ModelArts Studio

用户指南

文档版本	01
发布日期	2025-07-31

华为云计算技术有限公司

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。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 ModelArts Studio（MaaS）使用场景和使用流程 1
2 配置 ModelArts Studio（MaaS）访问授权 3
2.1 创建 IAM 用户并授权使用 ModelArts Studio（MaaS） 3
2.2 配置 ModelArts 委托授权以使用 ModelArts Studio（MaaS） 6
2.3 配置用户缺失的 ModelArts Studio（MaaS）相关服务权限 13
3 准备 ModelArts Studio（MaaS）资源 20
4 ModelArts Studio（MaaS）在线推理服务 21
4.1 在 ModelArts Studio（MaaS）模型广场查看预置模型 21
4.2 使用 ModelArts Studio（MaaS）部署模型服务 23
4.3 在 ModelArts Studio（MaaS）管理我的服务 26
4.3.1 在 ModelArts Studio（MaaS）启动/停止/定时启停/删除服务 26
4.3.2 在 ModelArts Studio（MaaS）扩缩容模型服务实例数 26
4.3.3 在 ModelArts Studio（MaaS）修改模型服务 QPS 27
4.3.4 在 ModelArts Studio（MaaS）升级模型服务 28
4.4 调用 ModelArts Studio（MaaS）部署的模型服务 28
4.5 ModelArts Studio（MaaS） API 调用规范 35
4.5.1 对话 Chat/POST 35
4.5.2 获取模型列表 Models/GET 41
4.5.3 错误码 42
4.6 使用 ModelArts Studio（MaaS）创建多轮对话 45
5 ModelArts Studio（MaaS）管理与统计 47
5.1 在 ModelArts Studio（MaaS）管理 API Key 47

1 ModelArts Studio（MaaS）使用场景和使用流程

ModelArts Studio大模型即服务平台（后续简称为MaaS服务），提供端到端的大模型生产工具链和昇腾算力资源，并预置了当前主流的第三方开源大模型，支持大模型数据生产、微调、提示词工程、应用编排等功能。用户可以基于MaaS平台开箱即用，对预置大模型进行二次开发，用于生产商用。

背景介绍

近年来，AI大模型凭借强大的自然语言理解、内容生成和决策辅助能力，正在成为企业数字化转型的重要推动力。越来越多的企业希望借助大模型优化业务流程，例如智能客服、数据分析、自动化报告生成等。然而，企业在尝试自主训练或微调大模型时，往往面临三大核心挑战：高昂的算力成本、复杂的技术门槛以及业务系统集成难题。由于大多数企业缺乏专业的AI团队，从零开始构建和优化模型变得异常困难，这直接导致了AI应用落地效率低下甚至项目失败。

针对这些痛点，MaaS提供了一站式解决方案：

- 工具链：提供可视化训练平台，降低技术门槛，使企业无需深厚AI背景即可完成模型定制。
- 资源共享：通过云端算力共享和预训练模型复用，帮助企业避免重复投资，显著降低算力成本。
- 场景化适配：基于行业需求提供预置模型模板，加速企业AI应用的落地部署。

应用场景

ModelArts Studio大模型即服务平台（MaaS）的应用场景：

- **业界主流开源大模型覆盖全**

  MaaS集成了业界主流开源大模型，含DeepSeek等模型系列，所有的模型均基于昇腾AI云服务进行全面适配和优化，使得精度和性能显著提升。开发者无需从零开始构建模型，只需选择合适的预训练模型直接应用，减轻模型集成的负担。

- **资源易获取，按需收费，按需扩缩，支撑故障快恢与断点续训**

  企业在具体使用大模型接入企业应用系统的时候，不仅要考虑模型体验情况，还需要考虑模型具体的精度效果，和实际应用成本。

  MaaS提供灵活的模型开发能力，同时基于昇腾云的算力底座能力，提供了若干保障客户商业应用的关键能力。

保障客户系统应用大模型的成本效率，按需收费，按需扩缩的灵活成本效益资源配置方案，有效避免了资源闲置与浪费，降低了进入AI领域的门槛。

架构强调高可用性，多数据中心部署确保数据与任务备份，即使遭遇故障，也能无缝切换至备用系统，维持模型训练不中断，保护长期项目免受时间与资源损耗，确保进展与收益。

- **大模型应用开发，帮助开发者快速构建应用**

  在企业中，项目级复杂任务通常需要理解任务并拆解成多个问题再进行决策，然后调用多个子系统去执行。MaaS基于多个优质昇腾云开源大模型，让大模型准确理解业务意图，分解复杂任务，沉淀出丰富的解决方案，帮助企业快速智能构建和部署大模型应用。

支持区域

仅香港区域支持使用MaaS。

使用流程

表 1-1 MaaS 使用流程

模块	操作	说明	相关文档
授权	配置访问授权	对于所有用户（包括个人用户），需要完成ModelArts委托授权才能使用MaaS服务，否则会造成您的操作出现不可预期的错误。	• 创建IAM用户并授权使用ModelArts Studio（MaaS） • 配置ModelArts委托授权以使用ModelArts Studio（MaaS）
在线推理服务	查看模型广场的预置模型	ModelArts Studio大模型即服务平台提供了丰富的开源大模型，在“模型广场”页面可以查看。模型详情页可以查看模型的详细介绍，根据这些信息选择合适的模型进行训练、推理，接入到企业解决方案中。	在ModelArts Studio（MaaS）模型广场查看预置模型
	部署模型服务	ModelArts Studio大模型即服务平台支持将模型广场的预置模型部署到计算资源上，便于在“模型体验”或其他业务环境中可以调用该模型。	使用ModelArts Studio（MaaS）部署模型服务
API调用	调用模型服务	在ModelArts Studio大模型即服务平台完成模型部署后，可以在其他业务环境中调用该模型服务进行预测。	调用ModelArts Studio（MaaS）部署的模型服务

2 配置 ModelArts Studio（MaaS）访问授权

2.1 创建 IAM 用户并授权使用 ModelArts Studio（MaaS）

配置ModelArts委托授权以使用ModelArts Studio（MaaS）章节中介绍的一键式自动授权方式创建的委托的权限比较大，基本覆盖了依赖服务的全部权限。如果华为云账号已经能满足您的要求，则不需要创建独立的IAM用户，您可以跳过本章节，不影响您使用MaaS服务的功能。

ModelArts作为一个完备的AI开发平台，支持用户对其进行细粒度的权限配置，以达到精细化资源、权限管理之目的。这类特性在大型企业用户的使用场景下很常见。如果需要对委托授权的权限范围进行精确控制，可以参考本章节进行MaaS服务的定制化委托授权。

本章节主要介绍如何给IAM用户下的子用户配置更细粒度的权限。

场景描述

统一身份认证（Identity and Access Management，简称IAM）是华为云提供权限管理的基础服务，可以帮助您安全地控制云服务和资源的访问权限。

IAM无需付费即可使用，您只需要为您账号中的资源进行付费。您注册华为云后，系统自动创建账号，账号是资源的归属以及使用计费的主体，对其所拥有的资源具有完全控制权限，可以访问华为云所有的云服务。更多信息，请参见**什么是IAM**。

授权流程

创建用户组并授权：如果企业中不需要每个人都注册账号，则可以由企业的管理员注册一个账号，在这个账号下创建用户组并分配权限，然后将创建的IAM用户根据不同的职能加入到不同的用户组中，分发给企业的人员使用。更多信息，请参见**创建用户组并授权**。

创建IAM用户并登录：创建一个IAM用户，并将其加入用户组中获得相应的权限。IAM用户登录**ModelArts Studio（MaaS）控制台**，使用权限范围内的资源。更多信息，请参见**创建IAM用户并登录**。

计费说明

授权是通过IAM（身份和访问管理）服务进行的，用于控制用户对ModelArts资源的访问权限。IAM服务本身是免费的，您无需为授权操作支付费用。

前提条件

- 仅管理员才可以创建IAM子用户。
- 给用户组授权之前，请先了解用户组可以添加的使用ModelArts及其依赖服务的权限，并结合实际需求进行选择，MaaS服务支持的系统权限，请参见**表2-1**。

表 2-1 服务授权列表

待授权的服务	授权说明	IAM权限设置	是否必选
ModelArts	授予子用户使用ModelArts服务的权限。 ModelArts CommonOperations没有任何专属资源池的创建、更新、删除权限，只有使用权限。推荐给子用户配置此权限。	ModelArts CommonOperations	必选
	如果需要给子用户开通专属资源池的创建、更新、删除权限，此处要勾选ModelArts FullAccess，请谨慎配置。	ModelArts FullAccess	可选 ModelArts FullAccess权限和ModelArts Common Operations权限建议二选一。
OBS对象存储服务	授予子用户使用OBS服务的权限。ModelArts的数据管理、开发环境、训练作业、模型推理部署均需要**通过OBS进行数据中转**。	OBS OperateAccess	必选
SWR容器镜像仓库	授予子用户使用SWR服务权限。ModelArts的**自定义镜像功能**依赖镜像服务SWR FullAccess权限。	SWR OperateAccess	必选

待授权的服务	授权说明	IAM权限设置	是否必选
CES云监控	授予子用户使用CES云监控服务的权限。通过CES云监控可以查看ModelArts的在线服务和对应模型负载运行状态的整体情况，并设置监控告警。	CES FullAccess	必选
SMN消息服务	授予子用户使用SMN消息服务的权限。SMN消息通知服务配合CES监控告警功能一起使用。	SMN FullAccess	必选
VPC虚拟私有云	子用户在创建ModelArts的专属资源池过程中，如果需要开启自定义网络配置，需要配置VPC权限。	VPC FullAccess	可选

配置 MaaS 基础操作权限

步骤1 创建用户组。

管理员登录IAM管理控制台，单击“用户组>创建用户组”。在“创建用户组”界面，输入“用户组名称”单击“确定”。

步骤2 配置用户组权限。

在用户组列表中，单击步骤1新建的用户组右侧的“授权”，在用户组“授权”页面，您需要配置的权限如下：

1. 配置ModelArts使用权限。选择系统策略，然后在搜索框搜索ModelArts。ModelArts FullAccess权限和ModelArts CommonOperations权限建议二选一。

   选择说明如下：

   - ModelArts CommonOperations没有任何专属资源池的创建、更新、删除权限，只有使用权限。推荐给子用户配置此权限。
   - 如果需要给子用户开通专属资源池的创建、更新、删除权限，此处要勾选ModelArts FullAccess，请谨慎配置。

   图 2-1 配置 ModelArts 使用权限

2. 配置其他依赖云服务的使用权限，此处以OBS为例，搜索OBS，勾选“OBS OperateAccess”。ModelArts训练作业中需要依赖OBS作为数据中转站，需要配置OBS的使用权限。

更多需要配置的云服务权限请参见**表2-1**，例如SWR等，重复操作此步骤。

3. 单击“下一步”，设置最小授权范围。单击“指定区域项目资源”，勾选待授权使用的区域，单击“确定”。
4. 提示授权成功，查看授权信息，单击“完成”。此处的授权生效需要15-30分钟。

步骤3 创建子用户账号。在IAM左侧菜单栏中，选择“用户”，单击右上角“创建用户”，在“创建用户”页面中，添加多个用户。请根据界面提示，填写必选参数，然后单击“下一步”。

步骤4 将子用户子账号加入用户组。在“加入用户组”步骤中，选择“用户组”，然后单击“创建用户”。系统将前面设置的多个用户加入用户组中。

步骤5 **用户登录**并验证权限。

新创建的用户登录**登录IAM管理控制台**，切换至授权区域，验证权限：

- 在“服务列表”中选择ModelArts，进入ModelArts主界面，选择不同类型的专属资源池，在页面单击“创建”，如果无法进行创建（当前权限仅包含ModelArts CommonOperations），表示“ModelArts CommonOperations”已生效。
- 在“服务列表”中选择除ModelArts外（假设当前策略仅包含ModelArts CommonOperations）的任一服务，如果提示权限不足，表示“ModelArts CommonOperations”已生效。
- 在“服务列表”中选择ModelArts，进入ModelArts主界面，单击“算法管理>创建算法”，如果可以成功访问对应的OBS路径，表示OBS权限已生效。
- 参考**表2-1**依次验证其他可选权限。

----结束

2.2 配置 ModelArts 委托授权以使用 ModelArts Studio（MaaS）

在使用ModelArts平台的MaaS服务时，权限管理是保障服务正常运行和数据安全的关键环节。ModelArts平台所有功能均依托IAM体系进行权限管控，服务管理员可借此对用户进行精细化权限设置。然而，部分用户在操作过程中，因未正确处理权限相关设置，出现了不可预期的错误，导致服务使用受阻。

无论是个人用户还是其他类型用户，都需要完成ModelArts委托授权，这是使用MaaS服务的必要前提，否则将导致操作出现错误。对于个人用户而言，无需考虑细粒度权限问题，完成ModelArts委托授权后，即可使用ModelArts。

场景描述

MaaS服务的访问授权是通过ModelArts统一管理的，当用户已拥有ModelArts的访问授权时，无需单独配置MaaS服务的访问授权，当用户没有ModelArts的访问授权时，则需要先完成配置才能正常使用MaaS服务。

ModelArts在任务执行过程中需要访问用户的其他服务，典型的就是训练过程中，需要访问OBS读取用户的训练数据。在这个过程中，就出现了ModelArts“代表”用户去访问其他云服务的情形。从安全角度出发，ModelArts代表用户访问任何云服务之前，均需要先获得用户的授权，而这个动作就是一个“委托”的过程。用户授权ModelArts再代表自己访问特定的云服务，以完成其在ModelArts平台上执行的AI计算任务。

ModelArts提供了一键式自动授权功能，用户可以在ModelArts的权限管理功能中，快速完成委托授权，由ModelArts为用户自动创建委托并配置到ModelArts服务中。

本章节主要介绍一键式自动授权方式。一键式自动授权方式支持给IAM子用户、联邦用户（虚拟IAM用户）、委托用户和所有用户授权。

约束与限制

- 华为云账号
  - 只有华为云账号可以使用委托授权，可以为当前账号授权，也可以为当前账号下的所有IAM用户授权。
  - 多个IAM用户或账号，可使用同一个委托。
  - 一个账号下，最多可创建50个委托。
  - 对于首次使用ModelArts的新用户，请直接新增委托即可。一般用户新增普通用户权限即可满足使用要求。如果有精细化权限管理的需求，可以自定义权限按需设置。
- IAM用户
  - 如果已获得委托授权，则可以在权限管理页面中查看到已获得的委托授权信息。
  - 如果未获得委托授权，当打开“访问授权”页面时，ModelArts会提醒您当前用户未配置授权，需联系此IAM用户的管理员账号进行委托授权。

计费说明

授权是通过IAM（身份和访问管理）服务进行的，用于控制用户对ModelArts资源的访问权限。IAM服务本身是免费的，您无需为授权操作支付费用。

计费主要与资源的使用相关，例如计算资源（vCPU、GPU、NPU）、存储资源（云硬盘、对象存储）等，详情请见**计费说明**。

前提条件

已注册华为账号并开通华为云，详情请见**注册华为账号并开通华为云**。

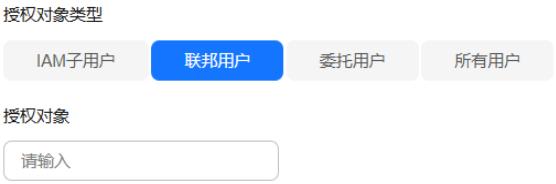
MaaS 委托授权

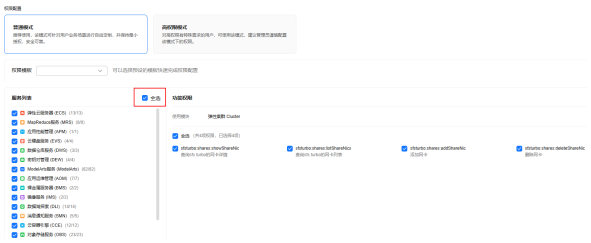
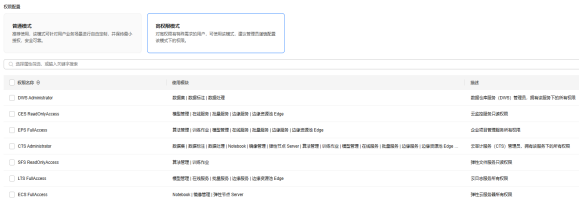
1. 登录**ModelArts管理控制台**，按照版本选择以下操作。
   - 新版本：在左侧导航栏选择“系统管理 > 权限管理”。
   - 旧版本：在左侧导航栏选择“全局配置”。
2. 单击“添加授权”，进入“访问授权”配置页面，根据**表2-2**参数说明进行配置。

   以IAM子用户添加授权为例，参考**表2-2**中“举例”列快速授权。

表 2-2 授权配置参数说明

参数	说明	举例
“授权对象类型”	包括IAM子用户、联邦用户、委托用户和所有用户。 • IAM子用户：由主账号在IAM中创建的用户，是服务的使用人员，具有独立的身份凭证（密码和访问密钥），根据账号授予的权限使用资源。IAM子用户相关介绍请参见**IAM用户介绍**。 • 联邦用户：又称企业虚拟用户。联邦用户相关介绍请参见**联邦身份认证**。 • 委托用户：IAM中创建的一个委托。IAM创建委托相关介绍请参见**创建委托**。 • 所有用户：该选项表示会将委托的权限授权到当前账号下的所有子账号、包括未来创建的子账号，授权范围较大，需谨慎使用。个人用户选择“所有用户”即可。	选择“IAM子用户”。

参数	说明	举例
“授权对象”	“授权对象类型”选择“所有用户”时不涉及此参数。 • IAM子用户：选择指定的IAM子用户，给指定的IAM子用户配置委托授权。 **图 2-2** 选择 IAM 子用户  • 联邦用户：输入联邦用户的用户名或用户ID。 **图 2-3** 选择联邦用户  • 委托用户：选择委托名称。使用账号A创建一个权限委托，在此处将该委托授权给账号B拥有的委托。在使用账号B登录控制台时，可以在控制台右上角的个人账号切换角色到账号A，使用账号A的委托权限。 **图 2-4** 委托用户切换角色  	选择指定的IAM子用户，给指定的IAM子用户配置委托授权。
“委托选择”	• 已有委托：列表中如果已有委托选项，则直接选择一个可用的委托为上述选择的用户授权。单击委托名称查看该委托的权限详情。 • 新增委托：如果没有委托可选，可以在新增委托中创建委托权限。对于首次使用ModelArts的用户，需要新增委托。	选择“新增委托”。

参数	说明	举例
“新增委托 > 委托名称”	系统自动创建委托名称，用户可以手动修改。 ModelArts自动生成委托命名规则： 1. 授权对象类型为“IAM子用户”时，默认委托名称为ma_agency_[授权对象名称]。 2. 授权对象类型为其他对象类型时，默认委托名称为modelarts_agency。 3. 如果上述规则生成的委托名称已存在，则名称新增四位随机码后缀。	-
“新增委托 > 权限配置 > 普通模式”	普通模式下配置权限，该模式可针对用户业务场景进行自由定制，并保持最小授权，安全可靠。 在服务列表右侧勾选“全选”。 **图 2-5** 普通模式 	在服务列表右侧勾选“全选”。
“新增委托 > 权限配置 >高权限模式 ”	高权限模式下配置权限，配置的权限范围较大，适用于有管理员权限需求的用户。 对高权限有特殊需求的用户，可使用该模式，建议管理员谨慎配置该模式下的权限。 **图 2-6** 高权限模式 	-

3. 勾选“我已经详细阅读并同意《ModelArts服务声明》”，单击“创建”，即可完成委托配置。

   登录**ModelArts Studio（MaaS）控制台**，如果页面上没有显示任何关于权限配置的提示信息，说明委托配置已经完成。

   也可前往“权限管理”页面**查看授权的权限列表**，查看已配置权限的权限详情。

查看授权的权限列表

用户可以在“权限管理”页面的授权列表中，查看已经配置的委托授权内容。在操作列单击“查看权限”，可以查看该授权的权限详情。

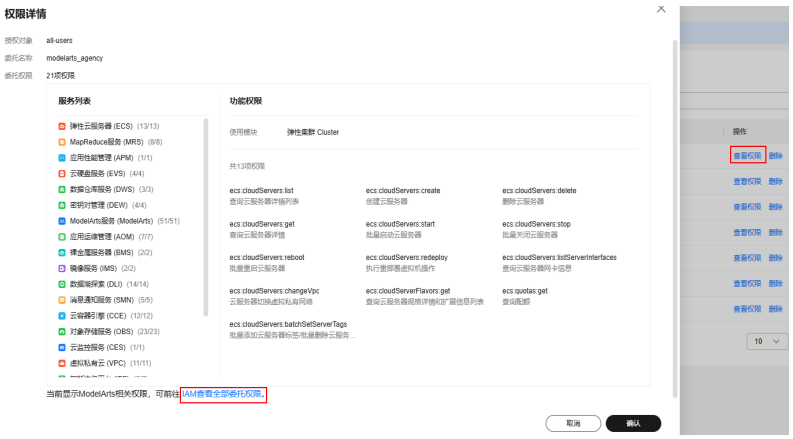
图 2-7 查看权限

图 2-8 普通模式权限详情示例

修改授权的权限范围

1. 在查看授权详情时，如果想要修改授权范围，可以在权限详情页单击“IAM查看全部委托权限”。

图 2-9 去 IAM 修改委托权限



2. 进入**登录IAM管理控制台**的“委托”页面，单击需要修改的委托名称，按需修改该委托的基本信息。“持续时间”可以选择永久、1天，或者自定义天数，例如30天。

图 2-10 手动创建的委托

3. 在“授权记录”页面单击“授权”，勾选要配置的策略，单击“下一步”设置最小授权范围，单击“确定”，完成授权修改。

   设置最小授权范围时，可以选择指定的区域，也可以选择所有区域，即不设置范围。

删除授权

为了更好地管理您的授权，您可以删除某一IAM用户的授权，也可批量清空所有用户的授权。删除操作无法恢复，请谨慎操作。

- **删除某一用户的授权**

  在“权限管理”页面，展示当前账号下为其IAM用户配置的授权列表，针对某一用户，您可以单击“操作”列的“删除”，输入“DELETE”后单击“确认”，可删除此用户的授权。删除生效后，此用户将无法继续使用ModelArts的相关功能。

- **批量清空所有授权**

  在“权限管理”页面，单击授权列表上方的“清空授权”，输入“DELETE”后单击“确认”，可删除当前账号下的所有授权。删除生效后，此账号及其所有IAM子用户将无法继续使用ModelArts的相关功能。

常见问题

- 首次使用ModelArts如何配置授权?

  直接选择“新增委托”中的“普通用户”权限即可，普通用户包括用户使用ModelArts完成AI开发的所有必要功能权限，如数据的访问、训练任务的创建和管理等。一般用户选择此项即可。

- 如何获取访问密钥AK/SK?

  如果在其他功能（例如访问模型服务等）中使用到访问密钥AK/SK认证，获取AK/SK方式请参考**如何获取访问密钥**。

- 如何删除已有委托?

  需要前往统一身份认证服务**登录IAM管理控制台**的委托页面删除。具体操作，请参见**删除或修改委托**。

- 进入**ModelArts管理控制台**的某个页面时，为什么会提示权限不足?

  可能原因是用户委托权限配置不足或模块能力升级，需要更新授权信息。根据界面操作提示追加授权即可。具体操作，请参见**配置用户缺失的ModelArts Studio（MaaS）相关服务权限**。

2.3 配置用户缺失的 ModelArts Studio（MaaS）相关服务权限

在使用ModelArts Studio（MaaS）服务时，您可能会遇到权限配置的问题。如果未正确配置或缺失相关权限，系统将提示您进行授权处理。如果未及时处理，部分功能将无法正常运行，影响您的使用体验。为确保服务的正常运行，请您及时检查并配置缺失的权限，以避免功能异常。

前提条件

已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。具体操作，请参见**注册华为账号并开通华为云**和**实名认证**。

计费说明

授权本身不收费，但使用过程中涉及的数据存储、模型导入以及部署上线等功能依赖OBS、SWR等服务会产生费用，详情请参见**计费概述**。

添加依赖服务授权

由于大模型即服务平台的数据存储、模型导入以及部署上线等功能依赖OBS、SWR等服务，需获取依赖服务授权后才能正常使用相关功能。

如果您未配置依赖服务授权，MaaS控制台顶部会出现获取依赖服务授权提示。

主用户：单击“此处”，跳转至**ModelArts管理控制台**的“权限管理”页面，添加依赖服务权限。具体操作，请参见**MaaS委托授权**。

子用户：联系管理员进行配置。

图 2-11 获取依赖服务授权提示

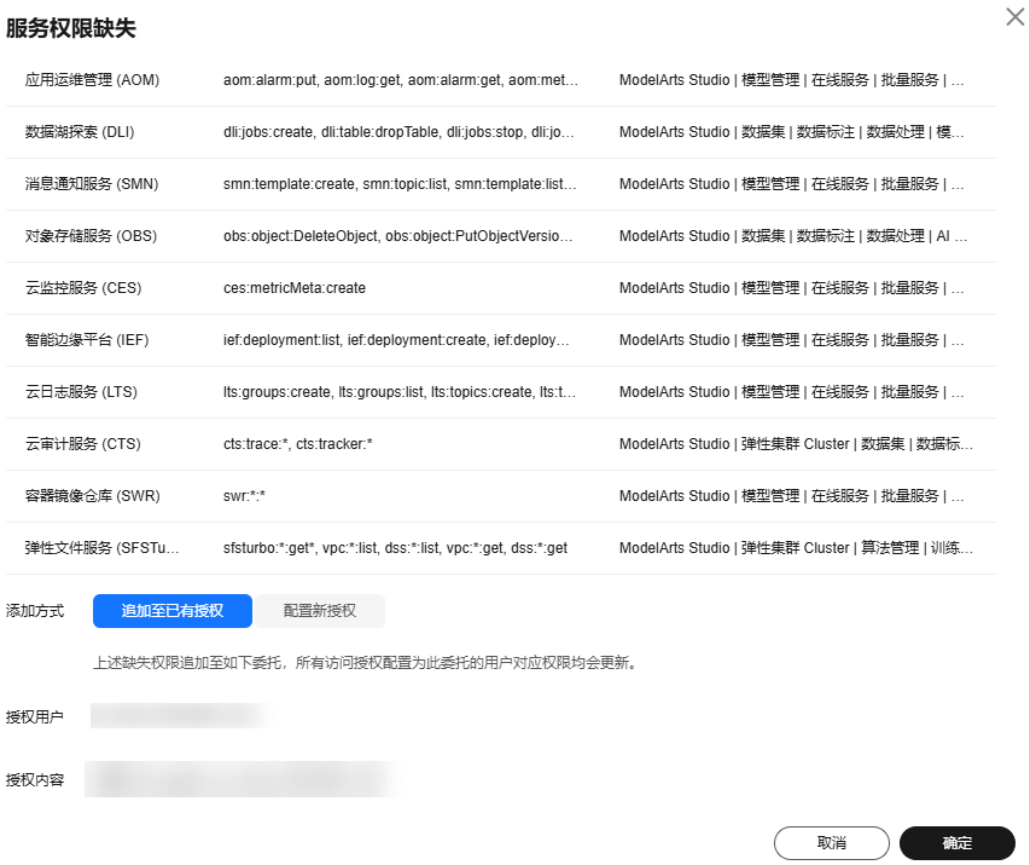
主用户添加缺失的权限

如果您的权限不足，MaaS控制台顶部会出现缺失部分服务权限提示。

图 2-12 缺失部分服务权限提示

您可以在弹出的提示语中单击“此处”，在“服务权限缺失”对话框，按需选择“追加至已有权限”或“配置新授权”，然后单击“确定”。

图 2-13 服务权限缺失



子用户添加缺失的权限

如果您的权限不足，MaaS控制台会出现“访问受限”对话框。请按照以下步骤创建自定义策略、为用户组添加自定义策略、查看缺失的服务权限并联系管理员进行配置。

图 2-14 访问授权对话框

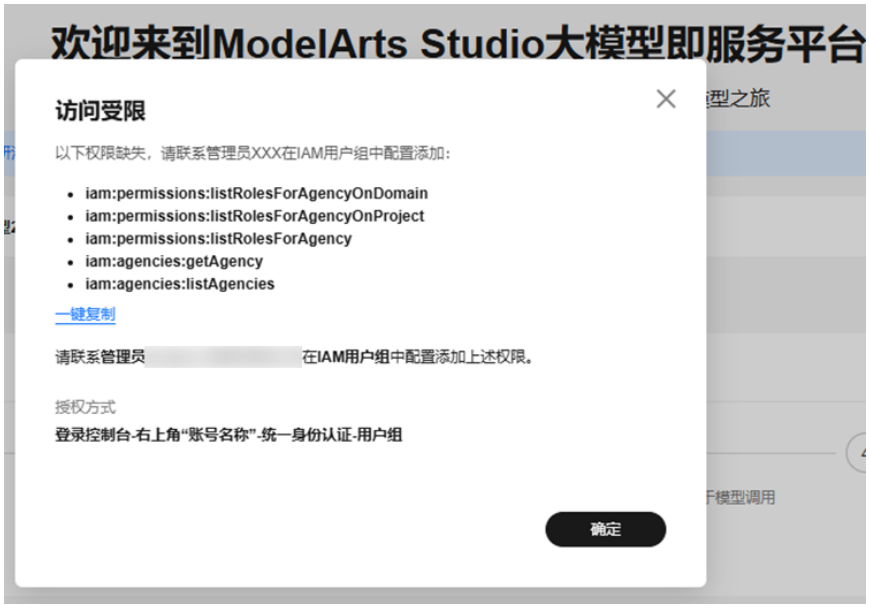
缺失权限的说明请参见**表2-3**。更多信息，请参见**授权项**。

表 2-3 缺失权限说明

权限	说明
iam:permissions:listRolesForAgencyOnDomain	查询全局服务中的委托权限。
iam:permissions:listRolesForAgencyOnProject	查询项目服务中的委托权限。
iam:permissions:listRolesForAgency	查询委托的所有权限。
iam:agencies:getAgency	查询委托详情。
iam:agencies:listAgencies	查询指定条件下的委托列表。

1. 创建自定义策略。
   a. 在“访问受限”对话框，单击“一键复制”，保存权限缺失内容，单击“确定”。

图 2-15 访问受限提示



b. 鼠标悬停至右上角账号处，单击“统一身份认证”。
c. 在**登录IAM管理控制台**左侧导航栏，选择“权限管理 > 权限”。
d. 在“权限”页面右上角，单击“创建自定义策略”。
e. 在“创建自定义策略”页面，配置相关信息，单击“确定”。

图 2-16 创建自定义策略

表 2-4 创建自定义策略参数说明

参数	说明	配置示例
策略名称	自定义策略名称	policykl631g
策略配置方式	单击JSON视图。	JSON视图

参数	说明	配置示例

策略内容	在Statement参数的[]中粘贴**步骤1.a**保存的权限策略。	{
		"Version": "1.1",
		"Statement": [
		{
		"Effect": "Allow",
		"Action": [
		
		"iam:permissions:listRolesForAgencyOnDomain",
		
		"iam:permissions:listRolesForAgencyOnProject",
		
		"iam:permissions:listRolesForAgency",
		"iam:agencies:getAgency",
		"iam:agencies:listAgencies"
		]
		}
		]
		}

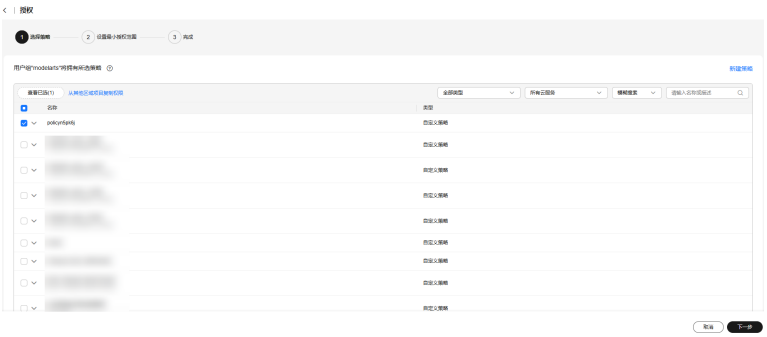
策略描述	自定义策略描述。	-
作用范围	默认为全局级服务。	全局级服务

2. 为用户组添加自定义策略。
   a. 在**登录IAM管理控制台**左侧导航栏，选择“用户组”。
   b. 在“用户组”页面，按需搜索目标用户组名称，在操作列单击“授权”。

图 2-17 授权用户组

   c. 在“授权”页面，选中**步骤1**创建的策略名称，单击“下一步”，按需选择授权范围方案，单击“确定”。

图 2-18 授权页面

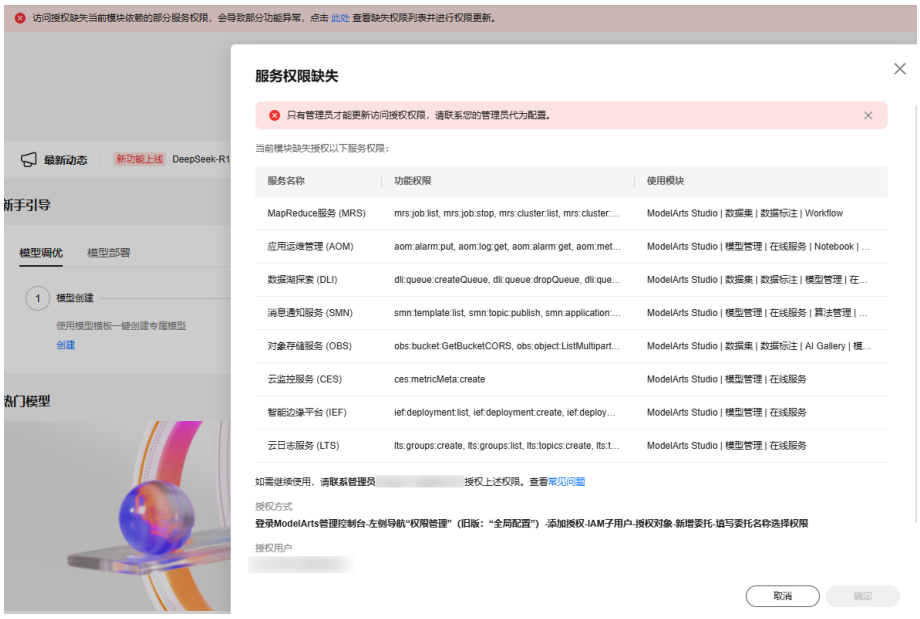


   d. 在“权限生效时间提醒”对话框，仔细阅读相关信息，然后单击“知道了”。

3. 查看并配置缺失的服务权限。

   a. 登录MaaS控制台，单击顶部提示中的“此处”，在“服务权限缺失”对话框，查看缺失的服务权限。

   图 2-19 服务权限缺失

   b. 联系管理员配置缺失的服务权限。具体操作，请参见**MaaS委托授权**。

常见问题

- 如何获取访问密钥AK/SK?

  如果在其他功能（例如访问模型服务等）中使用到访问密钥AK/SK认证，获取AK/SK方式请参考**如何获取访问密钥**。

- 如何删除已有委托?

  需要前往统一身份认证服务**登录IAM管理控制台**的委托页面删除。具体操作，请参见**删除或修改委托**。

3 准备 ModelArts Studio（MaaS）资源

在使用MaaS服务时，需要先完成资源池等准备工作。

准备资源池

在ModelArts Studio大模型即服务平台进行模型部署时，需要选择资源池。MaaS服务支持专属资源池。

专属资源池不与其他用户共享，资源更可控。在使用专属资源池之前，您需要先创建一个专属资源池，然后在AI开发过程中选择此专属资源池。MaaS服务可以使用在ModelArts Standard形态下创建的专属资源池用于模型训推。创建专属资源池的操作指导请参见**创建Standard专属资源池**。

说明

资源池必须和MaaS服务在同一个Region下，否则无法选择到该资源池。

4 ModelArts Studio（MaaS）在线推理服务

4.1 在 ModelArts Studio（MaaS）模型广场查看预置模型

ModelArts Studio大模型即服务平台提供了丰富的开源大模型，在“模型广场”页面可以查看。模型详情页可以查看模型的详细介绍，根据这些信息选择合适的模型进行训练、推理，接入到企业解决方案中。

前提条件

已注册华为账号并开通华为云，详情请见**注册华为账号并开通华为云**。

访问模型广场

1. 登录ModelArts Studio控制台，在顶部导航栏选择目标区域。
2. 在左侧导航栏，单击“模型广场”。
3. 在“模型广场”页面的“模型筛选”区域，按需选择模型系列、模型类型、支持作业和上下文长度进行筛选，或者直接输入模型名称进行搜索。
   关于模型系列的介绍，请参见**模型介绍**。

   表 4-1 模型筛选说明

筛选项	说明
模型系列	支持按照全部、DeepSeek等模型系列进行筛选。
模型类型	支持按照全部、文本生成等模型类型进行筛选。
支持作业	支持按照全部、部署、调优等支持作业进行筛选。
上下文长度	支持按照全部、16K以下、16K等上下文长度进行筛选。

4. 单击目标模型下方的“模型详情”，进入模型详情页查看模型的介绍、基本信息和版本信息。
5. 进入模型详情页，单击右侧“调优”、“压缩”或“部署”等，使用模型进行训推。

当按钮置灰时，表示模型不支持该任务。

模型介绍

表4-2列举了ModelArts Studio大模型即服务平台支持的模型清单。关于模型的详细信息请在“模型详情”页面查看。

表 4-2 模型广场的模型系列介绍

模型系列		模型类型	应用场景	支持语言	支持地域	模型介绍
DeepSeek	DeepSeek-R1	文本生成	对话问答、文本生成推理	中文、英文	香港	深度求索（DeepSeek）自主研发的DeepSeek-R1模型，基于核心技术突破，具备超长上下文理解与高效推理能力，支持多模态交互及API集成，可驱动智能客服、数据分析等场景应用，以行业领先的性价比加速企业智能化升级。
	DeepSeek-V3	文本生成	对话问答、翻译	中文、英文	香港	DeepSeek-V3是一个强大的混合专家 (MoE) 语言模型，开创了一种无辅助损失的负载平衡策略，并设置了多Token预测训练目标以获得更强大的性能。
	DeepSeek-R1-Distill-Qwen-14B	文本生成	对话问答、文本生成推理	中文、英文	香港	通过DeepSeek-R1的输出，蒸馏了Qwen-14B， 使得模型在多项能力上实现了对标OpenAI o1-mini的效果。DeepSeek-R1在数学、代码和推理任务中实现了与OpenAI-o1相当的性能。
	DeepSeek-R1-Distill-Qwen-32B	文本生成	对话问答、文本生成推理	中文、英文	香港	通过DeepSeek-R1的输出，蒸馏了Qwen-32B， 使得模型在多项能力上实现了对标OpenAI o1-mini的效果。DeepSeek-R1在数学、代码和推理任务中实现了与OpenAI-o1相当的性能。

4.2 使用 ModelArts Studio（MaaS）部署模型服务

在ModelArts Studio（MaaS）大模型即服务平台可以将模型广场的预置模型部署为我的服务，便于在其他业务环境中可以调用。

场景描述

从模型广场中选择一个模型进行部署，当模型部署完后会显示在“我的服务”列表中。

计费说明

在MaaS进行模型推理时，会产生计算资源和存储资源等费用。计算资源为运行模型服务的费用。存储资源包括数据存储到OBS的计费。使用消息通知服务会产生相关服务费用。详细计费说明请参考**ModelArts Studio（MaaS）模型推理计费项**。

约束限制

部署模型服务时，ModelArts Studio大模型即服务平台预置了推理的最大输入输出长度，详情如下表所示。

表 4-3 模型默认最大输入输出长度

模型	默认最大输入输出长度
DeepSeek-R1-Distill-Llama-70B-8K DeepSeek-R1-Distill-Qwen-14B-8K DeepSeek-R1-Distill-Qwen-32B-8K	8192
DeepSeek-R1-Distill-Qwen-32B-32K	32768

前提条件

已准备专属资源池，详细请参见**准备ModelArts Studio（MaaS）资源**。

部署模型服务

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”进入服务列表。
3. 在“在线推理”页面的“我的服务”页签，在右上角单击“部署模型服务”进入部署页面，完成创建配置。

表 4-4 部署模型服务参数说明

参数		说明
服务设置	服务名称	自定义部署模型服务的名称。 支持1~64位，以中文、大小写字母开头，只包含中文、大小写字母、数字、中划线、下划线的名称。
	描述	自定义部署模型服务的简介。支持256字符。
模型设置	部署模型	单击“选择模型”，选择“模型广场”下面的模型。
资源设置	资源池类型	仅支持专属资源池。专属资源池需单独创建，不与其他租户共享。
	实例规格	选择实例规格，规格中描述了服务器类型、型号等信息。仅显示模型支持的资源规格。
	实例数	设置服务器个数。
资源设置	流量限制（QPS）	设置待部署模型的流量限制QPS。 单位：次/秒 **说明** 在部署过程中出现错误码“ModelArts.4206”时，表示QPS请求数量达到限制，建议等待限流结束后再重启服务。
更多选项	事件通知	选择是否打开“事件通知”开关。 • 开关关闭（默认关闭）：表示不启用消息通知服务。 • 开关打开：表示订阅消息通知服务，当任务发生特定事件（如任务状态变化或疑似卡死）时会发送通知。此时必须配置“主题名”和“事件”。 – “主题名”：事件通知的主题名称。单击“创建主题”，前往消息通知服务中创建主题。 – “事件”：选择要订阅的事件类型。例如“运行中”、“已终止”、“运行失败”等。 **说明** • 需要为消息通知服务中创建的主题添加订阅，当订阅状态为“已确认”后，方可收到事件通知。订阅主题的详细操作请参见**添加订阅**。 • 使用消息通知服务会产生相关服务费用，详细信息请参见**计费说明**。

参数		说明
	自动停止	当使用付费资源时，可以选择是否打开“自动停止”开关。 • 开关打开：表示启用自动停止功能，此时必须配置自动停止时间，支持设置为“1小时”、“2小时”、“4小时”、6小时或“自定义”。启用该参数并设置时间后，运行时长到期后将会自动终止服务，准备排队等状态不扣除运行时长。 • 开关关闭（默认关闭）：表示服务将一直运行。

4. 参数配置完成后，单击“提交”。
   在“我的服务”列表中，当模型部署服务的“状态”变成“运行中”时，表示模型部署完成。

   说明

   资源池类型为“专属资源池”时，专属资源池的费用已在购买时支付，部署服务不再收费。

5. 模型部署完成后，可以进行API调用。具体操作，请参见**调用ModelArts Studio（MaaS）部署的模型服务**。

查看部署服务信息

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”进入服务列表。
3. 单击服务名称，进入部署模型服务详情页面，可以查看服务信息。
   - “详情”：可以查看服务的基本信息，包括服务、模型、资源等设置信息。
   - “资源监控”：可以查看服务的算力利用率、显存利用率和资源监控信息。

     表 4-5 资源监控参数说明

参数	说明
算力使用率	服务的算力使用情况。当请求率较低时，使用率会显示为0。
显存利用率	服务的显存使用情况。

   - “事件”：可以查看服务的事件信息。事件保存周期为1个月，1个月后自动清理数据。
   - “日志”：可以搜索和查看服务日志。

相关操作

- 在AI开发过程中，需要对服务的生命周期进行管理，对已部署的模型服务进行优化、升级模型服务等，详细请参考**在ModelArts Studio（MaaS）管理我的服务**。

- API调用请参考**调用ModelArts Studio（MaaS）部署的模型服务**。

4.3 在 ModelArts Studio（MaaS）管理我的服务

4.3.1 在 ModelArts Studio（MaaS）启动/停止/定时启停/删除服务

停止/启动部署服务

只有服务处在排队中、启动中、运行中、部署中、告警状态，才可执行停止操作；只有服务处在部署失败、已停止状态，才可执行启动操作。

- 停止部署服务
  a. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
  b. 在左侧导航栏，选择“在线推理”。
  c. 在“在线推理”页面的“我的服务”页签，在目标服务右侧，单击操作列的“停止”。
  d. 在“停止服务”对话框，单击“确定”。
- 启动部署服务
  a. 在“在线推理”页面的“我的服务”页签，在目标服务右侧，单击操作列的“启动”。
  b. 在“启动服务”对话框，仔细阅读提示信息，单击“确定”。
     服务状态为运行中时会产生费用。

删除部署服务

删除操作无法恢复，请谨慎操作。

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”进入服务列表。
3. 选择“我的服务”页签。
4. 选择待删除的服务，单击操作列的“更多 > 删除”，在弹窗中输入“DELETE”，单击“确定”，删除服务。

4.3.2 在 ModelArts Studio（MaaS）扩缩容模型服务实例数

在使用大型模型进行推理时，其业务需求会呈现出明显的峰谷波动。因此，模型服务必须具备灵活的扩缩容能力，以适应不同时间段内的用户负载变化，确保服务的高可用性和资源的高效利用。

ModelArts Studio大模型即服务平台支持手动扩缩容模型服务的实例数，该操作不会影响部署服务的正常运行。

前提条件

已经在ModelArts Studio（MaaS）**部署模型**。

约束限制

仅当模型服务处于这几个状态下才能扩缩容实例数：运行中、告警。

计费说明

- 扩容模型服务实例数后，在调用MaaS预置服务时，将根据实际使用的Tokens数量进行计费，详情请见**计费说明**。
- 扩容模型服务实例数后，在MaaS进行模型推理时，会产生计算资源和存储资源的累计值计费。计算资源为运行模型服务的费用。存储资源包括数据存储到OBS的计费。具体内容如请参见**ModelArts Studio（MaaS）模型推理计费项**。

扩缩容实例数

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”。
3. 在“在线推理”页面的“我的服务”页签，在目标模型服务右侧，单击操作列的“更多 > 扩缩容”，进入扩缩容页面。
4. 在“扩缩容”页面，根据业务需要增删模型服务的实例数，配置完成后，单击“确认”提交扩缩容任务。
5. 在“扩缩容服务”对话框，单击“确定”。

   在“我的服务”页签，单击服务名称，进入服务详情页，可以查看修改后的实例数是否生效。

4.3.3 在 ModelArts Studio（MaaS）修改模型服务 QPS

流量限制QPS是评估模型服务处理能力的关键指标，它指示系统在高并发场景下每秒能处理的请求量。这一指标直接关系到模型的响应速度和处理效率。不当的QPS配置可能导致用户等待时间延长，影响满意度。因此，能够灵活调整模型的QPS对于保障服务性能、优化用户体验、维持业务流畅及控制成本至关重要。

ModelArts Studio大模型即服务平台支持手动修改模型服务的实例流量限制QPS，该操作不会影响部署服务的正常运行。

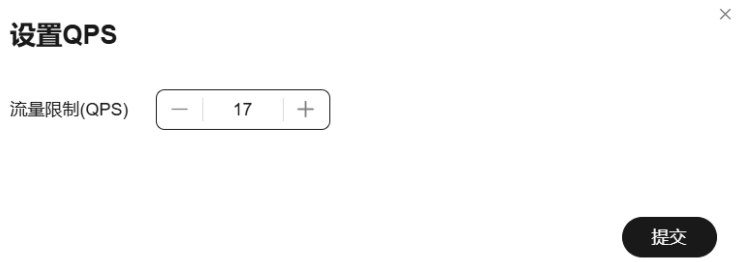
约束限制

仅当模型服务处于这几个状态下才能修改QPS：运行中、告警。

修改 QPS

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”。
3. 在“在线推理”页面的“我的服务”页签，在目标模型服务右侧，单击操作列的“更多 > 设置QPS”，在弹窗中修改数值，单击“提交”启动修改任务。

图 4-1 修改 QPS



在“我的服务”页签，单击服务名称，进入服务详情页，可以查看修改后的QPS是否生效。

4.3.4 在 ModelArts Studio（MaaS）升级模型服务

在AI开发过程中，服务升级包括对已部署的模型服务进行优化，以提高性能、增加功能、修复缺陷，并适应新的业务需求。更新模型版本作为服务升级的一部分，涉及用新训练的模型版本替换原来的模型，以提高预测的准确性和模型的环境适应性。

前提条件

已经在ModelArts Studio（MaaS）**部署模型**。

约束限制

仅当模型服务处于这几个状态下才能进行服务升级：运行中、告警。

服务升级

说明

- 服务升级不可逆。服务升级过程中，原部署服务将正常运行。
- 升级期间、升级完成后，仍然会按照该服务原计费方式产生费用。

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”。
3. 在“在线推理”页面的“我的服务”页签。
4. 在目标模型服务右侧，单击操作列的“更多 > 服务升级”。
5. 在“服务升级”对话框，选择需要升级的版本，然后单击“确认”。

4.4 调用 ModelArts Studio（MaaS）部署的模型服务

在ModelArts Studio大模型即服务平台部署成功的模型服务支持在其他业务环境中调用。本文以我的服务为例，调用部署的模型服务。您也可以调用预置服务-免费服务、预置服务-商用服务或预置服务-自定义接入点。

场景描述

在企业AI应用开发过程中，开发人员通常需要将训练好的模型部署到实际业务环境中。然而，传统方法需要手动配置环境、处理依赖关系、编写部署脚本，整个过程耗时且容易出错，且存在环境复杂、迁移困难、维护成本高、版本更新麻烦等问题。

ModelArts Studio（MaaS）大模型即服务平台提供了一站式解决方案，提供统一的API接口方便业务系统调用，并提供监控和日志功能便于运维管理。

计费说明

在调用模型推理服务的过程中，输入内容首先会被分词（tokenize），转换为模型可识别的Token。在调用MaaS预置服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见**计费说明**。

前提条件

在“在线推理”页面的“我的服务”页签，服务列表存在运行中、更新中或升级中的模型服务。具体操作，请参见**使用ModelArts Studio（MaaS）部署模型服务**。

步骤一：获取 API Key

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。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。更多信息，请参见**在ModelArts Studio（MaaS）管理API Key**。

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，单击“API Key管理”。
3. 在“API Key管理”页面，单击“创建API Key”，填写标签和描述信息后，单击“确定”。

   标签和描述信息在创建完成后，不支持修改。

   表 4-6 创建 API Key 参数说明

参数	说明
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。
描述	自定义API Key的描述，长度范围为1~100个字符。

4. 在“您的密钥”对话框，复制密钥并保存至安全位置。
5. 保存完毕后，单击“关闭”。

   单击“关闭”后将无法再次查看密钥。

步骤二：调用 MaaS 模型服务进行预测

1. 在**ModelArts Studio（MaaS）控制台**左侧导航栏，选择“在线推理”。
2. 在“在线推理”页面的“我的服务”页签，在目标服务右侧，单击操作列的“更多 > 调用说明”。
3. 在“调用说明”页面，选择接口类型，复制调用示例，修改接口信息和API Key后用于业务环境调用模型服务API。

   Rest API、OpenAI SDK的示例代码如下。

   - 使用普通requests包调用。

```
import requests
import json

if __name__ == '__main__':
    url = "https:/example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/v1/chat/
completions"
    api_key = "<your_apiKey>"  # 把<your_apiKey>替换成已获取的API Key。

    # Send request.
    headers = {
        'Content-Type': 'application/json',
        'Authorization': f'Bearer {api_key}'
    }
    data = {
        "model": "******",  # 调用时的模型名称。
        "max_tokens": 1024,  # 最大输出token数。
        "messages": [
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": "hello"}
        ],
        # 是否开启流式推理，默认为False,表示不开启流式推理。
        "stream": False,
        # 在流式输出时是否展示使用的token数目。只有当stream为True时该参数才会生效。
        # "stream_options": {"include_usage": True},
        # 控制采样随机性的浮点数，值较低时模型更具确定性，值较高时模型更具创造性。"0"表示贪
婪取样。默认为0.6。
        "temperature": 0.6
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)
    # Print result.
    print(response.status_code)
    print(response.text)
```

- 使用OpenAI SDK调用。

```
from openai import OpenAI

if __name__ == '__main__':
    base_url = "https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06******/v1"
    api_key = "<your_apiKey>"  # 把<your_apiKey>替换成已获取的API Key。

    client = OpenAI(api_key=api_key, base_url=base_url)

    response = client.chat.completions.create(
        model="******",
        messages=[
            {"role": "system", "content": "You are a helpful assistant"},
            {"role": "user", "content": "Hello"},
        ],
        max_tokens=1024,
        temperature=0.6,
        stream=False
    )
    # Print result.
        print(response.choices[0].message.content)
```

模型服务的API与vLLM相同，**表4-7**仅介绍关键参数，详细参数解释请参见vLLM官网。使用昇腾云909镜像的模型，开启流式输出时，需要新增stream_options参数，值为{"include_usage":true}，才会打印token数。

表 4-7 请求参数说明

参数	是否必选	默认值	参数类型	描述
url	是	无	Str	调用时的API地址。假设URL为https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/{endpoint}， 其中{endpoint}仅支持如下接口，详细介绍请参见**接口调用说明**。 • /v1/chat/completions • /v1/models
model	是	无	Str	调用时的模型名称。 在ModelArts Studio大模型即服务平台的“在线推理”页面，选择调用的模型服务，单击操作列的“更多 > 调用”，在调用页面可以获取“模型名称”。
messages	是	-	Array	请求输入的问题。
stream_options	否	无	Object	该参数用于配置在流式输出时是否展示使用的token数目。只有当stream为True的时候该参数才会激活生效。如果您需要统计流式输出模式下的token数目，可将该参数配置为stream_options={"include_usage":True}。
max_tokens	否	16	Int	每个输出序列要生成的最大Tokens数量。
top_k	否	-1	Int	控制要考虑的前几个Tokens的数量的整数。设置为“-1”表示考虑所有Tokens。 适当降低该值可以减少采样时间。
top_p	否	1.0	Float	控制要考虑的前几个Tokens的累积概率的浮点数。 取值范围：0~1 设置为“1”表示考虑所有Tokens。
temperature	否	0.6	Float	控制采样的随机性的浮点数。较低的值使模型更加确定性，较高的值使模型更加随机。“0”表示贪婪采样。
stop	否	None	None/Str/List	用于停止生成的字符串列表。返回的输出将不包含停止字符串。 例如，设置为["你"，"好"]时，在生成文本过程中，遇到“你”或者“好”将停止文本生成。
stream	否	False	Bool	是否开启流式推理。默认为“False”，表示不开启流式推理。

参数	是否必选	默认值	参数类型	描述
n	否	1	Int	返回多条正常结果。 ● 不使用beam_search场景下，n取值建议为1≤n≤10。如果n>1时，必须确保不使用greedy_sample采样，也就是top_k > 1，temperature > 0。 ● 使用beam_search场景下，n取值建议为1<n≤10。如果n=1，会导致推理请求失败。 **说明** n建议取值不超过10，n值过大会导致性能劣化，显存不足时，推理请求会失败。
use_beam_search	否	False	Bool	是否使用beam_search替换采样。 使用该参数时，如下参数必须按要求设置。 ● n：大于1 ● top_p：1.0 ● top_k：-1 ● temperature：0.0
presence_penalty	否	0.0	Float	presence_penalty表示会根据当前生成的文本中新出现的词语进行奖惩。取值范围[-2.0,2.0]。
frequency_penalty	否	0.0	Float	frequency_penalty会根据当前生成的文本中各个词语的出现频率进行奖惩。取值范围[-2.0,2.0]。
length_penalty	否	1.0	Float	length_penalty表示在beam search过程中，对于较长的序列，模型会给予较大的惩罚。 使用该参数时，必须添加如下三个参数，且必须按要求设置。 ● top_k：-1 ● use_beam_search：true ● best_of：大于1
ignore_eos	否	False	Bool	ignore_eos表示是否忽略EOS并且继续生成Token。

- 普通requests包、OpenAI SDK的返回示例如下所示：

```
{
   "id": "cmpl-29f7a172056541449eb1f9d31c*****",
   "object": "chat.completion",
   "created": 17231*****,
   "model": "******",
```

```
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "你好！很高兴能为你提供帮助。有什么问题我可以回答或帮你解决吗？"
      },
      "logprobs": null,
      "finish_reason": "stop",
      "stop_reason": null
    }
  ],
  "usage": {
    "prompt_tokens": 20,
    "total_tokens": 38,
    "completion_tokens": 18
  }
}
```

- 思维链模型的返回示例如下所示：

```
messages = [{"role": "user", "content": "9.11 and 9.8, which is greater?"}]
response = client.chat.completions.create(model=model, messages=messages)
reasoning_content = response.choices[0].message.reasoning_content
content = response.choices[0].message.content
print("reasoning_content:", reasoning_content)
print("content:", content)
```

表 4-8 返回参数说明

参数	参数类型	描述
id	Str	请求ID。
object	Str	请求任务。
created	Int	请求生成的时间戳。
model	Str	调用的模型名。
choices	Array	模型生成内容。
usage	Object	请求输入长度、输出长度和总长度。 ● prompt_tokens：输入Tokens数。 ● completion_tokens：输出Tokens数。 ● total_tokens：总Tokens数。 总Tokens数 = 输入Tokens数 + 输出Tokens数
reasoning_content	Str	当模型支持思维链时，模型的思考内容。对于支持思维链的模型，开启流式输出时，会首先在reasoning_content字段输出思考内容，然后在content中输出回答内容。
content	Str	模型的回答内容。

当调用失败时，可以根据错误码调整脚本或运行环境。

表 4-9 常见错误码

错误码	错误内容	说明
400	Bad Request	请求包含语法错误。
403	Forbidden	服务器拒绝执行。
404	Not Found	服务器找不到请求的网页。
500	Internal Server Error	服务内部错误。

接口调用说明

假设API地址为https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/{endpoint}，其中{endpoint}仅支持如下接口：

- /v1/chat/completions
- /v1/models

注意：

- /v1/models使用GET方法不需要请求体，而/v1/chat/completions需要POST请求方式和对应的JSON请求体。
- 通用请求头为Authorization: Bearer YOUR_API_KEY，对于POST请求，还需包含Content-Type: application/json。

表 4-10 接口说明

类型/接口	/v1/models	/v1/chat/completions
请求方法	GET	POST
用途	获取当前支持的模型列表。	用于聊天对话型生成调用。
请求体说明	无需请求体，仅需通过请求头传入认证信息。	• model：使用的模型标识。 • messages：对话消息数组，每条消息需要包含role（如 "user" 或 "assistant"）和content。 • 其他可选参数：例如temperature（生成温度）、max_tokens等，用于控制生成结果的多样性和长度。

类型/接口	/v1/models	/v1/chat/completions

请求示例	GET https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/v1/models HTTP/1.1
	Authorization: Bearer YOUR_API_KEY
		POST https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/v1/chat/completions HTTP/1.1
		Content-Type: application/json
		Authorization: Bearer YOUR_API_KEY
		
		{
		"model": "******",
		"messages": [
		{"role": "user", "content": "Hello, how are you?"}
		],
		"temperature": 0.7
		}

响应示例	{
	"data": [
	{
	"id": "******",
	"description": "最新一代大模型"
	},
	{
	"id": "******",
	"description": "性价比较高的替代方案"
	}
	]
	}
		{
		"id": "******",
		"object": "chat.completion",
		"choices": [
		{
		"index": 0,
		"message": {"role": "assistant", "content": "I'm doing well, thank you! How can I help you today?"}
		}
		],
		"usage": {
		"prompt_tokens": 15,
		"completion_tokens": 25,
		"total_tokens": 40
		}
		}

常见问题

在ModelArts Studio（MaaS）创建API Key后需要等待多久才能生效?

MaaS API Key在创建后不会立即生效，通常需要等待几分钟才能生效。

相关文档

- ModelArts Studio（MaaS）API调用规范
- 使用ModelArts Studio（MaaS）创建多轮对话

4.5 ModelArts Studio（MaaS）API 调用规范

4.5.1 对话 Chat/POST

MaaS平台提供功能丰富的在线推理能力，支持用户选取模型在专属实例上进行自部署。本文介绍对话Chat相关API的调用规范。

接口信息

表 4-11 接口信息

名称	说明	取值
API地址	调用模型服务的API地址。	https://api.modelarts-maas.com/v1/chat/completions
model参数	model参数调用名称。	在“调用说明”页面获取。更多信息，请参见**调用ModelArts Studio（MaaS）部署的模型服务**。

创建聊天对话请求

- 鉴权说明

  MaaS推理服务支持使用API Key鉴权，鉴权头采用如下格式：

  ```
  'Authorization': 'Bearer 该服务所在Region的ApiKey'
  ```

- 请求参数和响应参数说明如下：

  表 4-12 请求参数说明

参数名称	是否必选	默认值	参数类型	说明
model	是	无	Str	调用时的模型名称。取值请参见上方**表4-11**。
messages	是	-	Array	请求输入的问题，其中role为角色，content为对话内容。示例如下： `"messages": [ {"role": "system","content": "你是一个乐于助人的AI助手"}, {"role": "user","content": "9.11和9.8哪个大？"} ]` 更多信息，请参见**表4-13**。
stream_options	否	无	Object	该参数用于配置在流式输出时是否展示使用的Token数目。只有当“stream”为“True”时，该参数才会激活生效。如果您需要统计流式输出模式下的Token数目，可将该参数配置为stream_options={"include_usage":True}。更多信息，请参见**表4-14**。
max_tokens	否	无	Int	模型回复的最大长度。
top_k	否	-1	Int	控制要考虑的前几个Tokens的数量的整数。设置为“-1”表示考虑所有Tokens。 适当降低该值可以减少采样时间。

参数名称	是否必选	默认值	参数类型	说明
top_p	否	1.0	Float	控制要考虑的前几个Tokens的累积概率的浮点数。 取值范围：0~1 设置为“1”表示考虑所有Tokens。
temperature	否	1.0	Float	控制采样的随机性的浮点数。较低的值使模型更加确定性，较高的值使模型更加随机。“0”表示贪婪采样。
stop	否	None	None/Str/List	用于停止生成的字符串列表。返回的输出将不包含停止字符串。 例如，设置为["你"，"好"]时，在生成文本过程中，遇到“你”或者“好”将停止文本生成。
stream	否	False	Bool	是否开启流式推理。默认为“False”，表示不开启流式推理。
n	否	1	Int	返回多条正常结果。 ● 不使用beam_search场景下，n取值建议为1≤n≤10。如果n>1时，必须确保不使用greedy_sample采样，即top_k > 1，temperature > 0。 ● 使用beam_search场景下，n取值建议为1<n≤10。如果n=1，会导致推理请求失败。 **说明** ● n建议取值不超过10，n值过大会导致性能劣化，显存不足时，推理请求会失败。 ● DeepSeek-R1和DeepSeek-V3暂不支持设置n的值大于1。
use_beam_search	否	False	Bool	是否使用beam_search替换采样。 使用该参数时，如下参数必须按要求设置。 ● n：大于1 ● top_p：1.0 ● top_k：-1 ● temperature：0.0 **说明** DeepSeek-R1和DeepSeek-V3暂不支持设置n的值大于1。
presence_penalty	否	0.0	Float	表示会根据当前生成的文本中新出现的词语进行奖惩。取值范围[-2.0,2.0]。

参数名称	是否必选	默认值	参数类型	说明
frequency_penalty	否	0.0	Float	会根据当前生成的文本中各个词语的出现频率进行奖惩。取值范围[-2.0,2.0]。
length_penalty	否	1.0	Float	表示在beam search过程中，对于较长的序列，模型会给予较大的惩罚。 使用该参数时，必须添加如下三个参数，且必须按要求设置。 ● top_k：-1 ● use_beam_search：true ● best_of：大于1 **说明** DeepSeek-R1和DeepSeek-V3暂不支持设置length_penalty。
ignore_eos	否	False	Bool	表示是否忽略EOS并且继续生成Token。

表 4-13 请求参数 messages 说明

参数名称	是否必选	默认值	参数类型	说明
role	是	无	Str	当前支持的角色如下： ● user: 用户 ● assistant: 对话助手（模型） ● system：助手人设
content	是	无	Str	● 当role为system时：给AI模型设定的人设。 {"role": "system","content": "你是一个乐于助人的AI助手"} ● 当role为user时：用户输入的问题。 {"role": "user","content": "9.11和9.8哪个大？"} ● 当role为assistant时：AI模型输出的答复内容。 {"role": "assistant","content": "9.11大于9.8"}

表 4-14 请求参数 stream_options 说明

参数名称	是否必选	默认值	参数类型	说明
include_usage	否	true	Bool	流式响应是否输出Token用量信息： ● true：是，在每一个chunk会输出一个usage字段，显示累计消耗的Token统计信息。 ● false：否，不显示消耗的Token统计信息。

表 4-15 响应参数说明

参数名称	类型	说明
id	Str	该次请求的唯一标识。
object	Str	类型-chat.completion：多轮对话返回。
created	Int	时间戳。
model	Str	调用时的模型名称。
choices	Array	模型答复的内容，包含index和message两个参数，message中： ● content为模型的正式答复内容。 ● reasoning content为模型的深度思考内容（仅限DeepSeek系列模型）。
usage	Object	请求消耗的Token统计信息： ● 非流式请求默认返回。 ● 流式请求默认返回，在每一个chunk会输出一个usage字段，显示消耗的Token统计信息。 参数说明： ● prompt tokens：输入Token数量。 ● completion tokens: 输出Token数量。 ● total tokens：总Token数量。
prompt_logprobs	Float	对数概率。用户可以借此衡量模型对其输出内容的置信度，或者探索模型给出的其他选项。

请求示例

- Python请求示例：

```
# coding=utf-8

import requests
```

```
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/chat/completions"

    # Send request.
    headers = {
        'Content-Type': 'application/json',
        'Authorization': 'Bearer yourApiKey' # 把yourApiKey替换成真实的API Key。
    }
    data = {
        "model": "DeepSeek-R1",
        "max_tokens": 20,
        "messages": [
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": "你好"}
        ],
        # 是否开启流式推理, 默认为False, 表示不开启流式推理。
        "stream": False,
        # 在流式输出时是否展示使用的Token数目。只有当stream为True时改参数才会生效。
        # "stream_options": { "include_usage": True },
        # 控制采样随机性的浮点数，值较低时模型更具确定性，值较高时模型更具创造性。"0"表示贪婪取样。默认为1.0。
        "temperature": 1.0
    }
    resp = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

    # Print result.
    print(resp.status_code)
    print(resp.text)
```

- OpenAI SDK请求示例：

```
from openai import OpenAI

client = OpenAI(
  base_url = "https://api.modelarts-maas.com/v1",
  api_key = "yourApiKey"
)

completion = client.chat.completions.create(
  model="DeepSeek-R1",
  messages=[{"role":"user","content":"9.11和9.8哪个大"}],
  temperature=0.6,
  top_p=0.7,
  max_tokens=4096,
  stream=True
)

for chunk in completion:
  if chunk.choices[0].delta.content is not None:
    print(chunk.choices[0].delta.content, end="")
```

响应示例

```
{
   "id":"chat-71406e38b0d248c9b284709f8435****",
   "object":"chat.completion",
   "created":1740809549,
   "model":"DeepSeek-R1",
   "choices":[
      {
         "index":0,
         "message":{
            "role":"assistant",
            "content":"\n\n比较两个小数9.11和9.8的大小：\n\n1. **比较整数部分**：两者的整数部分都是9，相等。\n2. **比较十分位**： \n - 9.11的十分位是 **1** \n - 9.8可以看作是9.80，其十分位是 **8** \n - **8 > 1**，因此在这一位上已经能分出大小。\n\n**结论**： \n**9.8 > 9.11** \n（小数值比较时需对齐位数，直接比较对应数位的数字即可）",
            "reasoning_content":"嗯，我现在需要比较9.11和9.8哪个大。首先，我得回忆一下小数比较的方法。记得比较小数的时候，应该先比较整数部分，如果整数部分相同，再依次比较小数部分的十分位、百分位，直到
```


分出大小。\n\n这两个数的整数部分都是9，所以整数部分相同。接下来比较十分位。9.11的十分位是1，而9.8的十分位是8。这里可能会有问题，因为有时候可能会有同学直接把9.8当作9.80来看，或者考虑十分位上的数字大小对比。\n\n现在比较的话，9.8的十分位是8，而9.11的十分位是1，明显8比1大，所以这时候是不是应该认为9.8比9.11大呢？\n\n不过要注意到，有的同学可能误以为小数位数越多数值越大，但实际并非如此，比如0.9比0.8999要大，所以位数多不一定数值大。\n\n另外，可以把两个数的小数部分统一成相同的位数来比较。例如，9.8可以写成9.80，这样十分位是8，百分位是0，而9.11的十分位是1，百分位是1。那么在十分位的时候，8比1大，所以9.80（即9.8）大于9.11。\n\n因此，最终结论是9.8比9.11大。\n",
"tool_calls":[]
},
"logprobs":null,
"finish_reason":"stop",
"stop_reason":null
}
],
"usage":{
"prompt_tokens":21,
"total_tokens":437,
"completion_tokens":416
},
"prompt_logprobs":null
}

4.5.2 获取模型列表 Models/GET

本文介绍如何通过Models接口查询模型列表的API调用规范。

接口信息

表 4-16 接口信息

名称	说明	取值
API地址	调用模型服务的API地址。	https://api.modelarts-maas.com/v1/models

创建请求

- 鉴权说明

  MaaS推理服务支持使用API Key鉴权，鉴权头采用如下格式：

  ```
  'Authorization': 'Bearer 该服务所在Region的ApiKey'
  ```

- 响应参数说明

  表 4-17 响应参数

名称	类型	说明
object	string	类型-list：列出查询到的信息。

名称	类型	说明
data	Array	当前模型服务的模型信息，主要参数如下： • id：调用接口创建请求时使用的模型ID。 • object：模型类型。 • created：创建时间戳。

请求示例

```
import requests

url = "https://api.modelarts-maas.com/v1/models"

headers = {"Authorization": "Bearer yourApiKey"}

response = requests.request("GET", url, headers=headers)

print(response.text)
```

响应示例

```
{
  "object": "list",
  "data": [
    {
      "id": "DeepSeek-R1",
      "object": "model",
      "created": 0,
      "owned_by": ""
    },
    {
      "id": "DeepSeek-V3",
      "object": "model",
      "created": 0,
      "owned_by": ""
    }
  ]
}
```

4.5.3 错误码

在调用MaaS部署的模型服务时，可能出现的错误码及相关信息如下。

表 4-18 错误码

HTTP状态码	错误码	错误信息	说明
400	ModelArts.81001	Invalid request body.	解析body体失败，如JSON格式化失败、model参数为空。

HTTP状态码	错误码	错误信息	说明
400	ModelArts.81002	Failed to get the authorization header.	请求头中Authorization为空，或者Authorization格式不是Bearer开头。
401	ModelArts.81003	Invalid authorization header.	API Key解析失败。
401	ModelArts.81004	Invalid request because you do not have access to it.	未开通预置服务。
401	ModelArts.81005	The free quota has been used up.	免费额度已用完。
401	ModelArts.81006	The resource is frozen.	常驻模型已冻结。
401	ModelArts.81109	No permission query task %s	没有查询该视频生成任务的权限。
403	ModelArts.81011	May contain senstive.....	输入或者非流式输出风控。
404	ModelArts.81009	Invalid model.	请求体中的model参数传入的模型不存在。
404	ModelArts.81108	Task %s does not exist	任务不存在。
429	ModelArts.81101	Too many requests, exceeded rate limit is {rpm} times per minute.	RPM流控校验失败。
429	ModelArts.81103	Too many requests. exceeded rate limit is %s tokens per minute.	TPM流控校验失败。
403	ModelArts.81109	No permission query task %s	无权限查询此任务。
5XX	APIG.0203	"error_msg":"Backend timeout",error_code:APIG.0203	请求的服务响应超时。

HTTP状态码	错误码	错误信息	说明
400	"object": "error"	"object": "error", "message": "[{'type': 'missing', 'loc': ('body', 'model'), 'msg': 'Field required', 'input': {'max_tokens': 20, 'messages': [{'role': 'system', 'content': 'You are a helpful assistant.'}, {'role': 'user', 'content': '你好'}], 'stream': False, 'temperature': 1.0}}]", "type": "BadRequestError", "param": null, "code": 400	请求体中缺失必填参数。
400	"object": "error"	"object": "error", "message": "[{'type': 'extra_forbidden', 'loc': ('body', 'test'), 'msg': 'Extra inputs are not permitted', 'input': 15}]", "type": "BadRequestError", "param": null, "code": 400	请求体中包含不支持的额外请求参数。
400	"object": "error"	"object": "error", "message": "[{'type': 'json_invalid', 'loc': ('body', 273), 'msg': 'JSON decode error', 'input': {}, 'ctx': {'error': \"Expecting ',' delimiter\"}}]", "type": "BadRequestError", "param": null, "code": 400	请求体json格式错误。
400	"object": "error"	"object": "error", "message": "[{'type': 'missing', 'loc': ('body',), 'msg': 'Field required', 'input': None}]", "type": "BadRequestError", "param": null, "code": 400	无请求体。

HTTP状态码	错误码	错误信息	说明
400	"object": "error"	"object": "error", "message": "This model's maximum context length is 4096 tokens. However, you requested 8242 tokens (20 in the messages, 8222 in the completion). Please reduce the length of the messages or completion.", "type": "BadRequestError", "param": null, "code": 400	max_tokens设置超出模型支持的上限。
404	"object": "error"	"object": "error", "message": "The model `DeepSeek-R1` does not exist.", "type": "NotFoundError", "param": null, "code": 404	请求体中model参数填写错误。
404	APIG.0101	"error_msg": "The API does not exist or has not been published in the environment", "error_code": "APIG.0101", "request_id": "d0ddda0fcdd0cc23a1588fafe426****"	请求接口地址错误或不存在。
405	-	"detail":"Method Not Allowed"	采用了错误的请求方式。
429	APIG.0308	"error_msg": "The throttling threshold has been reached: policy ip over ratelimit,limit:5,time:1 minute"	达到APIG流量控制上限。

4.6 使用 ModelArts Studio（MaaS）创建多轮对话

本文介绍如何使用MaaS Chat API进行多轮对话。

MaaS服务端不会记录用户请求的上下文，用户每次发起请求时，需要将之前所有对话历史拼接好后，传递给Chat API。下文以一个Python代码为例进行说明，请您根据实际情况进行修改。

以下为Python的上下文拼接和请求示例代码：

```
from openai import OpenAI
client = OpenAI(api_key="MaaS API Key", base_url="https://xxxxxxxxxxxxxxxx")
```

```
# 首轮对话
messages = [{"role": "user", "content": "9.11和9.8哪个大？"}]
response = client.chat.completions.create(
    model="DeepSeek-R1",
    messages=messages
)
messages.append(response.choices[0].message)
print(f"Messages Round 1: {messages}")
# 第二轮对话
messages.append({"role": "user", "content": "他们相加等于多少"})
response = client.chat.completions.create(
    model="DeepSeek-R1",
    messages=messages
)
messages.append(response.choices[0].message)
print(f"Messages Round 2: {messages}")
```

首轮对话时，请求体中的messages为：

```
[
    {"role": "user", "content": "9.11和9.8哪个大？"}
]
```

在第二轮对话时，请求体中的messages构建步骤如下：

1. 将首轮对话中模型（role的值为"assistant"）的输出内容添加到messages结尾。
2. 将新的用户问题添加到messages结尾。
3. 最终传递给Chat API的请求体中的messages为：

   ```
   [
       {"role": "user", "content": "9.11和9.8哪个大？"},
       {"role": "assistant", "content": "9.8更大"},
       {"role": "user", "content": "他们相加等于多少"}
   ]
   ```

5 ModelArts Studio（MaaS）管理与统计

5.1 在 ModelArts Studio（MaaS）管理 API Key

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，保障服务访问的安全性和合法性。本文介绍如何创建和删除API Key。

场景描述

当用户使用MaaS部署的模型服务进行数据请求、模型推理等操作时，系统通过验证API Key来确认用户的身份与访问权限，只有具备有效API Key的用户才能成功调用模型服务，防止未经授权的访问。

- 首次接入服务：用户首次调用模型接口时需要创建API Key完成身份认证。
- 密钥丢失或泄露：原有API Key泄露或遗忘时，需要新建并替换旧密钥以保障安全。
- 定期轮换密钥：根据安全策略定期更新密钥，减少长期暴露的风险。

约束限制

- 数量限制：每个账户最多可同时存在30个有效API Key，超出后需要删除旧API Key才能新建。
- 不可找回：新建API Key后需立即保存，系统不会存储明文，丢失后无法恢复。
- 鉴权时效：删除API Key后，基于该Key的接口调用将立即失效。

前提条件

- 具有API Key管理权限。
- 了解目标模型服务的调用场景和权限需求。

创建 API Key

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。

1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。

2. 在左侧导航栏，单击“API Key管理”。
3. 在“API Key管理”页面，单击“创建API Key”，填写标签和描述信息后，单击“确定”。

   标签和描述信息在创建完成后，不支持修改。

表 5-1 创建 API Key 参数说明

参数	说明
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。
描述	自定义API Key的描述，长度范围为1~100个字符。

4. 在“您的密钥”对话框，复制密钥并保存至安全位置。
5. 保存完毕后，单击“关闭”。

   单击“关闭”后将无法再次查看密钥。创建成功后，在“API Key管理”页面，可以看到新建的API Key。

图 5-1 API Key

删除 API Key

当API Key数量达到上限，或者API Key发生泄露、遗忘等情况时，建议您及时删除不再使用的API Key。删除后该API Key将无法使用且无法找回。

1. 登录ModelArts Studio控制台，在顶部导航栏选择目标区域。
2. 在左侧导航栏，单击“API Key管理”。
3. 在“API Key管理”页面右侧，单击API Key右侧的“删除”。
4. 在“删除API Key”对话框，单击“确定”。

常见问题

1. 创建API Key后需要等待多久才能生效?

   API Key在创建后不会立即生效，通常需要等待几分钟才能生效。
2. API Key是否支持跨区域使用?

   API Key是区域级别的，不支持跨区域使用。例如，香港区域的API Key必须通过香港控制台创建，且仅能在该区域内调用服务。