

解决方案实践

快速搭建 DeepSeek 推理系统

文档版本 1.0.0
发布日期 2025-05-23



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



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

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.....	6
3.1 准备工作.....	6
3.2 快速部署.....	8
3.3 开始使用.....	15
3.4 快速卸载.....	21
4 附录.....	23
5 修订记录.....	24

1 方案概述

应用场景

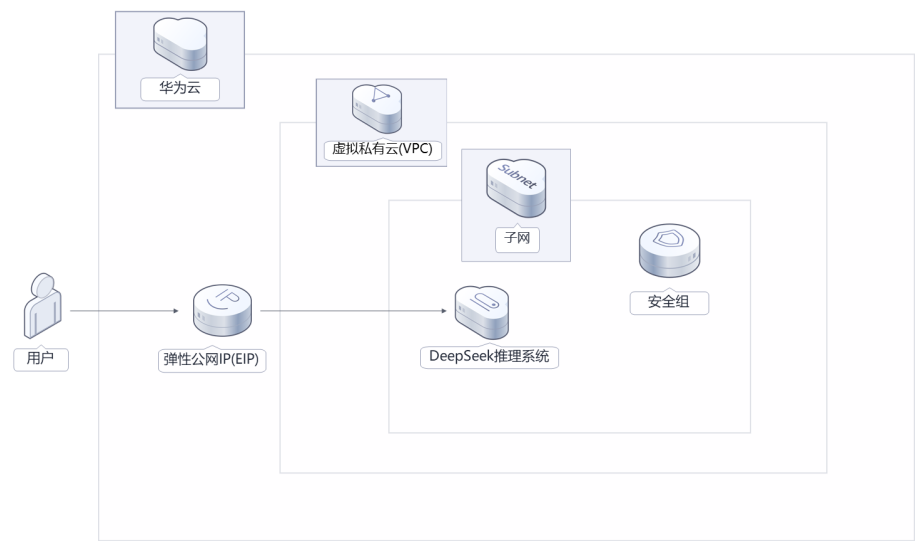
该解决方案帮助您在华为云Flexus云服务器X实例（弹性云服务器 ECS）上快速搭建 DeepSeek-R1轻量级推理系统。DeepSeek-R1是一个高性能的AI推理模型，专注于数学、代码和自然语言推理任务，通过Ollama在云服务器中部署DeepSeek-R1蒸馏版轻量模型，快速打造您的私人AI助手，主要适用如下场景：

- 自然语言处理（NLP）：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

- 在开始解决方案部署之前，请确认您已经拥有一个可以访问该区域的华为账号且已开通华为云。
- 如果选择计费模式选择“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”手动完成支付。
- 如果选用IAM委托权限部署资源，请确保使用的华为云账号有IAM的足够权限，具体请参考[创建rf_admin_trust委托（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，将采用当前登录用户的权限进行部署。

2 资源和成本规划

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[价格详情](#)，实际收费以账单为准。

表 2-1 资源和成本规划（按需计费）

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：亚太-新加坡	1	USD 0.00
子网 Subnet	building-a-deepseek-Inference-system-demo-subnet	- 子网网段：172.16.1.0/24 - 区域：亚太-新加坡	1	USD 0.00
安全组 SecurityGroup	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：119.8.185.245/32 - 区域：亚太-新加坡	1	USD 0.00

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：USD 0.13/小时 • 区域：亚太-新加坡 • 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	USD 93.66
弹性公网IP EIP	building-a-deepseek-Inference-system-demo-eip	• 按需计费：USD 0.12/GB • 区域：亚太-新加坡 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s	1	USD 0.12/GB
合计	-	-		USD 93.66 + 弹性公网IP EIP费用

表 2-2 资源和成本规划（包年包月）

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：亚太-新加坡	1	USD 0.00
子网 Subnet	building-a-deepseek-Inference-system-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：亚太-新加坡	1	USD 0.00

华为云服务	资源名称	配置示例	数量	每月预估花费
安全组 SecurityGroup	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: 119.8.185.245/32 • 区域: 亚太-新加坡	1	USD 0.00
Flexus云服务器X实例	building-a-deepseek-Inference-system-demo	• 包年包月 • 区域: 亚太-新加坡 • 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	USD 69.63
弹性公网IP EIP	building-a-deepseek-Inference-system-demo-eip	• 按需计费: USD 0.12/GB • 区域: 亚太-新加坡 • 线路: 动态BGP • 公网带宽: 按流量计费 • 带宽大小: 300Mbit/s	1	USD 0.12/GB
合计	-	-		USD 69.63 + 弹性公网IP EIP费用

3 实施步骤

- 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

当您使用首次使用华为时注册的账号，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[控制台管理](#)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 统一身份认证菜单



步骤2 进入“委托”菜单，搜索“rf_admin_trust”委托。

图 3-2 委托列表



- 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”，输入“RFS”，单击“完成”。

图 3-3 创建委托

委托 / 创建委托

★ 委托名称

rf_admin_trust

★ 委托类型

☐ 普通账号
将账号内资源的操作权限委托给其他华为云账号。

☒ 云服务
将账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

完成

取消

步骤4 单击“立即授权”。

图 3-4 委托授权

✓ 授权

✕

是否立即为当前创建的委托进行授权?

取消

立即授权

步骤5 在搜索框中输入”Tenant Administrator”并勾选搜索结果，单击“下一步”。

图 3-5 选择策略



步骤6 选择“所有资源”，并单击“确定”完成配置。

图 3-6 设置最小授权范围



步骤7 “委托”列表中出现“rf_admin_trust”委托则创建成功。

图 3-7 委托列表



----结束

3.2 快速部署

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。

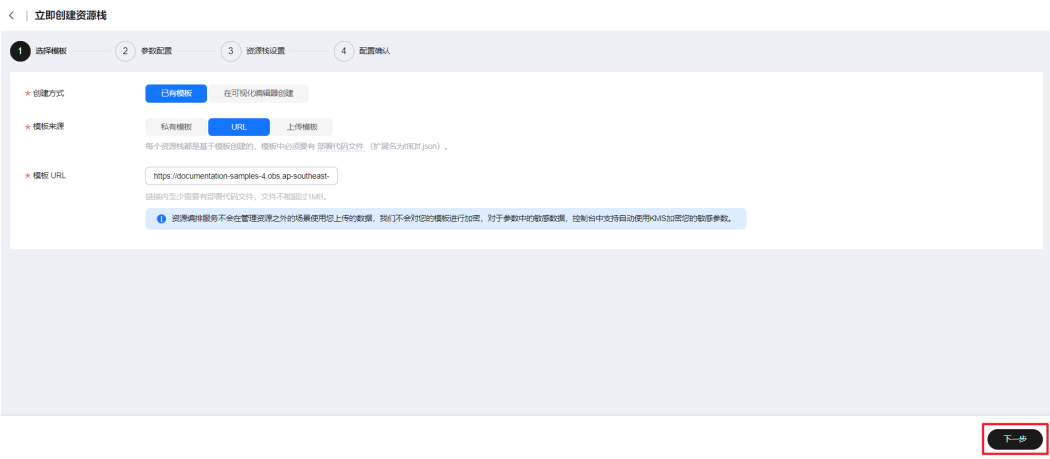
步骤1 登录华为云[解决方案实践](#)，选择“快速搭建DeepSeek推理系统”，以单击“一键部署（CPU版）”为例，跳转至解决方案创建资源栈界面。

图 3-8 解决方案实施库



步骤2 在选择模板界面中，单击“下一步”。

图 3-9 选择模板



步骤3 在配置参数界面中，参考表3-1（GPU版请参考表2）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

立即创建资源栈

✔ 创建成功

1 参数配置

3 资源栈设置

4 配置确认

* 资源栈名称

building-a-deepspeak-inference-system

资源栈名称以中文或英文开始，支持中文、大小写英文、数字、下划线，128个字符之内。资源栈名称不能重复。

描述

快速搭建DeepSeek推理系统（CPU版）

22/105

配置参数

请输入关键字搜索参数名称

🔍

🔗 新增模板要求时部分资源增加

参数名称	值	类型	描述
* vpc_name	building-a-deepspeak-inference-system-demo	string	虚拟私有云 VPC 名称。该模板使用新建 VPC，不为空参数。配置范围：1-64 个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。默认：building-a-dee-
* security_group_name	building-a-deepspeak-inference-system-demo	string	安全组名称。该模板新建安全组。安全组 ID 请参见资源组所属地域的资源组。配置范围：1-64 个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。默认：build-
* ecs_name	building-a-deepspeak-inference-system-demo	string	云服务器实例名称。不支持空格。配置范围：1-64 个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。默认：building-a-deepspeak-inference-system-d-
* distlib_model	DeepSeek-R1-Distill-Qwen-1.5B	string	DeepSeek-R1-Distill 模型。支持 DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B。默认 DeepSeek-R1-Distill-Qwe-
* ecs_flavor	x1.4g.4g	string	云服务器规格名称。支持弹性云服务器 ECS 及华为云 Flexus 云服务器 E 实例。Flexus 实例规格 ID 命名规则为 1.7n.7g，例如 C6a40.40规格 ID 为 1.2u.4g。默认：-
* ecs_password		string	云服务器密码。长度需 8-26 位，必须至少包含 6 个大写字母、小写字母、数字和特殊字符（如 %、*、+、@、_、/）中的 3 种，仅支持小写字母、数字、中划线（-）、.
* system_disk_size	40	number	云服务器系统盘大小。磁盘类型默认为通用型 SSD，单位：GB。配置范围为 40-1,024，不支持增量盘。默认：40。

上一步

下一步

表 3-1 参数填写说明（CPU 版）

参数名称	类型	是否可选	参数解释	默认值
vpc_name	string	必填	虚拟私有云 VPC名称，该模板使用新建VPC，不允许重名。取值范围：1-54个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。	building-a-deepseek-Inference-system-demo
security_group_name	string	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
ecs_name	string	必填	云服务器实例名称，不支持重名。取值范围：1-64个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。	building-a-deepseek-Inference-system-demo
distilled_model	string	必填	DeepSeek-R1-Distill模型。支持DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Qwen-8B。	DeepSeek-R1-Distill-Qwen-1.5B

参数名称	类型	是否可选	参数解释	默认值
ecs_flavor	string	必填	云服务器实例规格，支持弹性云服务器 ECS及华为云Flexus云服务器X实例。FlexusX实例规格ID命名规则为x1.?u.?g，例如2vCPUs4GiB规格ID为x1.2u.4g，具体FlexusX实例规格请参考控制台。ECS规格信息具体请参考官网 弹性云服务器规格清单 。1.5B推荐使用4vCPUs4GiB及以上规格；7B和8B推荐使用16vCPUs16GiB及以上规格。	x1.4u.4g
ecs_password	string	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。管理员账户默认root。	空
system_disk_size	number	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
charging_mode	string	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	string	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
charging_period	number	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

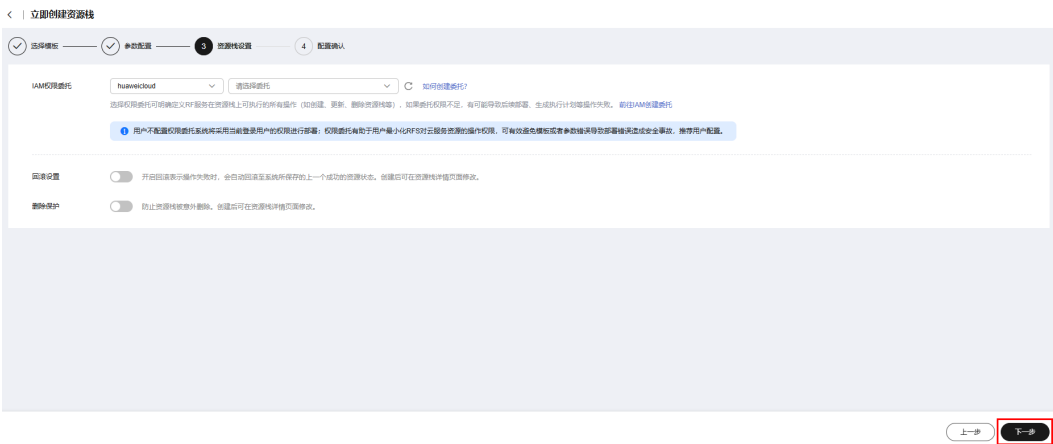
表 3-2 参数填写说明（GPU 版）

参数名称	类型	是否可选	参数解释	默认值
vpc_name	string	必填	虚拟私有云 VPC名称，该模板使用新建VPC，不允许重名。取值范围：1-54个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。	building-a-deepseek-Inference-system-demo
security_group_name	string	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
ecs_name	string	必填	云服务器实例名称，不支持重名。取值范围：1-64个字符，支持字母、数字、_（下划线）、-（中划线）、.（点）。	building-a-deepseek-Inference-system-demo
distilled_model	string	必填	DeepSeek-R1-Distill模型。支持DeepSeek-R1-Distill-Qwen-1.5B、DeepSeek-R1-Distill-Qwen-7B、DeepSeek-R1-Distill-Llama-8B、DeepSeek-R1-Distill-Qwen-14B。	DeepSeek-R1-Distill-Qwen-1.5B
dify_enable	string	必填	是否安装Dify。Dify提供了开箱即用的web应用程序来与模型交互。	enable
ecs_flavor	string	必填	弹性云服务器实例规格，须使用GPU加速型。具体规格信息请参考官网 弹性云服务器规格清单 （在一键部署前请确保该规格在该区域下可用）。	pi2.2xlarge.4
ecs_password	string	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。管理员账户默认root。	空
system_disk_size	number	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_mode	string	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	string	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
charging_period	number	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

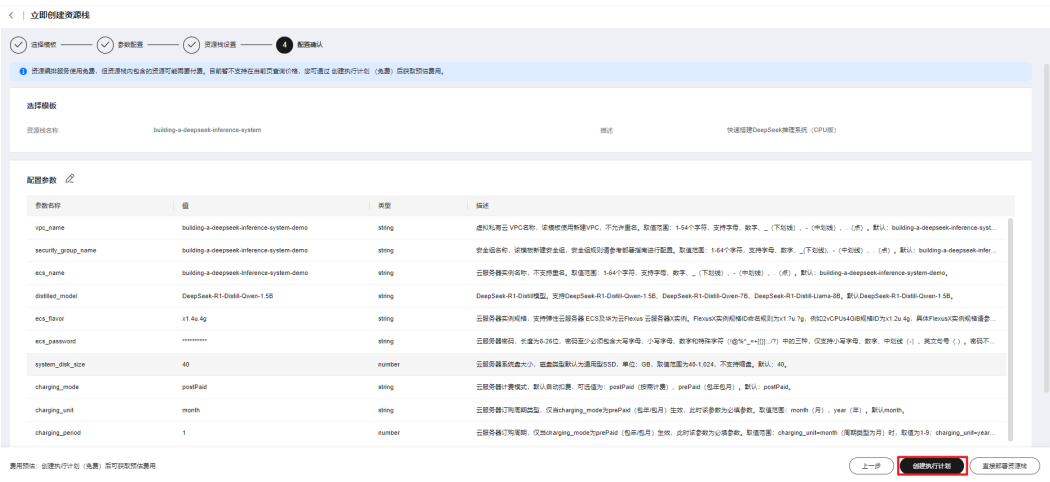
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

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



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

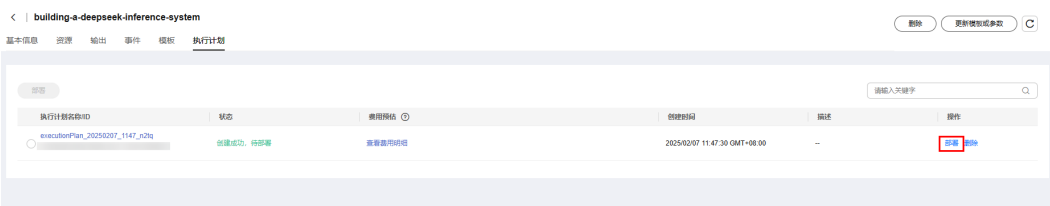


图 3-15 执行计划确认



- 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



----结束

3.3 开始使用

说明

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 选择远程登录

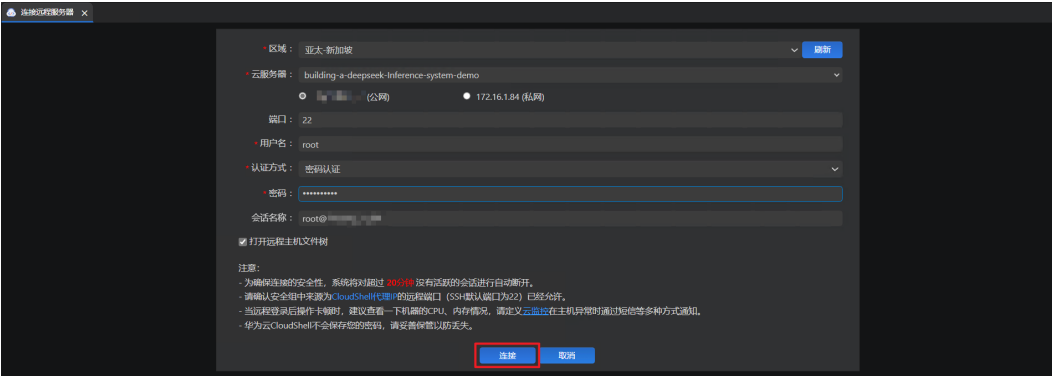


步骤2 选择使用CloudShell登录，在连接配置页面输入云服务器的密码，单击“连接”（部分区域不支持CloudShell，请使用VNC登录）。

图 3-18 选择 CloudShell 登录

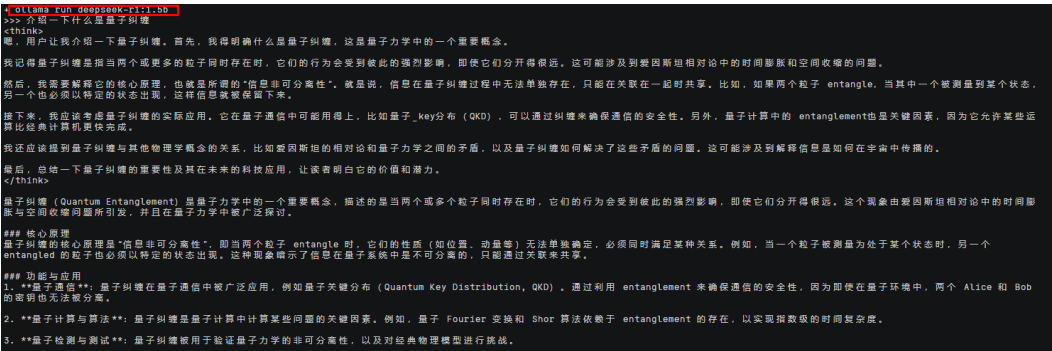


图 3-19 配置连接信息



步骤3 在命令行输入“ollama run deepseek-r1:\$distilled_model” distilled_model支持 1.5b、7b、8b，请对照3.2快速部署步骤3中填写的参数“distilled_model”的值填写（如：DeepSeek-R1-Distill-Qwen-1.5B 此处填写1.5b）执行命令即可测试预览。

图 3-20 效果测试



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，如下图所示获取该解决方案创建的弹性云服务器 ECS 的弹性公网 IP地址和私有 IP地址。

图 3-21 获取公网 IP



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



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



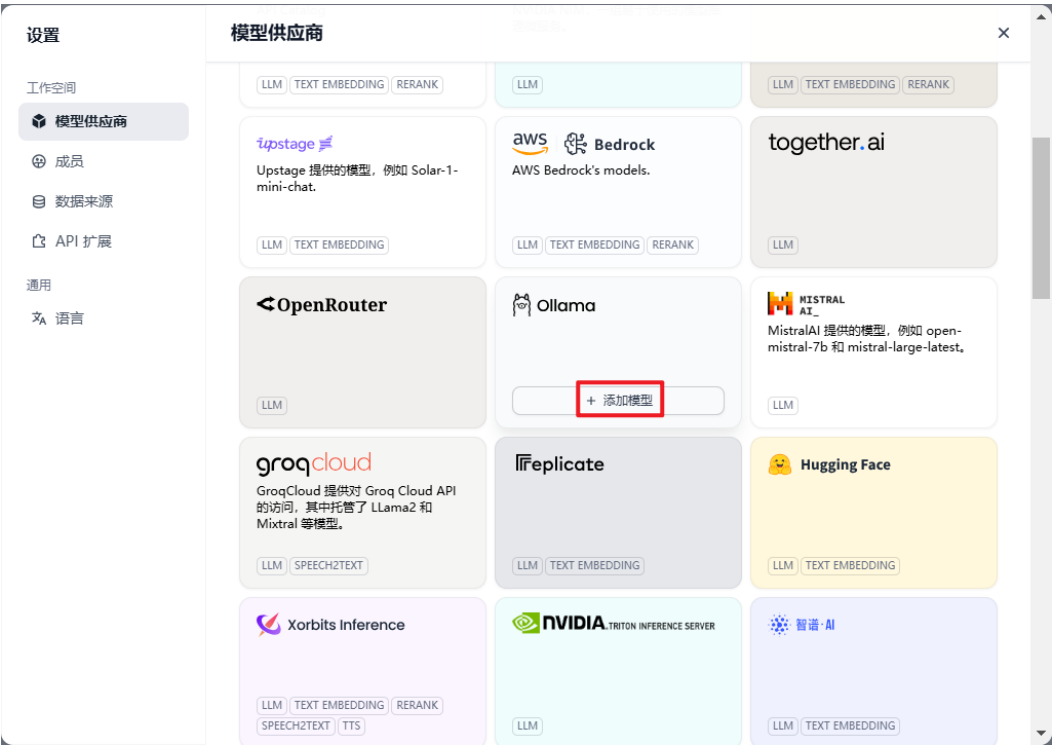
步骤4 单击右侧“用户名称”下拉并单击“设置”。

图 3-24 设置



步骤5 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-25 添加模型



步骤6 模型名称填写3.2快速部署步骤3中部署的模型，如“deepseek-r1:7b”，基础URL填写步骤1中获取的公网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



模型类型 *

☒ LLM

☐ Text Embedding

模型名称 *

deepseek-r1:7b

基础 URL *

http://172.16.1.101:11434

私网IP地址：端口

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 62,706

探索 工作台 知识库 工具

admin

全部 聊天助手 Agent 工作流

创建应用

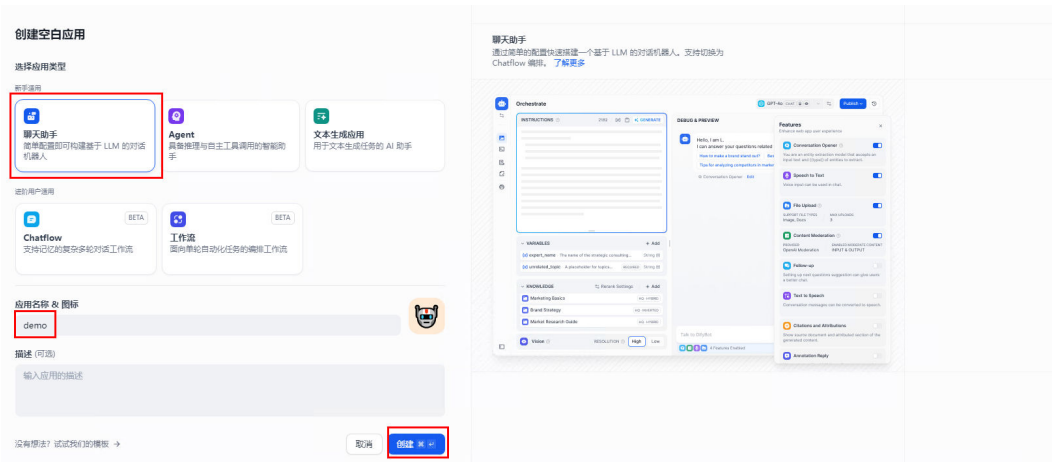
☒ 创建空白应用

☐ 从应用模板创建

☐ 导入 DSL 文件

未找到应用

图 3-28 创建应用



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。

图 3-29 调试与预览

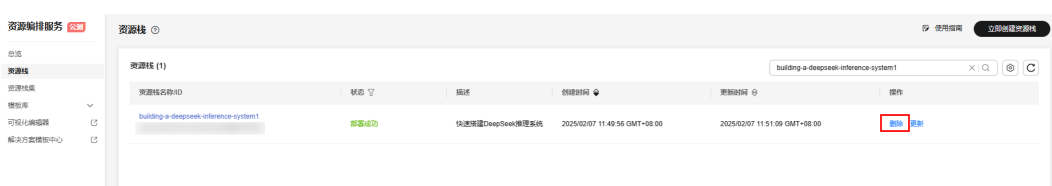


----结束

3.4 快速卸载

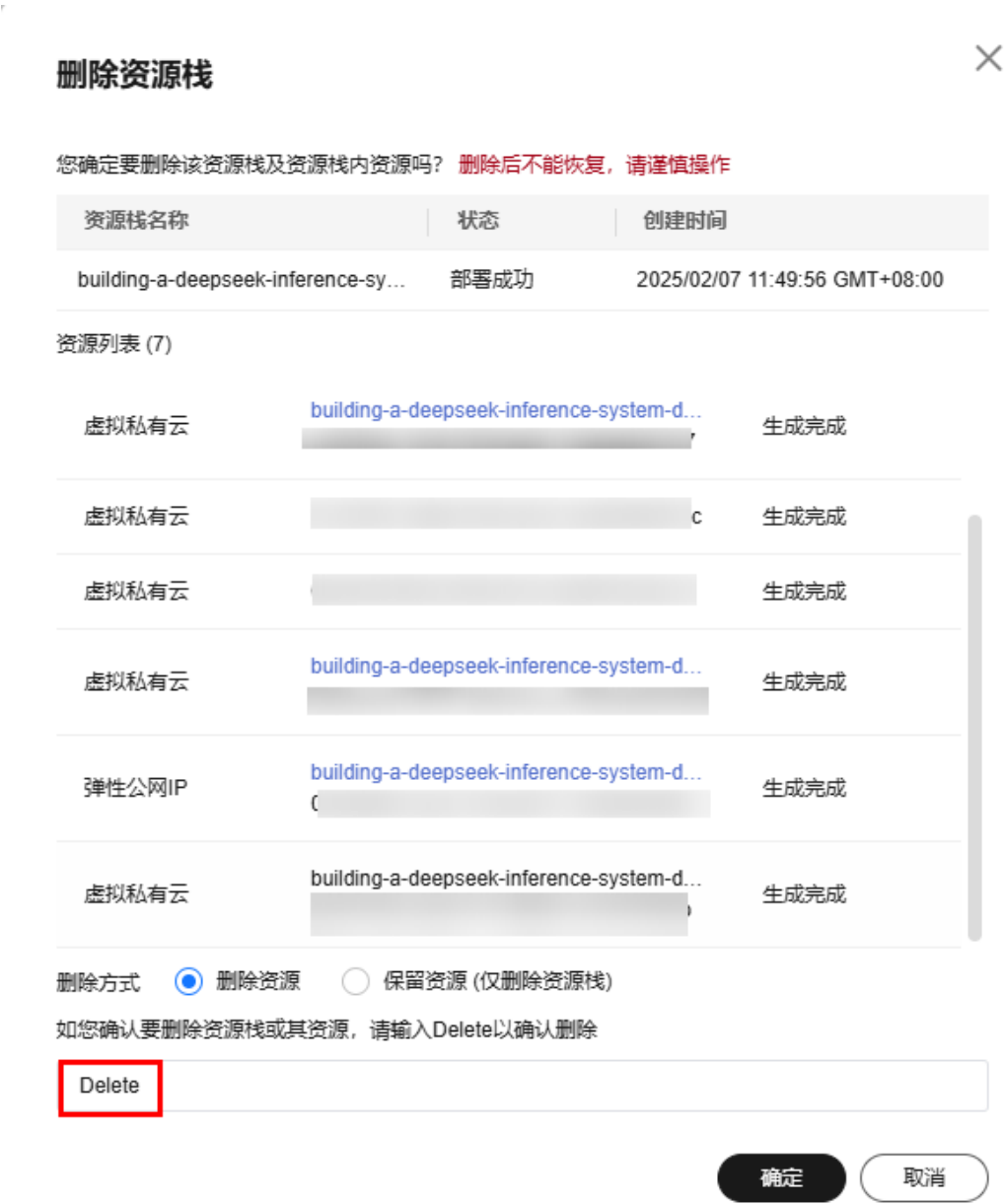
步骤1 登录[资源编排 RFS资源栈](#)，找到该解决方案创建的资源栈，单击资源栈名称右侧“删除”按钮。

图 3-30 一键卸载



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



----结束

4 附录

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。

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2025-05-23	第一次正式发布。