

解决方案实践

LiteLLM，统一的 AI 管理网关

文档版本 1.0
发布日期 2026-08-24



版权所有 © 华为云计算技术有限公司 2026。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云计算技术有限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，华为云计算技术有限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。

目 录

1 方案概述.....	1
2 资源和成本规划.....	5
3 实施步骤.....	12
3.1 准备工作.....	12
3.2 快速部署.....	18
3.3 开始使用.....	26
3.4 快速卸载.....	31
4 附录.....	33
5 修订记录.....	34

1 方案概述

应用场景

本方案通过华为云 Flexus X 实例上一键部署 LiteLLM LLM API 网关：LiteLLM 将 100+ 个 LLM 提供商（OpenAI、Anthropic、Azure、Cohere 等）统一为单个兼容 OpenAI 的 API，内置虚拟密钥管理、花费追踪、负载均衡和管理后台。

该方案通常应用到以下场景：

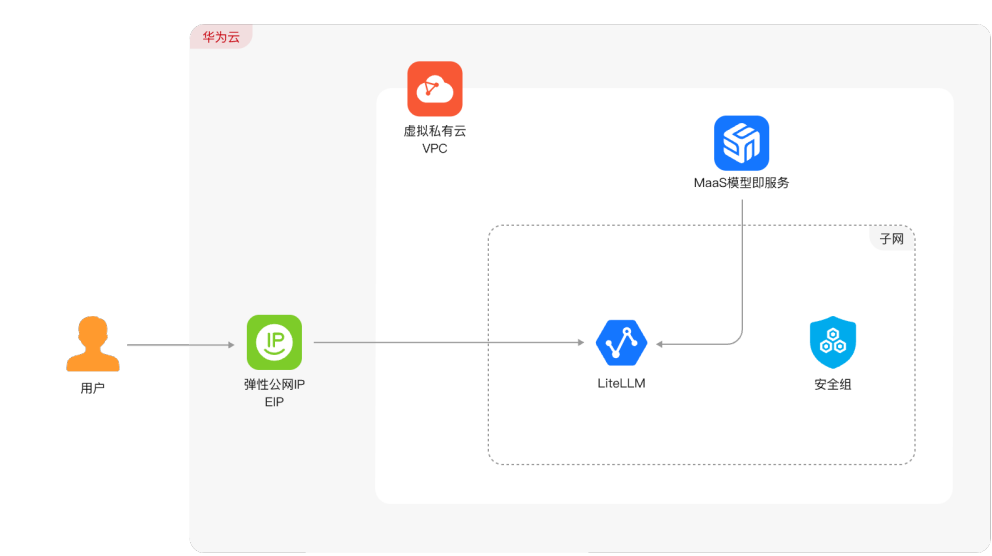
- **多业务线统一 LLM 管理：**通过 LiteLLM 虚拟密钥机制，为每个团队/项目创建独立 Key，设置预算上限，自动追踪用量。
- **高可用 LLM 基础设施：**将 AI 集成到核心业务流程、需要 SLA 保障的企业。
- **个人AI管理系统：**个人开发过程中使用 OpenAI 兼容 API 在多个提供商间切换，并在AI领域中进行开发与测试。

方案架构

该解决方案帮助您快速部署 LiteLLM。

单机版部署

图 1-1 单机版方案架构图

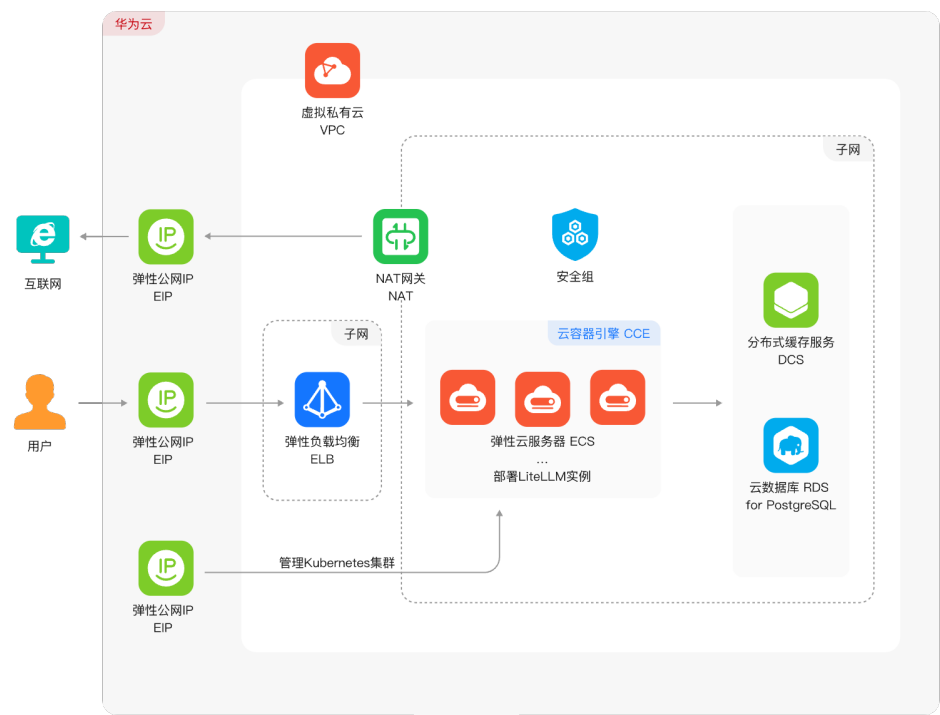


该解决方案将会部署如下资源：

- 创建一个虚拟私有云 VPC和子网，用于搭建云服务所需的网络基础
- 创建1台华Flexus云服务器X实例，用于部署LiteLLM
- 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

高可用版部署

图 1-2 CCE 高可用版方案架构图



在CCE容器中实现高可用部署：

- 创建一个虚拟私有云 VPC和子网，用于搭建云服务所需的网络基础
- 创建3个EIP分别绑定ELB、NAT、CCE实现网络双向访问
- 创建1个ELB并绑定EIP，自动将访问流量分发到各后端服务。这增强了应用系统的外部服务能力，实现应用级高可用容错
- 创建1个NAT网关并绑定EIP，配置SNAT规则使私有子网内所有实例在保持完全私密的同时可访问外网
- 创建3台FlexusX实例用于部署LiteLLM Proxy核心组件与服务
- 创建1个CCE Turbo集群，创建节点池并将三台ECS纳管为集群节点
- 创建1个兼容Redis的分布式缓存服务（DCS）实例，为用户提供高性价比、高性能的NoSQL数据库，同时保障数据流转过程中的一致性（用于全局限流和缓存同步）
- 创建1个支持跨AZ主备部署的RDS for PostgreSQL实例，确保容灾能力（用于存储API密钥和消费日志元数据）
- 创建4个配置指定规则的安全组保护云服务

方案优势

- **统一 API 网关**
通过与 OpenAI 兼容的端点，标准化对 100 多个 LLM 提供商的访问，简化集成并降低开发开销。

- **增强安全性和成本控制**

虚拟密钥管理可实现每把密钥的预算上限和实时使用情况的跟踪，从而改善财务监督和访问控制。

- **高可用性和可扩展性**

多提供商故障转移和负载均衡通过将请求分配到可用服务中，确保可靠的性能并最大限度地减少停机时间。

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。
- 本内容可能包含第三方开源软件、组件或框架的技术说明与配置示例，请您独立评估并自主决定是否使用，华为云不就第三方软件的功能、性能、安全性、兼容性或合规性作出任何保证。第三方软件知识产权归各自权利人所有，您应严格遵守适用的开源许可证条款（如 MIT、Apache 2.0、GPL 等），确保不侵犯任何第三方合法权益，并自行负责安全评估与漏洞管理。因您使用本方案提及的任何第三方软件产生的直接或间接损失（包括但不限于数据丢失、业务中断、安全事件、知识产权纠纷、合规处罚等），华为云及其关联方不承担任何责任。本方案对第三方软件的提及不构成认可或背书，亦不代表华为云与第三方权利人存在任何授权或合作关系。

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	配置示例	数量	一次完整流程预估（1小时）花费
虚拟私有云 VPC	- VPC网段：172.16.0.0/16 - 区域：华北-北京四	1	0.00元
子网 Subnet	- 子网网段：172.16.1.0/24 - 区域：华北-北京四	1	0.00元
安全组 SecurityGroup	- 允许ping：0.0.0.0/0 - 配置端口22允许Cloud Shell 登录：121.36.59.153/32 - 开放端口4000允许访问 LiteLLM 代理 API 和管理页面：0.0.0.0/0 - 开放端口3030显示日志部署进度：0.0.0.0/0 - 区域：华北-北京四	1	0.00元

华为云服务	配置示例	数量	一次完整流程预估（1小时）花费
Flexus云服务器X实例	• 按需计费：1.09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） x1.8u.16g 8核 16GB • 镜像：Ubuntu 24.04 server 64bit • 系统盘：通用型SSD 200GB	1	1.09
弹性公网IP EIP	• 按需计费：0.80元/GB • 区域：华北-北京四 • 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	-	-	1.09元 + 弹性公网IP EIP费用

 **注意**

预估费用仅供参考，实际费用取决于具体使用情况。详细价格请参见华为云官网。最终费用以实际账单为准。

表 2-2 资源和成本规划（包月计费）

华为云服务	配置示例	数量	每月预估花费
虚拟私有云 VPC	• VPC网段：172.16.0.0/16 • 区域：华北-北京四	1	0.00元
子网 Subnet	• 子网网段：172.16.1.0/24 • 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
安全组 SecurityGroup	- 允许ping: 0.0.0.0/0 - 配置端口22允许Cloud Shell 登录: 121.36.59.153/32 - 开放端口4000允许访问LiteLLM 代理 API 和 管理页面: 0.0.0.0/0 - 开放端口3030显示日志部署进度: 0.0.0.0/0 - 区域: 华北-北京四	1	0.00元
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 16GB - 镜像: Ubuntu 24.04 server 64bit - 系统盘: 通用型SSD 200GB	1	572.00元
弹性公网IP EIP	- 按需计费: 0.80元/GB - 区域: 华北-北京四 - 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	-	-	572.00元 + 弹性公网IP EIP费用

CCE 高可用部署

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	一次完整流程预估（1小时）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

华为云服务	配置示例	数量	一次完整流程预估 (1小时) 花费
子网 Subnet	- 区域：华北-北京四 - 子网网段： 192.168.200.0/24、 192.168.201.0/24、 192.168.202.0/24 - 网关：192.168.200.1、 192.168.201.1、 192.168.202.1	3	0.00
安全组 SecurityGroup	区域：华北-北京四	4	0.00
云容器引擎 CCE	- 区域：华北-北京四 - 类型：CCE Turbo - 规格：cce.s2.small (50 节点) - 集群 master 实例数： 3	1	2.91 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 vCPUs 16 GiB - 系统盘：通用型SSD 100GB - 镜像：Ubuntu 24.04 server 64bit	3	2.99 元
弹性公网IP EIP	- 区域：华北-北京四 - 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
弹性负载均衡 ELB	- 区域：华北-北京四 - 类型：独享型负载均衡 - 规格：弹性规格 应用型	1	0.15 元 + LCU费用 (按实际使用量收取LCU费用)
NAT网关 NAT	- 区域：华北-北京四 - 网关类型：公网 - 规格：小型 - 计费模式：按天	1	12.00 元

华为云服务	配置示例	数量	一次完整流程预估（1小时）花费
分布式缓存服务 Redis®*版	- 区域：华北-北京四 - 实例类型：Redis 7.0 (主备) - 规格：8G（基础版） 副本数：2	1	1.15 元
云数据库 RDS for PostgreSQL	- 区域：华北-北京四 - 数据库引擎：PostgreSQL(主备) - 数据库版本：16 - 规格：rds.pg.x1.2xlarge.2.ha 4 vCPUs 8 GB (通用型) - 存储：SSD云盘 100GB	1	2.12 元
MaaS tokens（可选）	- 计费模式：按Token计费 - 模型：DeepSeek-V3	-	输入：2.00 / 百万 tokens 输出：8.00 / 百万 tokens
合计	-	-	9.17 元 + 公网 NAT网关 12.00/天 + 应用型LCU费用 + 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：192.168.0.0/16	1	0.00
子网 Subnet	- 区域：华北-北京四 - 子网网段：192.168.200.0/24、192.168.201.0/24、192.168.202.0/24 - 网关：192.168.200.1、192.168.201.1、192.168.202.1	3	0.00

华为云服务	配置示例	数量	每月预估花费
安全组 SecurityGroup	区域：华北-北京四	4	0.00
云容器引擎 CCE	- 区域：华北-北京四 - 类型：CCE Turbo - 规格：cce.s2.small（50 节点） - 集群 master 实例数：3	1	1262.40 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 vCPUs 16 GiB - 系统盘：通用型SSD 100GB - 镜像：Ubuntu 24.04 server 64bit	3	1506.00 元
弹性公网IP EIP	- 区域：华北-北京四 - 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
弹性负载均衡 ELB	- 区域：华北-北京四 - 类型：独享型负载均衡 - 规格：弹性规格 应用型	1	400.00 元
NAT网关 NAT	- 区域：华北-北京四 - 网关类型：公网 - 规格：小型 - 计费模式：按天	1	306.00 元
分布式缓存服务 Redis®*版	- 区域：华北-北京四 - 实例类型：Redis 7.0（主备） - 规格：8G（基础版） 副本数：2	1	548.80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 - 数据库引擎：PostgreSQL(主备) - 数据库版本：16 - 规格：rds.pg.x1.2xlarge.2.ha 4 vCPUs 8 GB (通用型) - 存储：SSD云盘 100GB	1	965.00 元
MaaS tokens（可选）	- 计费模式：按Token计费 - 模型：DeepSeek-V3	-	输入：2.00 / 百万tokens 输出：8.00 / 百万tokens
合计	-	-	4988.20 元 + 弹性公网IP EIP费用 + MaaS tokens费用

须知

部署CCE容器高可用版需要用到3个弹性公网IP，分别用于绑定NAT网关为private subnet提供公网访问、绑定ELB连接CCE用于服务访问、绑定CCE集群的EIP是由于TerraForm访问k8s API Server需要（其中绑定CCE集群的EIP可待资源全部创建完成后，如果不需要API Server公网访问，可以释放）。

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。

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 进入[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”。

图 3-1 授权说明



----结束

获取华为云 MaaS API Key

- 步骤1 进入西南-贵阳一，登录MaaS控制台，在顶部导航栏中选择目标区域。
- 步骤2 在左侧导航栏，选择“管理与统计 > API Key 管理”。

图 3-2 选择 API Key 管理

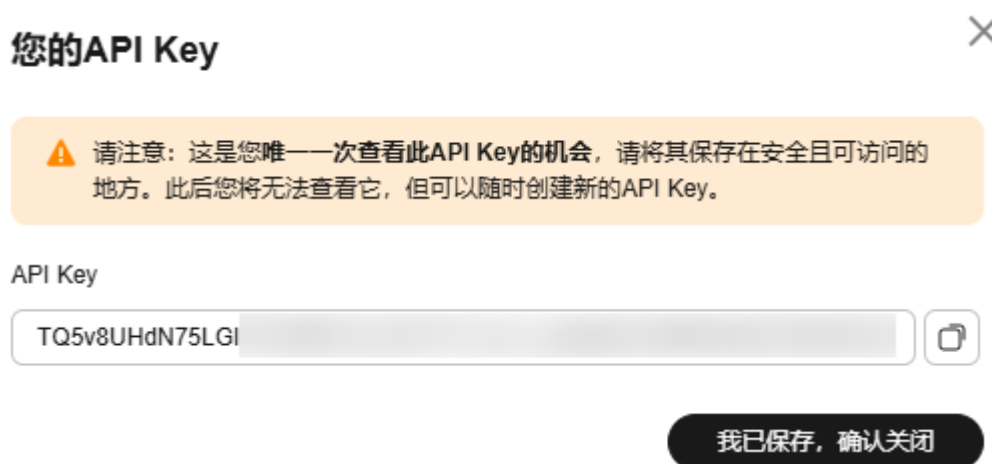


步骤3 在“API Key管理”页面，单击“创建API Key”，配置基础信息和权限设置，单击“创建”。

标签在创建完成后，不支持修改。

步骤4 在“您的API Key”对话框，仔细阅读提示信息，复制密钥并保存至安全位置，然后单击“我已保存，确认关闭”。

图 3-3 保存 API Key



----结束

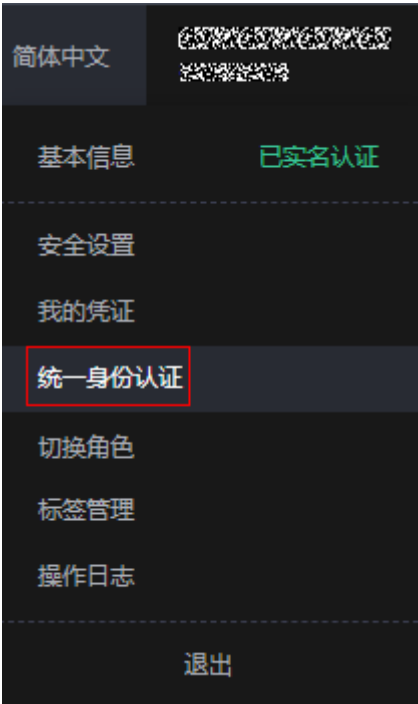
创建 rf_admin_trust 委托（可选）

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 3-4 控制台管理界面



图 3-5 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 3-6 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 3-7 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 3-8 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 3-9 选择策略



步骤6 选择“所有资源”，并单击“确定”完成配置。

图 3-10 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 3-11 委托列表



----结束

3.2 快速部署

本章节帮助用户高效地部署“LiteLLM，统一的AI管理网关”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。

操作步骤

步骤1 登录[华为云解决方案实践](#)，选择“LiteLLM，统一的AI管理网关”，单击“开始部署”，选择对应的版本（社区单机版和CCE容器高可用版），跳转至解决方案创建资源栈界面（以下以高可用版本为例）。

图 3-12 解决方案主页



步骤2 在选择模板界面中，单击“下一步”。

图 3-13 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”完成自定义参数填写，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 3-14 配置参数

立即创建资源栈

选择模板

参数配置

资源栈设置

配置确认

Provider: 华为云资源栈服务: 了解更多

资源栈名称

litellm-unified-ai-management-gateway

资源栈名称以中文或英文开始，支持中文、大小写英文、数字、下划线、中划线，128个字符之内，资源栈名称不能重复。

描述

LiteLLM，统一的 AI 管理网关（灰度功能）

23/105

配置参数

请输入关键字搜索参数名称

☒ 按模板要求对部分参数加锁

参数名称	值	类型	描述
* region_id	cn-nor-4	string	资源部署所在Region
* solution_name	ha-litellm	string	解决方案名称，命名规则{solution_name}-资源英文名称，例如：CCE集群名称为{solution_name}-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。
* bandwidth_size	300	string	弹性公网带宽大小，按流量计费方式按流量计费，取值范围：1-300Mbps，默认300Mbps。
* litellm_version	1.94.0	string	LiteLLM 网关镜像版本，该模板会自动识别并更新为 ghcr.io/bentallitellm/version 格式。
* litellm_master_key	*****	string	LiteLLM 主密钥，必须以 sk 开头，最大24个字符，这适用于外部调用代码的 Bearer Token。
* litellm_salt_key	*****	string	LiteLLM 加盐密钥，用于加密数据库中的敏感数据，取值范围：8-24个字符，可设置大、小写字母、数字或特殊字符（~@#%^_-=+[]{};.,/?）。中的三种。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。
* cce_cluster_flavor	cce.s2.small	string	CCE Turbo集群规格，集群创建后规格将不可再变更，可选值：cce.s2.small、cce.s2.medium、cce.s2.large、cce.s2.xlarge，具体请参考 弹性云服务器规格清单 。
* cce_node_pool_password	W564521	string	CCE集群节点密码，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（~@#%^_-=+[]{};.,/?）。中的三种。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。
* cce_node_pool_flavor	x1.8u.16g	string	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，具体华为云Flexus 云服务器X实例规格请参考 弹性云服务器规格清单 。
* cce_node_pool_count	3	number	CCE节点池中的初始节点数。

上一步

下一步

表 3-1 参数填写说明（社区单机版）

参数名称	类型	是否可选	参数解释	默认值
solution_name	string	必填	解决方案名称，命名规则 {solution_name}-资源英文名称，例如：CCE集群名称为 {solution_name}-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。	litellm-llm-gateway
ecs_flavor	string	非必填	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 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，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。	-

参数名称	类型	是否可选	参数解释	默认值
master_key	number	必填	LiteLLM 主密钥，用于外部调用代理的 Bearer Token，必须以 sk-开头，取值范围：8-24个字符，可设置大、小写字母、数字。	-
salt_key	string	必填	LiteLLM 加盐密钥，用于加密数据库中的敏感数据，取值范围：8-24个字符，可设置大、小写字母、数字或特殊字符（~!@#\$%^*-=_,?）	-
system_disk_size	string	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。	200
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 参数填写说明（CCE 容器高可用版）

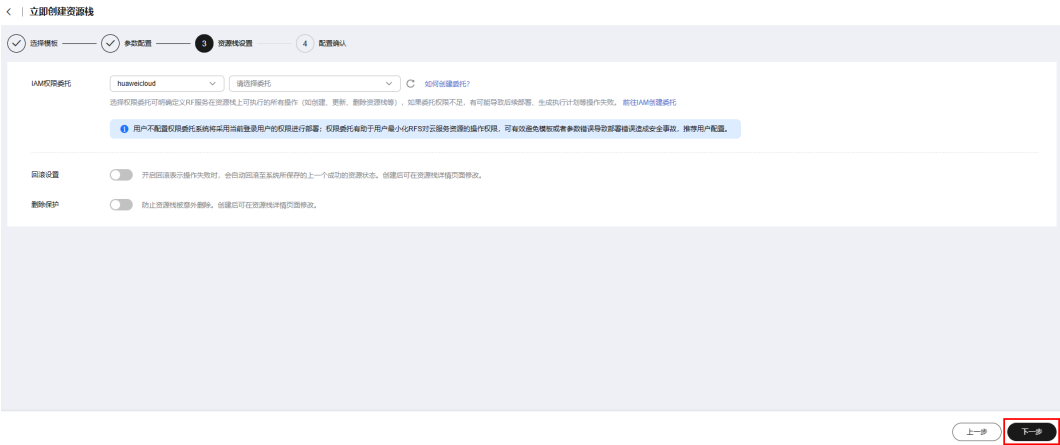
参数名称	类型	是否可选	参数解释	默认值
region_id	string	必填	资源部署所在Region	cn-north-4

参数名称	类型	是否可选	参数解释	默认值
solution_name	string	必填	解决方案名称，命名规则 {solution_name}-资源英文名称，例如：CCE集群名称为 {solution_name}-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。	ha-litellm
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。取值范围：1-300Mbit/s。	300
litellm_version	string	必填	LiteLLM 网关镜像版本。该模板会自动将其组装为 ghcr.io/berriai/litellm:{version} 格式。	1.94.0
litellm_master_key	string	必填	LiteLLM 主密钥，必须以 sk-开头。最大值24个字符，用于外部调用代理的 Bearer Token。	-
litellm_salt_key	string	必填	LiteLLM 加盐密钥，用于加密数据库中的敏感数据，取值范围：8-24个字符，可设置大、小写字母、数字或特殊字符（~!@#%^*-=,?）	-
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：8-24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 X实例控制台配置 ，弹性云服务器 ECS规格请参考 ECS控制台配置 。	x1.8u.16g
cce_node_pool_count	number	必填	CCE节点池初始节点数，最大值9。	3
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。 。其他规格请参考 RDS 控制台配置 。	rds.pg.x1.2xlarge.2.ha
pgsql_admin_password	string	必填	PostgreSQL 管理员密码。取值范围：8-24个字符。密码必须至少包含大写字母、小写字母和数字或特殊字符（~!^*-=_+, ）。。	-
redis_capacity	number	必填	分布式缓存服务 Redis版实例规格。可选值：1GB-64GB。	8
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（~!^*-=_+, ）。。	-
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
charging_unit	string	必填	订购周期类型，当charging_mode为prePaid时，该参数为必填参数。可选值为：month（月），year（年）。	month
charging_period	number	必填	计费周期，当charging_mode为prePaid时，该参数为必填参数。可选值为：1-3（year）、1-9（month）。	1

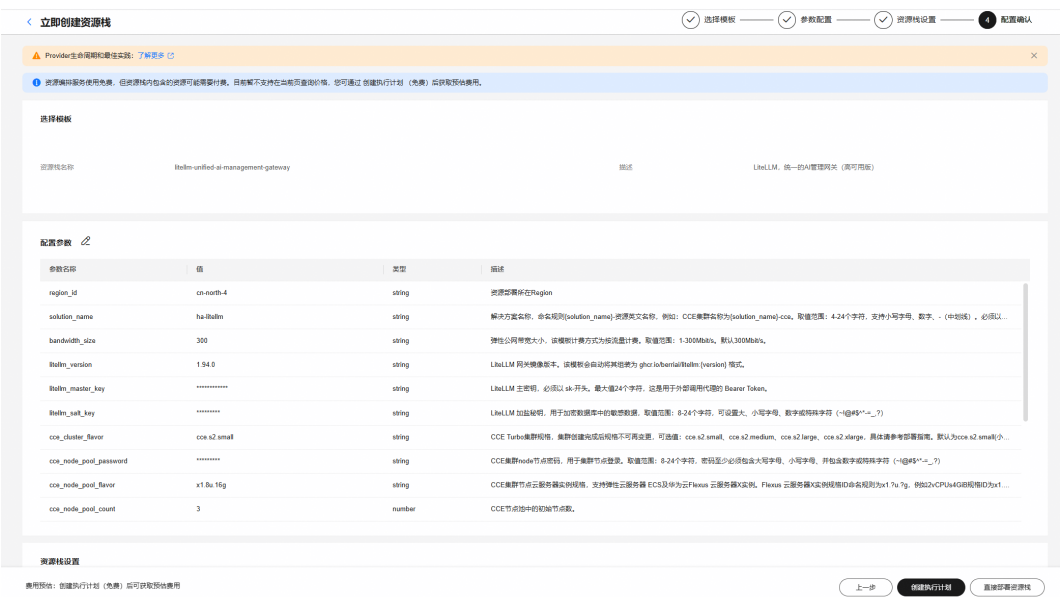
步骤4 （可选，如果使用华为主账号或admin用户组下的IAM子账户可不选委托）在资源设置界面中，在权限委托下拉框中选择“rf_admin_trust”委托，单击“下一步”。

图 3-15 资源栈设置



步骤5 在配置确认界面中，确认填写参数并单击“创建执行计划”。

图 3-16 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 3-17 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_20260820_1713_nidi

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 3-18 执行计划

litellm-unified-ai-management-gateway

删除

更新模板或参数

C

基本信息

资源

输出

事件

模板

执行计划

搜索

请输入关键字

执行计划名称ID	状态	操作	创建时间	描述	操作
executionPlan_20260820_1713_nidi	创建成功，待部署	查看费用明细	2026/08/20 17:14:22 GMT+08:00	--	部署

图 3-19 执行计划确认

执行计划

您确定要执行该计划吗？

执行计划名称	状态	创建时间
executionPlan_20260820_17...	创建成功...	2026/08/20 17:14:22 GMT+0...

确定执行后，资源栈会按照该计划更新，并且 会开通模板内的资源，根据资源付费要求，可能会产生费用。

执行

取消

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2-4）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，脚本开始执行，耐心等待10分钟左右（受网络波动影响）。

图 3-20 部署完成



步骤10 刷新页面，在“输出”页面查看访问链接。



----结束

3.3 开始使用

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。

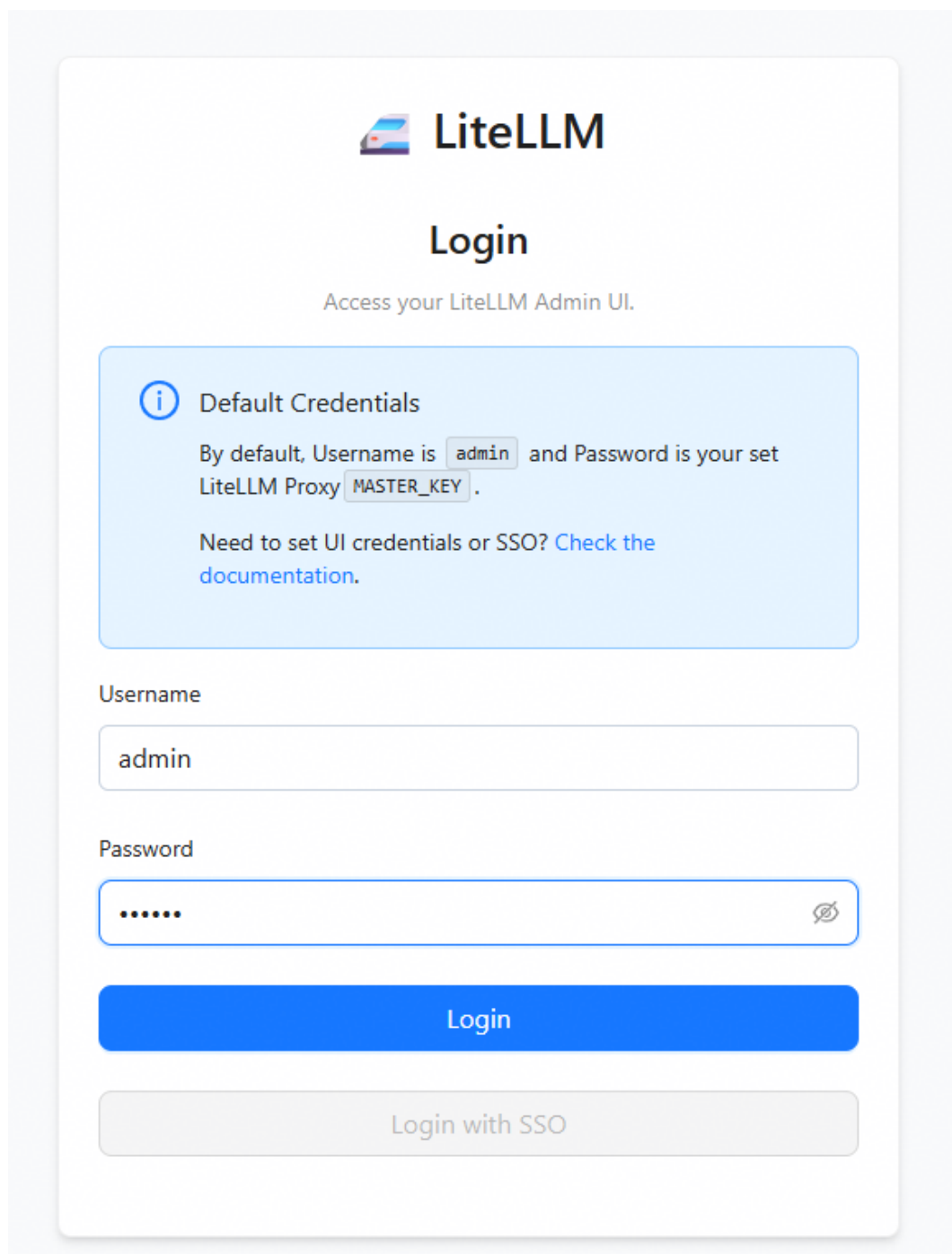
如果您的实例关联的安全组策略无法满足使用需求，比如需要添加、修改、删除某个TCP端口，请参考以下内容进行修改。

- 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。

使用后台 Web UI 进行模型添加

步骤1 根据“快速步骤”步骤五的相关提示进行登录，输入初始用户名admin与“快速部署”中设置的初始密码（master_key）。

图 3-21 登录 LiteLLM 管理 Web UI



 **LiteLLM**

Login

Access your LiteLLM Admin UI.

 **Default Credentials**

By default, Username is `admin` and Password is your set LiteLLM Proxy `MASTER_KEY`.

Need to set UI credentials or SSO? [Check the documentation.](#)

Username

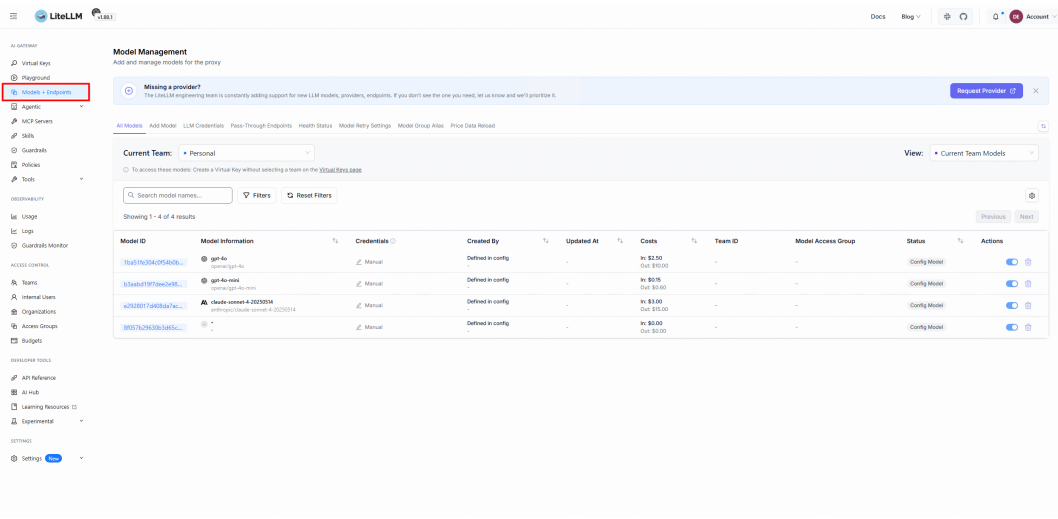
Password

说明

若忘记密码可前往服务器查询config.yaml文件中配置的master_key。

步骤2 单击左侧“Models+Endpoints”，进入到模型配置界面，存在一些系统自带的示例配置，可按需选择保存或者删除。

图 3-22 模型配置界面



步骤3 单击“Add Model”对以下参数进行配置。

图 3-23 添加模型

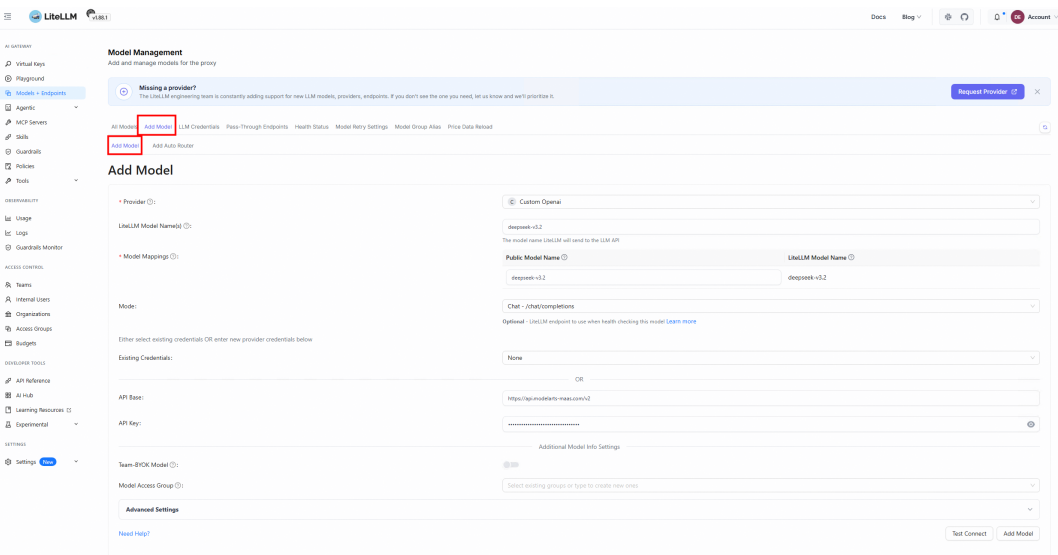


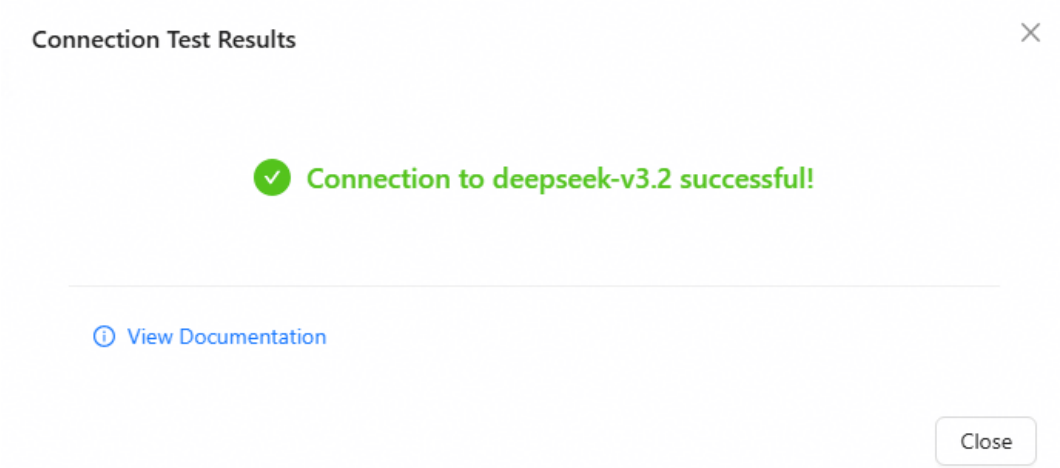
表 3-3 基础配置说明

参数	说明	默认值
参数	描述	示例
provider	模型提供商，如OpenAI、Bedrock等。	Custom OpenAI
LiteLLM Model Name	LiteLLM发送至API处的模型名称	deepseekv3.2
Model Mappings	模型映射	deepseekv3.2
Mode	接口模式	Chat - /chat/completions

参数	说明	默认值
API Base	API地址的URL	https://api.modelarts-maas.com/v2
API Key	API密钥	您获取的API Key

步骤4 单击“Test Connect”，当出现下图内容，即可单击“Add Model”添加模型。

图 3-24 Test Connect 测试成功



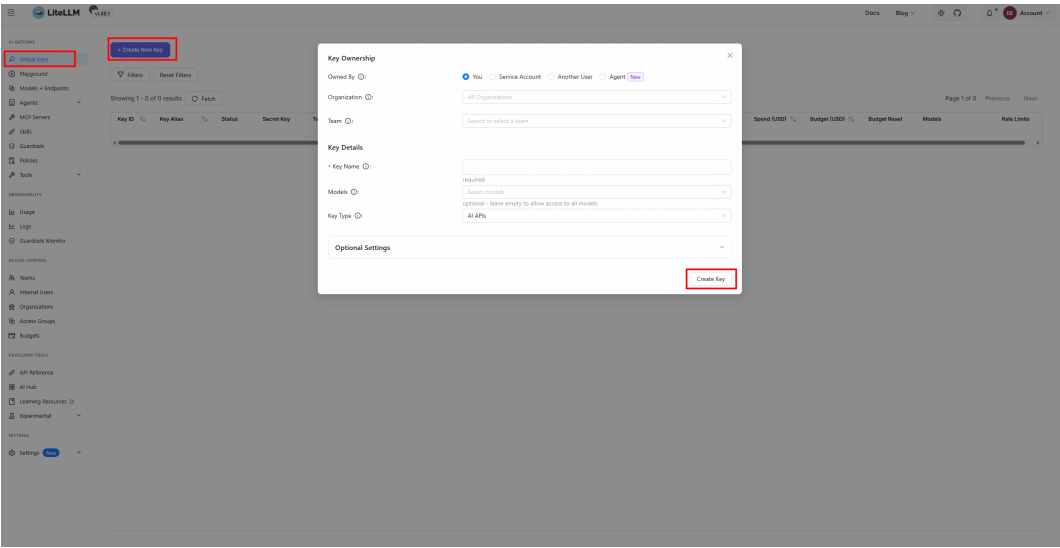
----结束

创建 Virtual Key 进行模型管理和验证

Virtual Key的拥有者可以是你自己、某个账号或某个具体用户。

步骤1 单击“Virtual Key”，选择“Create New Key”。

图 3-25 创建虚拟密钥



步骤2 根据下表，按需填入信息后点击确定，复制Virtual Key信息。

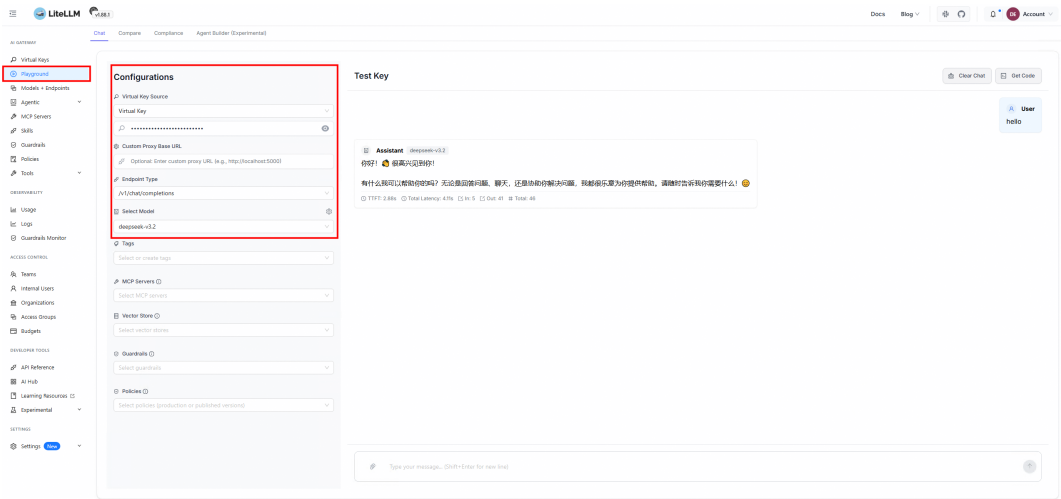
表 3-4 密钥相关信息

模块	参数名	选项 / 示例	含义说明
Key Ownership (密钥所有权)	Owned By	You / Service Account / Another User	密钥的归属主体，决定谁是密钥的所有者： You：当前登录用户，个人开发者场景常用 Service Account：服务账号，用于CI/CD、自动化脚本等 Another User：其他已存在的 LiteLLM 用户，用于分配给团队成员
	Team	Search or select a team	关联的团队，用于多租户 / 团队场景。密钥会继承团队的预算、模型权限等配置，实现按团队维度的成本与权限管理
Key Details (密钥详情)	Key Name	deepseek-v3.2	密钥的自定义别名，仅用于管理侧识别，方便区分不同用途的密钥（如项目名、应用名），不影响 API 调用
	Models	deepseek-v3.2	该密钥允许访问的模型列表，可多选。限制密钥只能调用指定模型，防止越权使用未授权的模型资源

模块	参数名	选项 / 示例	含义说明
	Key Type	LLM API	密钥的类型，定义其可访问的 API 端点类别： LLM API：通用大语言模型接口（如 chat/completions、completions） - 其他可选值通常包含 embedding、image、audio 等专用 API 类型，用于精细化控制

步骤3 单击“Playground”，在Virtual Key Source中选择Virtual Key，输入Virtual Key等相关信息。在对话框中输入“hello”测试是否能够正常回答您的信息。

图 3-26 测试验证



----结束



Virtual Key仅生成一次，请注意及时保存。
更多进阶使用教程，参考[LiteLLM官方仓库](#)。

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

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。

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。

资源编排服务 RFS：资源编排服务是应用编排服务新增的完全支持业界事实标准 Terraform（HCL + Provider）的终态编排引擎，用于管理系统资源及服务资源。资源编排服务采用业界开放生态HCL语法的模板，它按照模板自动化部署指定的云服务资源。

参考文档

[LiteLLM GitHub](#)

[LiteLLM Docs](#)

[华为云 MaaS 使用指南](#)

[华为云 RFS](#)

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2026-08-24	第一次正式发布。