随时自助获取、可弹性伸缩的计算服务,可帮助您打造安全、可靠、灵活、高效的应用环境。
- 虚拟私有云 VPC:是用户在华为云上申请的隔离的、私密的虚拟网络环境。用户可以基于VPC构建独立的云上网络空间,配合弹性公网IP、云连接、云专线等服务实现与Internet、云内私网、跨云私网互通,帮您打造可靠、稳定、高效的专属云上网络。
- 弹性公网IP EIP:提供独立的公网IP资源,包括公网IP地址与公网出口带宽服务。 可以与弹性云服务器、裸金属服务器、虚拟IP、弹性负载均衡、NAT网关等资源 灵活地绑定及解绑,提供访问公网和被公网访问能力。



表 5-1 修订记录

发布日期	修订记录
2025-02-07	第一次正式发布。
2025-02-19	上线华北-北京四、华南-广州GPU一键部 署。