解决方案实践

快速搭建 Dify-LLM 应用开发平台

文档版本	1.0.0
发布日期	2026-01-28

华为技术有限公司

商标声明


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

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（社区版单机部署）

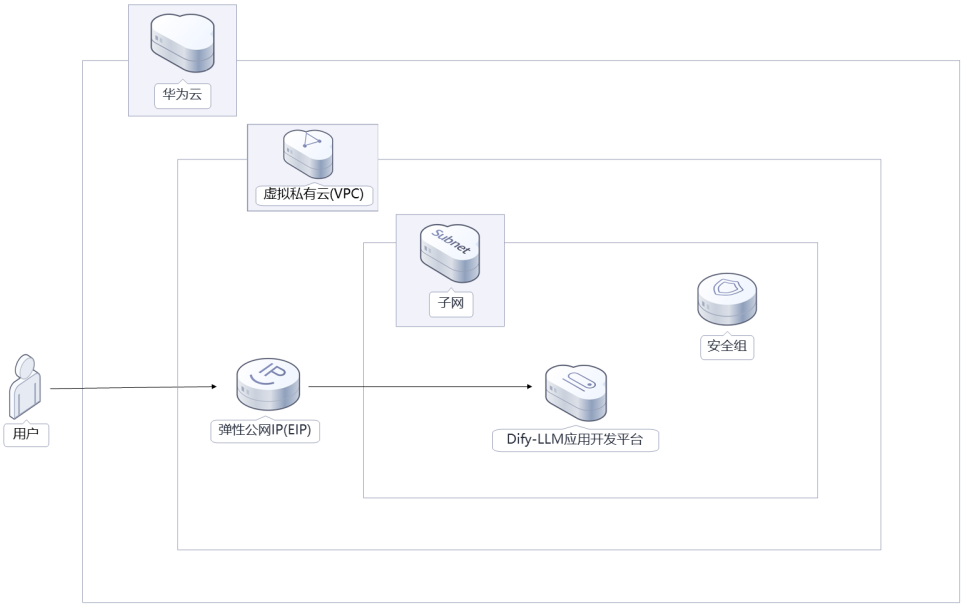


图 1-2 方案架构图（知识库搜索增强版）

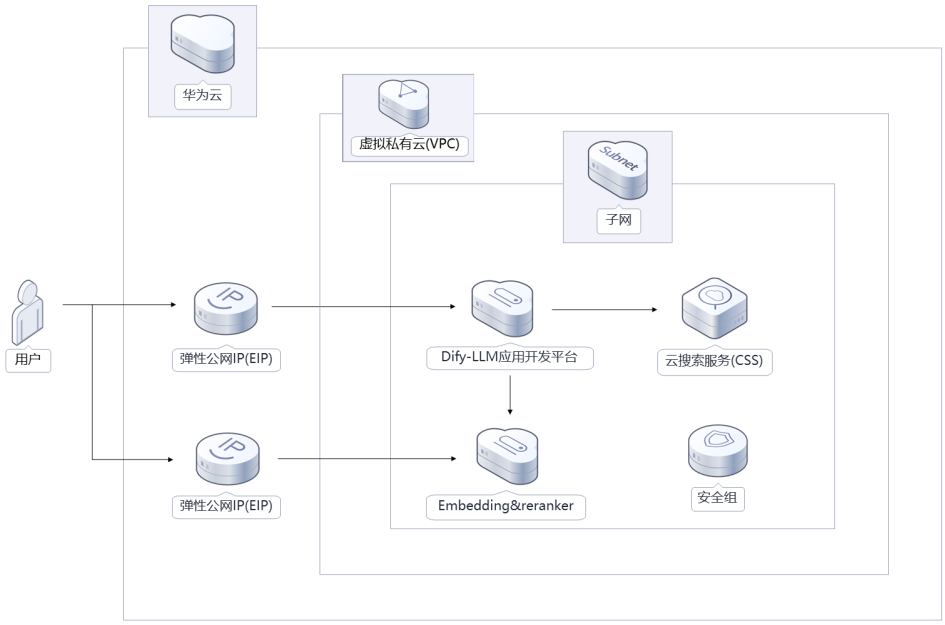
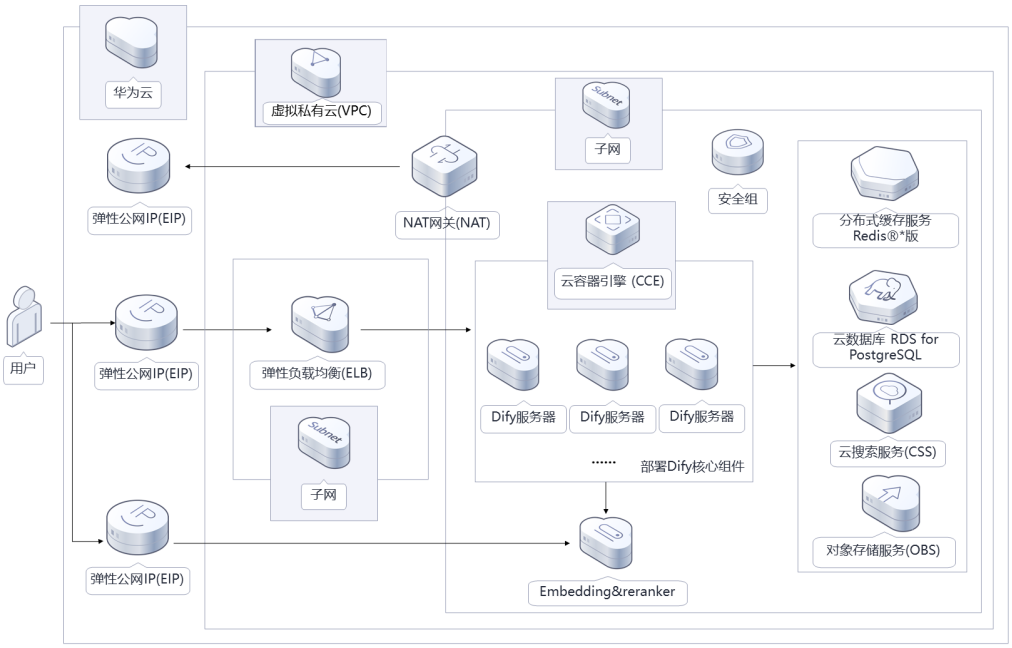


图 1-3 方案架构图（CCE 容器高可用版）



该解决方案将会部署如下资源：

社区版单机部署：

- 创建1台**华为云Flexus云服务器X实例**（FlexusX），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。

知识库搜索增强版：

- 创建1台FlexusX实例，用于搭建Dify-LLM应用开发平台。
- 创建1台FlexusX实例，用于部署Embedding（bge-m3）及Reranker（bge-reranker-v2-m3）模型。
- 创建2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**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容器高可用版：

- 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，用于安装部署Dify5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个华为云**Flexus云服务器X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

- 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。

- 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；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，实际以收费账单为准。

社区版单机部署

表 2-1 资源和成本规划（按需计费）

华为云服务	配置示例	数量	每月预估花费
虚拟私有云 VPC	• 区域：华北-北京四 • VPC网段：172.16.0.0/16	1	0.00元
子网 Subnet	• 区域：华北-北京四 • 子网网段：172.16.1.0/24 • 网关：172.16.1.1	1	0.00元
安全组 SecurityGroup	• 区域：华北-北京四 • 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	0.00元

华为云服务	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	• 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	1	683.28元
弹性公网IP EIP	• 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	1	0.80元/GB
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**683.28元 + 弹性公网IP EIP费用+ MaaS tokens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	数量	每月预估花费
虚拟私有云 VPC	• 区域：华北-北京四 • VPC网段：172.16.0.0/16	1	0.00元

华为云服务	配置示例	数量	每月预估花费
子网 Subnet	• 区域：华北-北京四 • 子网网段：172.16.1.0/24 • 网关：172.16.1.1	1	0.00元
安全组 SecurityGroup	• 区域：华北-北京四 • 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	0.00元
华为云Flexus云服务器X实例	• 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	1	467.00元
弹性公网IP EIP	• 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	1	0.80元/GB

华为云服务	配置示例	数量	每月预估花费
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**467.00元 + 弹性公网IP EIP费用+ MaaS tokens费用**

知识库搜索增强版

表 2-3 资源和成本规划（按需计费）

华为云服务	配置示例	数量	每月预估花费
虚拟私有云 VPC	• 区域：华北-北京四 • VPC网段：172.16.0.0/16	1	0.00元
子网 Subnet	• 区域：华北-北京四 • 子网网段：172.16.1.0/24 • 网关：172.16.1.1	1	0.00元
安全组 SecurityGroup	• 区域：华北-北京四	1	0.00元

华为云服务	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	• 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	1	683.28元
华为云Flexus云服务器X实例	• 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（开启）| x1e.16u.16g | 16核 | 16 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 100GB	1	1686.96元
云搜索服务 CSS	• 按需计费：1.33元/小时 • 区域：华北-北京四 • 计费模式：按需计费 • 规格： ess.spec-4u8g | 4 vCPUs | 8 GB • 节点存储总容量：超高I/O | 40GB • 集群类型：OpenSearch • 节点数：1	1	954.72元
弹性公网IP EIP	• 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	1	0.80元/GB

华为云服务	配置示例	数量	每月预估花费
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**3324.96元 + 弹性公网IP EIP费用+ MaaS tokens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	• 区域：华北-北京四 • VPC网段：172.16.0.0/16	1	0.00元
子网 Subnet	• 区域：华北-北京四 • 子网网段：172.16.1.0/24 • 网关：172.16.1.1	1	0.00元
安全组 SecurityGroup	• 区域：华北-北京四	1	0.00元
华为云Flexus云服务器X实例	• 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	1	467.00元

华为云服务	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	• 区域：华北-北京四 • 规格：Flexus云服务器X实例 | 性能模式（开启）| x1e.16u.16g | 16核 | 16 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 100GB	1	1150.00元
云搜索服务 CSS	• 区域：华北-北京四 • 计费模式：包年包月 • 规格： ess.spec-4u8g | 4 vCPUs | 8 GB • 节点存储总容量：超高I/O | 40GB • 集群类型：OpenSearch • 节点数：1	1	654.74元
弹性公网IP EIP	• 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	1	0.80元/GB
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**2271.74元 + 弹性公网IP EIP费用+ MaaS tokens费用**

CCE 容器高可用版

表 2-5 资源和成本规划（按需计费）

华为云服务	配置示例	数量	每月预估花费
虚拟私有云 VPC	• 区域：华北-北京四 • VPC网段：192.168.0.0/16	1	0.00
子网 Subnet	• 区域：华北-北京四 • 子网网段：192.168.1.0/24，192.168.2.0/24，192.168.3.0/24，192.168.4.0/24 • 网关：192.168.0.1，192.168.1.1，192.168.2.1，192.168.3.1	4	0.00
安全组 SecurityGroup	• 区域：华北-北京四	4	0.00
华为云Flexus云服务器X实例	• 按需计费：1.42元/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）| x1.16u.16g | 16核 | 16 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：高IO | 40GB • 数据盘：高IO | 100 GiB	3	3064.18元
华为云Flexus云服务器X实例	• 按需计费：4.53元/小时 • 区域：华北-北京四 • 规格：Flexus云服务器X实例 | 性能模式（开启）| x1e.32u.32g | 32核 | 32 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 40GB	1	3262.18元

华为云服务	配置示例	数量	每月预估花费
弹性公网IP EIP	• 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	3	0.80元/GB
对象存储服务 OBS	• 区域：华北-北京四 • 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/PUT 0.01元/万次，DELETE 免费 • 存储空间：0.1390 元/GB/月 • 流量费用： – 内/公网流入流量（数据上传到OBS）：0元 – 内网流出流量（通过ECS云服务器下载OBS的数据）：0元公网流出流量 / 00:00-08:00（闲时）：0.2500元/GB 公网流出流量 / 08:00-24:00（忙时）： 0.5000元/GB	1	详细请参考每月账单。计费说明参考**价格详情**
云容器引擎 CCE	• 按需计费：2.91元/小时 • 区域：华北-北京四 • 计费模式：按需计费 • 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	2095.20元

华为云服务	配置示例	数量	每月预估花费
分布式缓存服务 Redis®*版	● 按需计费：0.58元/小时 ● 区域：华北-北京四 ● 计费模式：按需计费 ● 规格：4G (基础版）| 副本数：2 ● 实例类型：Redis(主备)	1	414.72元
云数据库 RDS for PostgreSQL	● 按需计费：5.82元/小时 ● 区域：华北-北京四 ● 计费模式：按需计费 ● 规格：rds.pg.x1.2xlarge.4.ha | 8 vCPUs | 32 GB (独享型) ● 储存：SSD云盘 | 100GB ● 数据库引擎：PostgreSQL(主备)	1	4190.4元
云搜索服务 CSS	● 按需计费：3.98元/小时 ● 区域：华北-北京四 ● 计费模式：按需计费 ● 规格： ess.spec-4u8g | 4 vCPUs | 8 GB ● 节点存储总容量：超高I/O | 120GB ● 集群类型：OpenSearch ● 节点数：3	1	2864.16元
弹性负载均衡 ELB	● 区域：华北-北京四 ● 可用区数量：2 ● 计费模式：按需计费 ● 独享型负载均衡 ● 网络型 | 弹性规格、应用型 | 弹性规格 ● 按需计费：¥0.15/小时+应用型LCU费用：¥0.05/个·小时（按实际使用量收取LCU费用）	1	108元 + 应用型LCU费用
NAT网关 NAT	● 按需计费：12元/天 ● 区域：华北-北京四 ● 规格：小型 ● SNAT规则数：3	1	360元

华为云服务	配置示例	数量	每月预估花费
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**16358.84元 + 应用型LCU费用 + 弹性公网IP EIP费用 + 对象存储服务 OBS存储及流量费用+ MaaS tokens费用**

表 2-6 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	• 区域：华北-北京四 • VPC网段：192.168..0.0/16	1	0.00
子网 Subnet	• 子网网段：192.168.1.0/24，192.168.2.0/24，192.168.3.0/24，192.168.4.0/24 • 网关：192.168.0.1，192.168.1.1，192.168.2.1，192.168.3.1	4	0.00
安全组 SecurityGroup	• 区域：华北-北京四	4	0.00

华为云服务	配置示例	数量	每月预估花费
华为云Flexus云服务器X实例	• 按需计费：1.42元/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）| x1.16u.16g | 16核 | 16 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：高IO | 40GB • 数据盘：高IO | 100 GiB	3	2,091.00元
华为云Flexus云服务器X实例	• 计费模式：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（开启）| x1e.32u.32g | 32核 | 32 GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 40GB	1	2,188.00元
弹性公网IP EIP	• 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	3	0.80元/GB

华为云服务	配置示例	数量	每月预估花费
对象存储服务 OBS	• 区域：华北-北京四 • 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/PUT 0.01元/万次，DELETE 免费 • 存储空间：0.1390 元/GB/月 • 流量费用： – 内/公网流入流量（数据上传到OBS）：0元 – 内网流出流量（通过ECS云服务器下载OBS的数据）：0元公网流出流量 / 00:00-08:00（闲时）：0.2500元/GB 公网流出流量 / 08:00-24:00（忙时）： 0.5000元/GB	1	详细请参考每月账单。计费说明参考**价格详情**
云容器引擎 CCE	• 区域：华北-北京四 • 计费模式：包年包月 • 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	1262.40元
分布式缓存服务 Redis®*版	• 区域：华北-北京四 • 计费模式：包年包月 • 规格：4G (基础版) | 副本数：2 • 实例类型：Redis(主备)	1	277.60元

华为云服务	配置示例	数量	每月预估花费
云数据库 RDS for PostgreSQL	• 区域：华北-北京四 • 计费模式：包年包月 • 规格：rds.pg.x1.2xlarge.4.ha | 8 vCPUs | 32 GB (独享型) • 储存：SSD云盘 | 100GB • 数据库引擎：PostgreSQL(主备)	1	2860.00元
云搜索服务 CSS	• 区域：华北-北京四 • 计费模式：包年包月 • 规格： ess.spec-4u8g | 4 vCPUs | 8 GB • 节点存储总容量：超高I/O | 120GB • 集群类型：OpenSearch • 节点数：3	1	1964.22元
弹性负载均衡 ELB	• 区域：华北-北京四 • 可用区数量：2 • 计费模式：按需计费 • 独享型负载均衡 • 网络型 | 弹性规格、应用型 | 弹性规格 • 按需计费：¥0.15/小时+应用型LCU费用：¥0.05/个·小时（按实际使用量收取LCU费用）	1	108元 + 应用型LCU费用
NAT网关 NAT	• 区域：华北-北京四 • 计费模式：包年包月 • 规格：小型 • SNAT规则数：3	1	306.00元

华为云服务	配置示例	数量	每月预估花费
MaaS tokens 计费（可选）	• 计费模式：按Token计费 • 模型：DeepSeek-V3-64K • 输入价格：0.002元 / 千tokens • 输出价格：0.008 / 千tokens • 模型：bge-reranker-v2-m3/BGE-M3 • 输入价格：0.00007 / 千tokens • 联网搜索功能 • 价格：50元/千次	-	输入0.00207元 / 千tokens 输出0.008元 / 千tokens 联网搜索：50元/千次
合计	-	-	**11056.62元 + 应用型LCU费用+ 弹性公网IP EIP费用 + 对象存储服务 OBS存储及流量费用+ MaaS tokens费用**

3 实施步骤

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。

（可选）创建 rf_admin_trust 委托

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

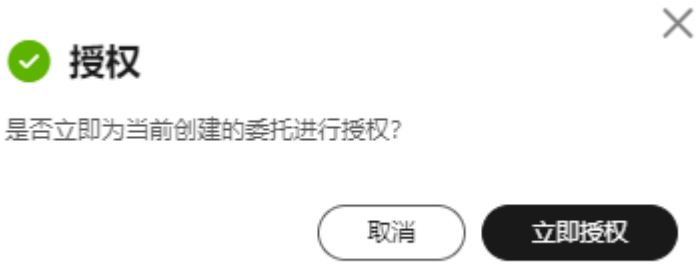
- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

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

步骤4 单击“立即授权”。

图 3-5 委托授权



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

（可选）登录 CCE 控制台授权

该章节只用于CCE高可用版本，该解决方案创建CCE集群需要您登录过CCE控制台并进行过相关授权，**否则首次开通CCE请参考如下步骤进行**授权 。

步骤1 使用华为账号登录**CCE控制台**。

步骤2 单击控制台左上角的  ，选择区域。

步骤3 在首次登录某个区域的CCE控制台时将跳出“授权说明”，请您在仔细阅读后单击“确定”。

----结束

获取 OBS 桶名（CCE 高可用版本）

CCE高可用版本obs桶须和方案部署region一致

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（参数配置）

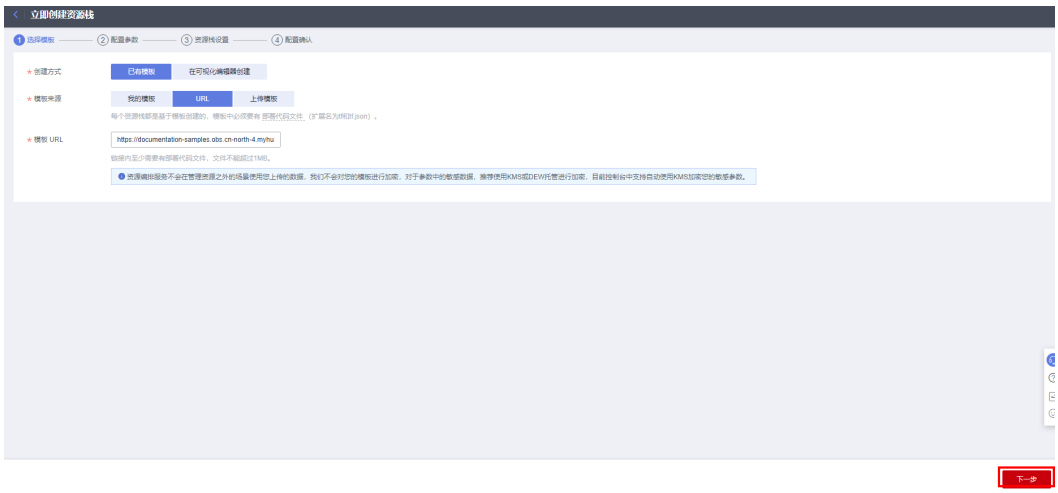
本章节帮助用户高效地部署“快速搭建Dify-LLM应用开发平台”解决方案。一键部署该解决方案时，参照本章节中的步骤和说明进行操作，即可完成快速部署。

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，以高可用部署为示例，单击“一键部署（CCE容器高可用版）”，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3-3**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（社区版单机部署）

参数名称	类型	是否可选	参数解释	默认值
dify_version	string	必填	Dify应用开发平台社区版版本，支持v1.8.1、v1.6.0、v1.4.1、v1.1.3及v0.15.8。	1.8.1
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
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。修改密码。管理员账户默认root。	空
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
charging_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

表 3-2 参数说明（知识库搜索增强版）

参数名称	类型	是否可选	参数解释	默认值
dify_version	string	必填	Dify应用开发平台社区版版本，支持v1.8.1、v1.6.0、v1.4.1、v1.1.3及v0.15.8。	1.8.1

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
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，支持数字、字母、中文、_(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，支持中文、英文字母、数字、_（下划线）、-（中划线）、.（点）。	dify-llm-application-development-platform-demo
css_name	string	必填	css实例名称。取值范围：4-32个字符，必须以字母开头，英文字母、数字、_（下划线）、-（中划线）。	dify-llm-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。弹性云服务器规格请参考官网**弹性云服务器规格清单**。	x1.8u.16g
embedding_reranker_flavor	string	必填	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x1e.?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。	x1e.16u.16g

参数名称	类型	是否可选	参数解释	默认值
ecs_password	string	必填	云服务器密码，取值范围：8-26个字符，密码至少包含大写字母、小写字母、数字和特殊字符（!@%^_=+[{}]:,.?）中的三种。管理员账户默认root。	空
css_password	string	必填	云搜索服务密码，取值范围：8-32个字符，密码至少包含大写字母、小写字母、数字和特殊字符（!@%^_=+[{}]:,.?）中的三种。管理员账户默认admin。	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
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	string	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。默认订购1月。	1

表 3-3 参数说明（CCE 容器高可用版）

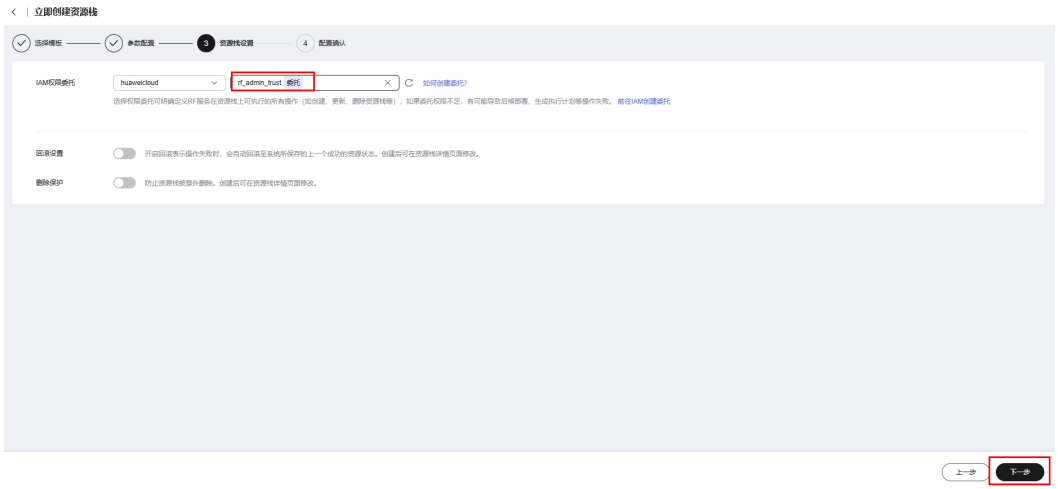
参数名称	类型	是否可选	参数解释	默认值
dify_version	string	必填	社区版Dify版本。可以选择1.7.1, 1.4.1, 0.15.8。默认1.7.1	1.7.1
resource_name_prefix	string	必填	资源名称前缀，命名规则{resource_name_prefix}-资源英文名称，例如：CCE集群名称为{resource_name_prefix}-cce。取值范围：1-28个字符，支持小写字母、数字、-（中划线）。必须以小写字母开头。禁止以中划线（-）开头。	ha-dify-app
bandwidth_size	string	必填	弹性公网带宽大小，该模板计费方式为按流量计费。单位：Mbit/s，取值范围：1-300Mbit/s。	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，具体请参考部署指南。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：8-26个字符，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。	空
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。弹性云服务器规格请参考官网**弹性云服务器规格清单**。	x1.16u.16g

参数名称	类型	是否可选	参数解释	默认值
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**实例类型	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（~!^*-=_+,）。	空
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。	空
redis_capacity	number	必填	分布式缓存服务 Redis版实例规格。可选值：1GB-64GB。	4
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（~!^*-=_+,）。	空
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（CCE高可用版本）**。	空
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，用于将生成的图像上传至OBS桶。参考**获取AK、SK密钥**。	空
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，用于将生成的图像上传至OBS桶。参考**获取AK、SK密钥**。	空

参数名称	类型	是否可选	参数解释	默认值
embedding_reranker_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x1e.?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 弹性云服务器规格请参考官网**弹性云服务器规格清单**。 可替换成GPU加速型获得更高性能。	x1e.32u.32g
ecs_password	string	可选	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
charging_mode	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
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

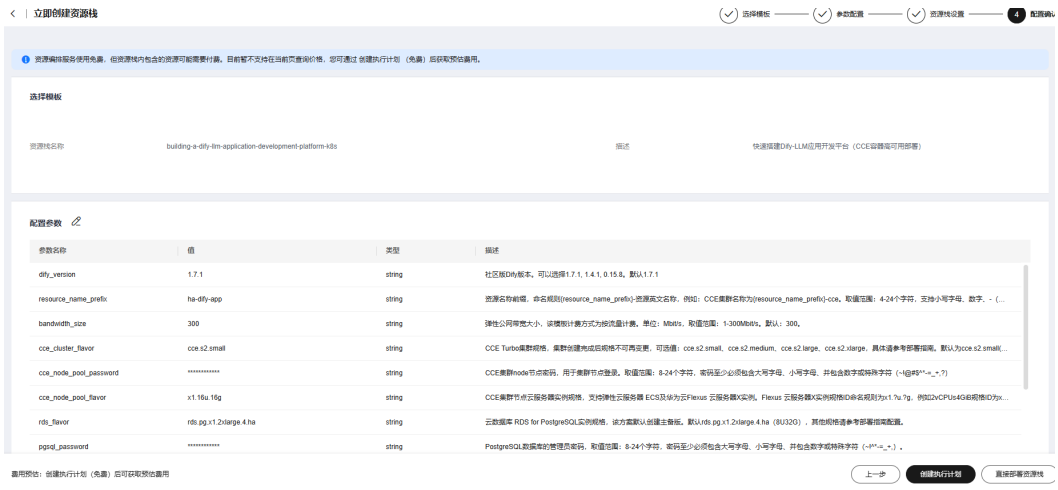
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

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

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

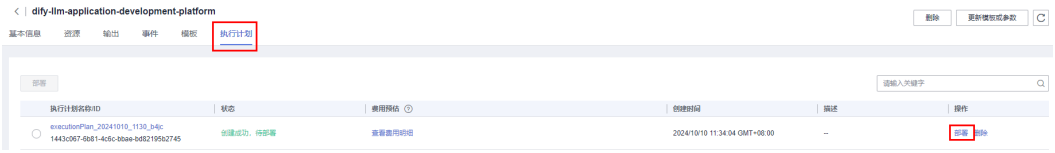
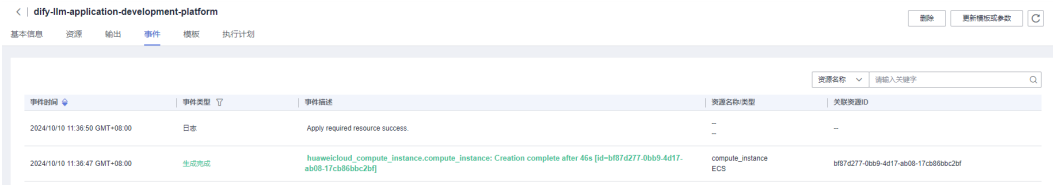


图 3-18 执行计划确认

步骤8 （可选）如果计费模式选择“包年包月”，在余额不充足的情况下（所需总费用请参考**2 资源和成本规划**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5-20分钟左右（受网络波动影响）。

图 3-20 说明

----结束

3.3 一键部署（快速选购）

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 进入**DeepSeek应用专场**，选择“智能问答”，单击“购买解决方案（高可用版）”跳转至解决方案购买页面。

图 3-21 购买解决方案

步骤2 进入解决方案购买页面，设置基础配置。

图 3-22 基础配置

表 3-4 基础配置说明

参数	说明
计费模式	根据业务特点选择适用的计费模式。 ● 包年/包月是预付费模式，按购买周期计费，适用于可预估资源使用周期的场景，价格比按需计费模式更优惠。 ● 按需计费是后付费模式，按资源的实际使用时长计费，可以随时开通、删除。 **说明** 仅对解决方案中支持该计费模式的服务或资源生效，例如选择了包年/包月计费模式，弹性公网IP仍会按实际使用的流量计费。详情请查看具体服务的配置项。
区域	请就近选择靠近您业务的区域，可以降低网络时延，提高访问速度。创建后无法更换区域，请谨慎选择。
解决方案名称	系统自动生成，建议自定义为方便您识别的解决方案名称。 支持添加描述，可填写解决方案的更多相关信息。
购买时长	单次购买最短为1个月，最长为3年。

步骤3 设置3台Flexus云服务器X实例配置（应用容器节点），用于部署Dify-LLM应用开发平台。

图 3-23 Flexus 云服务器 X 实例配置

表 3-5 Flexus 云服务器 X 实例配置说明

参数	说明
规格	请根据业务需要选择合适的规格。单击 ˅ 可调整CPU/内存配比、数据盘容量。 **注意** 选择含性能模式规格后，可以享受极致稳定性能SLA保障，但需要支付额外算力费用。请参见**Flexus X实例性能模式说明**。
密码	设置云服务器密码。长度为8~24位，密码至少必须包含大写字母、小写字母、数字和特殊字符（~!@#$^*-=_+,?）。用户名默认为root。
购买数量	不支持修改，默认购买数量为3。

步骤4 设置1台Flexus云服务器X实例配置（向量化/排序模型节点），用于部署Embedding、Reranker模型。

图 3-24 Flexus 云服务器 X 实例配置

表 3-6 Flexus 云服务器 X 实例配置说明

参数	说明
规格	请根据业务需要选择合适的规格。单击 可调整CPU/内存配比。
密码	设置云服务器密码。长度为8~24位，密码至少必须包含大写字母、小写字母、数字和特殊字符（~!@#^*=_+,?）。用户名默认为root。
购买数量	不支持修改，默认购买数量为1。

步骤5 设置云容器引擎 CCE配置。

图 3-25 云容器引擎 CCE 配置

表 3-7 云容器引擎 CCE 配置说明

参数	说明
集群类型	默认CCE Turbo集群：面向云原生2.0的新一代容器集群产品，计算、网络、调度全面加速。
集群master实例数	默认3实例：即步骤三创建的3台Flexus云服务器X实例。
集群规模	单次购买最小50节点，最大2000节点。即集群支持管理的最大节点数量，请根据业务场景选择。 **说明** 集群规模影响控制节点规格，创建完成后支持往高规格变更，不支持往低规格变更。
购买数量	不支持修改，默认购买数量为1。

步骤6 设置分布式缓存Redis配置。

图 3-26 分布式缓存 Redis 配置

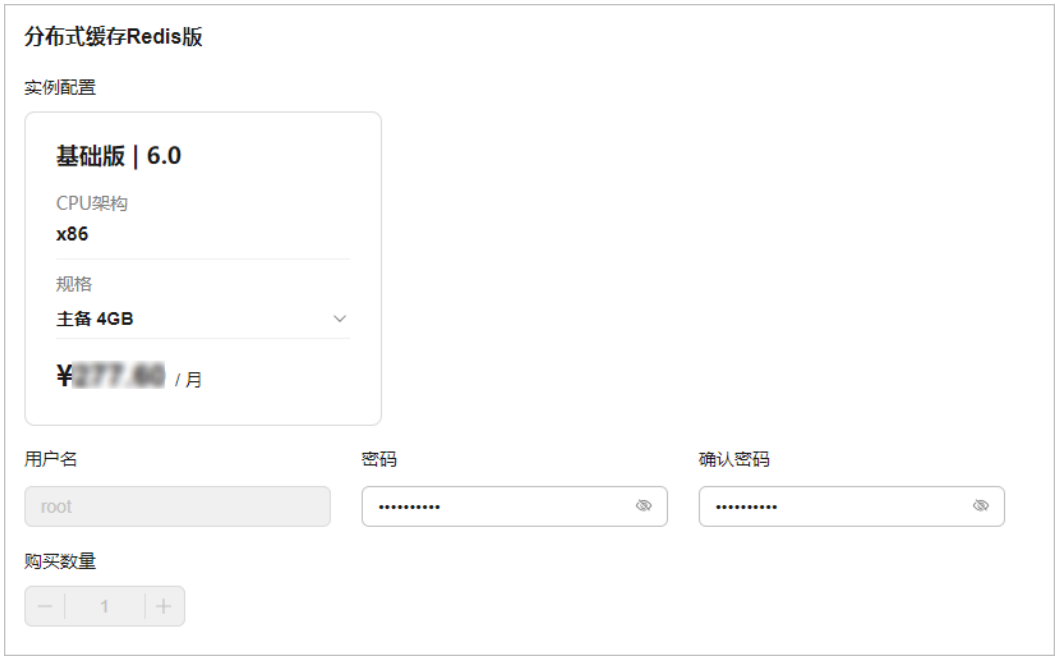


表 3-8 分布式缓存 Redis 配置说明

参数	说明
规格	请根据业务需要选择合适的规格。单击 ⌄ 可调整主备规格。
密码	设置用户密码。长度为8~24位，密码至少必须包含大写字母、小写字母、数字和特殊字符（ ~!^*-=_+, ）。用户名默认为root。
购买数量	不支持修改，默认购买数量为1。

步骤7 设置云数据库 RDS配置。

图 3-27 云数据库 RDS 配置

表 3-9 云数据库 RDS 配置说明

参数	说明
规格	请根据业务需要选择合适的规格。单击 ∨ 可调整性能规格、磁盘容量。
管理员账户名	设置管理员密码。长度为8~24位，密码至少必须包含大写字母、小写字母、数字和特殊字符（~!^*-=_+,）。管理员账户名默认为root。
用户名	设置PostgreSQL数据库的database用户密码。长度为8~24位，密码至少必须包含大写字母、小写字母、数字和特殊字符（~!^*-=_+,）。不能与用户名或倒序的用户名相同。用户名默认为postgres。
购买数量	不支持修改，默认购买数量为1。

步骤8 设置云搜索服务 CSS配置。

图 3-28 云搜索服务 CSS 配置

表 3-10 云搜索服务 CSS 配置说明

参数	说明
规格	请根据业务需要选择合适的规格。单击 可调整节点规格、数据盘容量。
购买数量	不支持修改，默认购买数量为3。

步骤9 设置弹性负载均衡 ELB配置。

图 3-29 弹性负载均衡 ELB 配置

表 3-11 弹性负载均衡 ELB 配置说明

参数	说明
计费模式	默认按需计费：即按资源的实际使用时长计费。
实例类型	默认独享型：适用于大流量高并发的业务场景，如大型网站、云原生应用、车联网、多可用区容灾应用。 **说明** 独享型负载均衡可以独享已创建的实例资源，同时分别提供了应用型（HTTP/HTTPS）和网络型（TCP/UDP/TLS）两种类型的实例规格。请参见**选型说明**。
购买数量	不支持修改，默认购买数量为1。

步骤10 设置弹性公网 IP配置。

图 3-30 弹性公网 IP 配置

表 3-12 弹性公网 IP 配置说明

参数	说明
线路	默认全动态BGP：可根据设定的寻路协议实时自动优化调整网络结构，以保证客户网络的持续稳定和高效运行。
公网带宽	默认按流量计费：按照实际使用的流量来计费。
带宽大小	根据业务需要，选择所需的带宽大小。单位：Mbit/s，取值范围：1~300Mbit/s。
购买数量	不支持修改，默认购买数量为3。

步骤11 设置公网 NAT网关配置。

图 3-31 公网 NAT 网关配置

表 3-13 公网 NAT 网关配置说明

参数	说明
规格	请根据业务需要选择合适的规格。 **说明** 包年/包月计费模式的NAT网关规格选定后，后续不支持降级。
购买数量	不支持修改，默认购买数量为1。

步骤12 设置对象存储服务 OBS配置。

图 3-32 对象存储服务 OBS 配置

表 3-14 对象存储服务 OBS 配置说明

参数	说明
OBS桶	支持选择已有的OBS桶。若无OBS桶，请先前往创建，创建说明请参见**官方文档步骤一**。 **说明** 创建桶时需与解决方案中的资源在同一区域。

参数	说明
访问密钥ID（AK）	访问密钥ID（AK），识别访问用户的身份，长度为20位，仅支持大写字母和数字，用于将生成的图像上传至OBS桶。请参见**获取访问密钥AK/SK**。
秘密访问密钥（SK）	秘密访问密钥（SK），对请求数据进行签名验证，长度为40位，仅支持大小写字母和数字，用于将生成的图像上传至OBS桶。请参见**获取访问密钥AK/SK**。

步骤13 查看页面右侧的“配置概要”，确认解决方案配置详情。

图 3-33 配置概要

步骤14 单击“一键部署”，系统将通过自动支付完成扣款，请确保账户余额充足。

注意

- 创建新账户或账户余额不足时，请进入“费用中心>**充值**”页面进行充值。具体请参见**账户充值**。
- 若选择了包年/包月计费模式，因账户余额不足导致自动支付失败，请进入“费用中心>**待支付订单**”页面，手动完成费用支付。

步骤15 一键部署后自动跳转至资源栈详情页面。

步骤16 待“事件”页面出现“Apply required resource success”，表示该解决方案已经部署完成。

图 3-34 部署完成

步骤17 刷新页面，在“输出”页面查看访问链接。

图 3-35 查看访问链接

说明

请耐心等待5~10分钟左右（受网络波动影响），待应用下载成功后，即可输入网址访问Dify开发平台，具体请参见**开始使用**。

----结束

3.4 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 3-36 设置管理员账户

步骤2 在登录界面，依次输入上一步骤中的“邮箱”“密码”登录Dify平台。

图 3-37 登录 Dify 平台

图 3-38 Dify 平台

----结束

说明

- 以下对接大模型方式**与MaaS服务对接**和**与一键部署DeepSeek对接**用户可根据自身情况选择使用。

与 MaaS 服务对接

步骤1 访问**ModelArts Studio**控制台在左侧导航栏中，选择“在线推理”进入“预置服务”服务列表，选择“商用服务”，如果未开通模型，请单击“开通服务”。

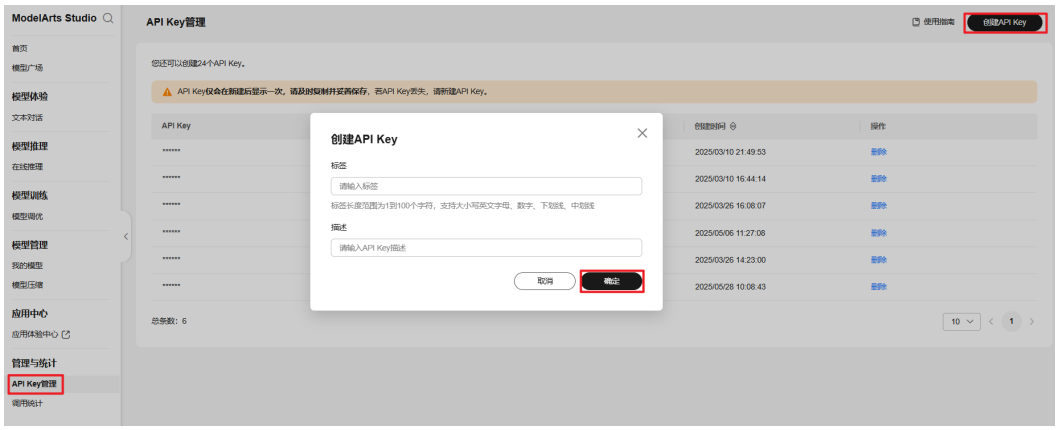
图 3-39 在线推理商用服务

步骤2 在调用MaaS部署的模型服务时，需要填写API Key用于接口的鉴权认证。在左侧导航栏，单击“API Key管理”（最多可创建30个密钥。**每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥**）。

图 3-40 API Key 管理

步骤3 在“API Key管理”页面，单击右上角“创建API Key”，填写标签（自定义API Key的标签，标签具有唯一性，不可重复。仅支持大小写英文字母、数字、下划线、中划线，长度范围为1~100个字符）和描述（自定义API Key的描述，长度范围为1~100个字符）信息后，单击“确定”。标签和描述信息在创建完成后，不支持修改。

图 3-41 创建 API Key



步骤4 对接Dify平台。打开您的Dify平台界面，单击右上角“**插件**”按钮。

图 3-42 插件安装

步骤5 单击“**探索Marketplace**”按钮，搜索“Maas”，选择**华为云Maas平台**，单击“**安装**”按钮，安装插件。

图 3-43 搜索 Maas 插件

图 3-44 安装插件

说明

已安装Maas插件，需要**更新MaaS插件至最新版本**。

步骤6 单击右上角用户名称，下拉并单击”**设置**”。

图 3-45 设置

步骤7 单击“**模型供应商**”，在右侧模型列表单击“**华为云MaaS平台**”列表中的“**设置**”。配置模型信息：**"API Key"**填写**步骤3**中创建的API Key，填写完成后单击“保存”，完成LLM模型的配置。

图 3-46 API KEY 设置

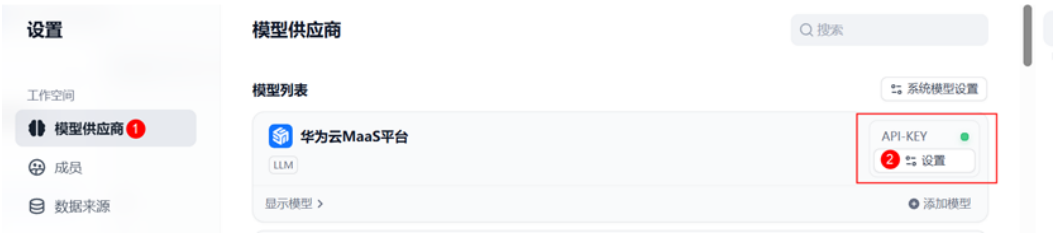


图 3-47 API KEY 填写

图 3-48 模型列表

----结束

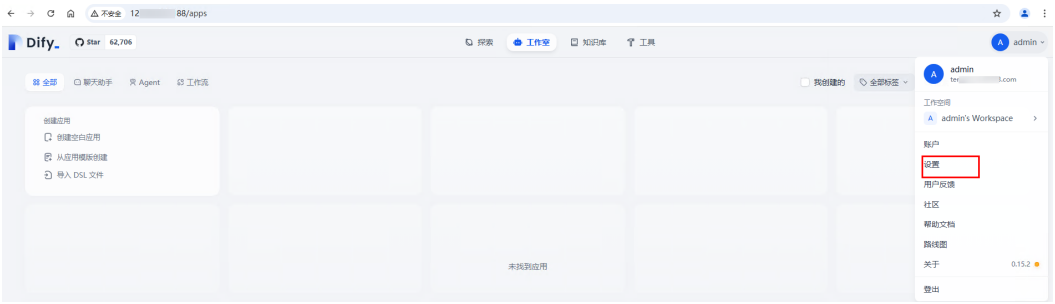
与一键部署 DeepSeek 对接

说明

该章节为成功部署**快速搭建DeepSeek推理系统**解决方案后，将DeepSeek大模型对接至Dify平台时参考使用。

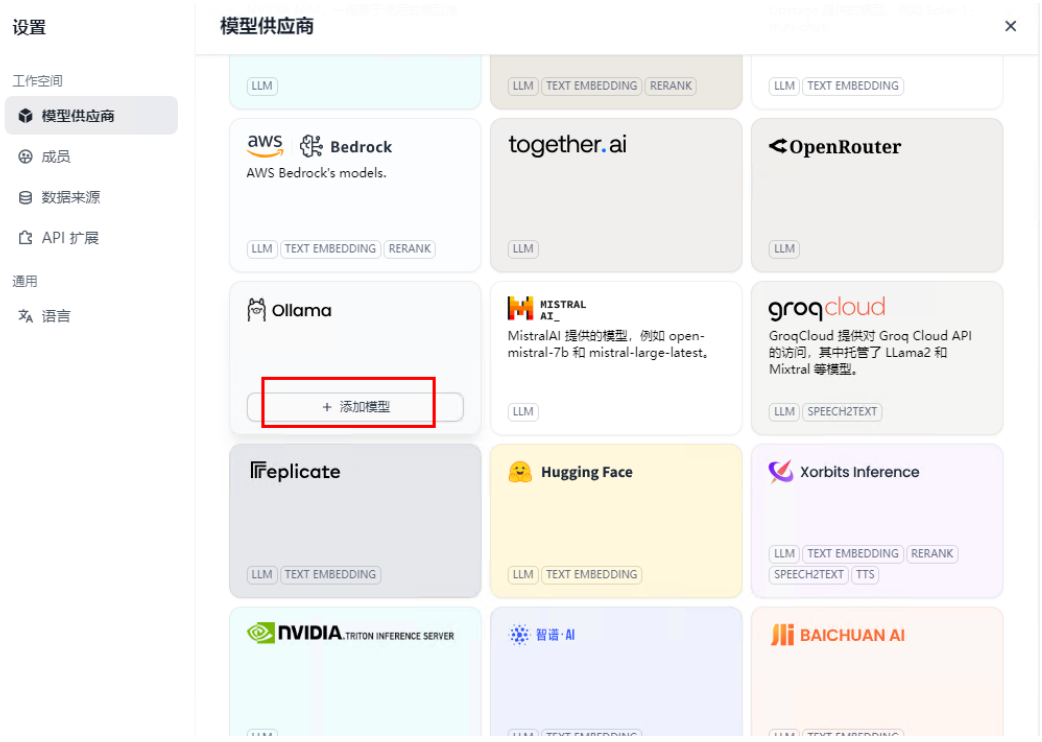
步骤1 右侧单击用户名称，下拉并单击“设置”。

图 3-49 设置



步骤2 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-50 添加 Ollama 模型



步骤3 模型名称填写**快速部署**中选择部署的模型，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通11434端口），单击右下角“保存”并关闭“设置”。

图 3-51 模型配置

----结束

创建知识库

说明

- 开通**Maas平台**bge-reranker-v2-m3和BGE-M3模型可直接创建高质量型知识库，步骤可参考**与MaaS服务对接**；未开通则需要提前添加Embedding和Reranker模型。
- 本解决方案一键部署（CCE容器高可用部署）提供可选的Embedding和Reranker模型服务器。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**，服务器URL填写私网IP地址（若不使用本方案部署的Embedding和Reranker模型，请填写公网IP）。

步骤1 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 3-52 创建知识库

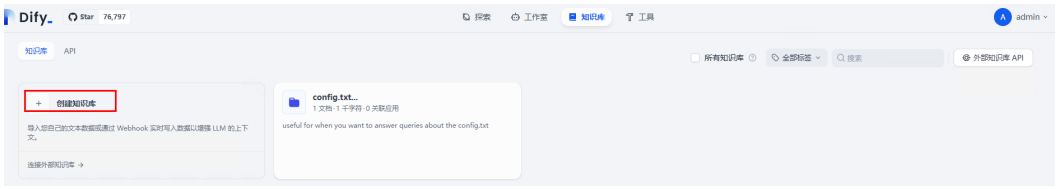
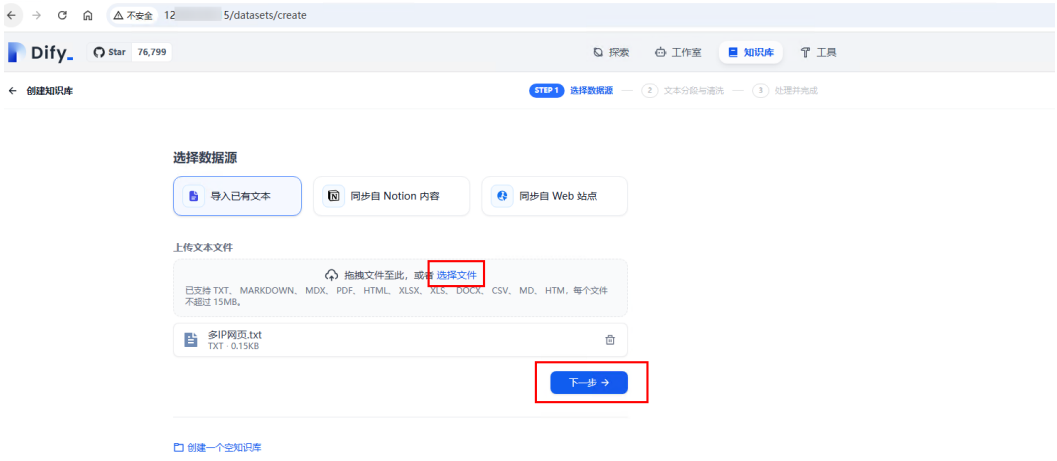


图 3-53 导入文件



步骤2 （**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。高质量型请参考以下**步骤3-5**。

图 3-54 配置经济型索引知识库

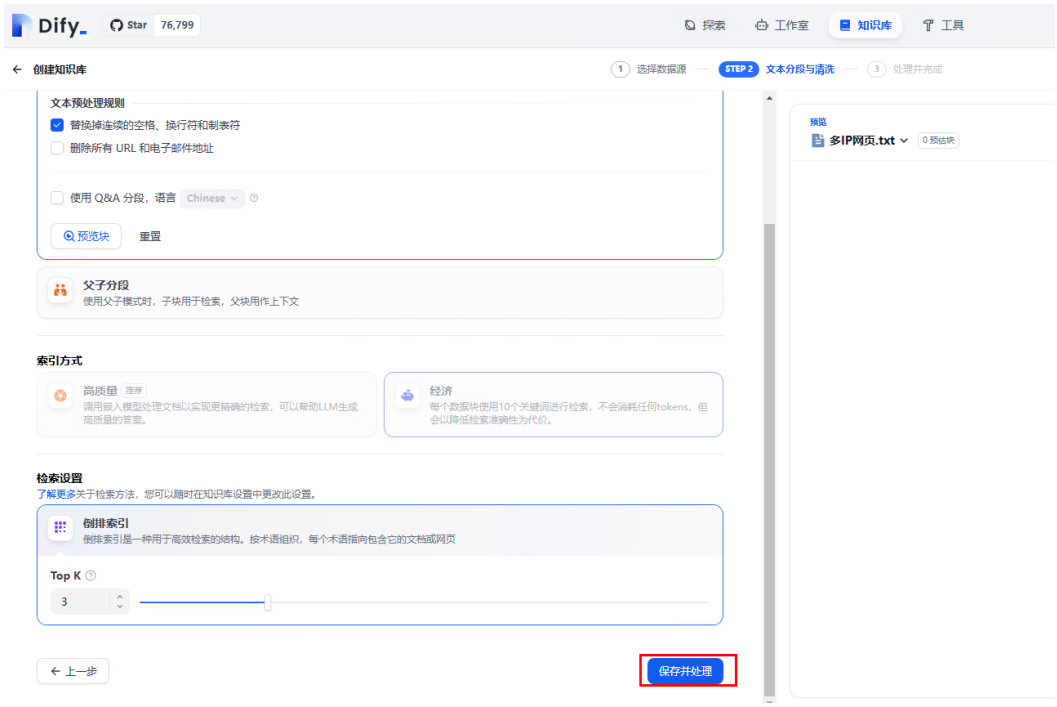
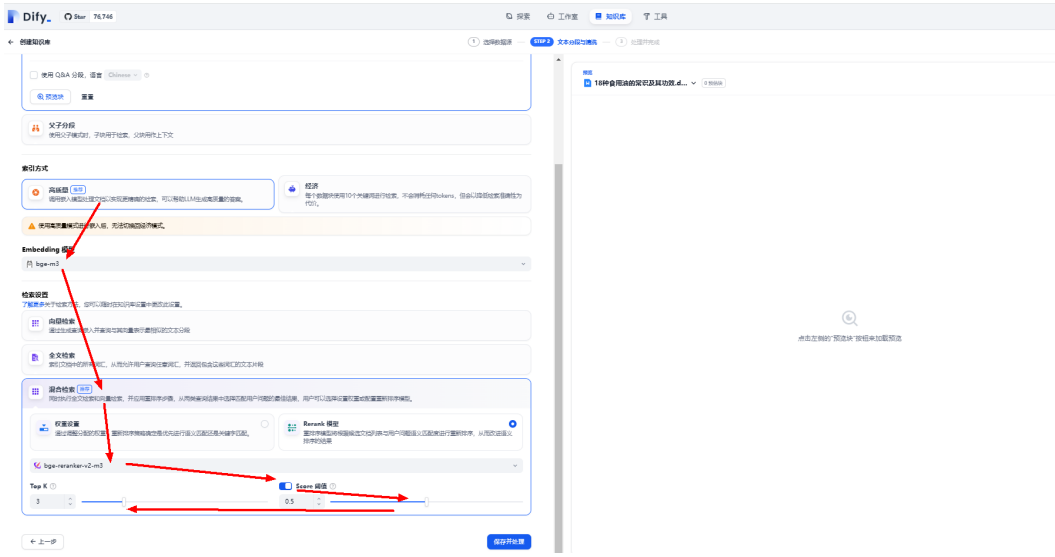


图 3-55 知识库配置完成

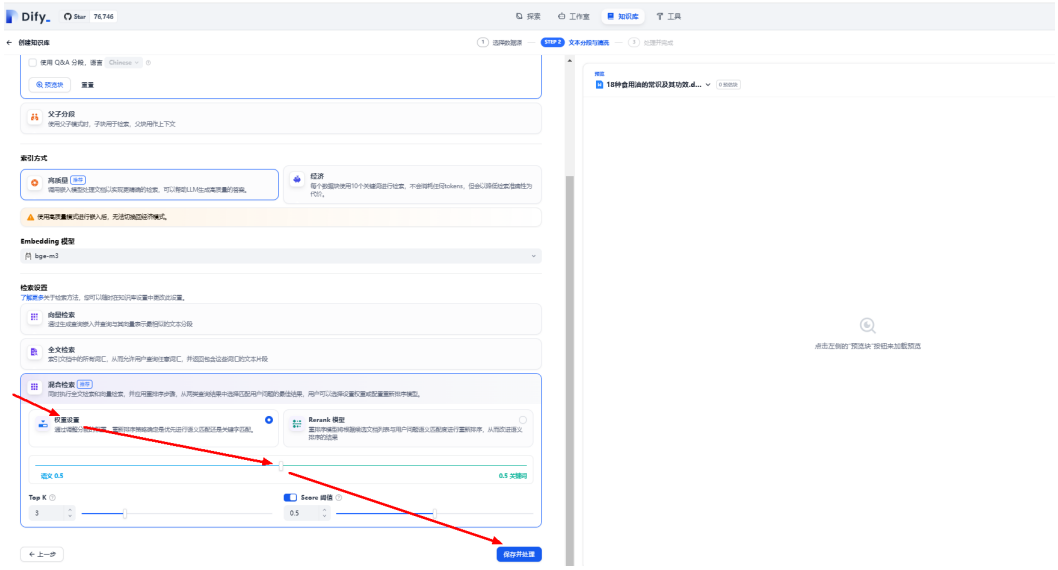
步骤3 （**高质量**）选择高质量模块，在Embedding模型栏目选择添加模型，在检索设置栏目，选择混合检索（可同时使用向量检索和关键词检索，并控制两者权重），在模型下拉窗口选择添加的bge-reranker-v2-m3模型，在下方开启Score阈值开关，并滑动下方的滑动按钮来选择阈值，推荐0.5，下方还有TOP K的选项，可以控制最终被使用的文档分片的最大数量。

图 3-56 配置高质量索引知识库



步骤4 （**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 3-57 配置语义等



步骤5 （**高质量**）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 3-58 嵌入完成

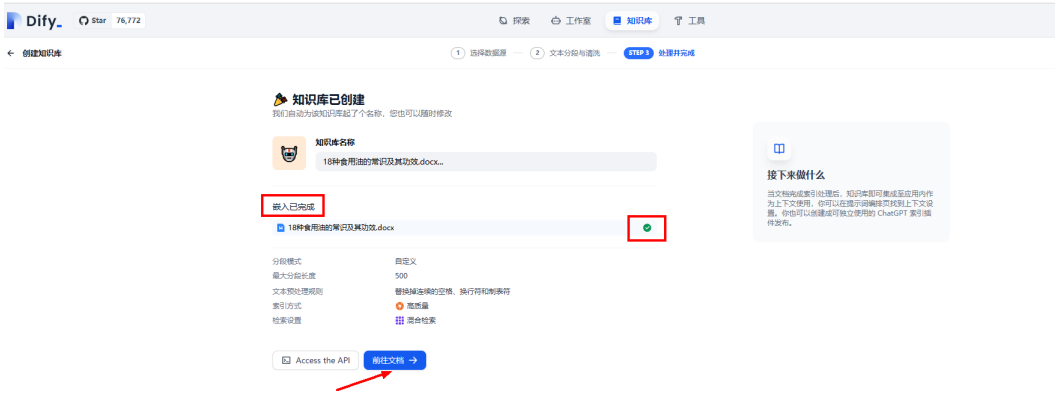


图 3-59 高质量型配置完成

----结束

说明

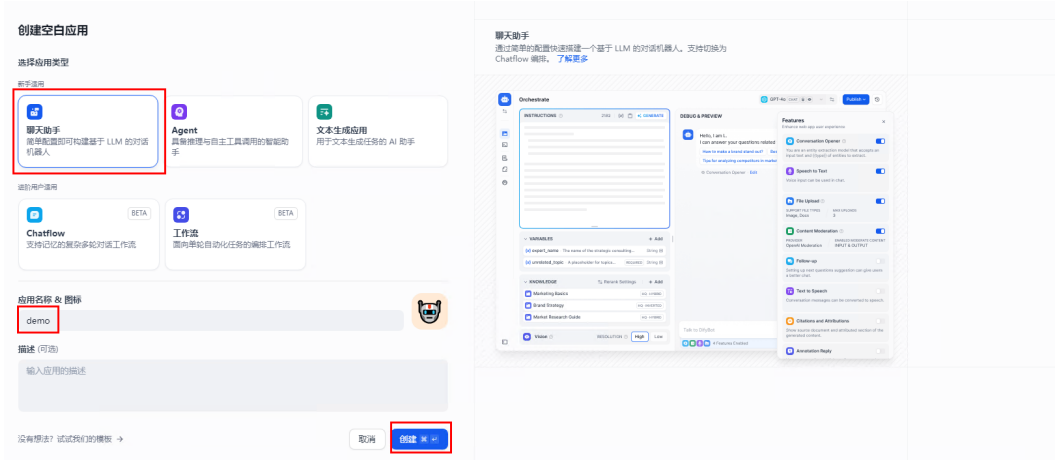
如果想要快速体验Dify应用平台，可参考**创建聊天助手**；如果想体验基于知识库检索和联网搜索服务的AI助手可以参考**创建工作流**。

创建聊天助手

步骤1 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 3-60 创建空白应用

图 3-61 创建聊天助手



步骤2 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 3-62 调试与预览

----结束

（可选）联网搜索

步骤1 在产品中搜索选择大模型即服务平台MaaS。

图 3-63 Maas 平台

步骤2 单击选择"**MCP广场**"。

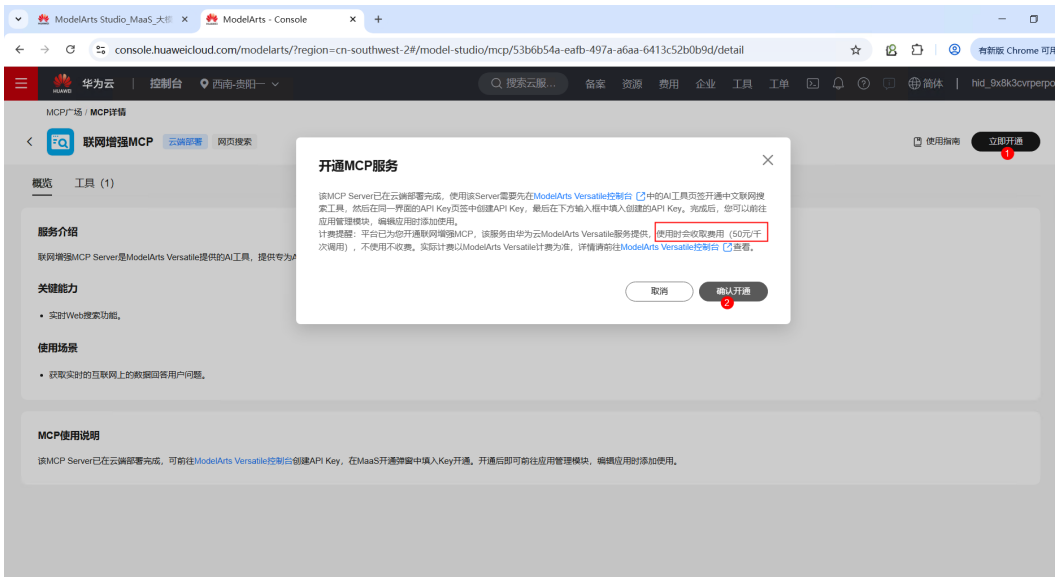
图 3-64 MCP 广场

步骤3 选择**联网增强MCP**。

图 3-65 联网增强 MCP

步骤4 依次单击“立即开通”、“确认开通”。

图 3-66 开通联网增强 MCP

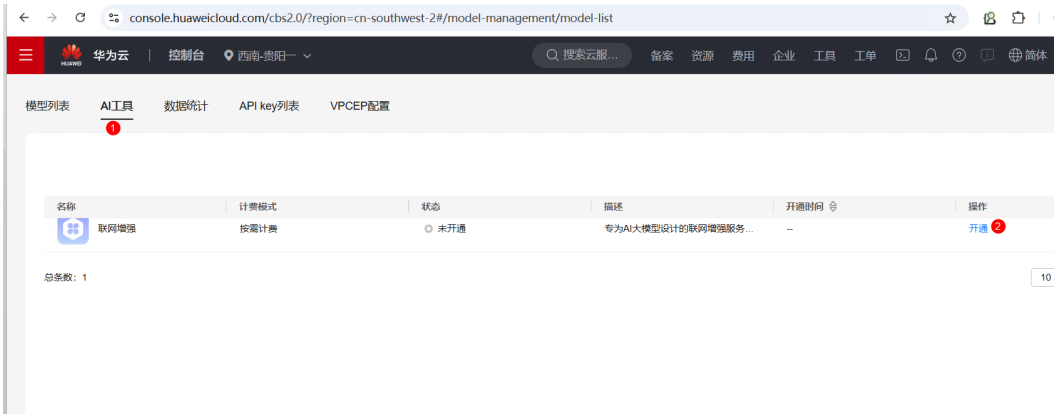


步骤5 服务开通后单击进入**ModelArts Versatile控制台**。

图 3-67 控制台设置

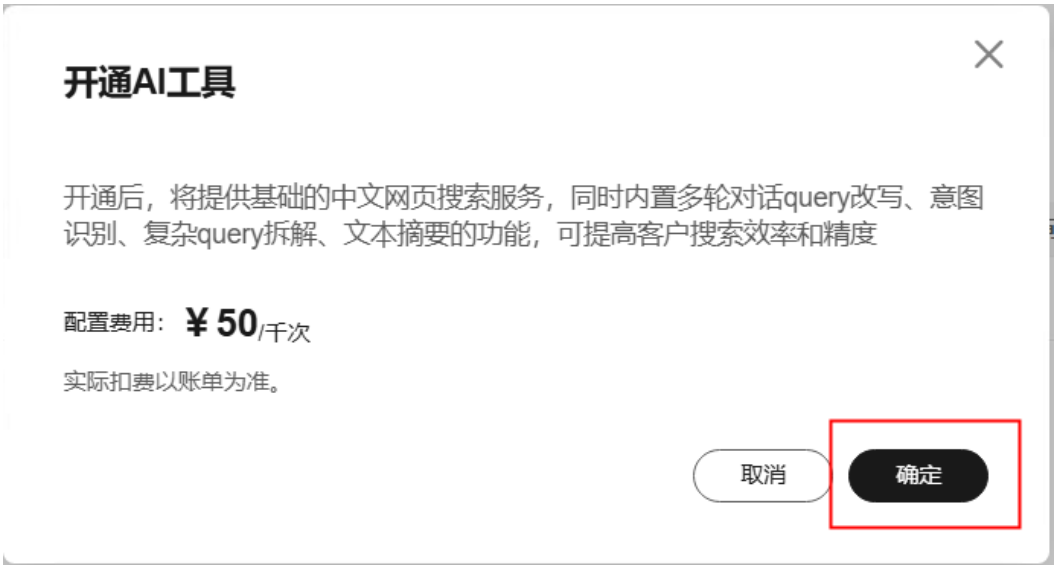
步骤6 控制台中单击AI工具，选择开通联网增强

图 3-68 开通联网增强



步骤7 确认开通工具

图 3-69 确认



步骤8 单击API调用，查看调用示例

图 3-70 API 调用

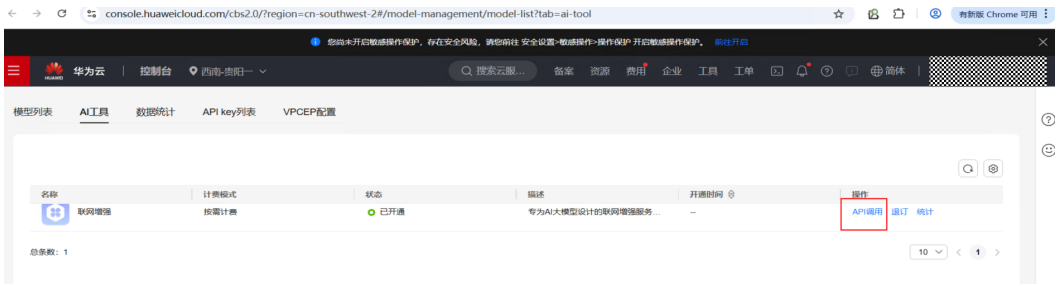
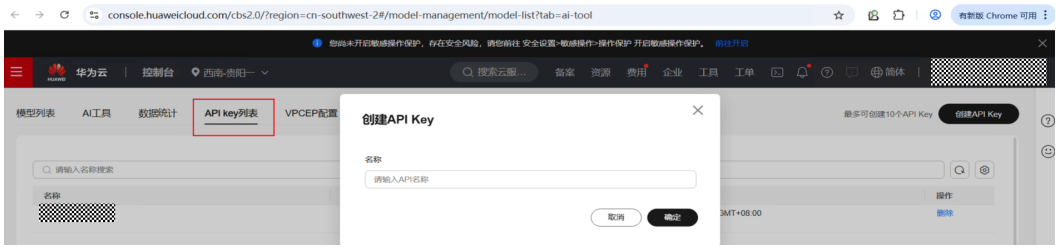


图 3-71 调用说明

步骤9 选择API key列表，单击"**创建API key**"按钮创建对应的API ley

图 3-72 创建 API key



----结束

创建工作流

步骤1 单击“工作室”，访问dify平台工作室。

图 3-73 工作室

步骤2 导入工作流。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，如下图所示：

```
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/%E6%99%BA%E8%83%BD%E5%8A%A9%E6%89%8B.yml
```

图 3-74 导入工作流

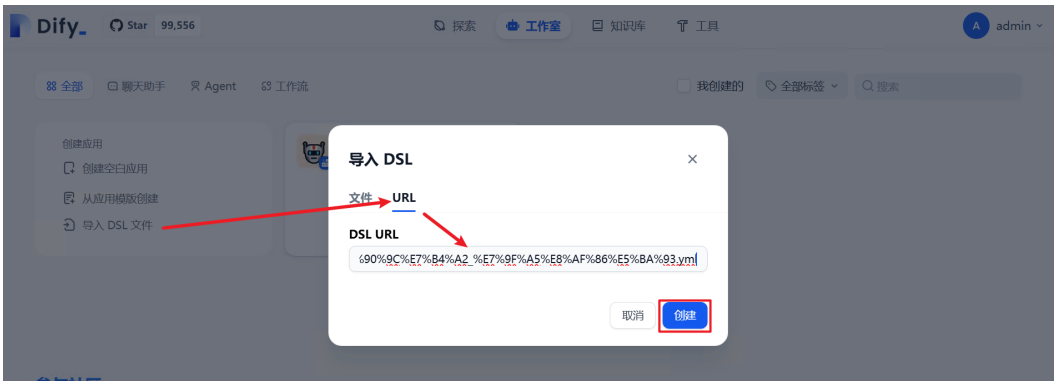
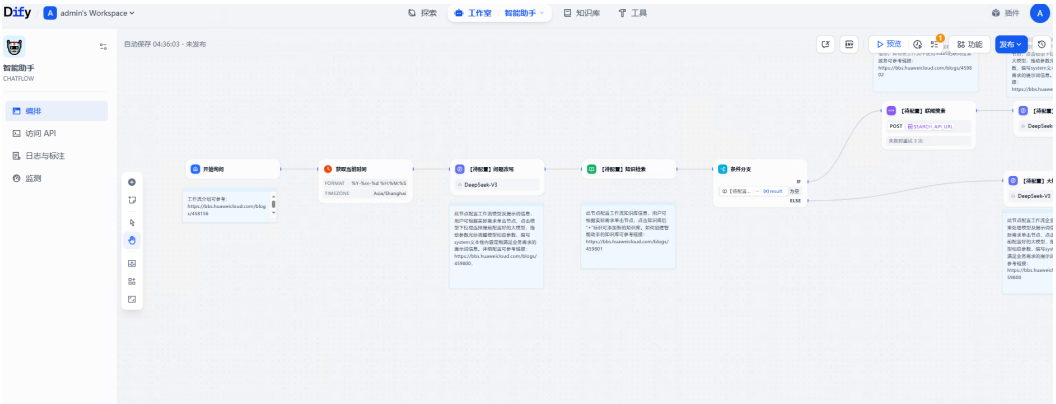


图 3-75 工作流模板



说明

- 首次导入，若未安装**华为云Maas平台**插件，弹窗提示安装插件，单击"**安装**"按钮完成安装

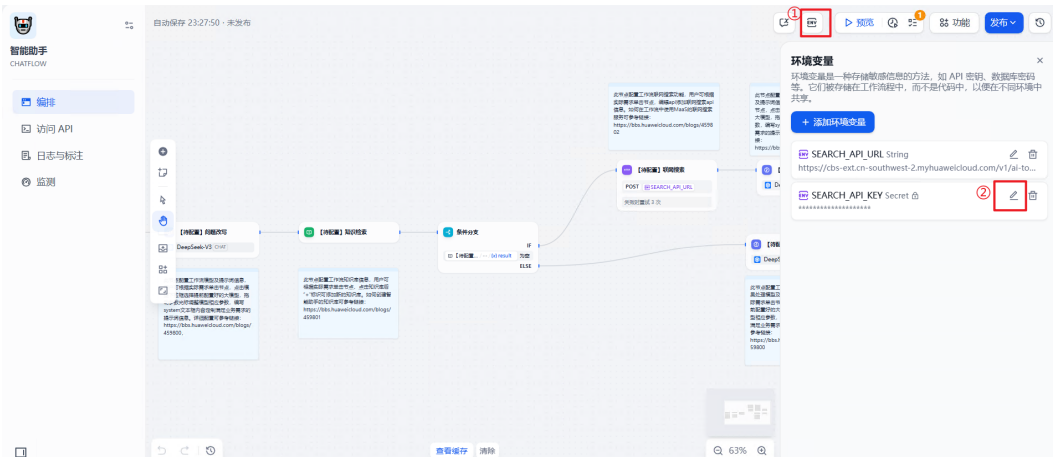
- 插件配置参考**与MaaS服务对接**中**步骤10-11**进行配置。

步骤3 配置知识库。添加知识库：单击“知识检索”节点，单击“+”按钮添加知识库，选择知识库，单击“添加”。

图 3-76 配置知识库

步骤4 联网搜索配置更改环境变量，"**SEARCH_API_KEY**"更改为编辑填写**（可选）联网搜索**的**步骤9**中的API key。

图 3-77 环境变量



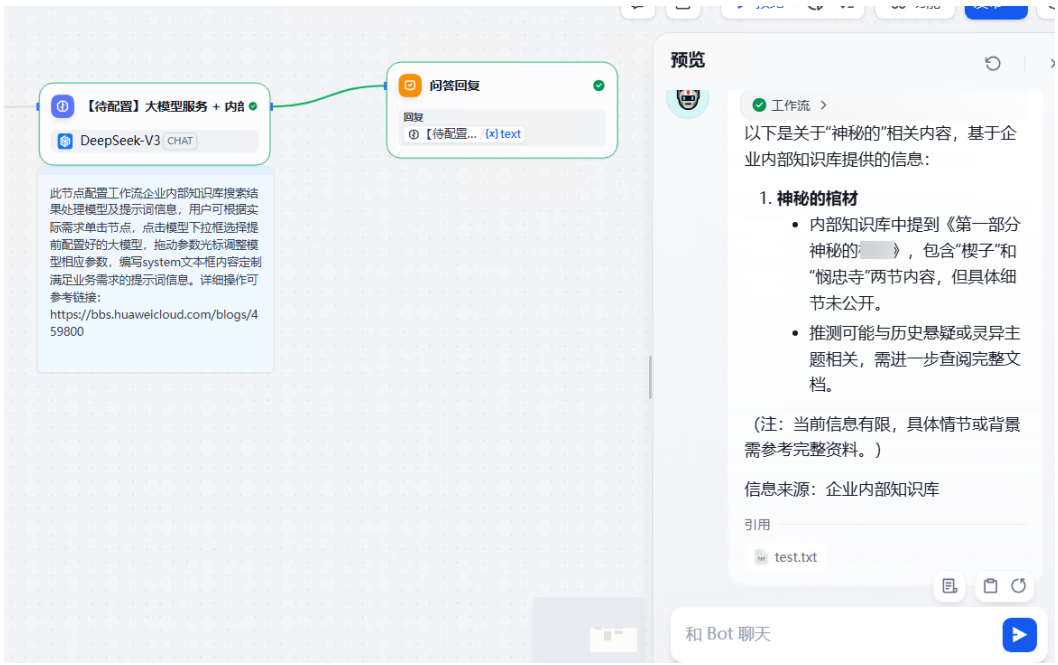
步骤5 调整模型。单击大模型名称，在右侧弹窗继续单击模型名称旁边的下拉列表，继续单击模型名称，在下拉列表中选择一个模型即可。其余大模型按照此步骤重复操作即可。

图 3-78 替换模型

步骤6 至此，工作流配置已完成，单击右上角“预览”进行对话。

图 3-79 联网搜索

图 3-80 知识检索



----结束

发布应用

步骤1 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 3-81 工作流列表

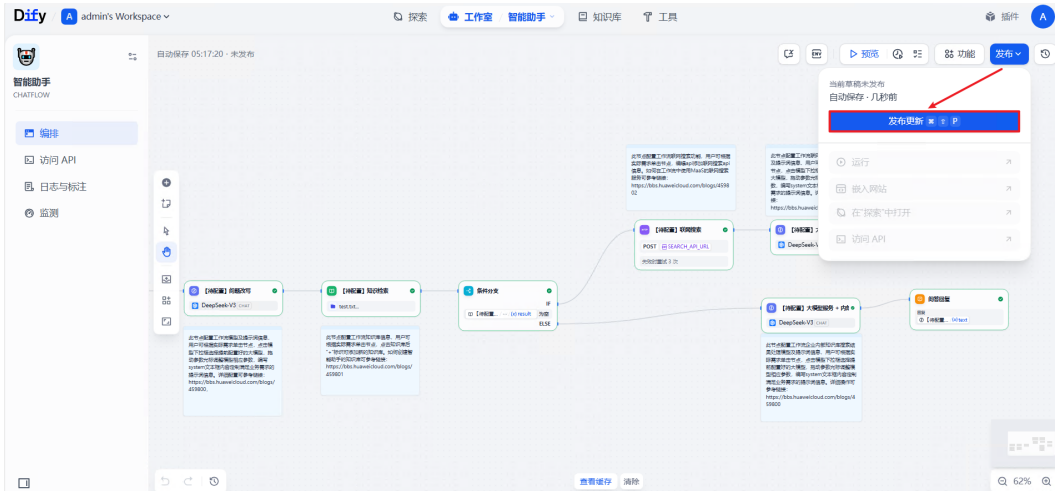


图 3-82 工作流访问方式

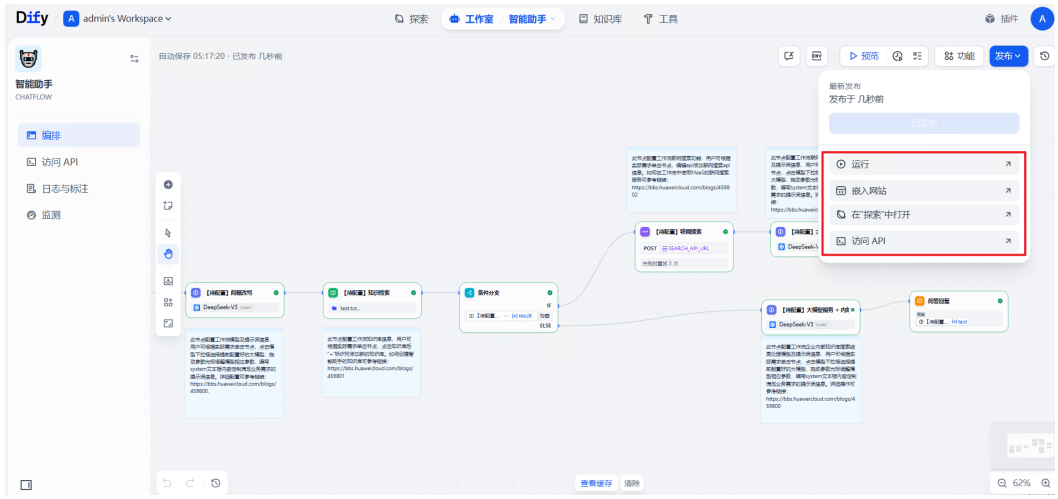


图 3-83 聊天应用访问方式

步骤2 发布完成后就可以使用此Agent应用/工作流，有以下三种使用方式：访问API、直接访问、嵌入网站、在“探索”中打开。

图 3-84 访问 API

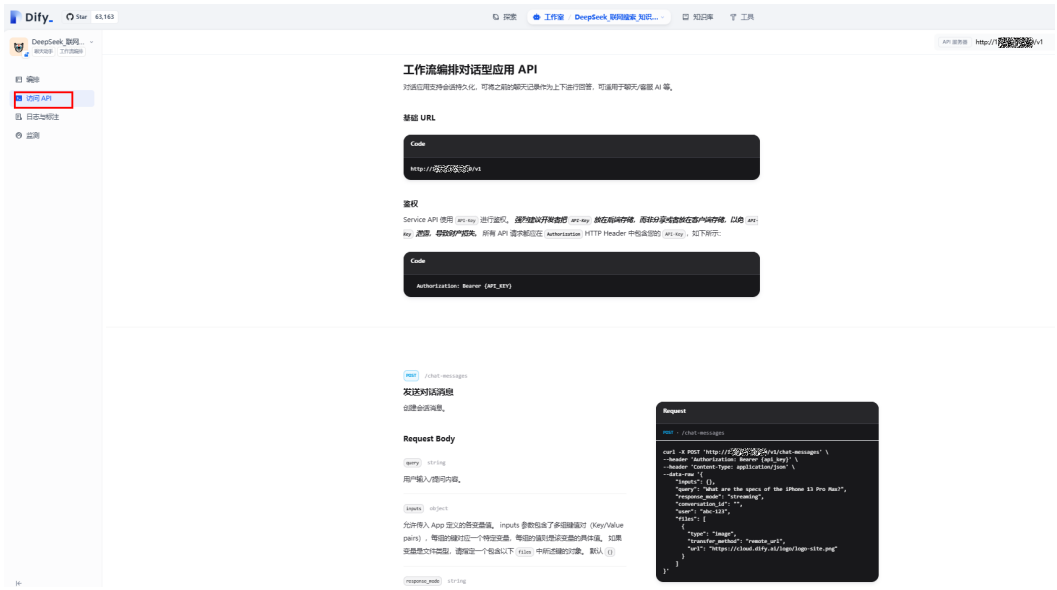


图 3-85 直接访问

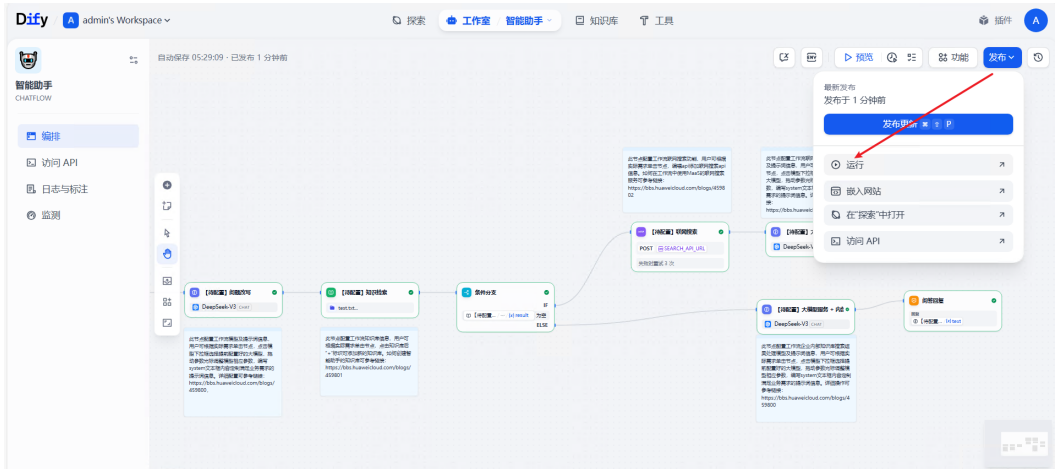


图 3-86 嵌入网站

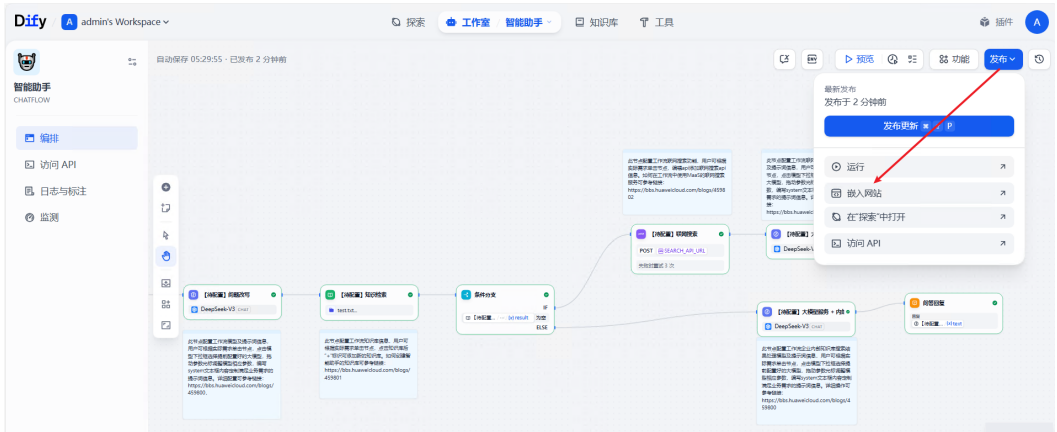
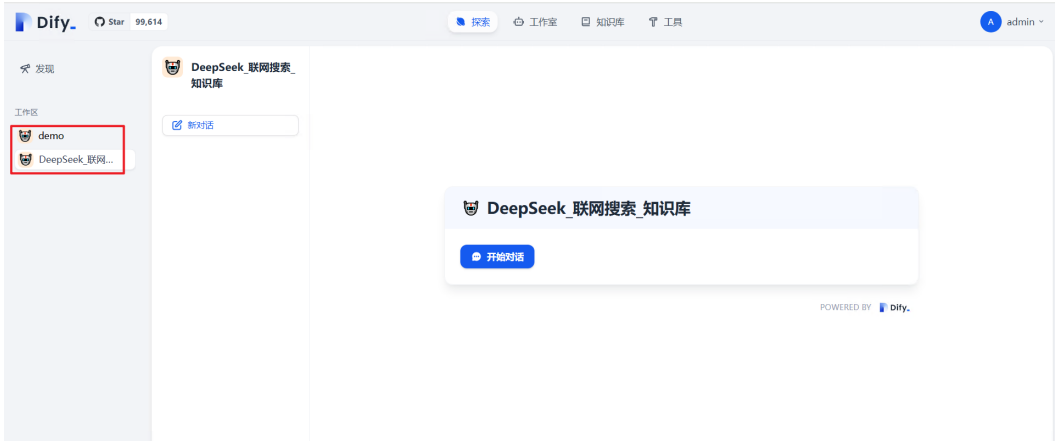


图 3-87 “探索”



----结束

Dify 配置公网域名（社区版单机部署）

您需要在域名解析系统中，添加一条A记录，值为Dify服务器的公网IP。本文档以华为云**云解析服务 DNS**为例。如在华为云购买域名，默认直接添加到公网域名。如果不是通过华为云购买的域名，可参考**创建公网域名**。

步骤1 进入**公网域名列表页面**，选择要使用的域名，单击“管理解析”。

图 3-88 管理解析

步骤2 单击“添加记录集”，填写配置信息如下

记录类型：“A – 将域名指向IPv4地址”

主机记录：解析域名的前缀。

例如创建的域名为“example.com”，其“主机记录”设置包括：

- www：用于网站解析，表示解析的域名为“www.example.com”。
- 空：用于网站解析，表示解析的域名为“example.com”。
  主机记录置为空，还可用于为空头域名“@”添加解析。
- abc：用于子域名解析，表示解析的域名为“example.com”的子域名“abc.example.com”。

- mail：用于邮箱解析，表示解析的域名为“mail.example.com”。
- *：用于泛解析，表示解析的域名为“*.example.com”，匹配“example.com”的所有子域名。

记录值：域名对应的IPv4地址。最多可以输入50个不重复地址，多个地址之间以换行符分隔。本文值为Dify的公网ip地址。

图 3-89 添加记录集

步骤3 登录**弹性云服务 ECS控制台**，选择部署Dify的服务器，单击“远程登录”。选择“VNC登录”，输入服务器密码，登录服务器。

图 3-90 远程登录

图 3-91 VNC 登录

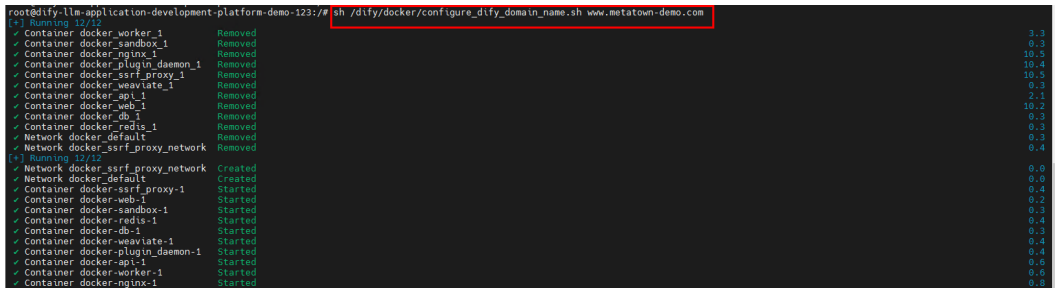
步骤4 将SSL证书文件命名为**dify.crt**、**dify.key**并上传至：**/dify/docker/nginx/ssl**

图 3-92 上传 SSL 证书

步骤5 修改环境变量，在命令行执行如下命令，解析的域名为步骤二获取，例如**www.example.com**：

```
sh /dify/docker/configure_dify_domain_name.sh ${解析的域名}
```

图 3-93 配置环境变量



步骤6 浏览器即可通过配置的域名访问Dify平台。

图 3-94 访问 Dify 平台

----结束

说明

拓展应用请参考：

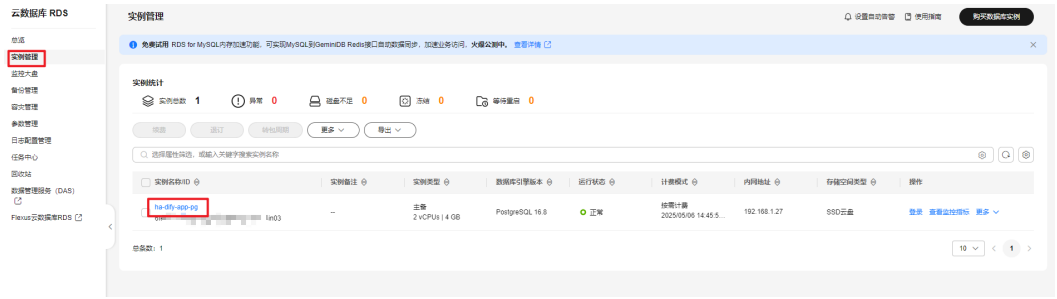
- **华为云ModelArts Studio，助力快速搭建专属大模型**
- **探索Dify：开启AI应用开发的新篇章**

3.5 快速卸载

步骤1 高可用部署版本删除堆栈前请确保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（非CCE容器高可用版可跳过直接进入步骤4）。

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 3-95 数据库实例



步骤3 依次单击“智能DBA助手>会话管理”，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”“dify_plugin”，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 3-96 实时会话

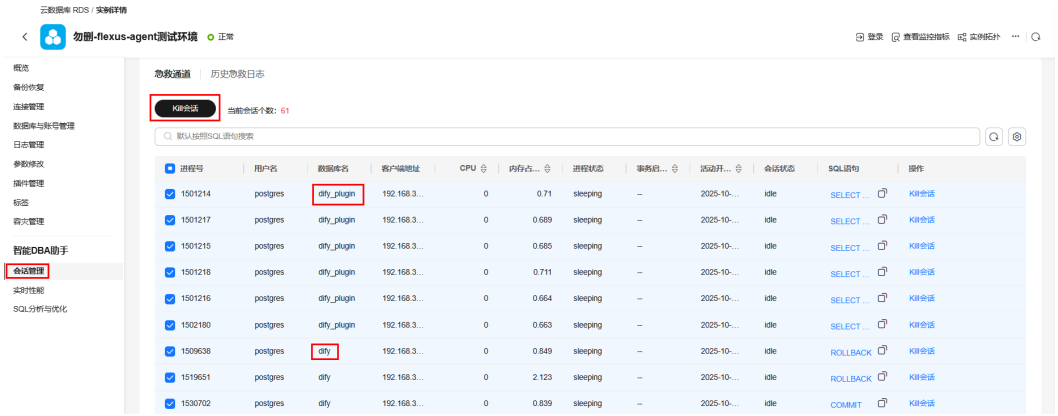
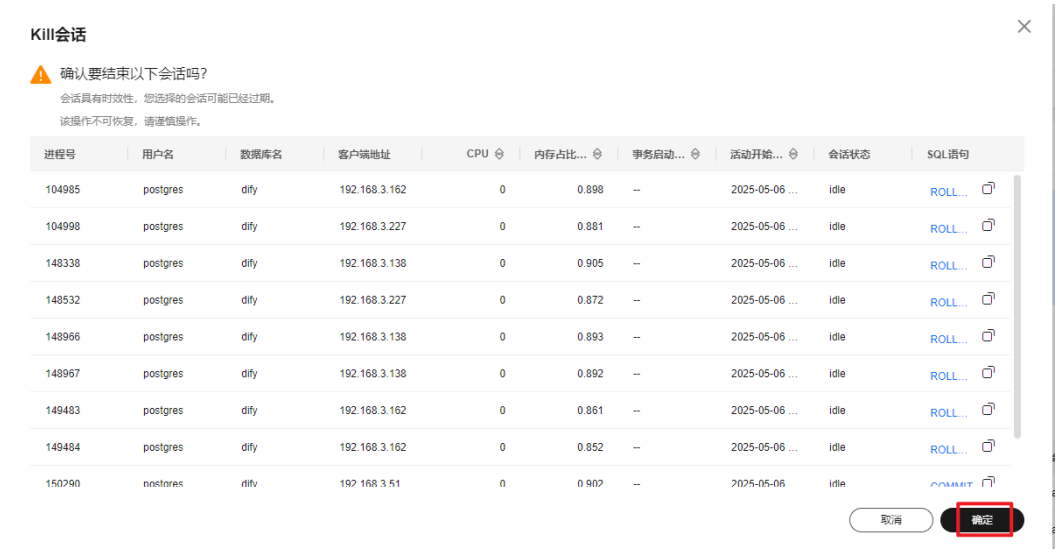


图 3-97 Kill 进程



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 3-98 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 3-99 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
2025-09-01	更新CCE容器高可用
2025-05-30	单机部署支持云搜索及Embedding&Reranker模型
2025-03-12	支持CCE容器高可用部署
2024-11-07	第一次正式发布