解决方案实践

快速部署 Embedding 及 Reranker 模型

文档版本	1.0
发布日期	2025-09-08

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

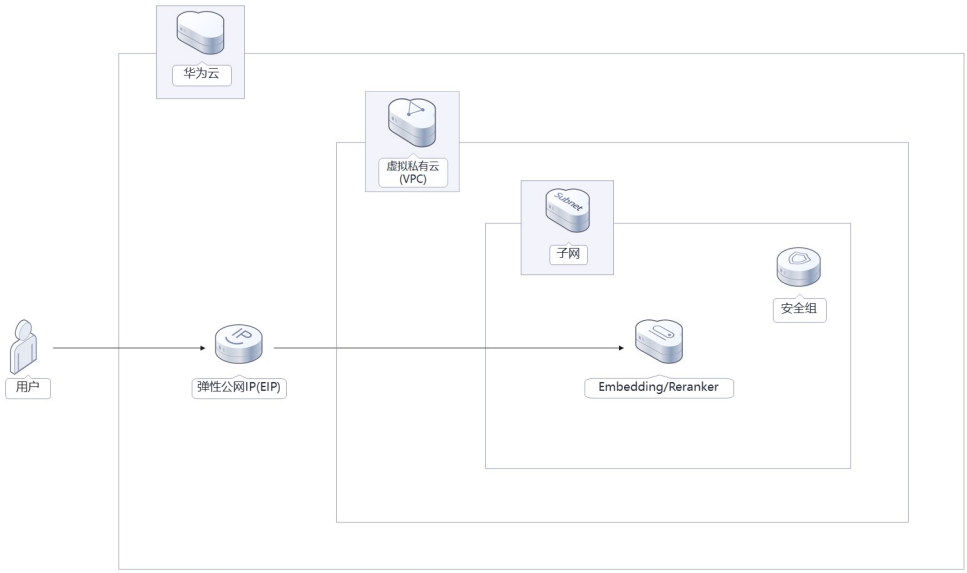
本方案旨在通过华为云Flexus云服务器X实例高效部署和运行**Embedding（bge-m3）**及**Reranker（bge-reranker-v2-m3）**模型。bge-m3模型是一种先进的文本嵌入模型，能够将文本转换为高维向量，从而实现高效的文本相似度计算、分类等任务。bge-reranker-v2-m3是一个轻量级的重排序模型，具有强大的多语言能力，易于部署，推理速度快。借助于Flexus云服务器的强大性能和弹性扩展能力，用户可以轻松地在云端部署此模型，并根据实际需求灵活调整资源。

- 文本相似度计算：适用于搜索引擎、推荐系统等领域，帮助提升搜索结果的相关性和推荐准确性。
- 文本分类与聚类：广泛应用于内容管理、舆情分析等场景，支持自动化的信息分类与主题发现。
- 自然语言处理任务：如情感分析、意图识别等，助力企业更好地理解和利用非结构化数据。
- 搜索引擎优化：在大型搜索引擎中，重排序模型可以帮助优化搜索结果，确保用户看到的信息是最相关和最有价值的。
- 问答系统：在问答系统中，重排序模型可以帮助确定哪些答案是最准确和最相关的，从而提高问题解决的质量。

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）上快速部署Embedding（bge-m3）及Reranker（bge-reranker-v2-m3）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**），用于部署Embedding（bge-m3）及Reranker（bge-reranker-v2-m3）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

- 高效

  内置 bge-m3及bge-reranker-v2-m3模型实现高效的文本相似度计算、分类等任务，重排序模型，推理速度快。
- 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Embedding bge-m3及Reranker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。

CPU 版

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	deploying-embedding-and-reranker-models-demo	• VPC网段：172.16.0.0/16 • 区域：西南-贵阳一	1	0.00元
子网 Subnet	deploying-embedding-and-reranker-models-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：西南-贵阳一	1	0.00元
安全组 SecurityGroup	deploying-embedding-and-reranker-models-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：西南-贵阳一	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
Flexus云服务器X实例	deploying-embedding-and-reranker-models-demo	• 按需计费：1.34/小时 • 区域：西南-贵阳一 • 规格：Flexus云服务器X实例 | 性能模式（开启）| x1e.16u.16g | 16核 | 16GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 40GB	1	967.82元
弹性公网IP EIP	deploying-embedding-and-reranker-models-demo-eip	• 按需计费：0.80元/GB • 区域：西南-贵阳一 • 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	-	-		**967.82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	deploying-embedding-and-reranker-models-demo	• VPC网段：172.16.0.0/16 • 区域：西南-贵阳一	1	0.00元
子网 Subnet	deploying-embedding-and-reranker-models-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：西南-贵阳一	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
安全组 SecurityGroup	deploying-embedding-and-reranker-models-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：西南-贵阳一	1	0.00元
Flexus云服务器X实例	deploying-embedding-and-reranker-models-demo	• 包年包月 • 区域：西南-贵阳一 • 规格：Flexus云服务器X实例 | 性能模式（开启）| x1e.16u.16g | 16核 | 16GB • 镜像：Ubuntu 22.04 server 64bit • 系统盘：通用型SSD | 40GB	1	650.40元
弹性公网IP EIP	deploying-embedding-and-reranker-models-demo-eip	• 按需计费：0.80元/GB • 区域：西南-贵阳一 • 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	-	-		**650.40元 + 弹性公网IP EIP费用**

GPU 版

表 2-3 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	deploying-embedding-and-reranker-models-demo	• VPC网段：172.16.0.0/16 • 区域：西南-贵阳一	1	0.00元
子网 Subnet	deploying-embedding-and-reranker-models-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：西南-贵阳一	1	0.00元
安全组 SecurityGroup	deploying-embedding-and-reranker-models-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：西南-贵阳一	1	0.00元
弹性云服务器 ECS	deploying-embedding-and-reranker-models-demo	• 按需计费：6.64/小时 • 区域：西南-贵阳一 • 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 • 系统盘：通用型SSD | 60GB	1	4780.8元

华为云服务	资源名称	配置示例	数量	每月预估花费
弹性公网IP EIP	deploying-embedding-and-reranker-models-demo-eip	• 按需计费：0.80元/GB • 区域：西南-贵阳一 • 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	-	-		**4780.8元 + 弹性公网IP EIP费用**

表 2-4 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	deploying-embedding-and-reranker-models-demo	• VPC网段：172.16.0.0/16 • 区域：西南-贵阳一	1	0.00元
子网 Subnet	deploying-embedding-and-reranker-models-demo-subnet	• 子网网段：172.16.1.0/24 • 区域：西南-贵阳一	1	0.00元
安全组 SecurityGroup	deploying-embedding-and-reranker-models-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：西南-贵阳一	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	deploying-embedding-and-reranker-models-demo	• 包年包月 • 区域：西南-贵阳一 • 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 • 系统盘：通用型SSD | 60GB	1	3199.31元
弹性公网IP EIP	deploying-embedding-and-reranker-models-demo-eip	• 按需计费：0.80元/GB • 区域：西南-贵阳一 • 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	-	-		**3199.31元 + 弹性公网IP EIP费用**

3 实施步骤

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

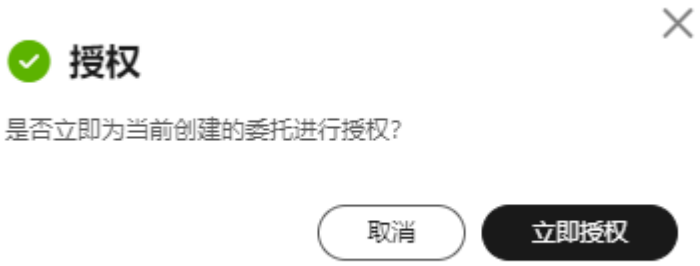
- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

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”，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

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Embedding及Rerank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。

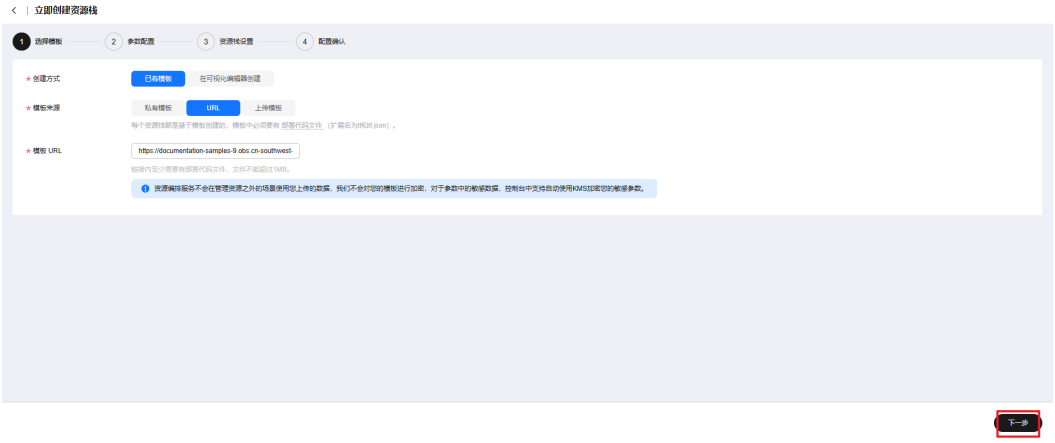
操作步骤

步骤1 登录**华为云解决方案实践**，选择“快速部署Embedding及Rerank模型”，支持区域下拉选择部署的区域（以贵阳一为例），选择需要部署的版本，以CPU版为示例，单击“一键部署（CPU版）”，跳转至解决方案创建资源栈界面。

图 3-9 解决方案实施库

步骤2 在选择模板界面中，单击“下一步”。

图 3-10 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（CPU版）**”完成全部自定义参数填写，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 3-11 配置参数

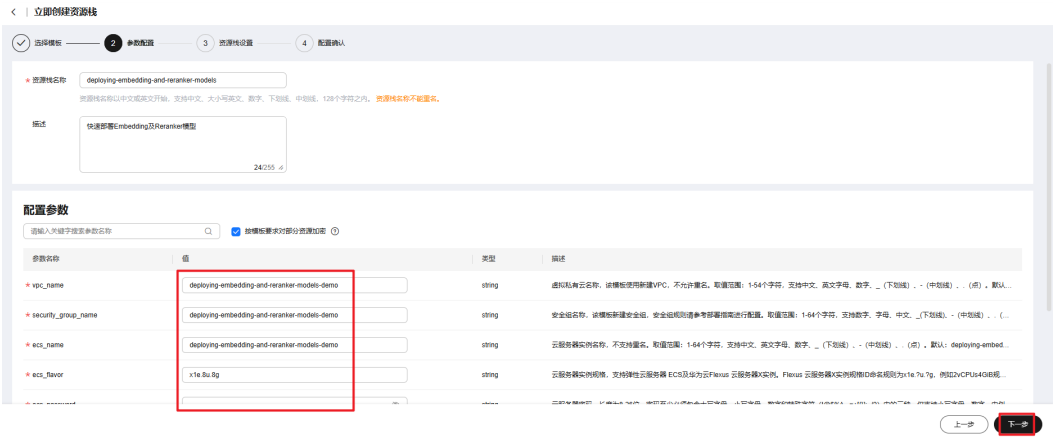


表 3-1 参数填写说明（CPU 版）

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）。	deploying-embedding-and-reranker-models-demo
security_group_name	string	必填	安全组名称，该模板新建安全组，安全组规则请参考部署指南进行配置。取值范围：1-64个字符，支持数字、字母、中文、_(下划线)、-（中划线）、.（点）。	deploying-embedding-and-reranker-models-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）。	deploying-embedding-and-reranker-models-demo
ecs_flavor	string	必填	云服务器实例规格，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，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规格请参考部署指南配置。**弹性云服务器规格清单**。	x1e.16u.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。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

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

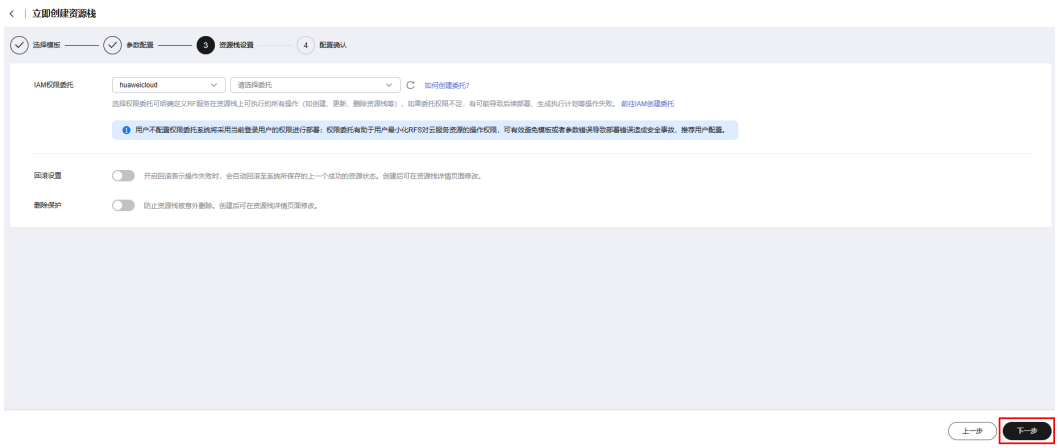
表 3-2 参数填写说明（GPU 版）

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）。	deploying-embedding-and-reranker-models-demo
security_group_name	string	必填	安全组名称，该模板新建安全组，安全组规则请参考部署指南进行配置。取值范围：1-64个字符，支持数字、字母、中文、_(下划线)、-（中划线）、.（点）。	deploying-embedding-and-reranker-models-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）。	deploying-embedding-and-reranker-models-demo
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16GB及以上。弹性云服务器ECS规格请参考部署指南配置。**弹性云服务器规格清单**。	pi2.2xlarge.4

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，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。	60
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

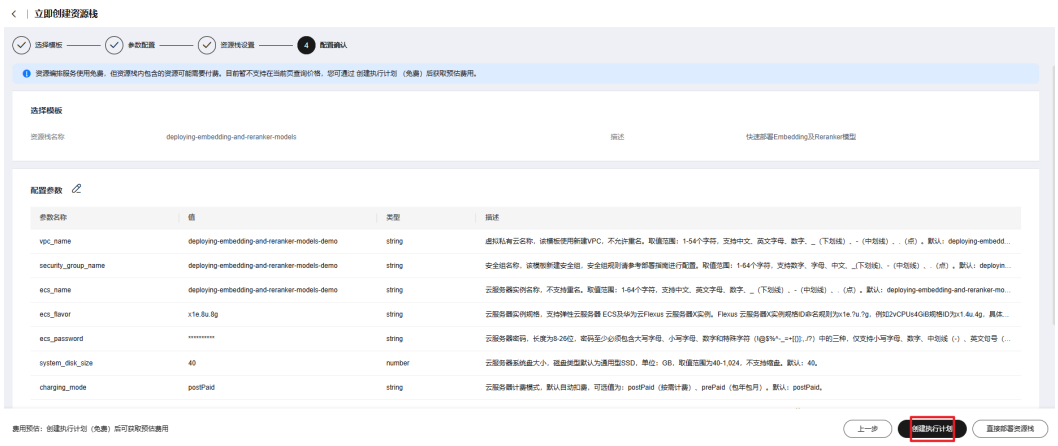
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

图 3-12 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 3-13 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 3-14 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 3-15 执行计划

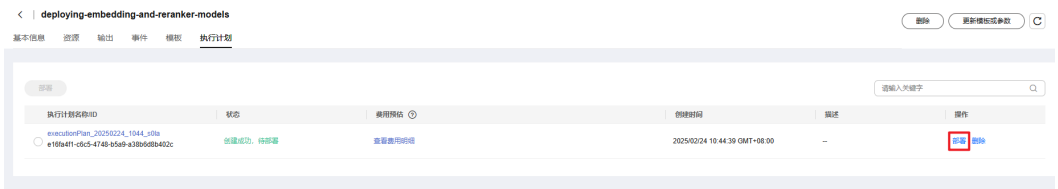
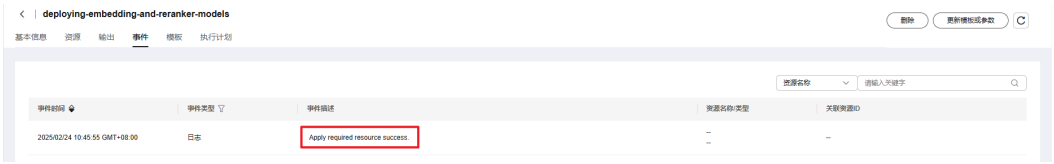


图 3-16 执行计划确认

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，部署Embedding/Reranker模型脚本开始执行，耐心等待10分钟左右（受网络波动影响）。

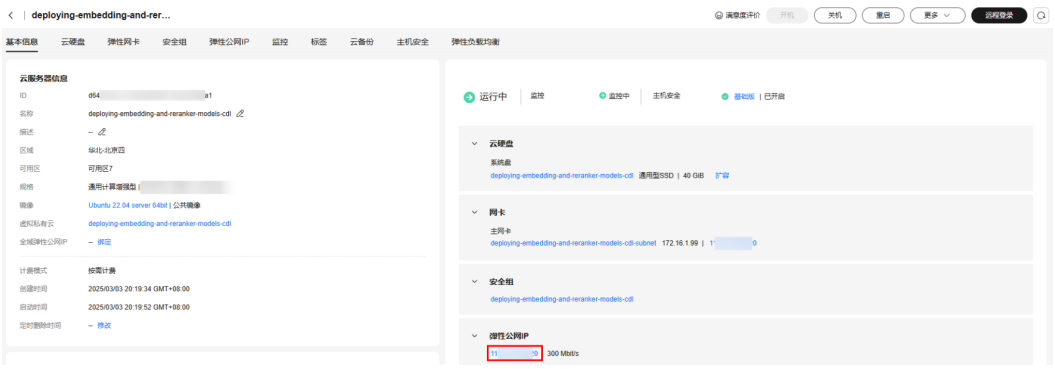
图 3-17 部署完成



步骤10 单击“资源”查看已创建好的资源，如下图所示，单击“蓝色资源名称”跳转至弹性云服务器详情页面，复制获取弹性云服务器绑定的弹性公网IP。

图 3-18 弹性云服务器详情页

图 3-19 获取弹性公网 IP



----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。

该解决方案自动放通Ollama API网络端口 11434，xinference 服务端口9997。

该解决方案部署成功后，环境初始化及应用安装预计5~10分钟不等，受网络、带宽影响，部署时间会有波动，部署完成方可正常使用。

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。

须知

在使用本方案之前需要部署Dify开发平台，如果您没有部署请参考**快速搭建Dify-LLM应用开发平台**部署。

步骤1 浏览器输入http://[弹性公网IP]，访问您已部署的Dify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 3-20 Dify 开发平台

步骤2 依次输入上一步骤中的“邮箱”、“密码”登录Dify平台。

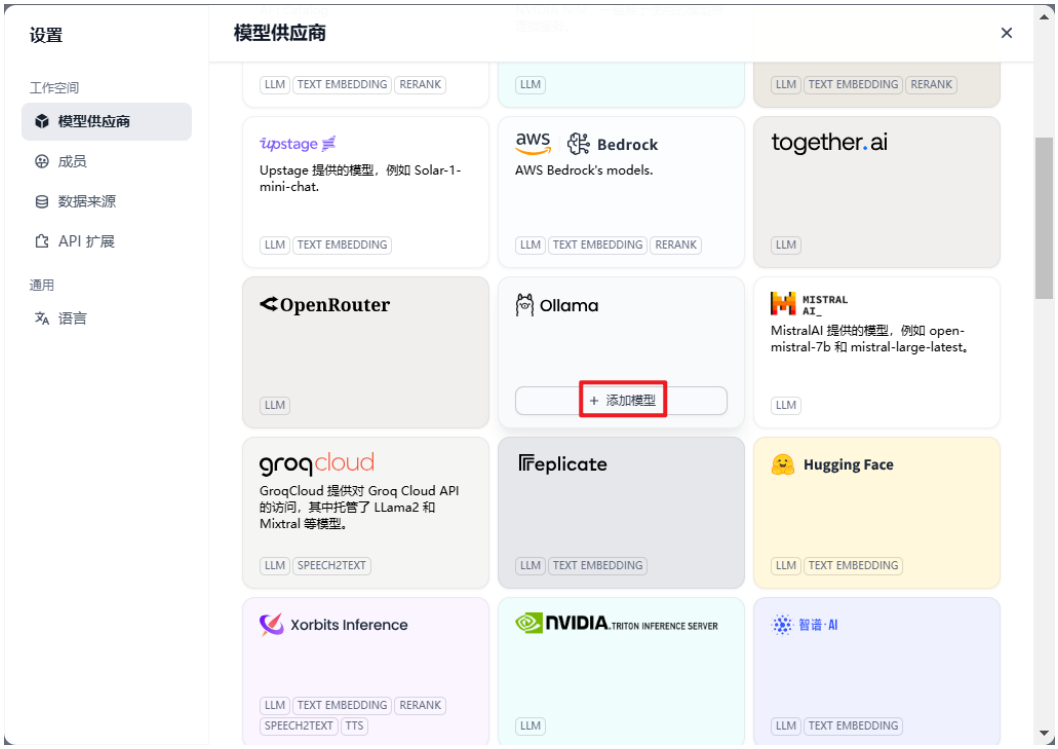
图 3-21 登录 Dify 平台

步骤3 单击右侧“用户名称”下拉并单击“设置”。

图 3-22 设置

步骤4 单击左侧“模型供应商”，在Ollama下单击“添加模型”。

图 3-23 添加模型

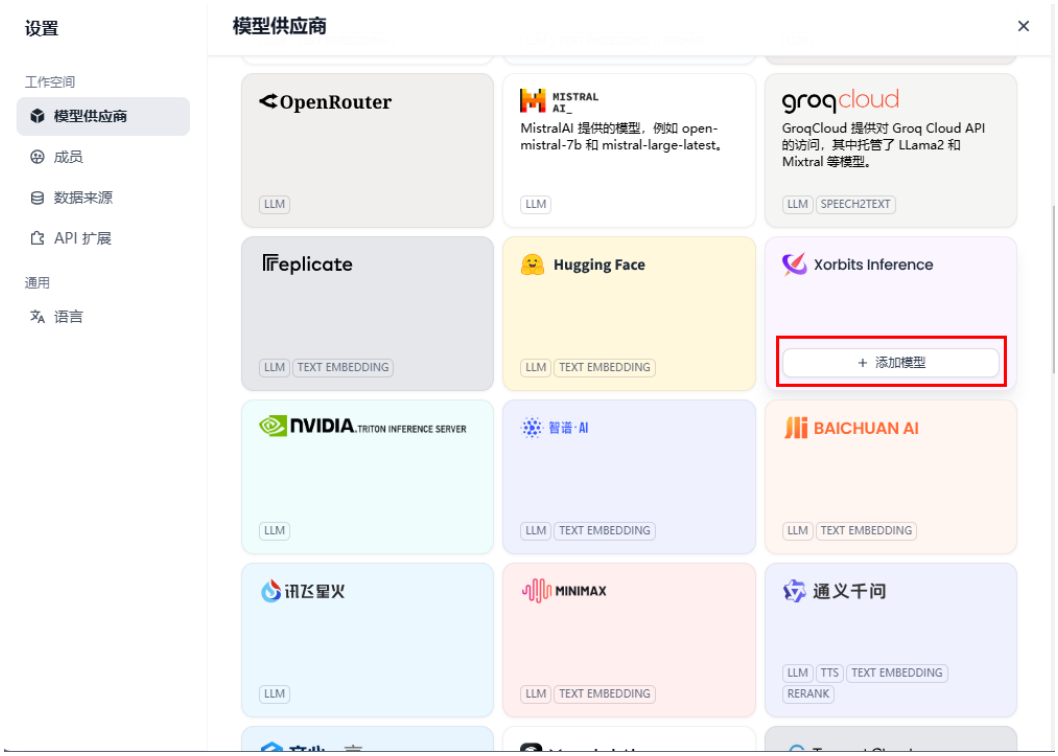


步骤5 模型类型选择“Text Embedding”，模型名称填写“bge-m3”，基础URL填写**快速部署步骤10**中获取的公网IP地址，端口号11434，单击右下角“保存”。

图 3-24 添加 Ollama

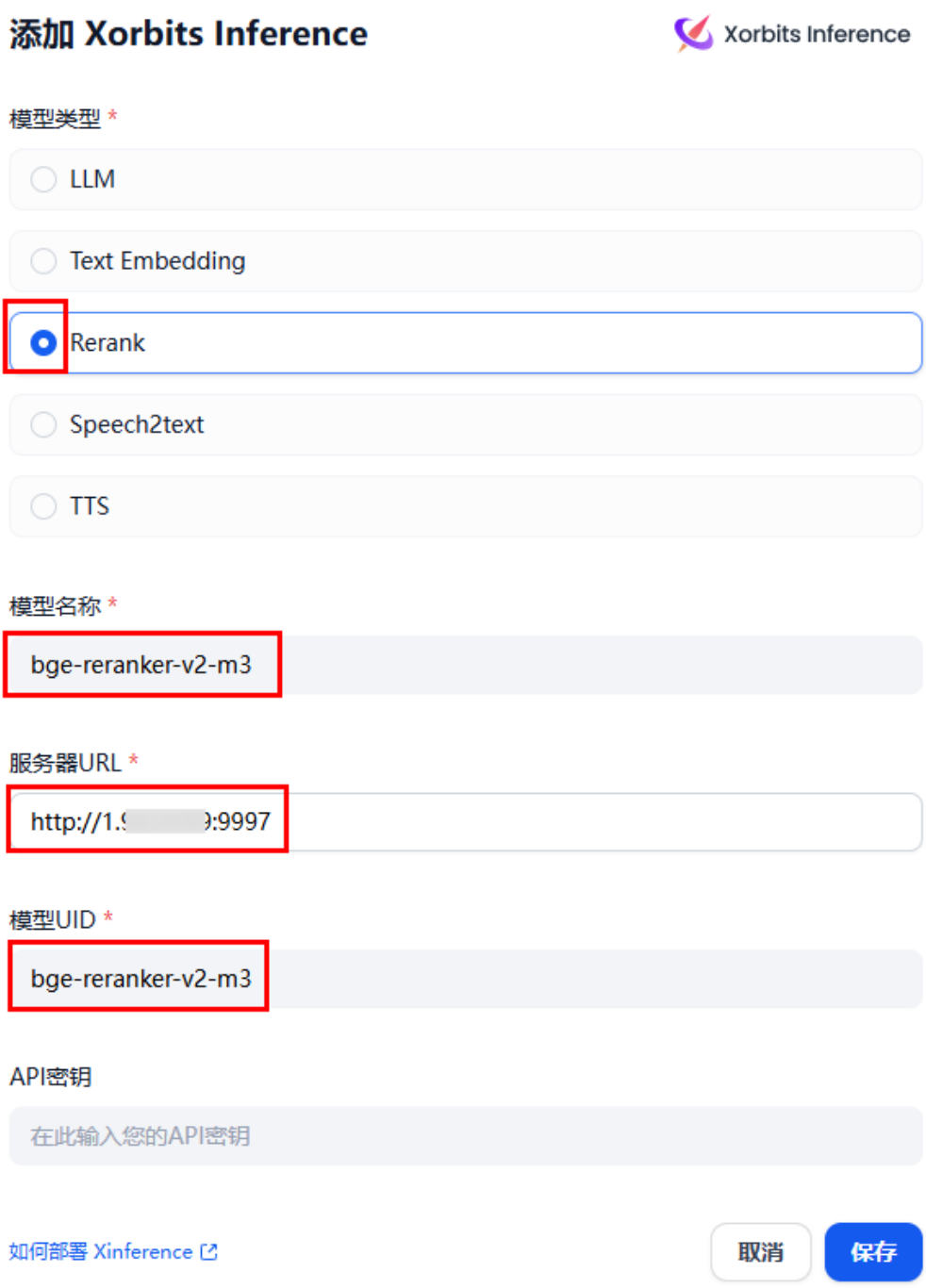
步骤6 单击左侧“模型供应商”，在Xorbits Inference下，单击“添加模型”。

图 3-25 添加 xorbits Inference



步骤7 模型类型选择 Rerank，模型名称与模型UID均填写“bge-reranker-v2-m3”服务器URL填写**快速部署步骤10**中获取的公网IP地址，端口号9997，单击右下角“保存”。

图 3-26 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 3-27 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 3-28 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
2025-02-24	第一次正式发布。
2025-03-04	升级云服务器默认规格
2025-09-08	支持GPU一键部署模板