CECSTACK 5.3.0

大模型推理服务 LAS 用户指南

文档密级：公开

文档版本：01

发布日期：2025-01-23

中电云计算技术有限公司

前言

概述

本文档主要介绍大模型推理服务 LAS 的使用步骤。

读者对象

本文档适用于以下读者：

- 产品用户
- 技术支持工程师
- 系统管理员

本书约定

符号标志约定

本书采用各种醒目标志来表示在操作过程中应该特别注意的地方，这些标志的意义如下：

警告	该标志后的注释需给予格外关注，不当的操作可能会对人身造成伤害。
注意	提醒操作中应注意的事项，不当的操作可能会导致数据丢失或者设备损坏。 “注意”不涉及人身伤害。
说明	对正文的重点信息进行必要的补充说明。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。
提示	配置、操作、或使用产品的技巧、窍门。

修订记录

文档版本	发布时间	修订说明
01	2025-01-23	第一次正式发布。

目 录

1 产品简介

1.1 什么是大模型推理服务LAS

大模型高效推理服务平台提供快速、高效和可扩展的推理能力，输出标准大模型推理服务 API。通过集成多种优化技术，如模型压缩、量化和剪枝，显著减少了模型的体积和计算需求，从而加快了推理速度并降低了运营成本。LAS 支持自动批处理请求，进一步提升处理效率，并通过弹性扩展机制，能够根据实时负载动态调整资源，确保服务的稳定性和可靠性。此外，平台的多模型支持功能使其能够适应各种应用场景，包括自然语言处理、计算机视觉等多模态任务。通过异构计算资源管理，平台能够充分利用 GPU、CPU 等不同类型的硬件资源，为大模型的推理提供强大的动力。整体而言，大模型高效推理服务平台为企业和研究者提供了一个强大的基础设施，以实现 AI 模型的高效部署和应用。

1.2 主要功能

大模型推理服务 API 通过提供现成的接口，极大地降低了 RAG（Retrieval-Augmented Generation，检索增强生成）和 Agent（智能代理）应用的开发和部署门槛。这些 API 使得开发者能够轻松地将大型预训练模型集成到各种应用中，而无需从头开始构建复杂的模型和推理系统。以下是这种 API 为 RAG 和 Agent 应用带来的一些关键优势：

快速开发

现成的 API 允许开发者快速实现 RAG 和 Agent 应用的核心功能，如文本生成、问答系统、对话管理等，从而缩短开发周期。

降低技术门槛

开发者无需深入了解深度学习、自然语言处理等领域的复杂技术细节，只需通过简单的 API 调用来实现高级的 AI 功能。

成本效益

使用标准化 API 避免了构建和维护大型模型所需的高昂成本，包括计算资源、存储和专业人才。

灵活性和可定制性

API 通常提供了丰富的配置选项，使得开发者可以根据具体的应用场景调整模型的行为和输出。

易于集成

标准化的 API 易于与现有的软件架构和工作流程集成，无论是在本地还是在云环境中。

性能保证

大模型推理服务 API 背后的基础设施由专业的团队维护，确保了高性能和高可靠性。

持续更新和优化

随着 AI 技术的发展，API 提供者会不断更新和优化模型，开发者可以无缝获得最新的功能和改进。

2 使用指南

2.1 模型服务

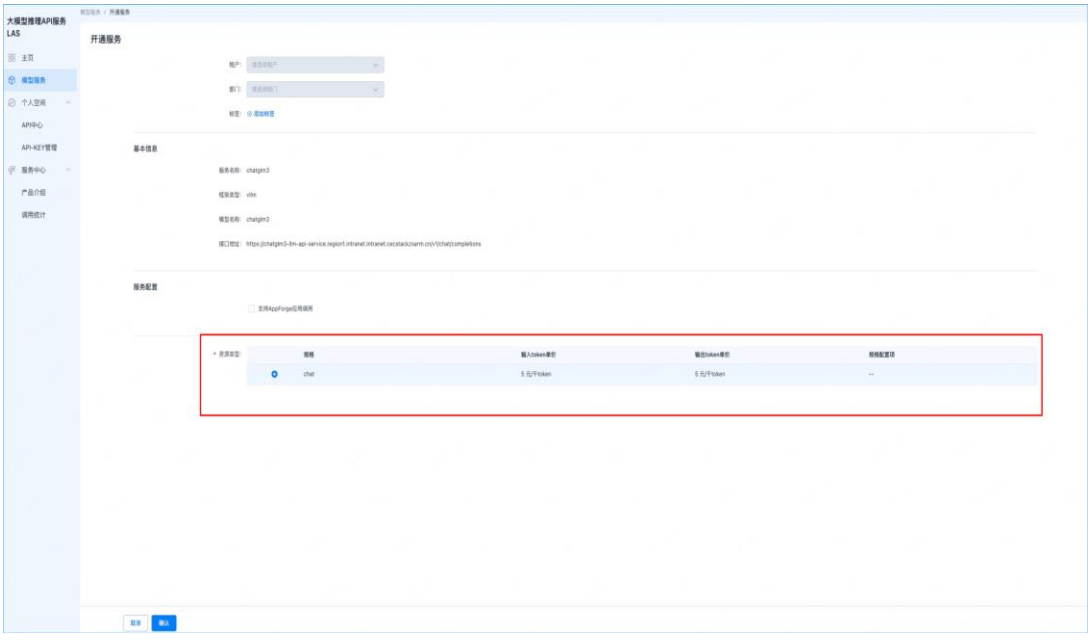
2.1.1 服务开通

(1) 点击“云服务 > LLM API Service”进入大模型推理服务。

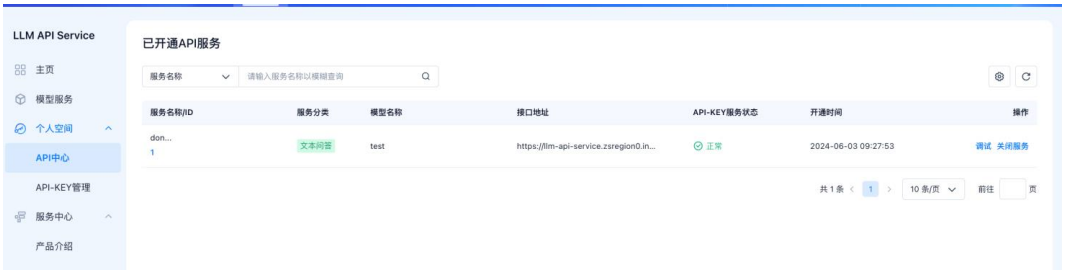
(2) 点击“模型服务”，在已开通的模型服务卡片中点击“开通服务”。

(3) 在确认页面点击确认按钮即可开通。

如需在 AppForge 里使用模型推理服务，可勾选“支持 AppForge 应用调用”选项以及计费标准。



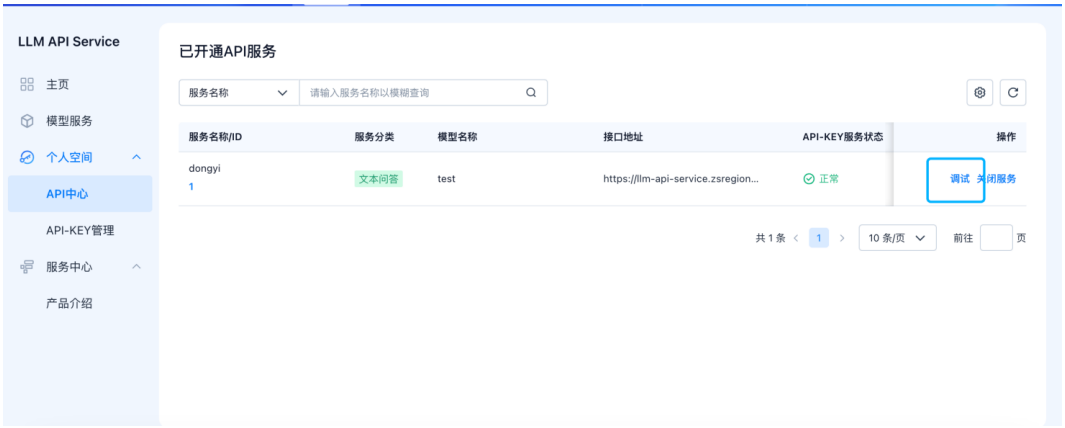
(4) 服务开通后，可在 API 中心查看到服务。



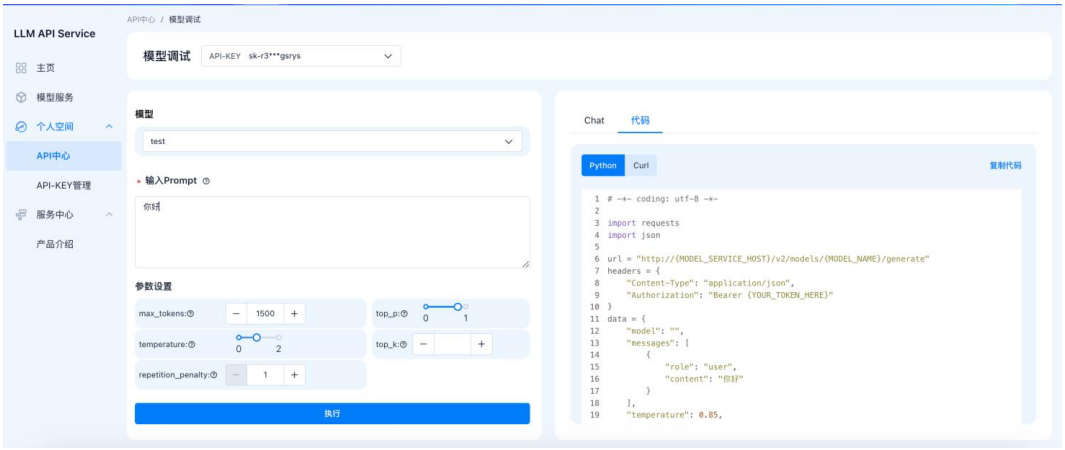
2.2 API中心

2.2.1 调试

(1) 对于已开通的推理服务，在 API 中心页面，点击调试可进入调试控制台。



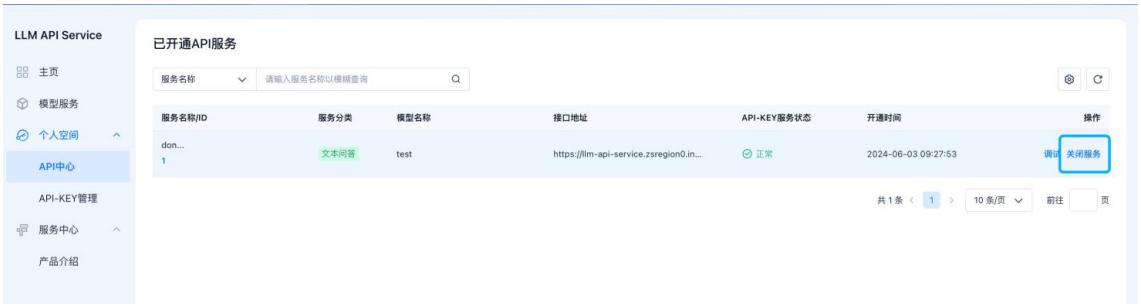
(2) 在调试页面可知识模型调用 API Key，切换模型，输入提示词，以及进行推理请求参数调整。同时在页面右侧可同时生成，基于 Python 和 Curl 的代码样例。



(3) 点击执行后，可发送请求，请求结果可在右侧“Chat”页面查看。

2.2.2 关闭服务

对于已开通 API 的服务，可在列表页点击“关闭服务”来停止服务，已停止的服务无法继续交互。



2.3 API-Key管理

2.3.1 创建 API-Key

2.3.2 查看已存在的 API-Key

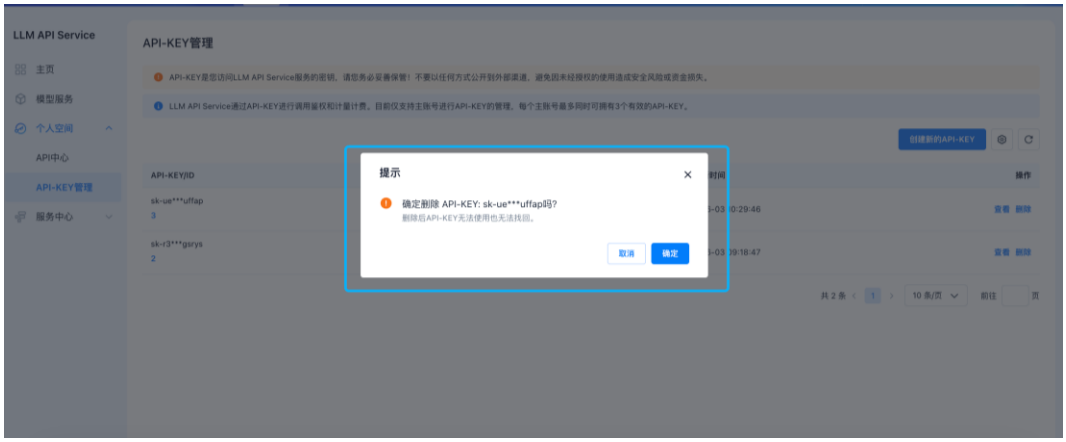
(1) 点击已创建 API KEY 列表的“查看”链接，可查看 API Key 信息。

(2) 在弹出窗口中可复制 key 的内容。

2.3.3 删除 API-Key

(1) 在 API-Key 的列表页，点击相应 key 的“删除”链接，可删除已有的 key。

(2) 点击确认后可将 key 删除，删除后的 key 无法再用于鉴权。



2.4 服务中心

服务中心页面主要提供了产品的介绍。

2.5 服务中心-调用统计

查看开通的模型，以及使用的 api-key，查看调用情况。