

解决方案实践

快速搭建 Dify-LLM 应用开发平台

文档版本 2.0.1
发布日期 2025-09-02



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.....	4
3 实施步骤.....	20
3.1 准备工作.....	20
3.2 快速部署.....	24
3.3 开始使用.....	35
3.4 快速卸载.....	48
4 附录.....	51
5 修订记录.....	53

1 方案概述

应用场景

该解决方案帮助您快速部署单机版、高可用版Dify LLM应用开发平台，同时支持将在Dify应用开发平台创建的文档知识库挂载华为云[对象存储服务 OBS](#)桶。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

该解决方案帮助您快速部署Dify LLM应用开发平台。

图 1-1 方案架构图（单机部署）

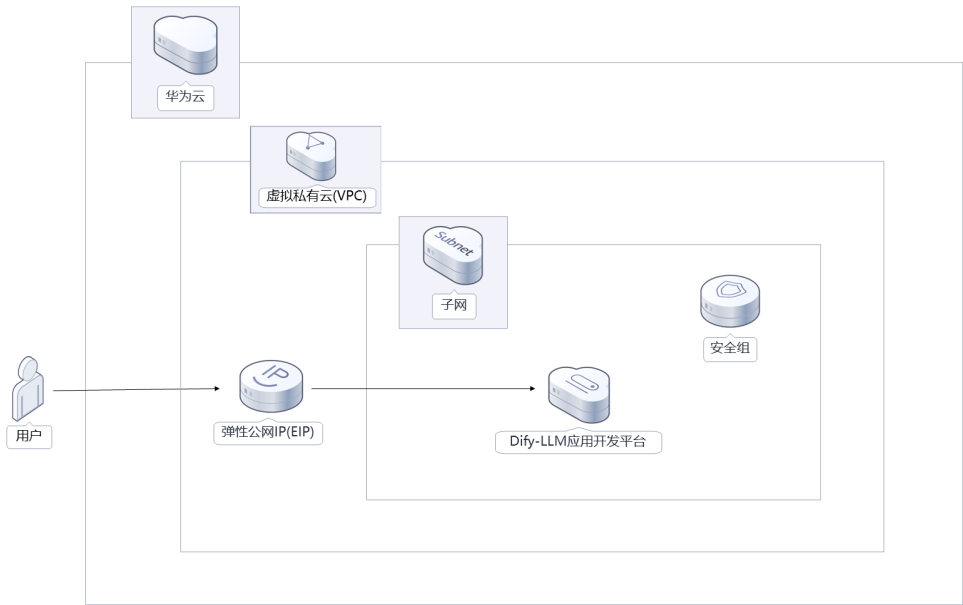
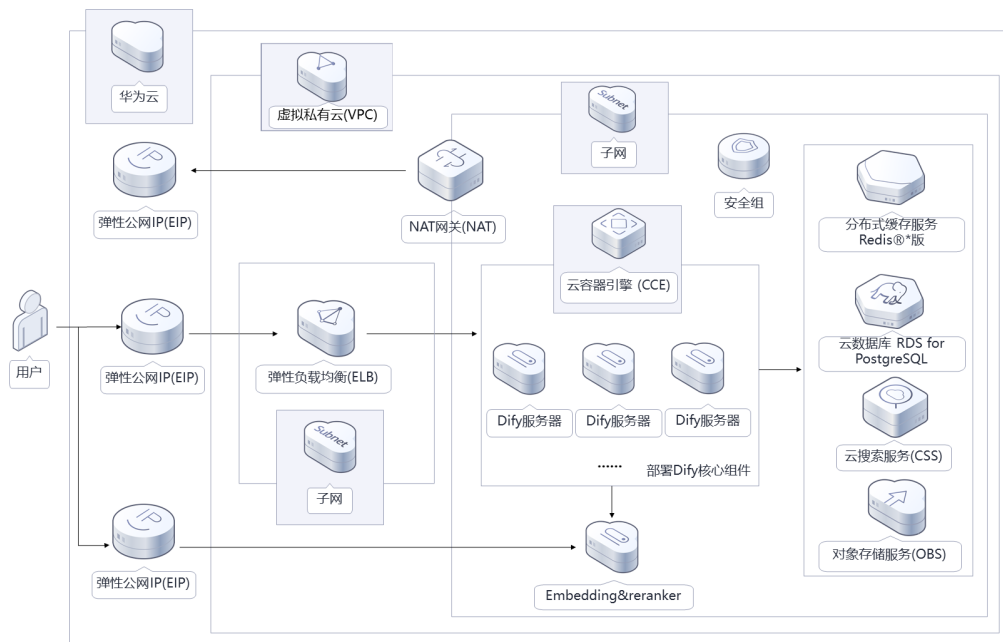


图 1-2 方案架构图（高可用部署）



该解决方案将会部署如下资源：

云服务器单机部署：

- 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FlexusX实例，提供访问公网和被公网访问能力。
- 创建1个安全组，通过配置安全组规则，为云服务器提供安全防护。

CCE容器高可用部署：

- 创建3个弹性公网IP EIP，提供访问公网和被公网访问能力。
- 创建1个**弹性负载均衡 ELB**，并绑定EIP，将访问流量自动分发到不同后端服务，扩展应用系统对外的服务能力，实现强大的应用容错性能。
- 创建1个**NAT网关 NAT**，并绑定EIP。配置SNAT规则，提供安全可靠的公网NAT网关和私网NAT网关服务，保护私有网络信息不对外暴露。
- 创建3台FlexusX实例，使用Helm包部署Dify（5个核心插件）。
- 创建1个**云容器引擎 CCE Turbo**集群，创建节点池并将上述3台FlexusX实例纳管为集群的Node节点。
- 创建1台FlexusX实例，用于部署Embedding（bge-m3）及Reranker（bge-reranker-v2-m3）模型。
- 使用**对象存储服务OBS**服务，用于将Dify的知识库挂载在对象存储服务 OBS桶上。
- 创建1个**分布式缓存服务Redis®*版**，兼容Redis，为用户提供高性能、低成本NoSQL数据库，同时数据流转过程中数据的一致性。
- 创建1个**云数据库 RDS for PostgreSQL**实例，主备分区部署，具备跨可用区故障容灾的能力。
- 创建1个**云搜索服务 CSS** OpenSearch集群，提供在线分布式搜索及语义搜索等功能。

- 创建4个安全组，通过配置安全组规则，为云服务提供安全防护。

方案优势

- 成本优化
提供高性价比的云服务器，按需选择资源规格、支持自动扩展，减少资源闲置，优化成本投入，进一步降低客户的运营成本。
- 高可用性
通过云容器引擎 CCE、云数据库 RDS for PostgreSQL、云搜索服务 CSS OpenSearch部署应用，更好地托管与简化维护应用实例，确保系统的高性能和可扩展性。
- 一键部署
一键轻松部署，即可完成云服务资源的创建及Dify-LLM应用开发平台的搭建。

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。

云服务器单机部署

表 2-1 资源和成本规划（按需计费）

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
虚拟私有云 VPC	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - VPC网段：172.16.0.0/16	1	USD 0.00
子网 Subnet	dify-llm-application-development-platform-demo-subnet	- 区域：亚太-曼谷 - 子网网段：172.16.1.0/24 - 网关：172.16.1.1	1	USD 0.00
安全组 SecurityGroup	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - 允许ping：0.0.0.0/0 - 开放端口80允许访问dify应用：0.0.0.0/0 - 开放端口80允许访问dify应用：0.0.0.0/0	1	USD 0.00

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - 按需计费：USD 0.27/小时 - 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	1	USD 195.98
弹性公网IP EIP	dify-llm-application-development-platform-demo-eip	- 区域：亚太-曼谷 - 按需计费：USD 0.11/GB/小时 - 线路：动态BGP - 公网带宽：按流量计费 - 带宽大小：300Mbit/s	1	USD 0.11/GB/小时
合计	-	-	-	USD 195.98 + 弹性公网IP EIP流量费用

表 2-2 资源和成本规划（包年包月）

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
虚拟私有云 VPC	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - VPC网段：172.16.0.0/16	1	USD 0.00
子网 Subnet	dify-llm-application-development-platform-demo-subnet	- 区域：亚太-曼谷 - 子网网段：172.16.1.0/24 - 网关：172.16.1.1	1	USD 0.00

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
安全组 SecurityGroup	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - 允许ping：0.0.0.0/0 - 开放端口22允许Cloud Shell 登录：121.36.59.153/32 - 开放端口80允许访问dify应用：0.0.0.0/0 - 开放端口80允许访问dify应用：0.0.0.0/0	1	USD 0.00
华为云Flexus云服务器X实例	dify-llm-application-development-platform-demo	- 区域：亚太-曼谷 - 计费模式：包月 - 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	1	USD 144.44
弹性公网IP EIP	dify-llm-application-development-platform-demo-eip	- 区域：亚太-曼谷 - 按需计费：USD 0.11/GB/小时 - 线路：动态BGP - 公网带宽：按流量计费 - 带宽大小：300Mbit/s	1	USD 0.11/GB/小时
合计	-	-	-	USD 144.44 + 弹性公网IP EIP流量费用

CCE 容器高可用部署

表 2-3 资源和成本规划（按需计费）

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
虚拟私有云 VPC	ha-dify-app-vpc	- 区域：亚太-曼谷 - VPC网段：192.168.0.0/16	1	USD 0.00
子网 Subnet	ha-dify-app-public-subnet ha-dify-app-private-subnet subnet-eni-1 subnet-eni-2	- 区域：亚太-曼谷 - 子网网段：192.168.0.0/24, 192.168.1.0/24, 192.168.2.0/24, 192.168.3.0/24 - 网关：192.168.0.1, 192.168.1.1, 192.168.2.1, 192.168.3.1	4	USD 0.00
安全组 SecurityGroup	ha-dify-app-sg ha-dify-app-cluster-cce-control-hedut ha-dify-app-cluster-cce-eni-hedut ha-dify-app-cluster-cce-node-hedut	- 区域：亚太-曼谷 - 允许vpc内网访问：192.168.0.0/16 - 其他规则见具体安全组	4	USD 0.00

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	ha-dify-app-pool-xxxxx	- 区域：亚太-曼谷 - 按需计费：USD 0.48/小时 - 规格：Flexus云服务器X实例 性能模式（关闭） x1.16u.16g 16核 16 GB - 镜像：CCE_images_Ubuntu22-Node-24.11.0-common-amd64-364e1f82-20241120165013 - 系统盘：通用SSD 40GB - 数据盘：高IO 100 GiB	3	USD 1,045.01
弹性云服务器ECS	ha-dify-app-embedding-reranker	- 区域：亚太-曼谷 - 按需计费：USD 1.42/小时 - 规格：通用计算增强型 32vCPUs 64GiB c7n.8xlarge.2 - 镜像：Ubuntu 24.04 server 64bit - 系统盘：通用型SSD 40GB	1	USD 1022.4

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
弹性公网IP EIP	ha-dify-app-bandwidth	• 区域：亚太-曼谷 • 按需计费：USD 0.11/GB/小时 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s	3	USD 0.33/GB/小时

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
对象存储服务 OBS	用户自定义	- 区域：亚太-曼谷 - 存储空间：数据存储（多AZ存储） - 默认存储类别：标准存储 - 桶策略：私有 - 请求费用： GET 0.0040USD/万次 PUT 0.0100USD/万次 DELETE 0.0004USD/千次 - 存储空间：0.0250USD/GB/月 - 流量费用： - 内/公网流入流量（数据上传到 OBS）：免费 - 内网流出流量（通过 ECS 云服务器下载 OBS 的数据）：免费 - 公网流出流量 / 00:00-08:00（闲时）：不涉及	1	详细请参考每月账单。计费说明参考 价格详情

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
		公网流出流量 / 08:00-24:00（忙时）：不涉及		
云容器引擎 CCE	ha-dify-app-cluster	• 区域：亚太-曼谷 • 按需计费：USD 0.39/小时 • 类型：Turbo • 规格：cce.s2.small（50节点） • 集群 master 实例数：3 • 集群 node 实例数：3 • 类型：CCE	1	USD 280.80
分布式缓存服务 Redis®*版	ha-dify-app-redis	• 区域：亚太-曼谷 • 按需计费：USD 0.13/小时 • 计费模式：按需计费 • 规格：4G 副本数：2 • 实例类型：Redis(主备)	1	USD 92.16

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
云数据库 RDS for PostgreSQL	ha-dify-app-pg	- 区域：亚太-曼谷 - 按需计费：USD 1.17/小时 - 规格：rds.pg.x1.2xlarge.4.ha 8 vCPU 32 GB (独享型) - 储存：SSD云盘 100GB - 数据库引擎：PostgreSQL 16.8	1	USD 842.4
云搜索服务 CSS	ha-dify-app-css	- 区域：亚太-曼谷 - 按需计费：USD 0.78/小时 - 计费模式：按需计费 - 规格：ess.spec-4u8g 4 vCPUs 8 GB - 节点存储总容量：40GB 超高I/O - 集群类型：OpenSearch	3	USD 563.33
弹性负载均衡 ELB	ha-dify-app-lb	- 区域：亚太-曼谷 - 按需计费：USD 0.05/小时 - 类型：独享型负载均衡 - 网络类型：私网	1	USD 38.16

华为云服务	资源名称（默认值）	配置示例	数量	每月预估花费
NAT网关 NAT	ha-dify-app-nat	- 区域：亚太-曼谷 - 按需计费：USD 2.44/天 - 规格：小型 - SNAT规则数：3	1	USD 73.14
合计		-	-	USD 3957.4 + 弹性公网IP EIP 流量费用 + 对象存储服务OBS存储及流量费用

表 2-4 资源和成本规划（包年包月）

华为云服务		配置示例	数量	每月预估花费
虚拟私有云 VPC	ha-dify-app-vpc	- 区域：亚太-曼谷 - VPC网段：192.168.0.0/16	1	USD 0.00
子网 Subnet	ha-dify-app-public-subnet ha-dify-app-private-subnet subnet-eni-1 subnet-eni-2	- 区域：亚太-曼谷 - 子网网段：192.168.0.0/24, 192.168.1.0/24, 192.168.2.0/24, 192.168.3.0/24 - 网关：192.168.0.1, 192.168.1.1, 192.168.2.1, 192.168.3.1	4	USD 0.00

华为云服务		配置示例	数量	每月预估花费
安全组 SecurityGroup	ha-dify-app-sg ha-dify-app-cluster-cce-control-hedut ha-dify-app-cluster-cce-eni-hedut ha-dify-app-cluster-cce-node-hedut	- 区域：亚太-曼谷 - 允许vpc内网访问：192.168.0.0/16 - 其他规则见具体安全组	4	USD 0.00
华为云Flexus云服务器X实例	ha-dify-app-pool-xxxxx	- 区域：亚太-曼谷 - 计费模式：包月 - 规格：Flexus云服务器X实例 性能模式（关闭） x1.16u.16g 16核 16 GB - 镜像：CCE_images_Ubuntu22-Node-24.11.0-common-amd64-364e1f82-20241120165013 - 系统盘：通用SSD 40GB - 数据盘：高IO 100 GiB	3	USD 770.58

华为云服务		配置示例	数量	每月预估花费
弹性云服务器 ECS	ha-dify-app-embedding-reranker	- 区域：亚太-曼谷 - 计费模式：包月 - 规格：通用计算增强型 32vCPUs 64GiB c7n.8xlarge.2 - 镜像：Ubuntu 24.04 server 64bit - 系统盘：通用型SSD 40GB	1	USD 747.77
弹性公网IP EIP	ha-dify-app-bandwidth	- 区域：亚太-曼谷 - 按需计费：USD 0.11/GB/小时 - 线路：动态BGP - 公网带宽：按流量计费 - 带宽大小：300Mbit/s	3	USD 0.33/GB/小时

华为云服务		配置示例	数量	每月预估花费
对象存储服务 OBS	用户自定义	• 区域：亚太-曼谷 • 存储空间：数据存储（多AZ存储） • 默认存储类别：标准存储 • 桶策略：私有 • 请求费用： GET 0.0040USD/万次 PUT 0.0100USD/万次 DELETE 0.0004USD/千次 • 存储空间：0.0250USD/GB/月 • 流量费用： - 内/公网流入流量（数据上传到 OBS）：免费 - 内网流出流量（通过 ECS 云服务器下载 OBS 的数据）：免费 - 公网流出流量 / 00:00-08:00（闲时）：不涉及公网流出流量 /	1	详细请参考每月账单。计费说明参考 价格详情

华为云服务		配置示例	数量	每月预估花费
		08:00-24:00（忙时）：不涉及		
云容器引擎 CCE	ha-dify-app-cluster	- 区域：亚太-曼谷 - 计费模式：包月 - 类型：Turbo - 规格：cce.s2.small（50节点） - 集群 master 实例数：3 - 集群 node 实例数：3 - 类型：CCE	1	USD 212.76
分布式缓存服务 Redis® 版	ha-dify-app-redis	- 区域：亚太-曼谷 - 按需计费：USD 0.13/小时 - 计费模式：按需计费 - 规格：4G 副本数：2 - 实例类型：Redis(主备)	1	USD 92.16

华为云服务		配置示例	数量	每月预估花费
云数据库 RDS for PostgreSQL	ha-dify-app-pg	- 区域：亚太-曼谷 - 计费模式：包月 - 规格：rds.pg.x1.2xlarge.4.ha 8 vCPU 32 GB (独享型) - 储存：SSD云盘 100GB - 数据库引擎：PostgreSQL 16.8	1	USD 626.52
云搜索服务 CSS	ha-dify-app-css	- 区域：亚太-曼谷 - 计费模式：包月 - 计费模式：按需计费 - 规格：ess.spec-4u8g 4 vCPUs 8 GB - 节点存储总容量：40GB 超高I/O - 集群类型：OpenSearch	3	USD 431.88
弹性负载均衡 ELB	ha-dify-app-lb	- 区域：亚太-曼谷 - 按需计费：USD 0.05/小时 - 类型：独享型负载均衡 - 网络类型：私网	1	USD 38.16

华为云服务		配置示例	数量	每月预估花费
NAT网关 NAT	ha-dify-app-nat	- 区域：亚太-曼谷 - 按需计费：USD 2.44/天 - 规格：小型 - SNAT规则数：3	1	USD 73.14
合计		-	-	USD 2992.97 + 弹性公网IP EIP 流量费用 + 对象存储服务OBS存储及流量费用

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 控制台管理界面

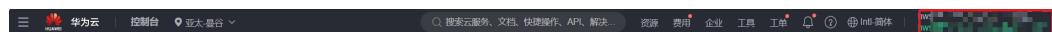
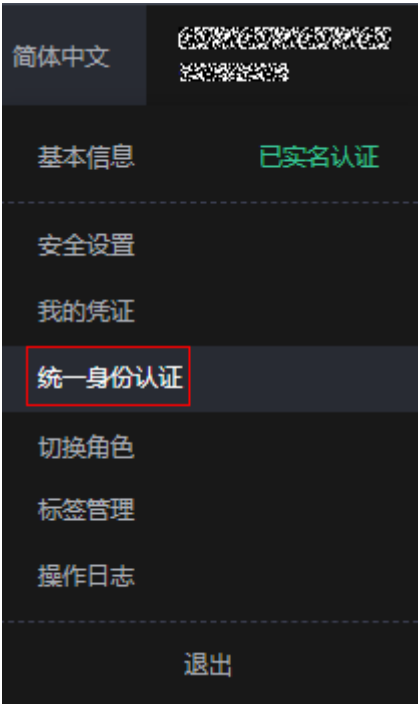


图 3-2 统一身份认证菜单



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”，输入“RFS”，单击“完成”。

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 3-5 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 3-6 选择策略



步骤6 选择“所有资源”，并单击“确定”完成配置。

图 3-7 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 3-8 委托列表



----结束

获取 OBS 桶名

- 步骤1 准备一个OBS桶：（如果已有，可跳过此步骤）登录华为云[对象存储服务控制台](#)，单击“创建桶”进入obs桶创建界面，参考[官方文档步骤一](#)创建OBS桶。
- 步骤2 获取OBS桶名：找到已有的OBS桶，单击桶名进去桶详情界面。

图 3-9 进入桶详情页面



步骤3 复制桶名：单击桶名旁边的复制按钮即可复制。

图 3-10 复制 OBS 桶名



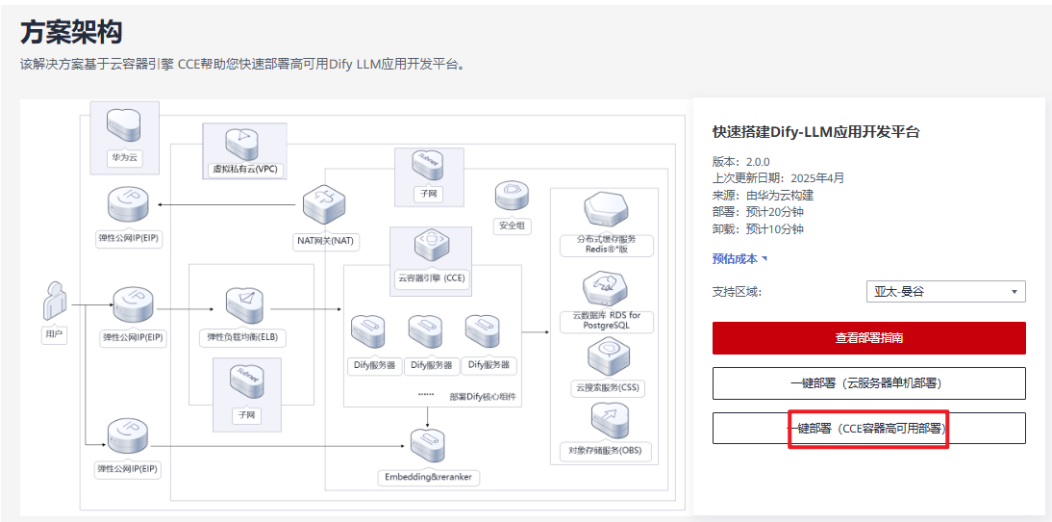
----结束

3.2 快速部署

本章节帮助用户高效地部署“快速搭建Dify-LLM应用开发平台”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。

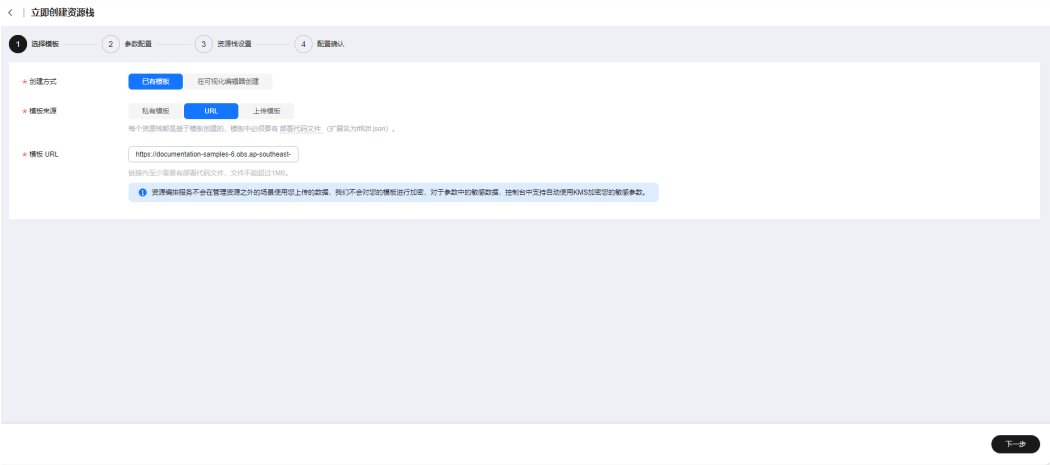
步骤1 登录[华为云解决方案实践](#)，选择“快速搭建Dify-LLM应用开发平台”，选择需要部署的版本，以高可用部署为例，单击“一键部署（CCE容器高可用部署）”，跳转至解决方案创建资源栈界面。

图 3-11 解决方案实施库



步骤2 在选择模板界面中，单击“下一步”。

图 3-12 选择模板



步骤3 在配置参数界面中，参考表3-2完成自定义参数填写，单击“下一步”。

图 3-13 配置参数



表 3-1 参数说明（云服务器单机部署）

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
secgroup_name	string	必填	安全组名称，该模板新建安全组，安全组规则请参考 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、_(下划线)、-(中划线)、.(点)。	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，支持字母、数字、_(下划线)、-(中划线)、.(点)。	dify-llm-application-development-platform-demo
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。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。密码不能包含用户名或用户名倒序。管理员账户默认root。	空
system_disk_size	number	必填	云服务器系统盘大小，磁盘类型默认为高IO，单位：GB，取值范围为40-1,024，不支持缩盘。	100

参数名称	类型	是否可选	参数解释	默认值
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	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 参数说明（CCE 容器高可用部署）

参数名称	类型	是否可选	参数解释	默认值
dify_version	string	必填	社区版Dify版本。可以选择1.7.1, 1.4.1, 0.15.8。默认1.7.1	1.7.1

参数名称	类型	是否可选	参数解释	默认值
resource_name_prefix	string	必填	资源名称前缀，命名规则 {resource_name_prefix}-资源英文名称，例如：CCE集群名称为 {resource_name_prefix}-cce。取值范围：4-24个字符，支持小写字母、数字、-（中划线）。必须以小写字母开头。禁止以中划线（-）开头。默认 ha-dify-app。	ha-dify-app
bandwidth_size	string	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。默认：300。	300
cce_cluster_flavor	string	必填	CCE Turbo集群规格，集群创建完成后规格不可再变更，可选值：cce.s2.small、cce.s2.medium、cce.s2.large、cce.s2.xlarge，具体请参考 创建CCE集群参数 flavor 。默认为 cce.s2.small(小规模多控制节点CCE集群，最大50节点)。	cce.s2.small
cce_node_pool_password	string	必填	CCE集群node节点密码，用于集群节点登录。取值范围：8-24个字符，密码至少必须包含大写字母、小写字母、并包含数字或特殊字符（~!@#\$%^&*_+=,?）	

参数名称	类型	是否可选	参数解释	默认值
cce_node_pool_flavor	string	必填	CCE集群节点云服务器实例规格，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。请使用3vCPUs6GiB及以上规格，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 ECS规格请参考官网 弹性云服务器规格清单 。默认：x1.16u.16g	x1.16u.16g
rds_flavor	string	必填	云数据库 RDS for PostgreSQL实例规格，该方案默认创建主备版。默认 rds.pg.x1.2xlarge.4.ha（8U32G），其他规格请参考官网云数据库 RDS for PostgreSQL实例规格 。	rds.pg.x1.2xlarge.4.ha
pgsql_password	string	必填	PostgreSQL数据库的管理员密码，取值范围：8-24个字符，密码至少必须包含大写字母、小写字母、并包含数字或特殊字符（~!^*-=_+,）。。	
pgsql_user_password	string	必填	PostgreSQL数据库的database用户密码。取值范围：8-24个字符，密码至少必须包含大写字母、小写字母、并包含数字或特殊字符（~!^*-=_+,），不能与用户名或倒序的用户名相同。	
redis_capacity	number	必填	分布式缓存服务 Redis版实例规格。可选值：1GB-64GB。	4

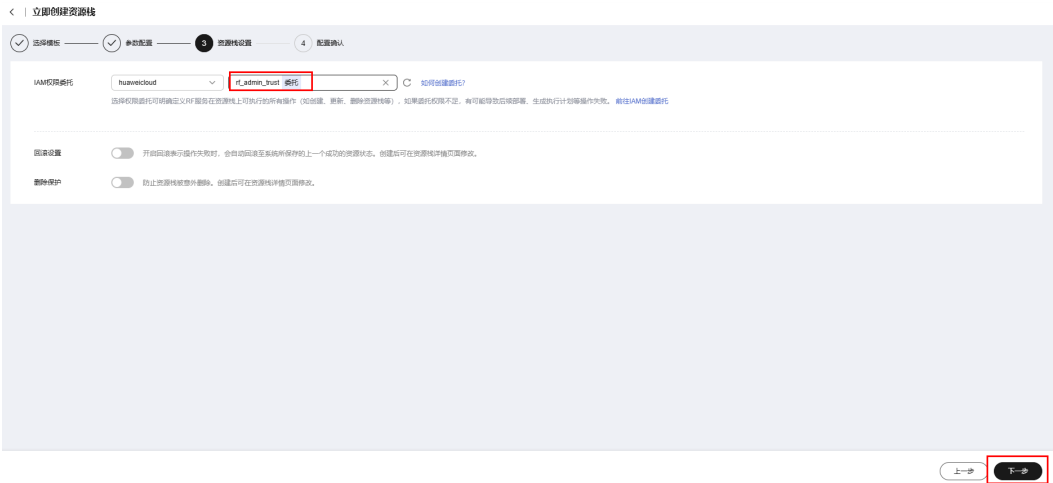
参数名称	类型	是否可选	参数解释	默认值
redis_password	string	必填	redis数据库密码。取值范围：8-24个字符，密码至少必须包含大写字母、小写字母、并包含数字或特殊字符（~!^*-=_+，）	
obs_bucket	string	必填	已有对象存储服务OBS桶名称，桶所属区域必须与一键部署选择的区域保持一致。用于存储Dify WebUI上传的知识库文件。获取请参考 获取OBS桶名 。	
access_key	string	必填	访问密钥ID（AK），识别访问用户的身份，取值范围：20，仅支持大写字母和数字，用于将知识库文件上传至OBS桶。参考 获取AK、SK密钥 。	
secret_key	string	必填	秘密访问密钥（SK），对请求数据进行签名验证，取值范围：40，仅支持大小写字母和数字，用于将知识库文件上传至OBS桶。获取请参考 获取AK、SK密钥 。	

参数名称	类型	是否可选	参数解释	默认值
embedding_re ranker_flavor	string	可选	（可选，置空不创建）部署Embedding和Reranker模型的云服务器规格，支持弹性云服务器 ECS（含GPU服务器）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4vCPUs4GiB规格ID为x1.4u.4g。建议使用8vCPUs8GiB及以上规格，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 ECS规格请参考官网 弹性云服务器规格清单 。可替换成GPU加速型获得更高性能。默认值：c7n.8xlarge.2。	c7n.8xlarge.2
ecs_password	string	可选	部署Embedding和Reranker模型的云服务器密码，取值范围：8-24个字符，密码至少包含大写字母、小写字母、并包含数字或特殊字符（~!@#\$%^&*_+=,;?）。	
charging_mod e	string	必填	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	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。可选值为：month（月），year（年）。	month

参数名称	类型	是否可选	参数解释	默认值
charging_period	number	必填	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。当charging_unit=month（周期类型为月）时，取值范围：1-9；当charging_unit=year（周期类型为年）时，取值范围：1-3。	1

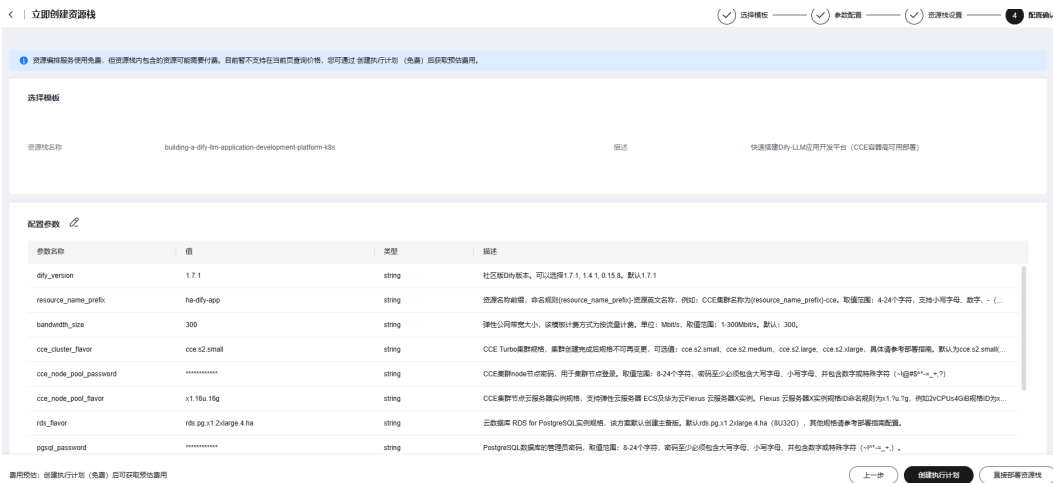
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

图 3-14 委托设置



步骤5 在配置确认界面中，单击“创建执行计划”。

图 3-15 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 3-16 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_20250312_2153_artl

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 3-17 执行计划

building-a-dify-llm-application-development-platform-k8s

删除

更新详情或事件

C

基本信息

资源

输出

事件

模板

执行计划

筛选

请输入关键字

Q

执行计划名称ID	状态	操作预览	创建时间	描述	操作
executionPlan_20250409_1129_11bp 85c7575b-546b-4077-a07b-2d81bba673fc	创建成功，待部署	查看操作预览	2025-04-09 11:29:37 GMT+08:00	-	部署 执行

图 3-18 执行计划确认



- 步骤8** （可选）如果计费模式选择“包年包月”，在余额不足的情况下（所需总费用请参考表2-4中对应一键部署云服务所需的包年包月费用表）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 3-19 部署完成



- 步骤10** 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 3-20 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。

Dify 基础使用

- 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 3-21 设置管理员账户

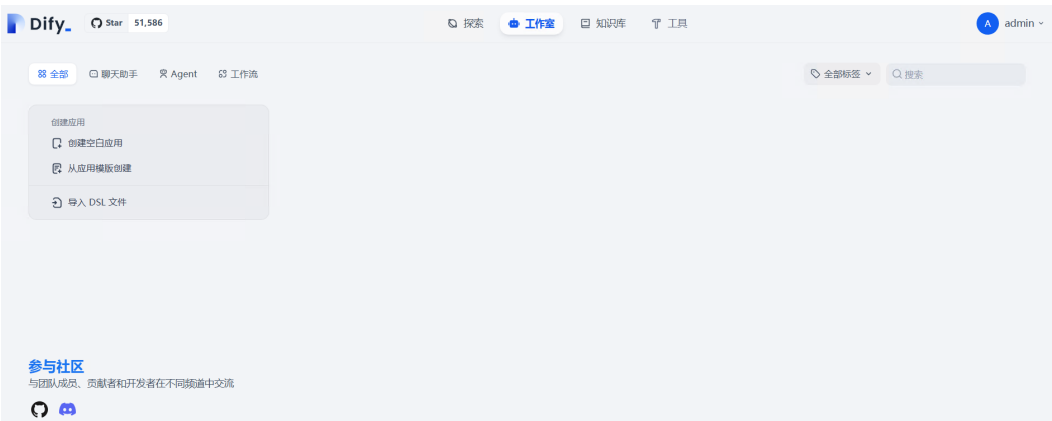


步骤2 浏览器进入您部署的Dify平台界面，依次输入上一步骤中的“邮箱”“密码”登录Dify平台。

图 3-22 登录 Dify 平台



图 3-23 Dify 平台



----结束

与一键部署 DeepSeek 模型对接

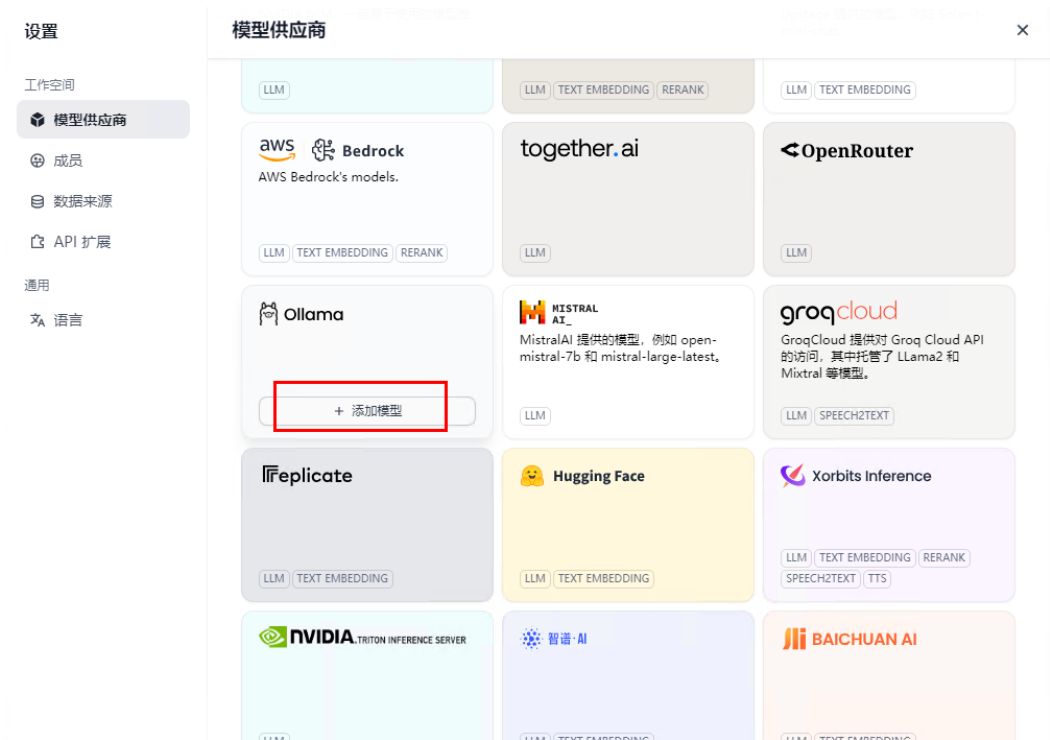
步骤1 右侧单击“用户名称”下拉并单击“设置”。

图 3-24 设置



步骤2 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-25 添加 ollama 模型



步骤3 已部署模型名称填写，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通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 3-26 模型配置

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 3-27 创建空白应用

Dify

Star 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

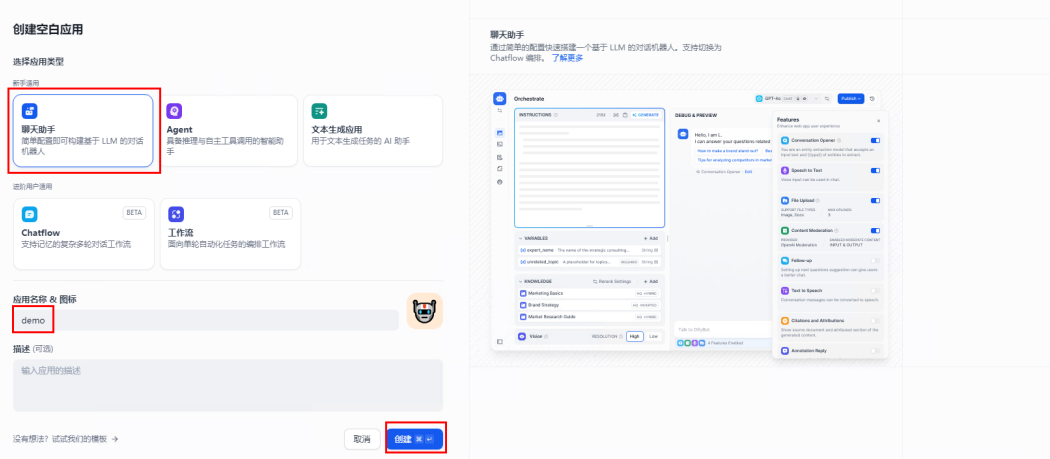
创建空白应用

从应用模板创建

导入 DSL 文件

没有找到应用

图 3-28 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 3-29 调试与预览

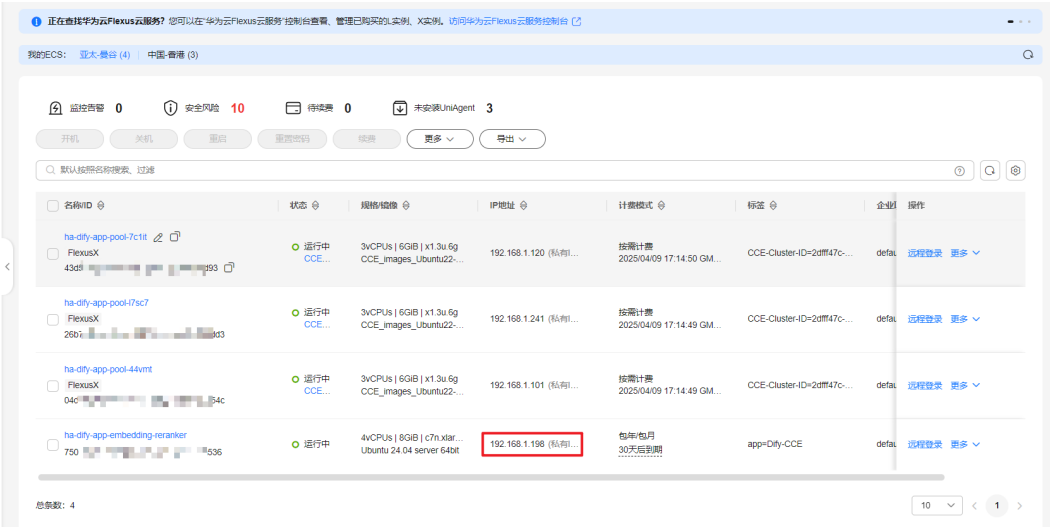


----结束

CCE 容器高可用版

- 步骤1 登录Dify开发平台，请参考[Dify基础使用](#)。
- 步骤2 添加模型。模型对接可参考[与一键部署DeepSeek模型对接](#)。
- 步骤3 添加Embedding和Reranker模型。登录[ECS控制台](#)，找到解决方案创建的Embedding&Reranker模型服务器，复制私网IP地址。后续步骤可参考[快速部署Embedding及Reranker模型部署指南](#)[开始使用](#)步骤3-7（服务器URL填写私网IP地址）。

图 3-30 服务器私网 IP



步骤4 创建知识库。在Dify平台页面，依次单击“知识库>创建知识库”上传文本文件后单击“下一步”。

图 3-31 创建知识库

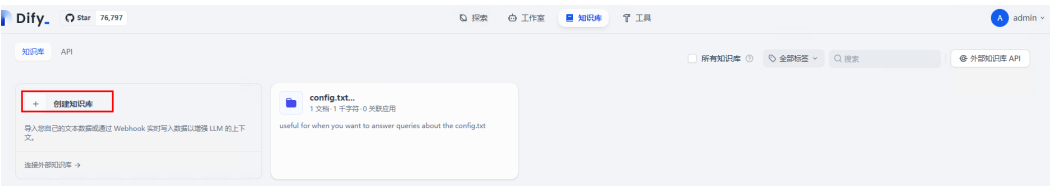
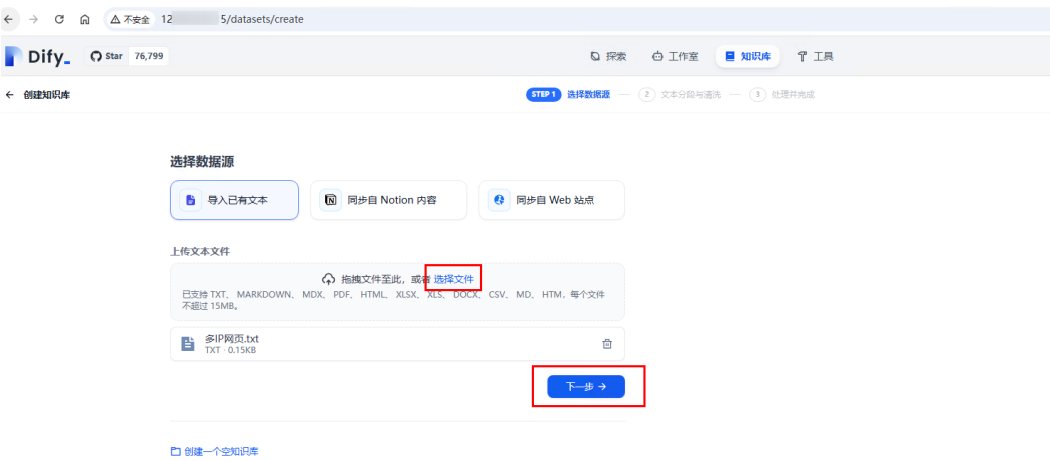


图 3-32 导入文件



步骤5 （经济型）按需求配置知识库，若没有特殊需求默认即可。单击“保存与处理”，待页面提示嵌入已完成表示配置完成。

图 3-33 配置经济型索引知识库

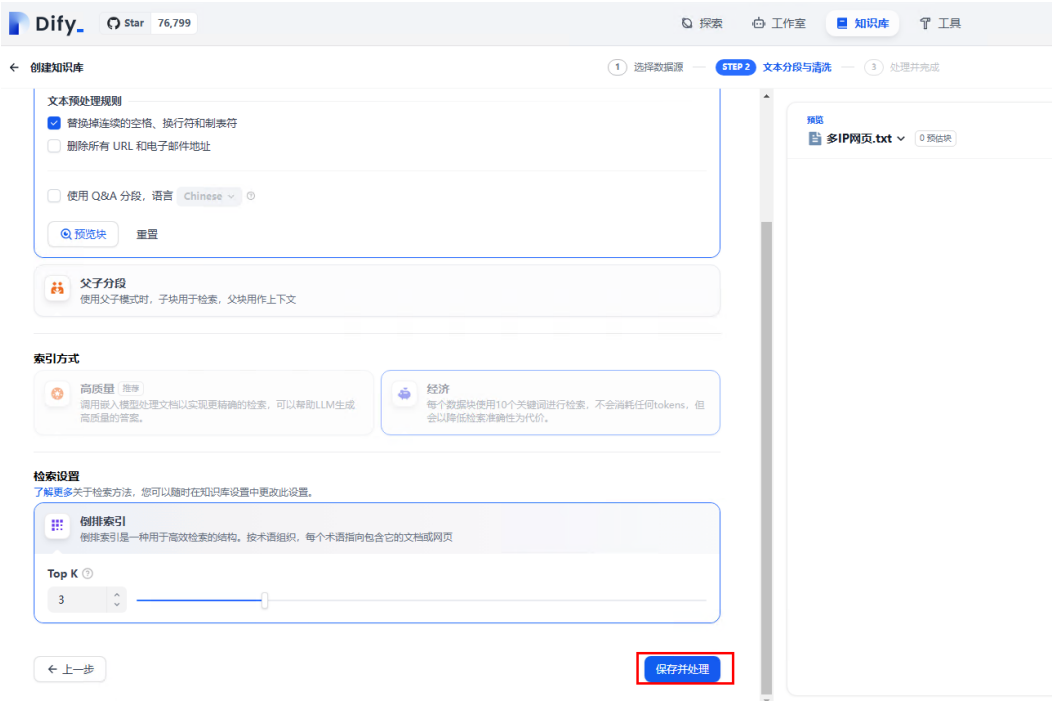
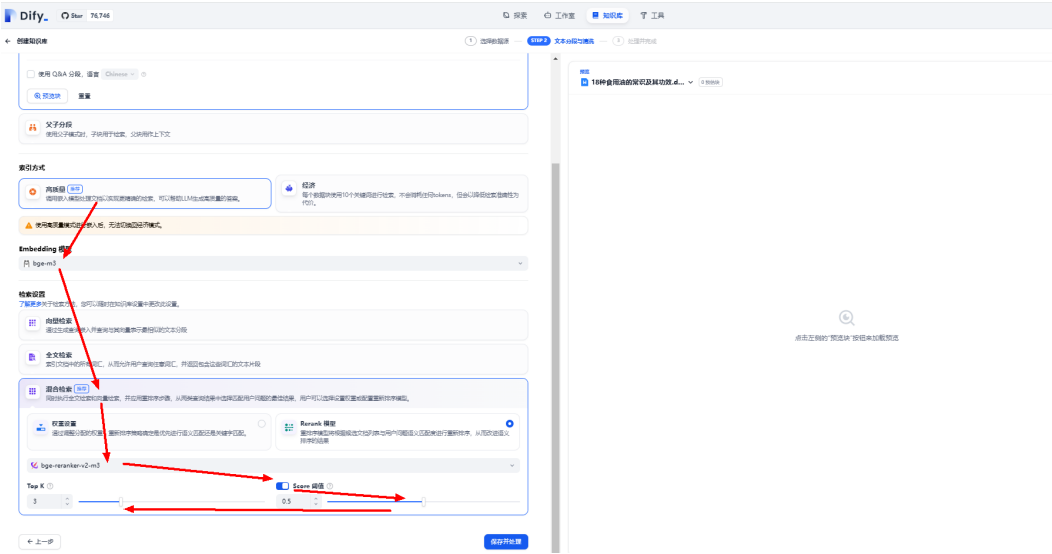


图 3-34 知识库配置完成



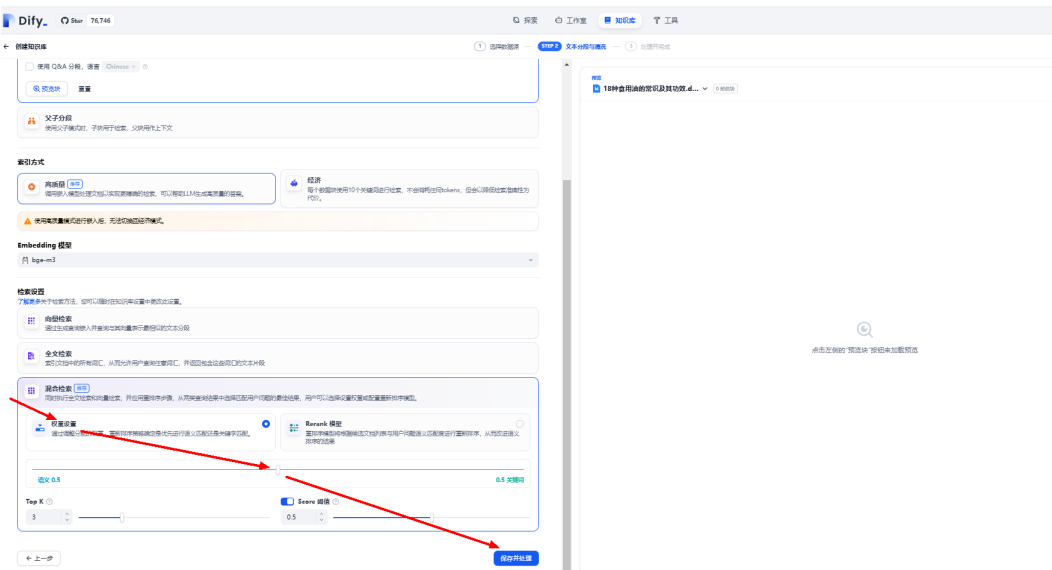
步骤6 （高质量）选择高质量模块，在Embedding模型栏目选择添加的bge-m3模型，在检索设置栏目，选择混合检索（可同时使用向量检索和关键词检索，并控制两者权重），在模型下拉窗口选择添加的bge-reranker-v2-m3模型，在下方开启Score阈值开关，并滑动下方的滑动按钮来选择阈值，推荐0.5，下方还有TOP K的选项，可以控制最终被使用的文档分片的最大数量。

图 3-35 配置高质量索引知识库



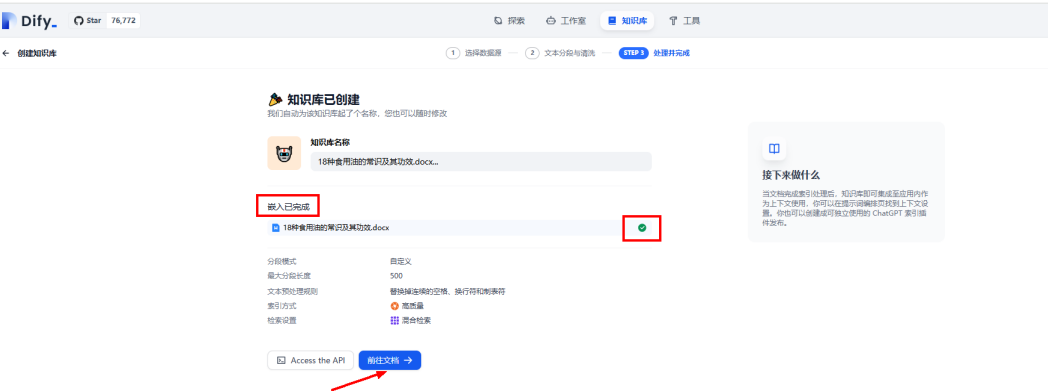
步骤7 （高质量）继续选择权重设置，调整滑动按钮，推荐选择语义（向量检索）0.5，关键词（关键词检索）0.5，然后单击保存，至此，高质量知识库创建完成。

图 3-36 配置语义等



步骤8 待显示嵌入已完成，文档对应状态变成绿色对钩，则继续单击“前往文档”，可以看到导入的知识库文档为可用状态。

图 3-37 嵌入完成



步骤9 访问创建的OBS桶，知识库的文件会保存在upload_files文件夹下，表示配置成功。在创建知识库的时候才会将用户的原始数据上传到OBS上，同时会将原始文件的名称修改成UUID的形式。

图 3-38 知识库文档



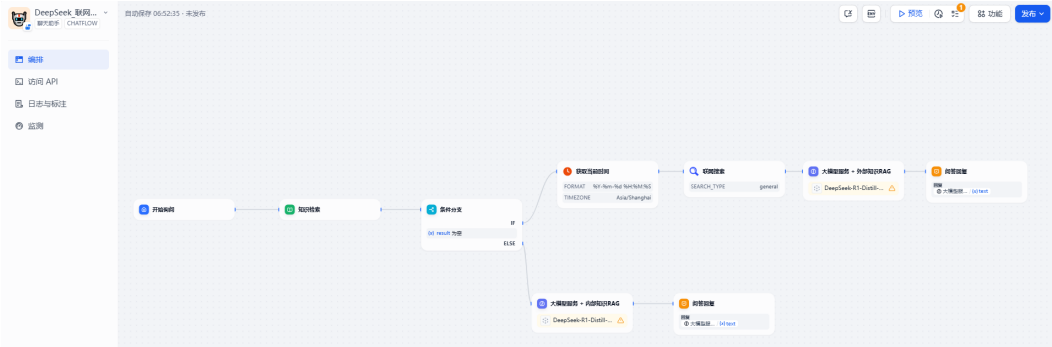
步骤10 创建Dify工作流。在工作室页面，单击“导入DSL文件”，在弹出的页面中选择“URL”，复制下面的地址，粘贴到DSL URL路径里，单击“创建”，即可导入工作流：

https://documentation-samples.obs.cn-north-4.myhuaweicloud.com/solution-as-code-publicbucket/solution-as-code-moudle/building-a-dify-llm-application-development-platform/workflow/DeepSeek_%E8%81%94%E7%BD%91%E6%90%9C%E7%B4%A2_%E7%9F%A5%E8%AF%86%E5%BA%93.yml

图 3-39 导入 DSL 文件



图 3-40 工作流 Demo

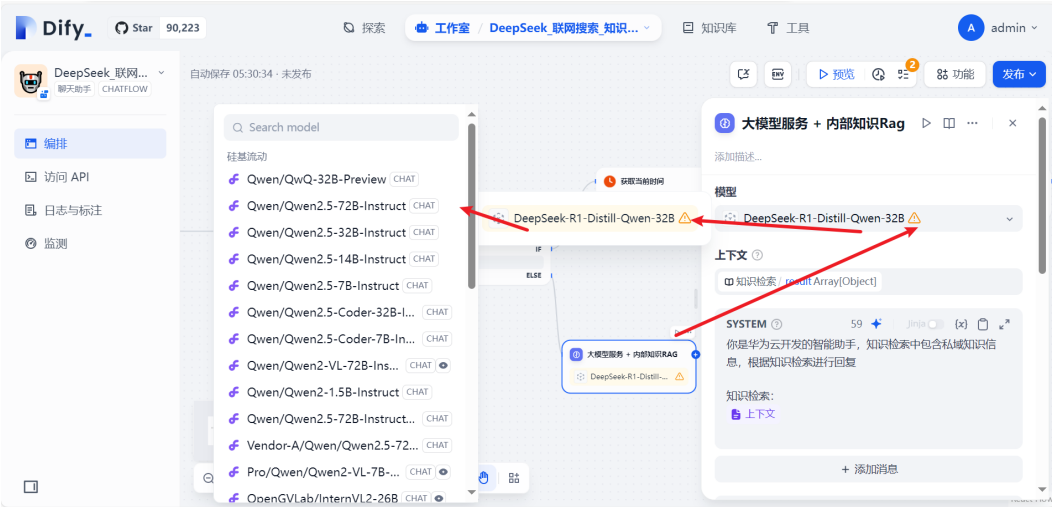


步骤11 若大模型出现如下提示，请修改为已有的模型。

图 3-41 模型不存在提示



图 3-42 修改模型



步骤12 单击工具查找搜索插件，单击APIKey 授权配置，并添加URL：http://searxng:8080并保存。如下图所示，单击保存后，即可设置成功。

图 3-43 工具设置

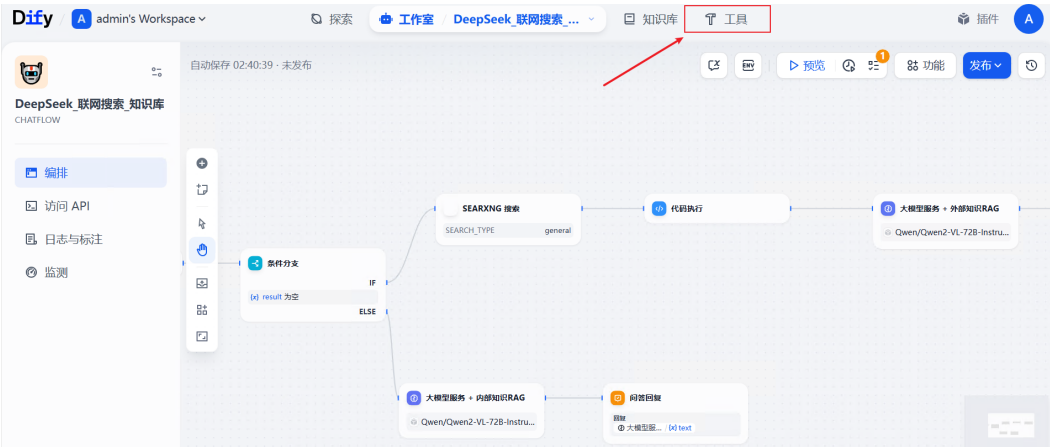


图 3-44 SearXNG 授权

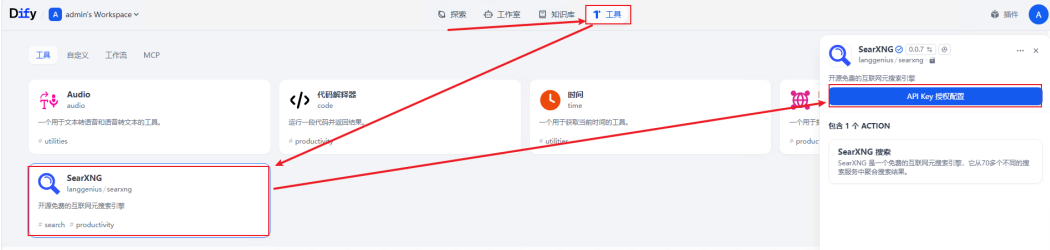
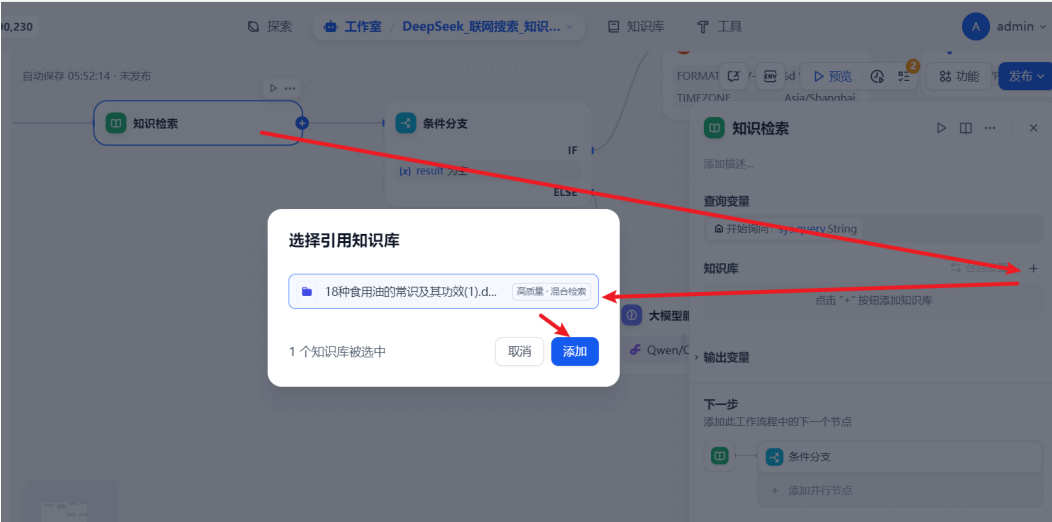


图 3-45 填写 URL



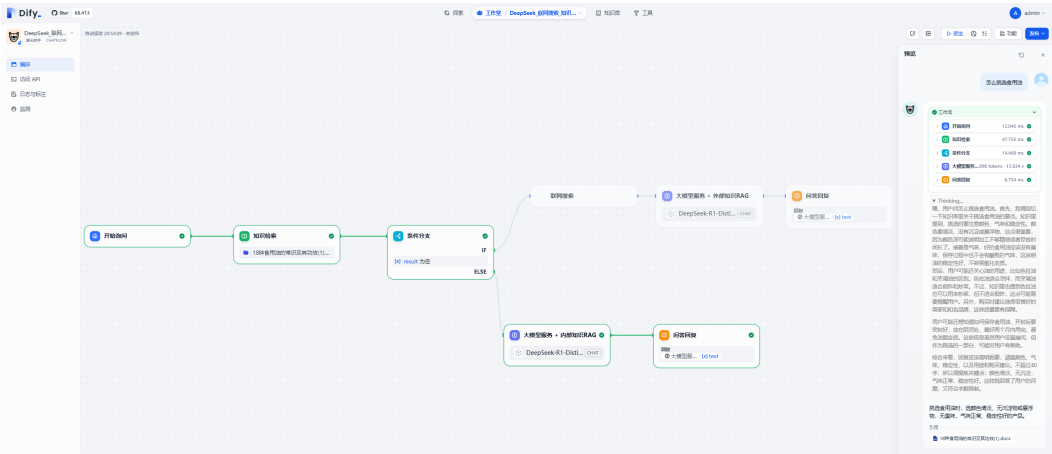
步骤13 添加知识库。单击“知识检索”节点，选择“+”添加知识库。

图 3-46 添加知识库



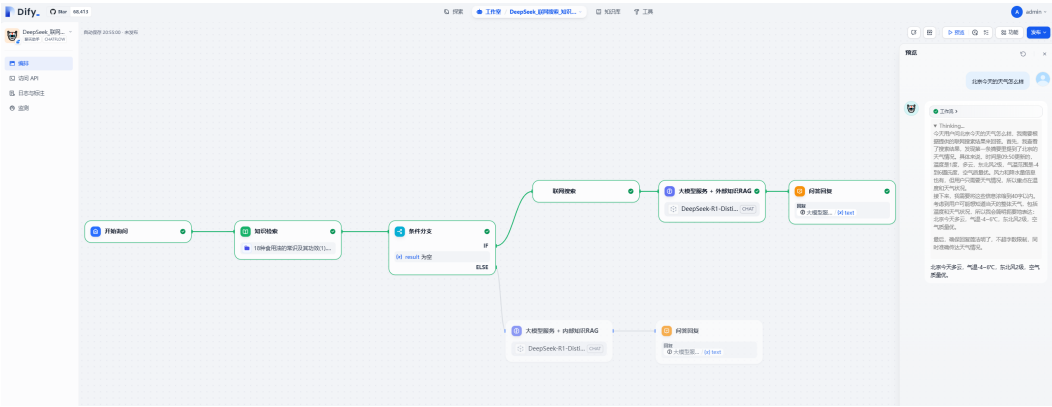
- 步骤14 至此联网搜索 workflow 配置已完成，可以进行测试并发布。
- 步骤15 在工作流页面，单击“预览”按钮，在弹出的聊天框输入问题进行测试。测试从知识库中进行内容检索：输入知识库相关内容，查看是否通过知识库检索。

图 3-47 知识库检索



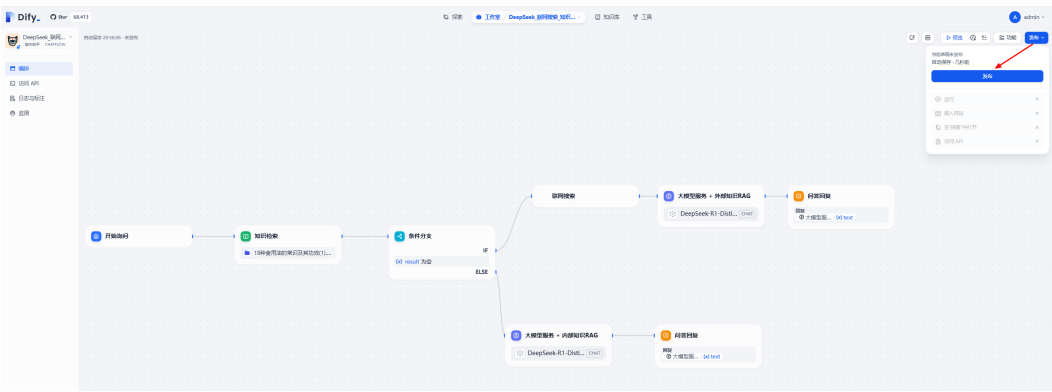
- 步骤16 测试从联网搜索中进行内容检索：输入知识库外的相关内容，查看是否通过联网搜索进行检索。

图 3-48 联网搜索



步骤17 在工作流页面的右上角单击“发布”按钮，再单击“发布”，即可完成工作流发布。

图 3-49 工作流列表



步骤18 发布完成后就可以使用此agent应用了，有以下三种使用方式：访问API、直接访问、嵌入网站。

图 3-50 访问 API

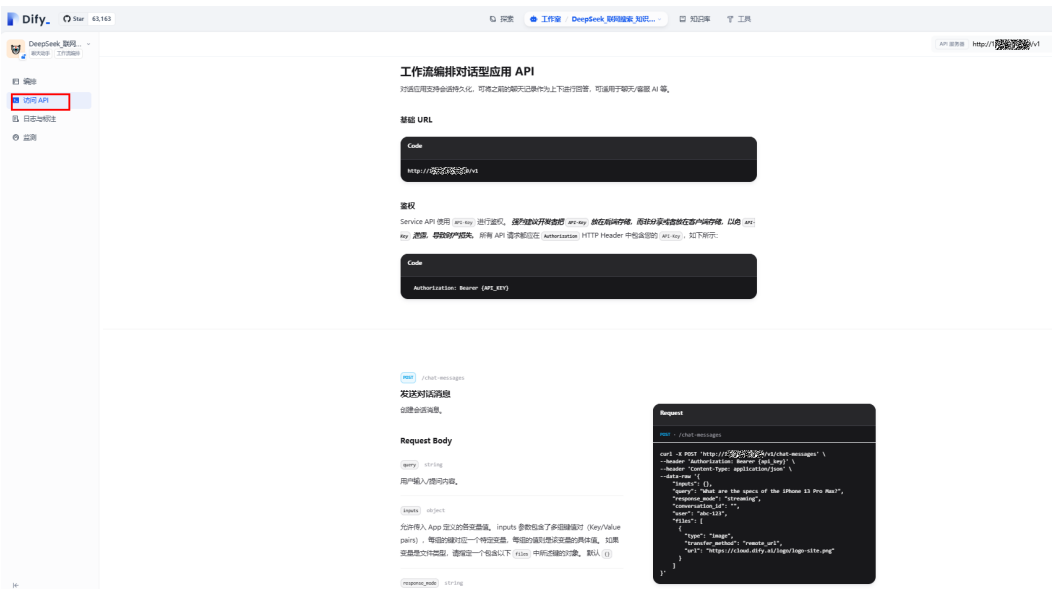


图 3-51 直接访问

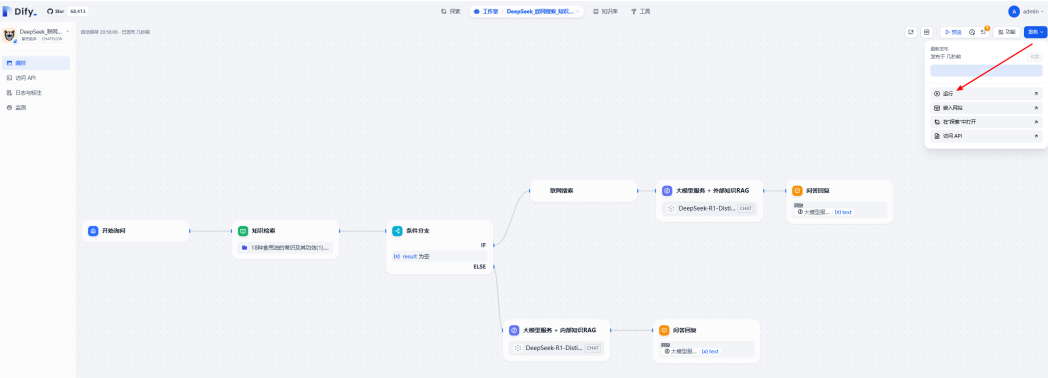
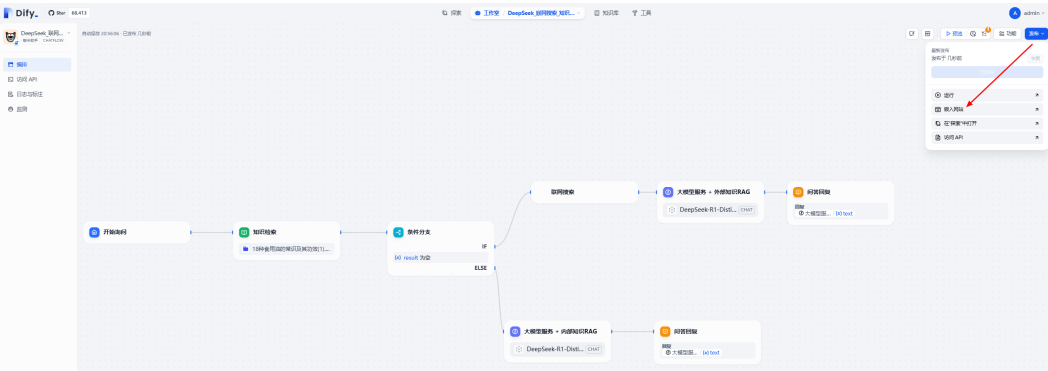


图 3-52 嵌入网站

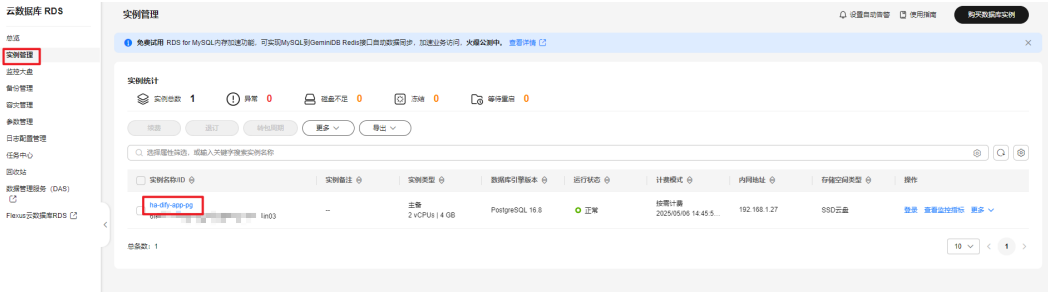


----结束

3.4 快速卸载

- 步骤1 删除堆栈前请确保Dify对外服务已停止，如数据库连接存在的话请先断开进程后再尝试删除堆栈，数据库进程终止操作请参考以下步骤2-3。
- 步骤2 进入[RDS控制台](#)，单击数据库实例名称，进入数据库详情页面。

图 3-53 数据库实例



- 步骤3 依次单击“智能DBA助手>实时诊断”，可以看到当前数据库所有实时会话。选中所有数据库名为“dify”，用户名为“postgres”的所有进程，单击“Kill会话”并在弹窗中单击“确定”终止连接。

图 3-54 实时会话

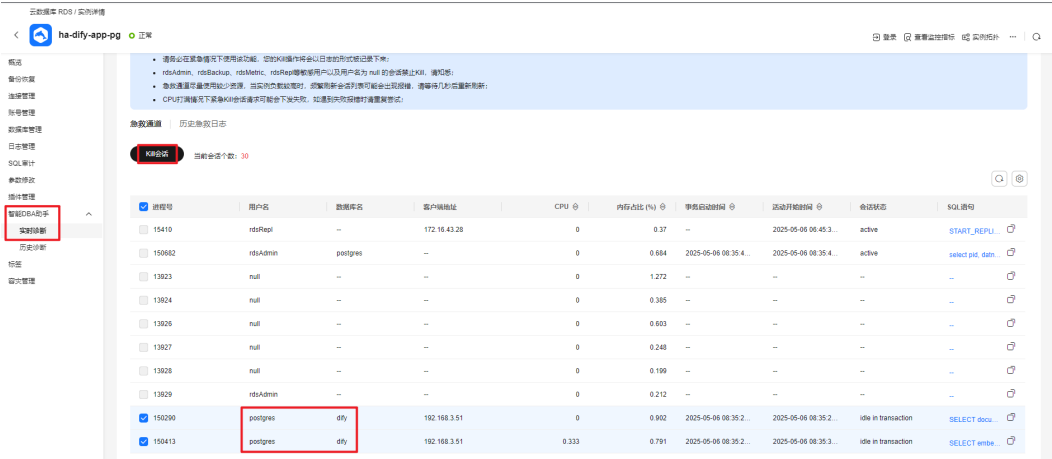
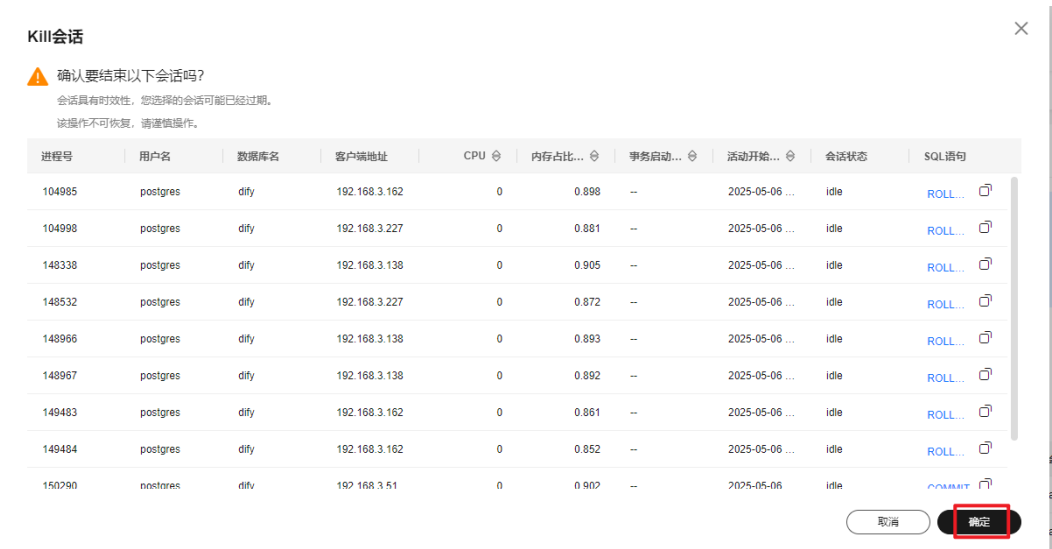


图 3-55 Kill 进程



步骤4 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 3-56 一键卸载



步骤5 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 3-57 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。
- 云容器引擎 CCE：云容器引擎（Cloud Container Engine，简称CCE）提供高度可扩展的、高性能的企业级Kubernetes集群。借助云容器引擎，您可以在华为云上轻松部署、管理和扩展容器化应用程序。
- 云搜索服务 CSS：云搜索服务（Cloud Search Service，简称CSS），为您提供托管的分布式搜索引擎服务，完全兼容开源Elasticsearch搜索引擎，支持结构化、非结构化文本的多条件检索、统计、报表。云搜索服务的使用流程和数据库类似。
- 分布式缓存服务 DCS：分布式缓存服务（Distributed Cache Service，简称DCS）为您提供即开即用、安全可靠、弹性扩容、便捷管理的在线分布式缓存能力，兼容Redis[®]，提供单机、主备、集群等丰富的实例类型，满足用户高并发及快速数据访问的业务诉求。
- 云数据库 RDS for PostgreSQL：PostgreSQL是一个开源对象云数据库管理系统，并侧重于可扩展性和标准的符合性，被业界誉为“先进的开源数据库”。云数据库 RDS for PostgreSQL面向企业复杂SQL处理的OLTP在线事务处理场景，支持NoSQL数据类型（JSON/XML/hstore），支持GIS地理信息处理，在可靠性、数据完整性方面有良好声誉，适用于互联网网站、位置应用系统、复杂数据对象处理等应用场景。
- NAT网关 NAT：NAT网关（NAT Gateway）提供公网NAT网关和私网NAT网关。公网NAT网关为VPC内的云主机提供SNAT和DNAT功能，可轻松构建VPC的公网出入口。私网NAT网关为VPC内的云主机提供网络地址转换服务，使多个云主机

可以共享私网IP访问用户本地数据中心或其他VPC，并支持云主机面向私网提供服务。

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2025-09-02	更新CCE容器高可用
2025-04-30	支持CCE容器高可用部署。
2025-02-11	第一次正式发布。