

解决方案实践

快速部署 Qwen3 系列模型

文档版本 1.0.0
发布日期 2025-04-29



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..... 9

3.3 开始使用..... 14

3.4 快速卸载..... 19

4 附录..... 21

5 修订记录..... 22

1 方案概述

应用场景

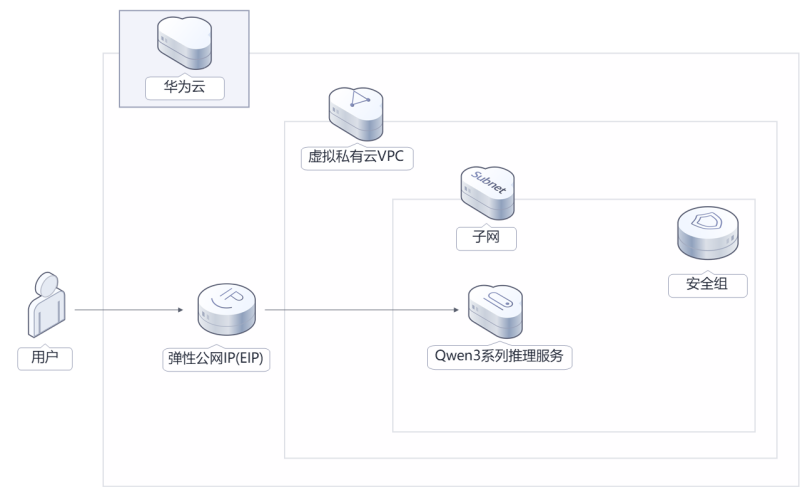
Qwen3是 Qwen 系列中最新一代大型语言模型。推理能力大幅增强，在数学、代码生成、常识逻辑推理等方面表现优异，支持 100 多种语言和方言，具有强大的多语言指令跟踪和翻译能力，擅长创意写作、角色扮演、多轮对话和指令遵循，提供更自然、更具吸引力和身临其境的对话体验。相比以往模型体积更小，性能极具竞争力，即使是像 Qwen3-4B 这样的微型模型也能与 Qwen2.5-72B-Instruct 的性能相媲美。通过 Ollama在云服务器中部署参数量14b或以下的某个Qwen3大模型，快速打造您的私人AI助手，主要适用如下场景：

- 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**弹性云服务器 ECS（GPU加速型）**实例上快速部署Qwen3系列模型。

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**弹性云服务器 ECS（GPU加速型）**，用于部署Qwen3系列模型。
- 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

- 性能提升
Qwen 3是 Qwen 系列中最新一代大型语言模型，相比以往大模型，体积更小的同时，性能极具竞争力。
- 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Qwen3系列大模型的部署。

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。

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	qwen3-large-language-model-family-demo	- VPC网段：172.16.0.0/16 - 区域：华北-北京四	1	0.00元
子网 Subnet	qwen3-large-language-model-family-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 SecurityGroup	qwen3-large-language-model-family-demo	- 允许ping：0.0.0.0/0 - 开放端口22允许 Cloud Shell 登录：121.36.59.153/32 - 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	每月预估花费
弹性云服务器 ECS	qwen3-large-language-model-family-demo	• 按需计费：7.42 元/小时 • 区域：华北-北京四 • 规格：GPU加速型 pi2.2xlarge.4 8vCPUs 32GiB GPU显卡: 1 * NVIDIA Tesla T4 / 1 * 16GiB • 镜像：Ubuntu 22.04 server 64bit with Tesla Driver 470.223.02 and CUDA 11.4 • 系统盘：通用型 SSD 100GB	1	5342.40元
弹性公网IP EIP	qwen3-large-language-model-family-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		-		5342.40元 + 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	qwen3-large-language-model-family-demo	• VPC网段：172.16.0.0/16 • 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	每月预估花费
子网 Subnet	qwen3-large-language-model-family-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 SecurityGroup	qwen3-large-language-model-family-demo	- 允许ping：0.0.0.0/0 - 开放端口22允许Cloud Shell 登录：121.36.59.153/32 - 区域：华北-北京四	1	0.00元
弹性云服务器 ECS	qwen3-large-language-model-family-demo	- 包年包月 - 区域：华北-北京四 - 规格：GPU加速型 pi2.2xlarge.4 8vCPUs 32GiB GPU显卡: 1 * NVIDIA Tesla T4 / 1 * 16GiB - 镜像：Ubuntu 22.04 server 64bit with Tesla Driver 470.223.02 and CUDA 11.4 - 系统盘：通用型 SSD 100GB	1	3584.79元
弹性公网IP EIP	qwen3-large-language-model-family-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		-		3584.79元 + 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[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 控制台管理界面



图 3-2 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 3-3 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”，选择“RFS”，单击“下一步”。

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

1 选择策略

2 设置委托授权范围

3 完成

委托“rf_admin_trust”将拥有指定策略

策略已选 11

从其他区域策略库复制策略

全部策略

所有云服务

Tenant Administrator

名称

类型

系统角色

☒

Tenant Administrator

全部云服务管理权限（即IAM管理权限）

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

1 选择策略

2 设置委托授权范围

3 完成

根据当前您选择的策略，系统推荐以下授权范围方案，更便于您最小化授权，可进行编辑。了解如何根据您应用场景选择合适授权范围方案

选择授权范围方案

☒ 所有资源

授权后，IAM用户可以根据权限使用帐号中所有资源，包括企业项目、区域项目和全局服务资源。

展开其他方案

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

委托 0						创建委托
删除 您还可以创建48个委托。						全部类型 请输入委托名称进行搜索 Q
☐	委托名称ID 正	委托对象 正	委托时长 正	创建时间 正	描述 正	操作
☐	rt_admin_host	云服务 RFS	永久	2023/05/31 11:07:56 GMT+08:00	-	授权 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“快速部署Qwen3系列模型”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“快速部署Qwen3系列模型”，支持区域下拉选择部署的区域（以北京四为例），单击“一键部署（GPU部署）”，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

该解决方案帮助您在华为云弹性云服务器 ECS（GPU加速型）实例上快速部署Qwen3系列模型。

```
graph LR
    User((用户)) --> EIP[弹性公网IP]
    EIP --> VPC[虚拟私有云VPC]
    VPC --> Subnet[子网]
    Subnet --> SG[安全组]
    SG --> Qwen3[Qwen3系列推理服务]
```

快速部署Qwen3系列模型

版本: 1.0.0
上次更新日期: 2025年4月
来源: 由华为云构建
部署: 预计10分钟
卸载: 预计5分钟

预估成本

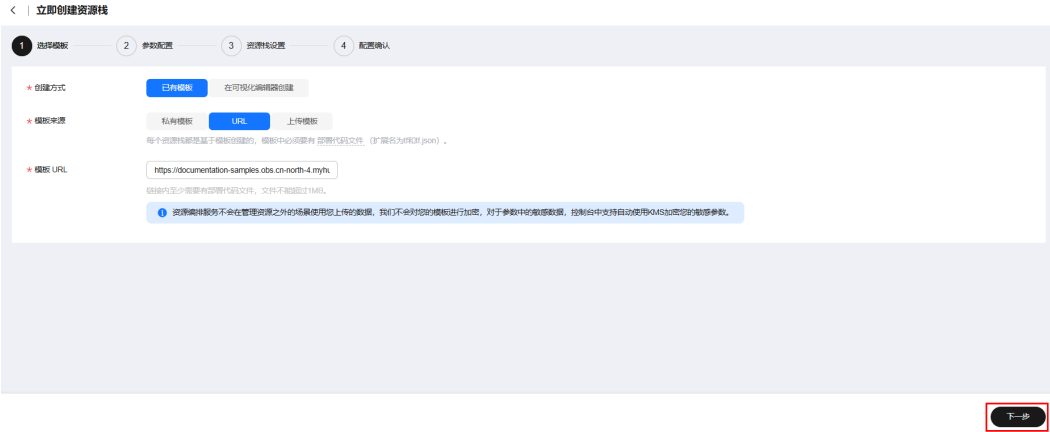
支持区域: 华北-北京四

[查看部署指南](#)

[一键部署（GPU部署）](#)

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

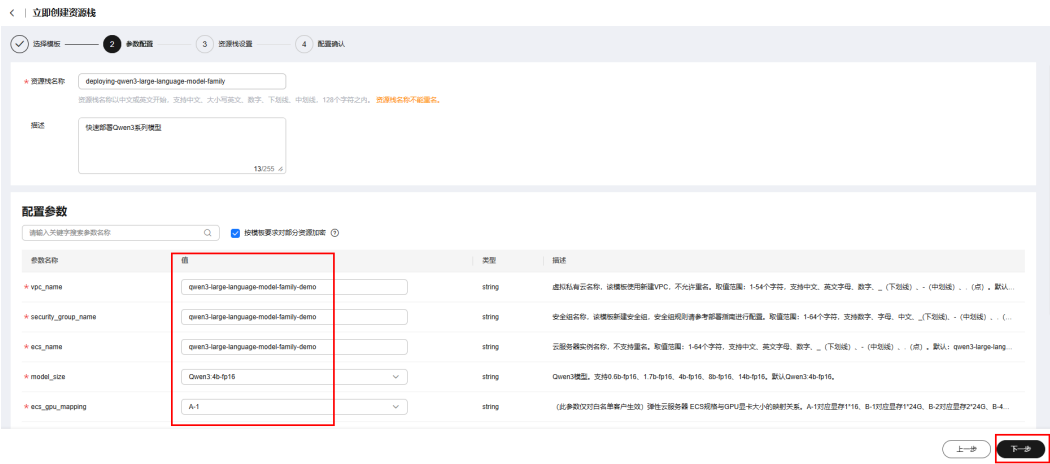


表 3-1 参数填写说明

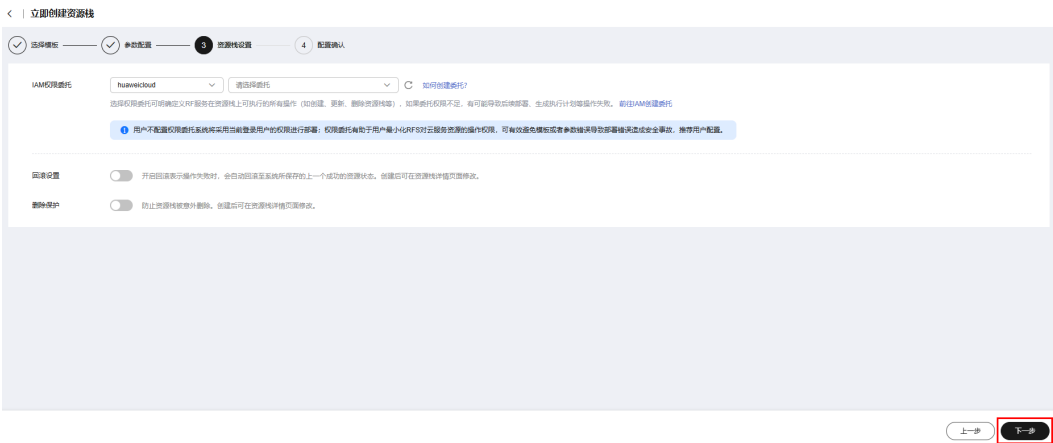
参数名称	类型	是否可选	参数解释	默认值
vpc_name	string	必填	虚拟私有云名称，该模板使用新建VPC，不允许重名。取值范围：1-54个字符，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	qwen3-large-language-model-family-demo
security_group_name	string	必填	安全组名称，该模板新建安全组，请参考 安全组规则修改（可选） 进行配置。取值范围：1-64个字符，支持字母、数字、中文、下划线（_）、中划线（-）、英文句号（.）。	qwen3-large-language-model-family-demo

参数名称	类型	是否可选	参数解释	默认值
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，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	qwen3-large-language-model-family-demo
model_size	string	必填	Qwen3模型。支持0.6b-fp16、1.7b-fp16、4b-fp16、8b-fp16、14b-fp16。	Qwen3:4b-fp16
ecs_gpu_mapping	string	必填	（此参数对白名单客户生效）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。	A-1
ecs_flavor	string	必填	弹性云服务器实例规格，须使用GPU加速型，推荐显存大小0.6b-fp16：4GB及以上，1.7b-fp16、4b-fp16：8GB及以上，8b-fp16：16GB及以上，14b-fp16：32GB及以上。此参数优先级高于{ecs_gpu_mapping}。规格信息具体请参考官网 弹性云服务器规格清单 。	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。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_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。	100
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

参数名称	类型	是否可选	参数解释	默认值
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

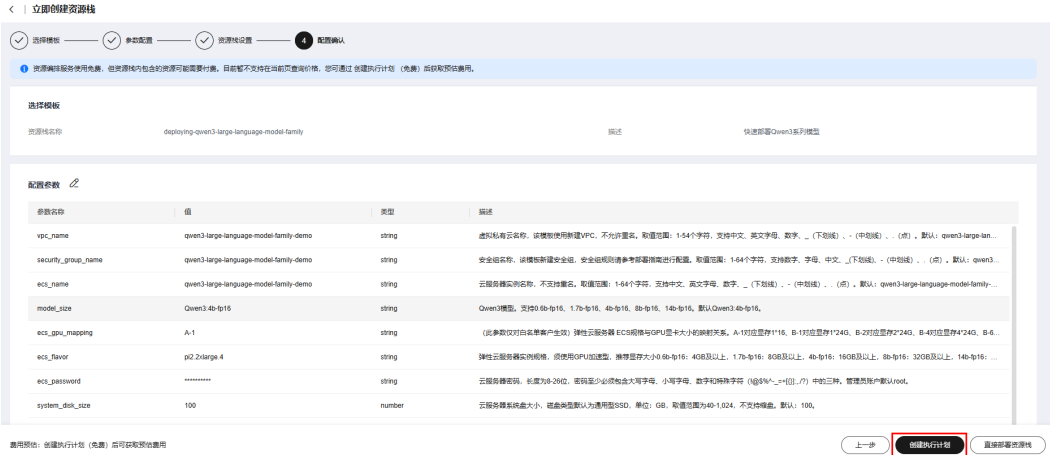
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

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

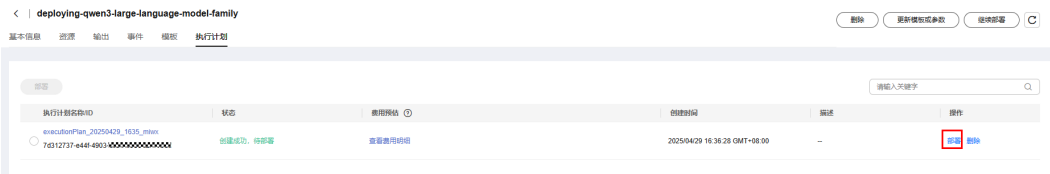
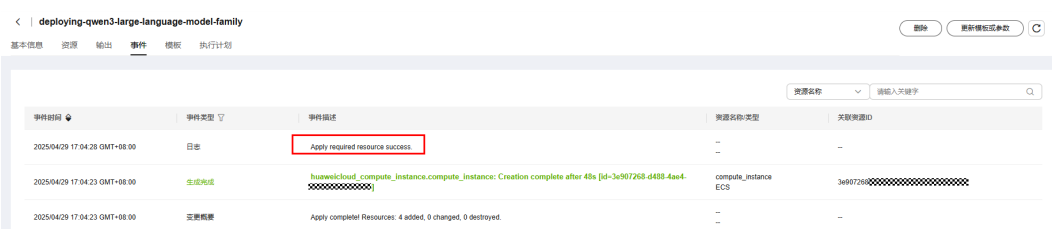


图 3-15 执行计划确认



- 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（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



----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。

使用步骤（以对接到 Dify 平台为例）

注意：本方案只基于Ollama部署模型，不涉及安装Dify，安装Dify请参考[快速搭建 Dify-LLM应用开发平台](#)。

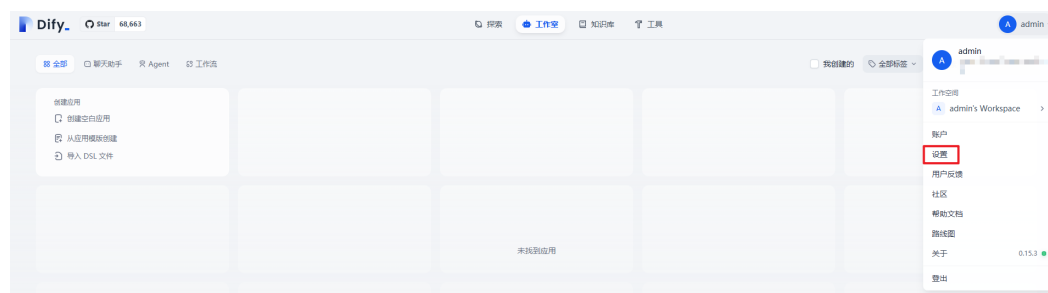
步骤1 登录[弹性云服务器 ECS控制台](#)，如下图所示获取[3.2快速部署步骤3](#)中的部署的弹性云服务器的弹性公网 IP地址。

图 3-17 获取公网 IP



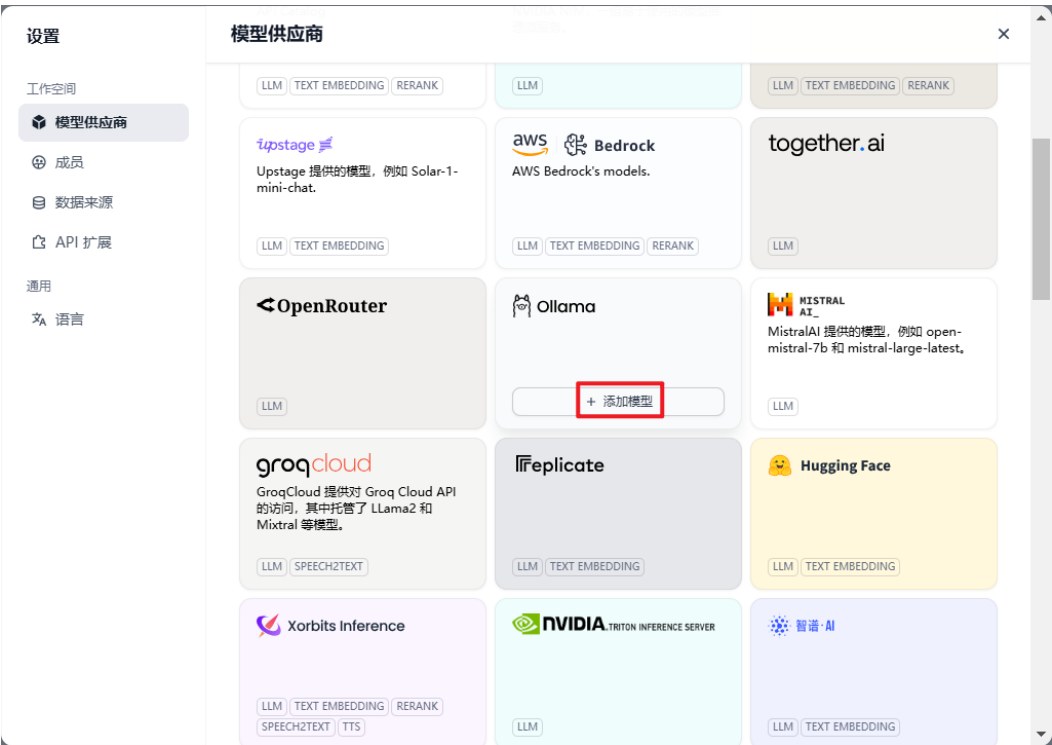
步骤2 登录Dify平台，单击右侧“用户名称”下拉并单击“设置”。

图 3-18 设置



步骤3 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-19 添加模型



步骤4 模型名称填写**3.2快速部署步骤3**中实际部署的模型对应到Ollama平台的名称，如“qwen3:8b-fp16”，基础URL填写**步骤1**中获取的公网IP地址，端口号11434（事先在安全组放通Dify平台IP，端口11434，参考[修改安全组规则](#)），单击右下角“保存”并关闭“设置”。

图 3-20 添加 Ollama

添加 Ollama



模型类型 *

☒ LLM

☐ Text Embedding

模型名称 *

qwen3:8b-fp16

基础 URL *

公网IP:11434

http://[redacted]:11434

模型类型 *

对话

模型上下文长度 *

4096

最大 token 上限 *

4096

是否支持 Vision

☐ 是

☒ 否

[如何集成 Ollama](#)

移除

取消

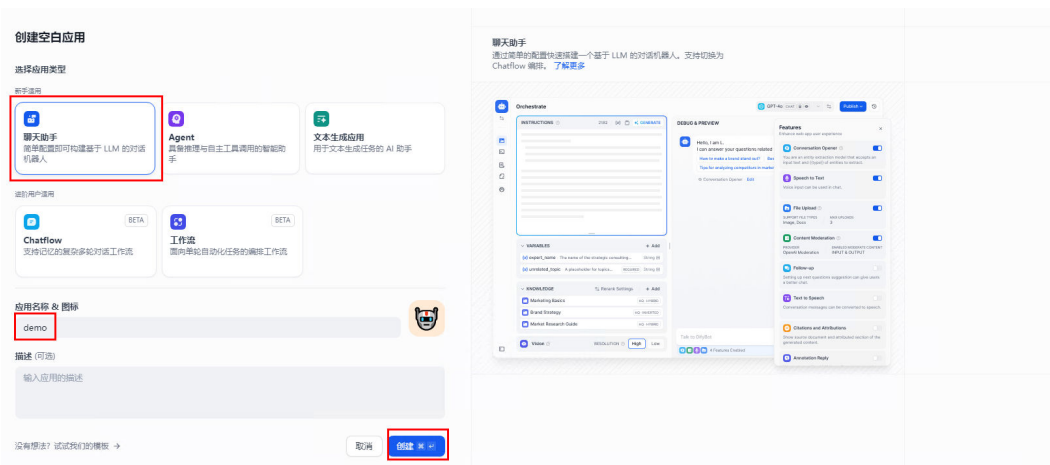
保存

步骤5 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 3-21 创建空白应用



图 3-22 创建应用



步骤6 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 3-23 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。

图 3-24 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 3-25 删除资源栈确认



----结束

4 附录

名词解释

- 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
2025-04-29	第一次正式发布。