

解决方案实践

快速搭建 Dify-LLM 应用开发平台

文档版本 1.0
发布日期 2025-02-26



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	5
3.1 准备工作.....	5
3.2 快速部署.....	8
3.3 开始使用.....	15
3.4 快速卸载.....	20
4 附录	22
5 修订记录	23

1 方案概述

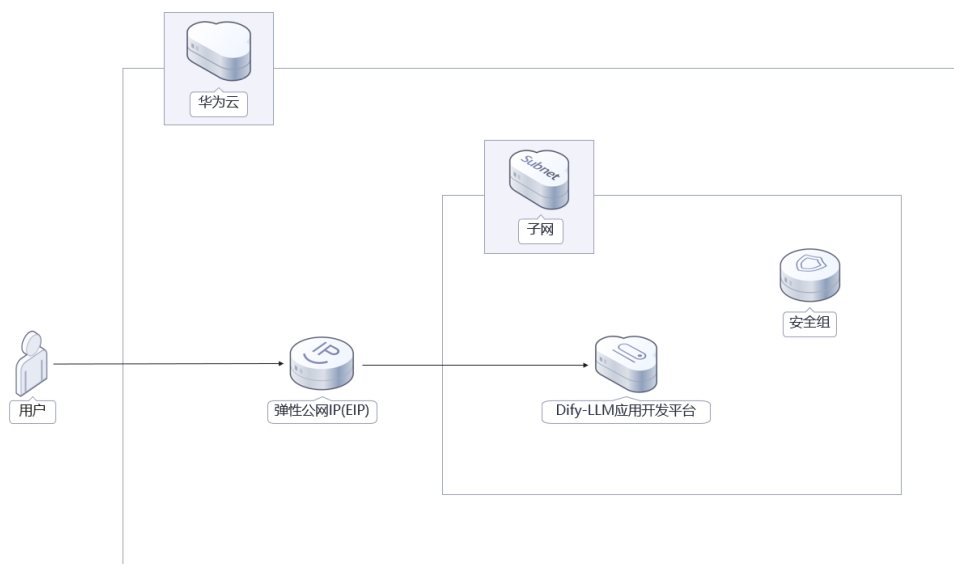
应用场景

该解决方案基于Flexus云服务器X实例帮助您快速部署Dify LLM应用开发平台。Dify是一款开源的大语言模型(LLM)应用开发平台。它融合了后端即服务(Backend as Service)和LLMOps的理念,使开发者可以快速搭建生产级的生成式AI应用。

方案架构

该解决方案基于Flexus云服务器X实例帮助您快速部署Dify LLM应用开发平台。

图 1-1 方案架构图



该解决方案将会部署如下资源:

- 创建1台**华为云Flexus云服务器X实例**,用于搭建Dify-LLM应用开发平台。
- 创建1个**弹性公网IP EIP**并关联华为云Flexus云服务器X实例,提供访问公网和被公网访问能力。

- 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

- 开箱即用
后端即服务，帮助开发者快速搭建生产级的生成式AI应用，部署完成即可使用。
- 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，即可完成云服务器的创建和Dify-LLM应用开发平台的搭建。

约束与限制

- 部署该解决方案之前，您需要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，请根据[2 资源和成本规划](#)中预估价格，确保余额充足。

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	配置示例	每月预估花费
华为云Flexus云服务器X实例	<ul style="list-style-type: none">● 按需计费● 区域：华北-北京四● 规格：Flexus云服务器X实例 性能模式（关闭） x1.8u.16g 8核 16 GB● 镜像：Ubuntu 22.04 server 64bit● 系统盘：高IO 100GB● 购买量：1	683.28元
弹性公网IP EIP	<ul style="list-style-type: none">● 区域：华北-北京四● 计费模式：按需计费● 线路：动态BGP● 公网带宽：按流量计费● 带宽大小：300Mbit/s● 购买量：1	0.80元/GB
合计	-	683.28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	配置示例	每月预估花费
华为云Flexus云服务器X实例	<ul style="list-style-type: none">● 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） x1.8u.16g 8核 16 GB● 镜像：Ubuntu 22.04 server 64bit● 系统盘：高IO 100GB● 购买量：1	467.00元
弹性公网IP EIP	<ul style="list-style-type: none">● 区域：华北-北京四● 计费模式：按需计费● 线路：动态BGP● 公网带宽：按流量计费● 带宽大小：300Mbit/s● 购买量：1	0.80元/GB
合计	-	467.00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

* 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

* 持续时间

描述

0/255

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

委托“rf_admin_trust”将资源委托策略

策略已选(1) 从其他区域资源目录复制策略

名称	类型
<input checked="" type="checkbox"/> Tenant Administrator	系统角色
<input type="checkbox"/> 全部云服务的管理员 (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

根据当前您选择的策略，系统会显示以下授权范围方案，更便于您最小化授权，可进行选择。了解如何根据您的应用场景选择最佳的授权范围方案

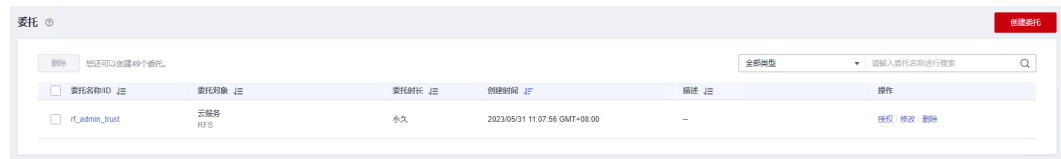
选择授权范围方案

所有资源
授权后，IAM用户可以按照权限使用帐号中所有资源，包括企业项目。区域项目和企业服务资源。

[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



----结束

3.2 快速部署

本章节主要帮助用户快速部署“快速搭建Dify-LLM应用开发平台”解决方案

表 3-1 参数说明

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）。	dify-llm-application-development-platform-demo
secgroup_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（.）。	dify-llm-application-development-platform-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，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	dify-llm-application-development-platform-demo

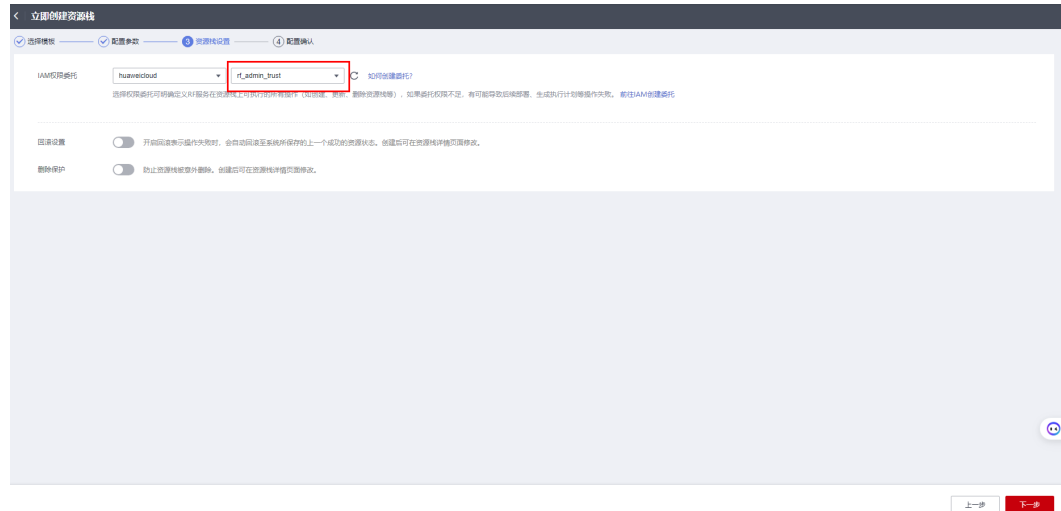
参数名称	类型	是否可选	参数解释	默认值
flexus_flavor	string	必填	云服务器实例规格，支持弹性云服务器ECS及华为云Flexus 云服务器X实例。Flexus 云服务器X实例规格ID命名规则为x1.?u.?g，例如2vCPUs4GiB规格ID为x1.2u.4g，具体华为云Flexus 云服务器X实例规格请参考控制台。弹性云服务器规格名称，具体请参考官网 弹性云服务器规格清单 。	x1.8u.16g
ecs_password	string	必填	云服务器密码，长度为8-26位，密码至少必须包含大写字母、小写字母、数字和特殊字符（!@\$\$%^&*_=-+[]{}:;./?）中的三种，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

参数名称	类型	是否可选	参数解释	默认值
ecs_volume_size	number	必填	云服务器系统盘大小，磁盘类型默认为高IO，单位：GB，取值范围为40-1,024，不支持缩盘。	100
bandwidth_size	number	必填	弹性公网带宽大小，该模板计费方式为按流量计费。单位：Mbit/s，取值范围：1-300Mbit/s。	300
charging_mode	string	必填	计费模式，默认自动扣费，取值为prePaid（包年包月）或postPaid（按需计费）。	postPaid
charge_period_unit	string	必填	计费周期单位，当计费方式设置为prePaid，此参数是必填项。有效值为：month（包月）和year（包年）。	month
charge_period	number	必填	计费周期，当计费模式设置为prePaid，此参数是必填项。可选值为：1-3（year）、1-9（month）。	1

步骤1 登录[华为云解决方案实践](#)，选择“快速搭建Dify-LLM应用开发平台”，单击“一键部署”，跳转至解决方案创建资源栈界面。

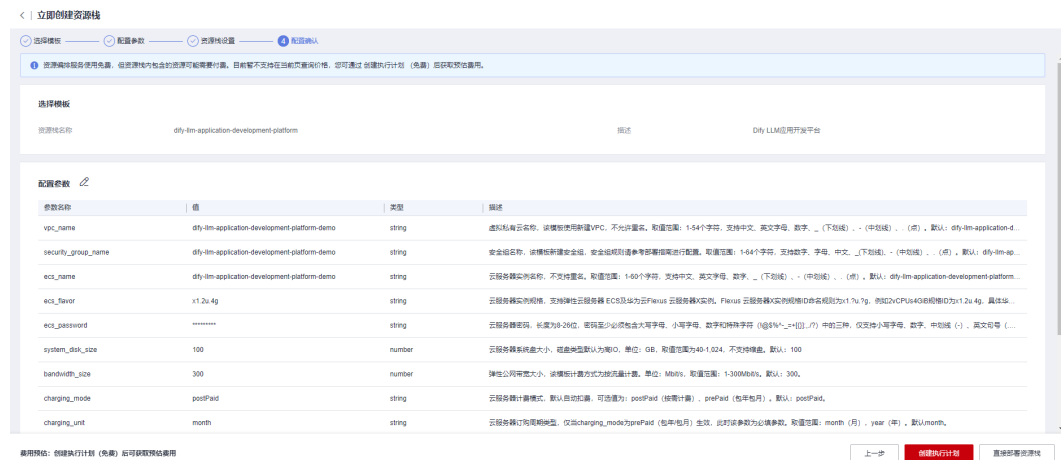
步骤4 在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托（可不选），单击“下一步”。

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

创建执行计划

通过执行计划，可以预览您的资源变更信息。

* 执行计划名称: executionPlan_20241010_1130_b4jc

描述: 请输入对执行计划的描述 (0/255)

确定 取消

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

执行计划名称ID	状态	操作按钮	创建时间	描述	操作
executionPlan_20241010_1130_b4jc 14430974681466c0bae048219562746	创建成功, 待部署	查看应用部署	2024/10/10 11:34:04 GMT+08:00	-	部署

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”, 表示该解决方案已经部署完成。

图 3-16 部署完成



步骤10 刷新页面, 在“输出”中查看Dify-LLM应用开发平台访问说明。堆栈部署成功后, Dify应用搭建脚本开始执行, 耐心等待5-10分钟左右(受网络波动影响)。

图 3-17 说明



----结束

3.3 开始使用

安全组规则修改（可选）

须知

- 该解决方案使用80端口用来访问Dify，默认全放通，请参考[修改安全组规则](#)，配置IP地址白名单。
- 该解决方案使用22端口用来以SSH方式远程登录云服务器，若需远程登录云服务器，请参考[修改安全组规则](#)，配置IP地址白名单，以便能正常访问服务。
- 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之后方可正常访问。

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。

步骤1 登录开发平台：输入[快速部署步骤10](#)的访问地址，即可浏览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 3-18 设置管理员账户

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

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

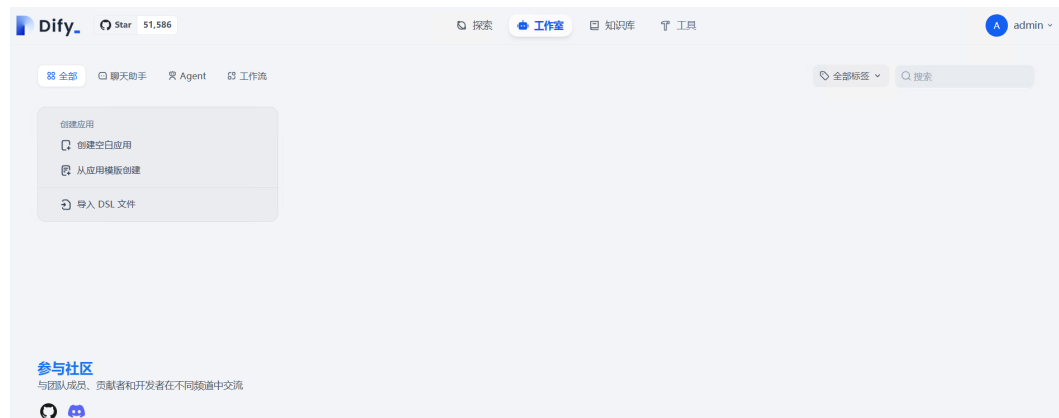
启动 Dify 社区版之前，请阅读 GitHub 上的 开源协议

步骤2 浏览器进入您部署的Dify平台界面，依次输入上一步骤中的“邮箱”“密码”登录Dify平台。

图 3-19 登录 Dify 平台



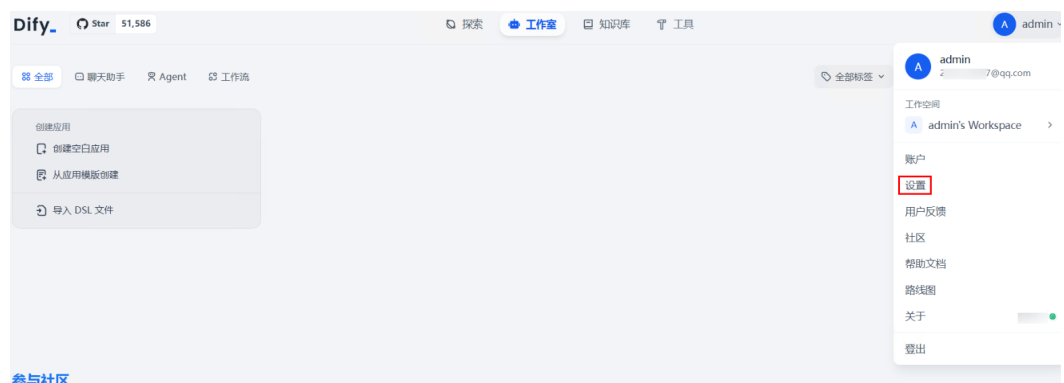
图 3-20 Dify 平台



----结束

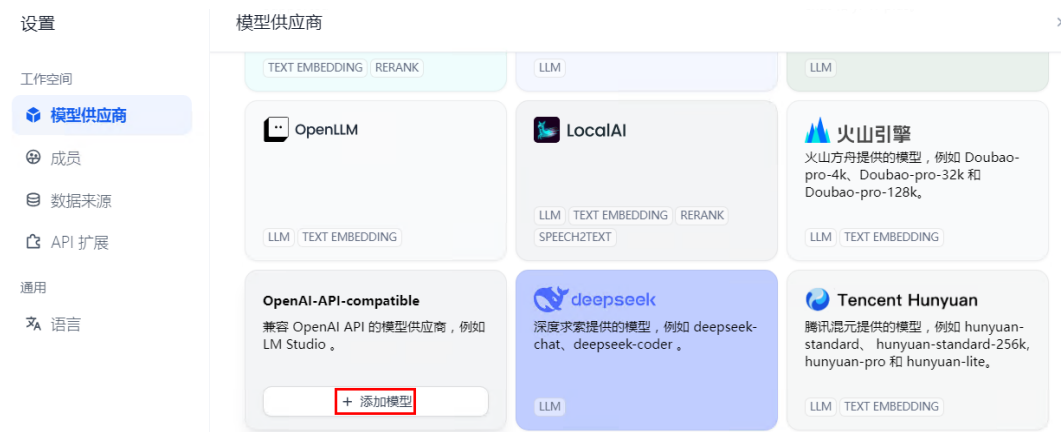
与 MaaS 服务对接

步骤1 右侧单击“用户名称”下拉并单击“设置”。



步骤2 进入模型供应商页签，选择OpenAI-API-compatible供应商。

图 3-21 选择供应商



步骤3 模型信息配置，按下图所示，请参考[华为云ModelArts Studio](#)，[助力DeepSeek模型](#)指导文档，获取对应信息。

图 3-22 模型信息配置

模型名称 *

DeepSeek-R1-Distill-Qwen-32B

API Key

API endpoint URL *

4.myhuaweicloud.com/v1/infers/ /v1/

Completion mode

对话 ×

模型上下文长度 *

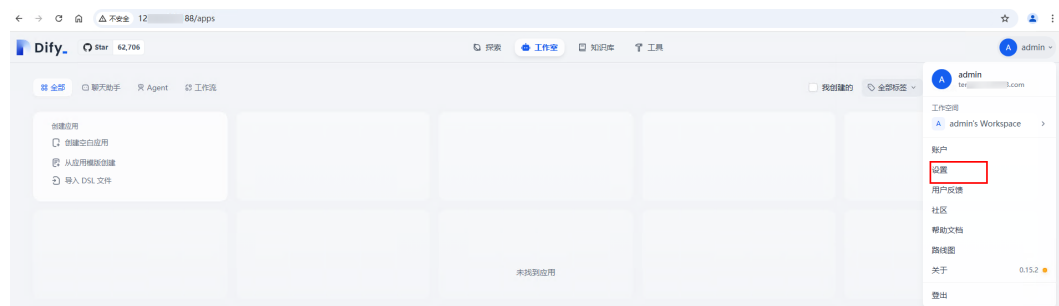
取消 保存

----结束

与一键部署 DeepSeek 对接

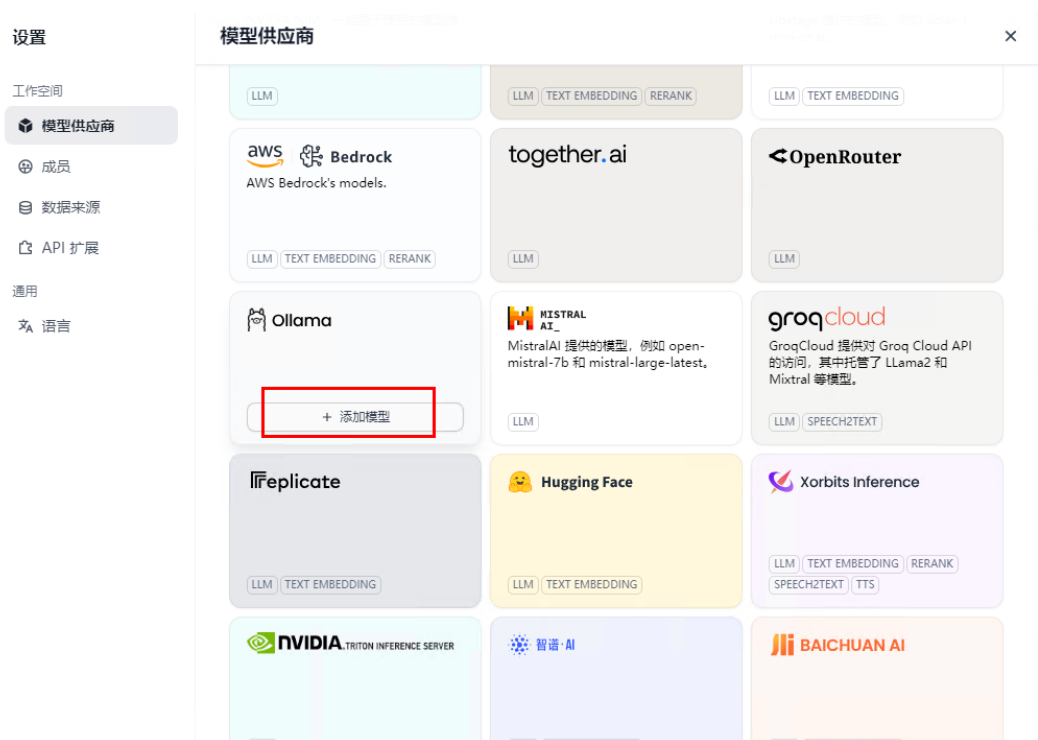
步骤1 右侧单击“用户名称”下拉并单击“设置”。

图 3-23 设置



步骤2 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-24 添加 ollama 模型



步骤3 模型名称填写2快速部署中部署的模型，如“deepseek-r1:7b”，基础URL填写中获取的私网IP地址（如果部署的Dify应用和DeepSeek-R1蒸馏版模型不在同一服务器且不在同一VPC下，需填写DeepSeek-R1蒸馏版模型所在服务器的公网IP），端口号11434（使用公网连接时，模型服务器所在安全组需放行11434端口，请参考添加安全组规则。），单击右下角“保存”并关闭“设置”。

图 3-25 模型配置

添加 Ollama

模型类型 *

LLM 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4 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 3-26 创建空白应用

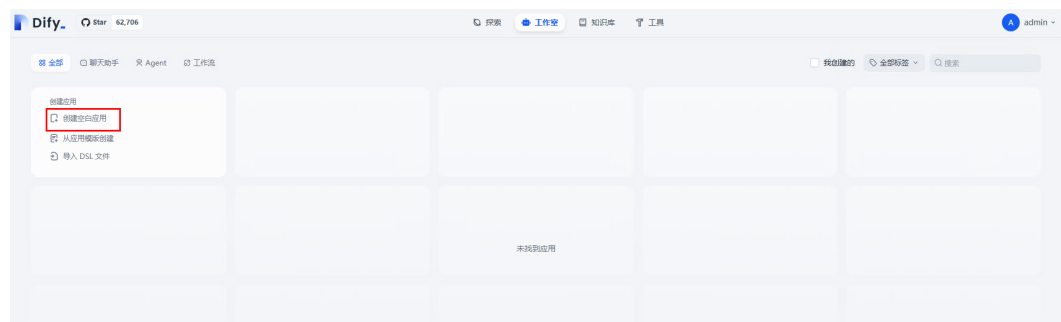
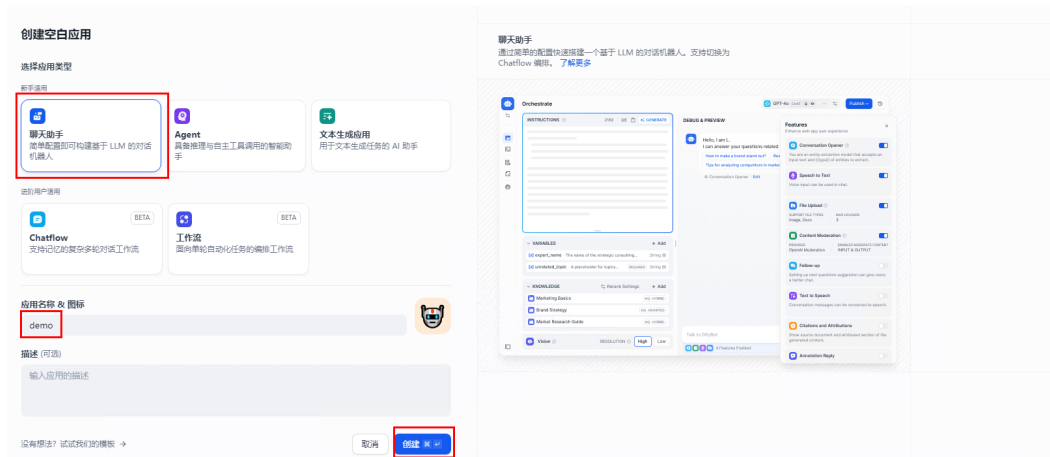


图 3-27 创建应用



步骤5 单击“编排”，在右下角“和机器人聊天”中输入内容即可调试预览。

图 3-28 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 3-29 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 3-30 删除资源栈确认



----结束

4 附录

名词解释

- 华为云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
2024-11-07	第一次正式发布。
2025-02-17	更新Dify 0.15.3版本。
2025-02-25	开始使用增加MAAS/一键部署DeepSeek对接指导
2025-03-04	升级云服务器默认规格