解决方案实践

快速搭建 New API 大模型网关

文档版本	1.0.0
发布日期	2025-09-22

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

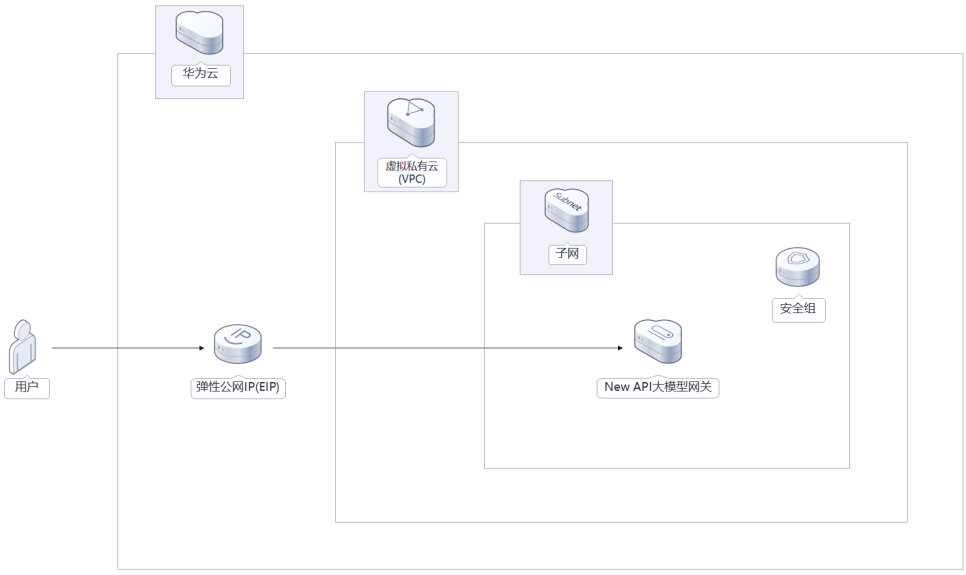
应用场景

该解决方案帮助您快速搭建单机版New API大模型网关，**New API** 是一个基于 One API 二次开发的开源项目，专注于 AI 模型接口管理与分发系统，支持将多种大模型转为 OpenAI 格式调用，并兼容 Midjourney Proxy、Suno、Rerank 等接口，主要用于企业内部或个人学习场景。

方案架构

该解决方案帮助您快速搭建New API大模型网关。

图 1-1 方案架构图（社区版单机部署）



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New API大模型网关。
- 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。

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。
- 开箱即用

  New API采用现代化的微服务架构，强调高可用性和易部署性，适用于内容创作、教育与科研等领域，部署后开箱即用。
- 一键部署

  一键轻松部署，即可完成云服务资源的创建及New API大模型网关的搭建。

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。

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
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	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	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
合计	-	-	**467.00元 + 弹性公网IP EIP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

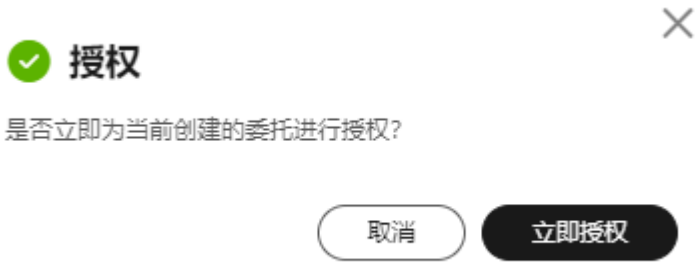
- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

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

3.2 快速部署

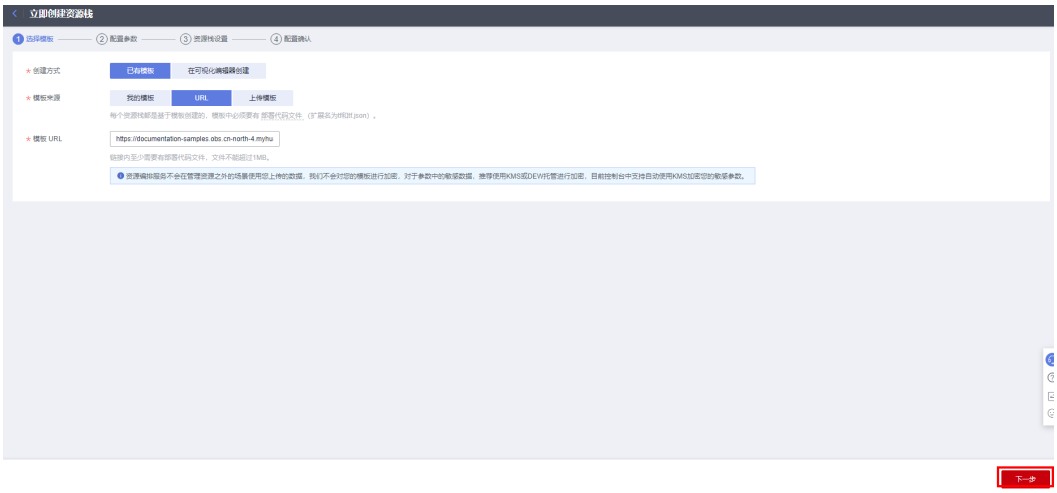
本章节帮助用户高效地部署“快速搭建New API大模型网关”解决方案。一键部署该解决方案时，参照本章节中的步骤和说明进行操作，即可完成快速部署。

步骤1 登录**华为云解决方案实践**，选择“快速搭建New API大模型网关”，单击“一键部署”，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 参数说明**完成自定义参数填写，单击“下一步”。

图 3-11 配置参数

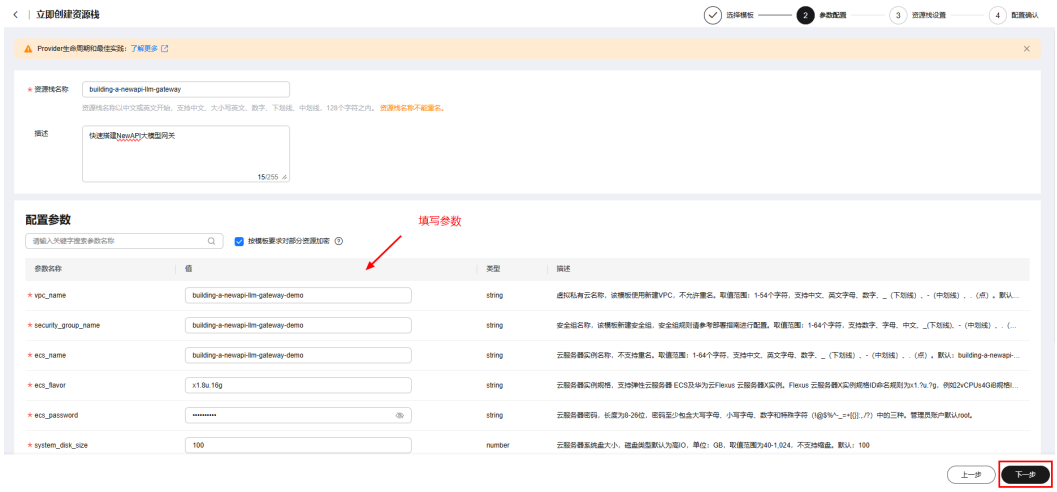


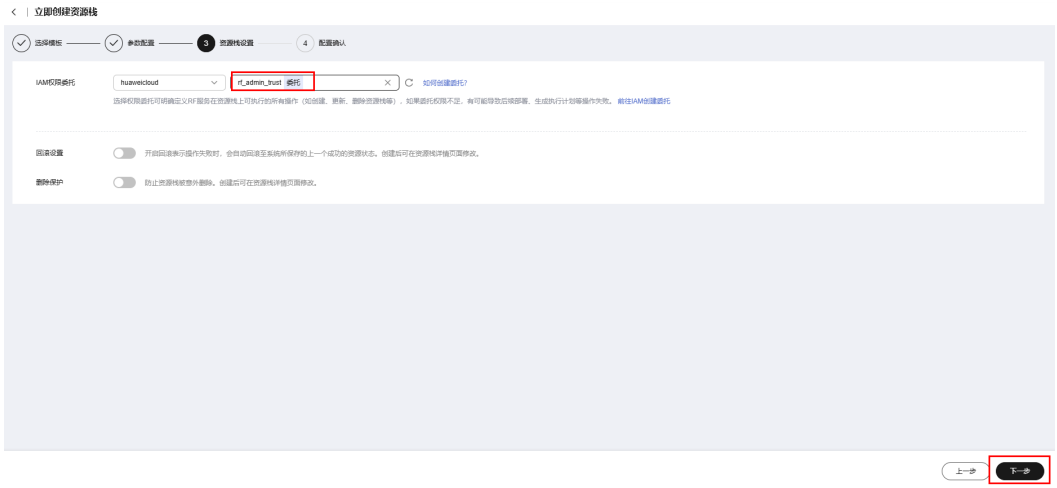
表 3-1 参数说明

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）。	building-a-New API-llm-gateway-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（.）。	building-a-New API-llm-gateway-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）。	building-a-New API-llm-gateway-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

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。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

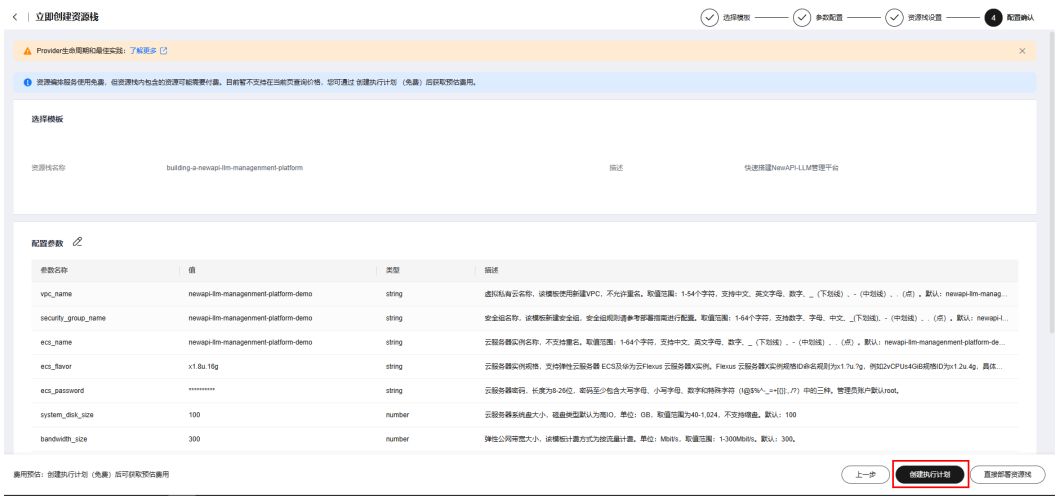
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

图 3-12 委托设置



步骤5 在配置确认界面中，单击“创建执行计划”。

图 3-13 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

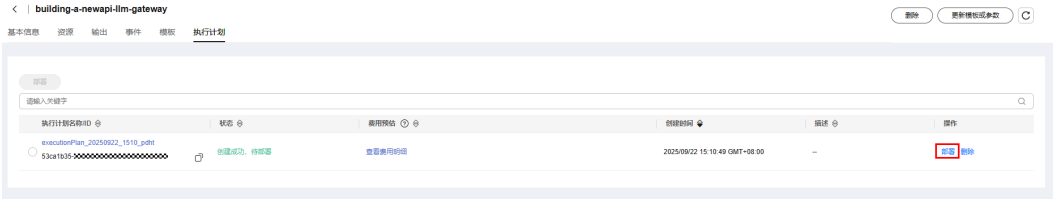
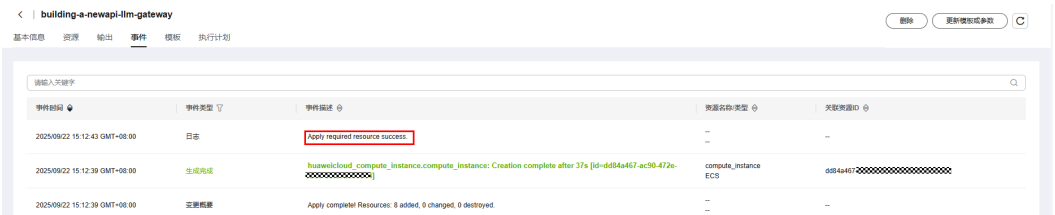


图 3-16 执行计划确认

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**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 3-17 部署完成



步骤10 在“输出”中查看访问说明。堆栈部署成功后，New API应用部署脚本开始执行，耐心等待10分钟左右（受网络波动影响）。

图 3-18 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。

New API 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，即可访问平台。首次登录需完成系统初始化，根据提示依次执行即可，使用模式按实际需求选择。

图 3-19 数据库检查

图 3-20 设置管理员账户

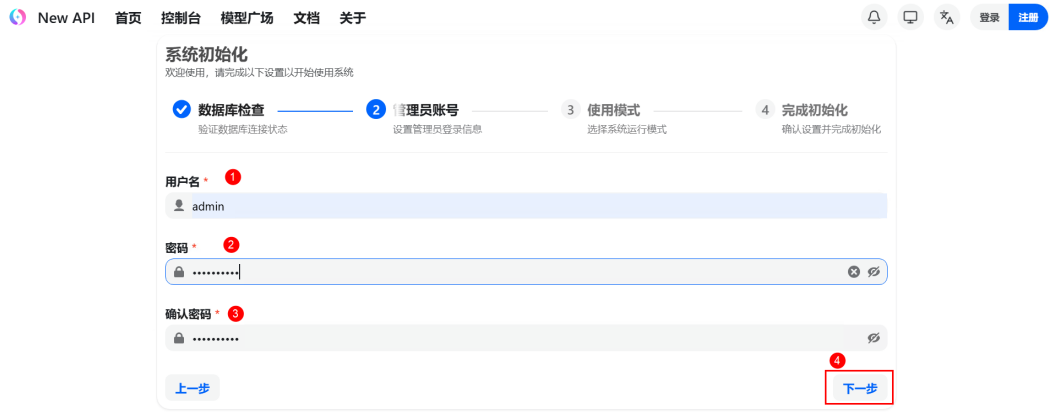


图 3-21 使用模式（按实际需求选择模式）

图 3-22 完成初始化

步骤2 初始化完成后自动跳转首页，单击右上角“登录“，依次输入上一步骤中的“用户名”“密码”登录New API网关平台。

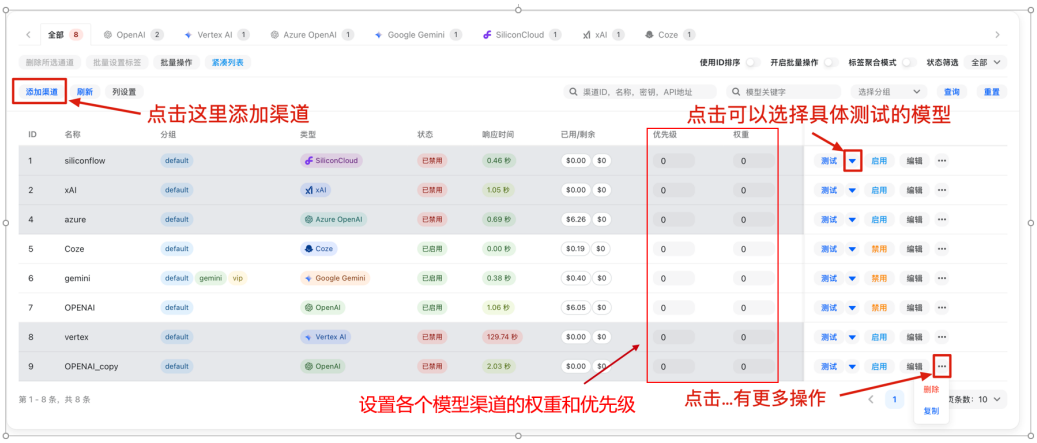
图 3-23 单击登录

图 3-24 登录

步骤3 配置令牌、渠道，如下图所示。

图 3-25 添加令牌

图 3-26 添加渠道



更多详细说明请阅读官方指南：https://docs.newapi.pro/

----结束

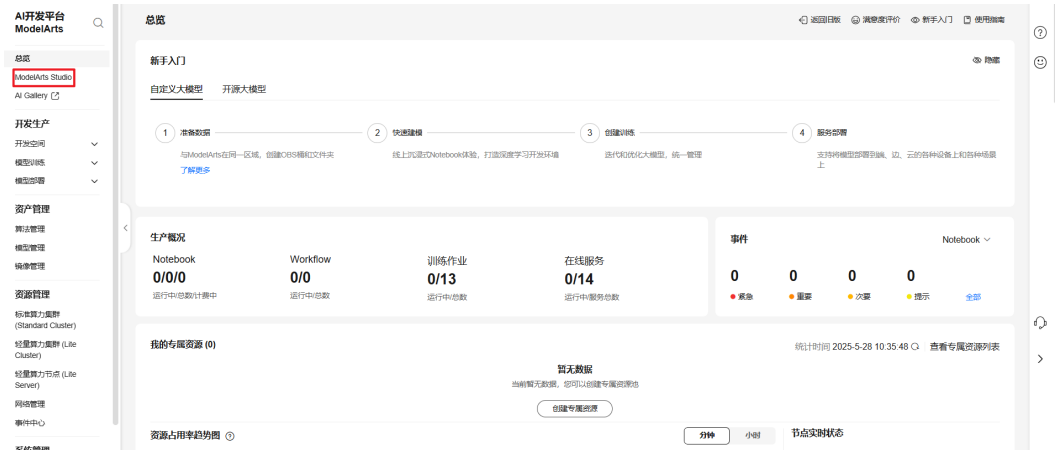
说明

- 以下对接大模型方式**与MaaS服务对接**和#new-api_06/section151218231252用户可根据自身情况选择使用。

与 MaaS 服务对接

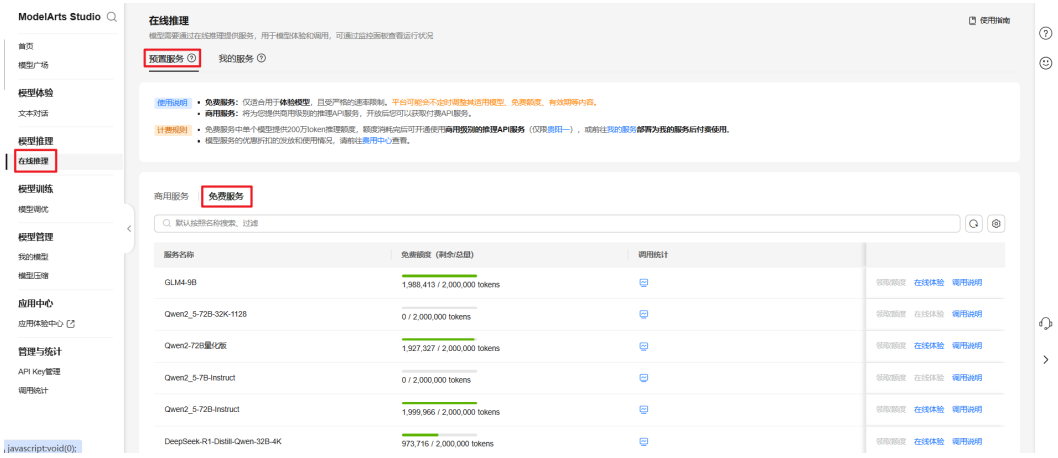
步骤1 登录**ModelArts Studio** 平台，本文以部署华东二的DeepSeek-R1-Distill-Qwen-32B-4K为例。

图 3-27 ModelArts Studio



步骤2 在ModelArts Studio左侧导航栏中，选择“在线推理”进入“预置服务”服务列表，选择“免费服务”。

图 3-28 免费服务

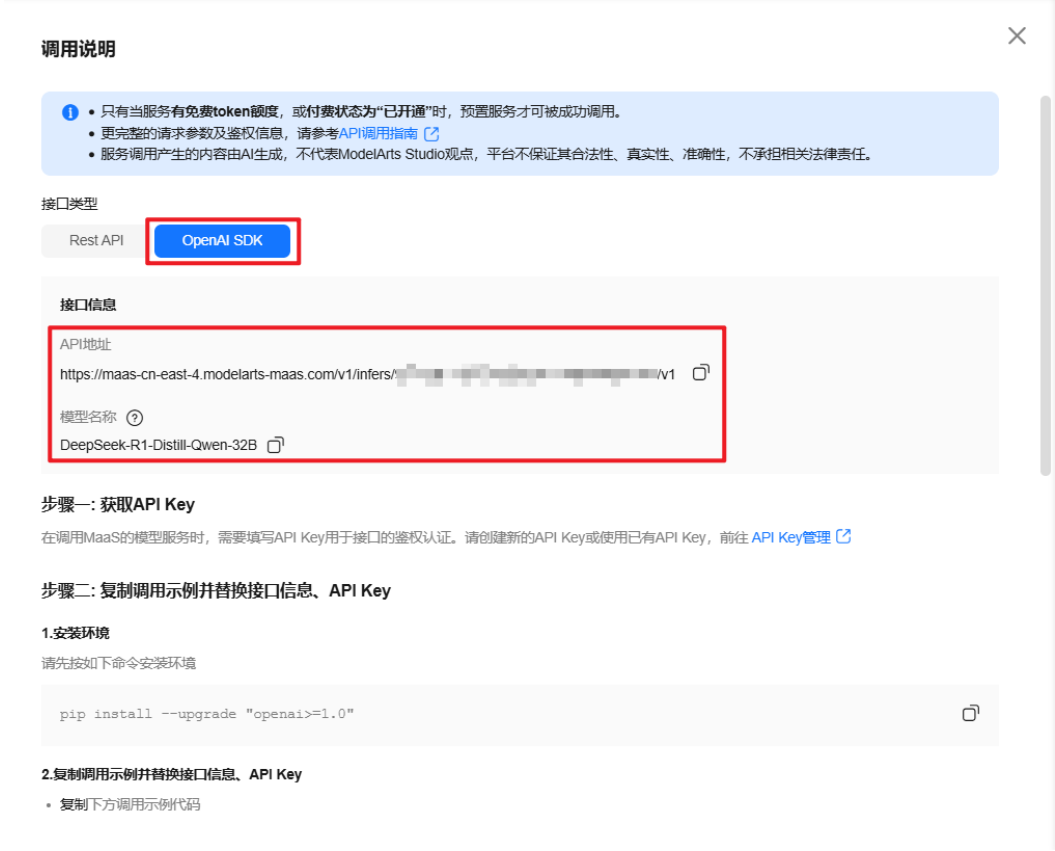


步骤3 领取免费调用额度。在免费服务列表，选择所需的服务，单击右侧操作列的“领取额度”。当领取置灰时，表示该服务的免费额度已领取。

图 3-29 领取额度

步骤4 成功领取后，在免费服务列表，选择所需的服务，单击“调用说明”，在调用弹窗中接口类型选择“OpenAI SDK”获取API地址和模型名称。

图 3-30 调用说明



步骤5 免费服务中单个模型提供200万token推理额度，额度消耗完后可开通使用商用级别的推理API服务（仅限**贵阳一**），或前往**我的服务**部署为我的服务后付费使用。

图 3-31 商用服务

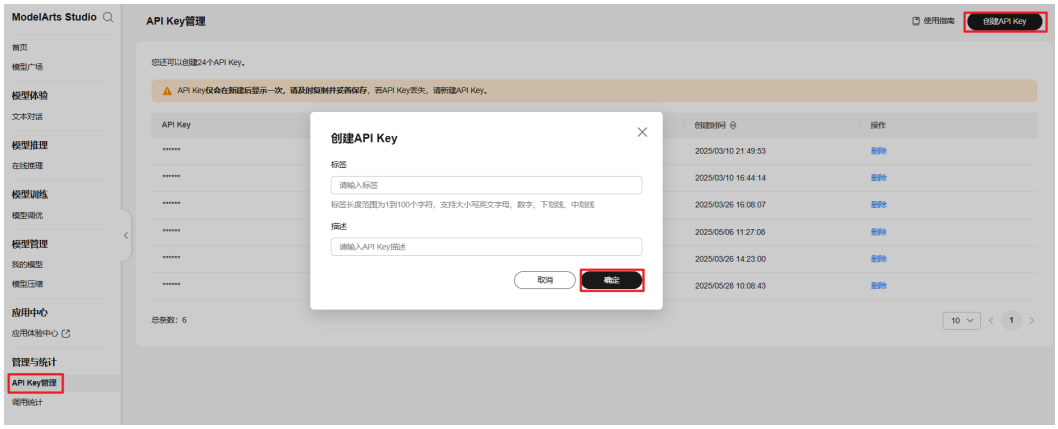
图 3-32 调用说明

步骤6 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 3-33 API Key 管理

步骤7 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 3-34 创建 API Key



步骤8 对接New API网关。打开New API网关平台，依次单击“渠道管理“、“添加渠道“，如**New API基础使用图8 添加渠道**所示，在弹窗内依次填写信息，如下图，完成后单击”提交”。

图 3-35 添加华为云 MaaS 渠道

图 3-36 测试渠道

----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**

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 3-37 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 3-38 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
2025-09-22	第一次正式发布