

ModelArts Studio (MaaS)

用户指南

文档版本 01
发布日期 2025-10-27



版权所有 © 华为云计算技术有限公司 2025。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。

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	4
2.1 创建 IAM 用户并授权使用 ModelArts Studio (MaaS)	4
2.2 配置 ModelArts 委托授权以使用 ModelArts Studio (MaaS)	8
2.3 配置用户缺失的 ModelArts Studio (MaaS) 相关服务权限	14
3 准备 ModelArts Studio (MaaS) 资源	21
4 ModelArts Studio (MaaS) 在线推理服务	22
4.1 在 ModelArts Studio (MaaS) 模型广场查看预置模型	22
4.2 在 ModelArts Studio (MaaS) 预置服务中开通商用服务	25
4.3 在 ModelArts Studio (MaaS) 创建自定义接入点	28
4.4 使用 ModelArts Studio (MaaS) 部署模型服务	32
4.5 在 ModelArts Studio (MaaS) 管理我的服务	36
4.5.1 在 ModelArts Studio (MaaS) 启动/停止/删除服务	36
4.5.2 在 ModelArts Studio (MaaS) 扩缩容模型服务实例数	36
4.5.3 在 ModelArts Studio (MaaS) 修改模型服务 QPS	37
4.5.4 在 ModelArts Studio (MaaS) 升级模型服务	38
4.6 调用 ModelArts Studio (MaaS) 部署的模型服务	39
4.7 ModelArts Studio (MaaS) API 调用规范	49
4.7.1 对话 Chat/POST	49
4.7.2 获取模型列表 Models/GET	59
4.7.3 错误码	60
4.8 使用 ModelArts Studio (MaaS) 创建多轮对话	63
5 ModelArts Studio (MaaS) 在线体验	64
5.1 在 ModelArts Studio (MaaS) 体验文本对话	64
6 ModelArts Studio (MaaS) 模型管理	67
6.1 在 ModelArts Studio (MaaS) 创建模型	67
7 ModelArts Studio (MaaS) 管理与统计	72
7.1 在 ModelArts Studio (MaaS) 管理 API Key	72
7.2 查看 ModelArts Studio (MaaS) 调用数据和监控指标	74
7.2.1 在 ModelArts Studio (MaaS) 查看在线推理的调用数据和监控指标	74
8 ModelArts Studio (MaaS) 业务最佳实践	81

8.1 使用 ModelArts Studio (MaaS) DeepSeek API 搭建 AI 应用..... 81

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算力资源，并预置了当前主流的第三方开源大模型。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，通常面临三大核心挑战：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Qwen、DeepSeek等模型系列，所有的模型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，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

仅中国-香港、中东-利雅得区域支持使用MaaS。

使用流程

下表展示了MaaS的核心使用流程。

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
	开通预置服务的商用服务	MaaS预置服务的商用服务为企业用户提供高性能、高可用的推理API服务，支持按Token用量计费的模式。该服务适用于需要商用级稳定性、更高调用频次和专业支持的场景。	在ModelArts Studio (MaaS) 预置服务中开通商用服务
	创建自定义接入点	MaaS支持用户创建自定义接入点，对模型进行限流设置，通过model参数进行调用，实现不同业务场景或模型版本的分流与精细化管理。	在ModelArts Studio (MaaS) 创建自定义接入点
	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，便于在其他业务环境中调用该模型。	使用ModelArts Studio (MaaS) 部署模型服务

模块	操作	说明	相关文档
在线体验	模型在线体验	您可以使用自定义接入点进行文本对话在线体验。	在ModelArts Studio (MaaS) 体验文本对话
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
模型管理	创建模型	ModelArts Studio提供了开源大模型，您可以使用这些基础模型，结合自定义的模型权重文件，创建个人专属的模型。创建成功的模型可以进行推理等操作。	在ModelArts Studio (MaaS) 创建模型
管理与统计	查看服务的调用数据和监控指标	MaaS提供调用统计功能，支持查看我的服务、预置服务的商用服务、自定义接入点在指定时间段内的调用数据和监控指标详情，包括总调用次数、总调用失败次数、调用总Tokens数、输入Tokens数、输出Tokens数、端到端时延等信息，并以分钟为最小时间粒度展示数据趋势，帮助您了解服务的使用情况和性能变化，从而更有效地进行模型评估、问题定位、故障排除和性能优化。	查看ModelArts Studio (MaaS) 调用数据和监控指标

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。

操作场景

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 and 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	策略类型	是否必选
Model Arts	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	系统策略	必选
	如果需要给予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	系统策略	可选 ModelArts FullAccess权限和ModelArts CommonOperations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	系统策略	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	系统策略	必选

待授权的服务	授权说明	IAM权限设置	策略类型	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	系统策略	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	系统策略	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	系统策略	可选
统一身份认证服务 IAM	用于检测是否有缺失委托的权限。	iam:permissions:listRolesForAgencyOnDomain iam:permissions:listRolesForAgencyOnProject iam:permissions:listRolesForAgency iam:agencies:getAgency iam:agencies:listAgencies	自定义策略	必选 如果未配置，IAM子用户进入MaaS控制台时会出现权限缺失的弹窗提示用户配置。

配置 MaaS 基础操作权限

步骤1 创建用户组。

1. 管理员登录[IAM管理控制台](#)，在左侧导航栏选择“用户组”。
2. 在“用户组”页面右上角，单击“创建用户组”。在“创建用户组”页面，输入“用户组名称”和“描述”，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。在筛选框选择系统策略，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。

- ModelArts FullAccess: 如果需要给予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. 参照表2-1，配置其他依赖云服务的使用权限。
此处以OBS为例，在搜索框搜索OBS OperateAccess并勾选。在MaaS创建自定义模型时，需要在对象存储服务OBS中创建OBS桶，用于存放模型权重文件、训练数据集或者存放永久保存的日志。

图 2-2 配置 OBS OperateAccess 权限



重复操作此步骤，勾选中所有必选的权限，可选权限请按需选择。IAM权限与其他权限配置不同，需要创建自定义策略、为用户组添加自定义策略，详情请参见[场景三：子用户添加缺失的权限](#)。

3. 勾选完所需权限后，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 将上一步创建的子用户账号加入用户组。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 使用子用户账号登录华为云并验证权限。更多信息，请参见[用户登录](#)。

新创建的用户登录[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。

操作场景

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](#)。

配置 MaaS 委托授权

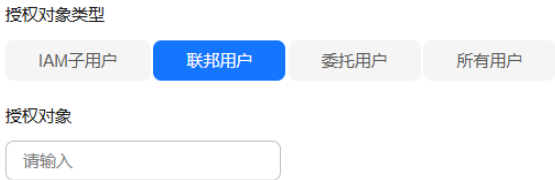
1. 登录[ModelArts管理控制台](#)，按照版本选择以下操作。

- 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 2-3 选择 IAM 子用户  图 2-3 展示了选择 IAM 子用户的界面。在“授权对象类型”部分，有四个按钮：IAM子用户（高亮）、联邦用户、委托用户和所有用户。在“授权对象”部分，有一个下拉菜单。 - 联邦用户：输入联邦用户的用户名或用户ID。 图 2-4 选择联邦用户  图 2-4 展示了选择联邦用户的界面。在“授权对象类型”部分，有四个按钮：IAM子用户、联邦用户（高亮）、委托用户和所有用户。在“授权对象”部分，有一个输入框，提示为“请输入”。 - 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ModelArts管理控制台时，可以在ModelArts管理控制台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 2-5 委托用户切换角色  图 2-5 展示了委托用户切换角色的界面。在顶部，有一个搜索框（标注1）。在“基本信息”部分，有一个“切换角色”按钮（标注2）。在“切换角色”按钮下方，有一个“切换其他角色”按钮。	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“权限模板”下拉框选择“MaaS (ModelArts Studio) ”。 图 2-6 普通模式 	选择“MaaS (ModelArts Studio) ”。
“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 2-7 高权限模式 	-

- 勾选“我已经阅读并同意《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。

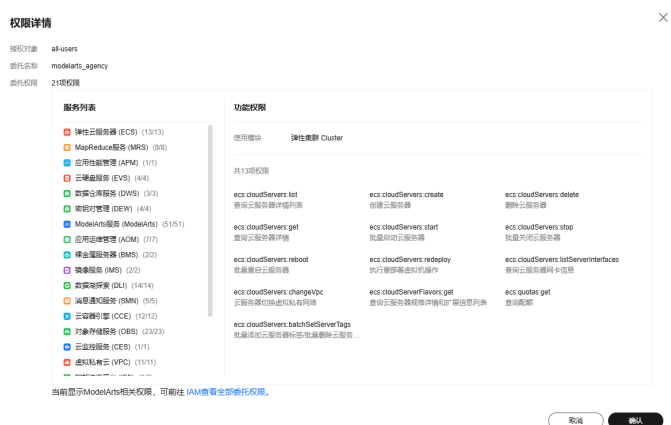
查看授权的权限列表

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 2-8 查看权限



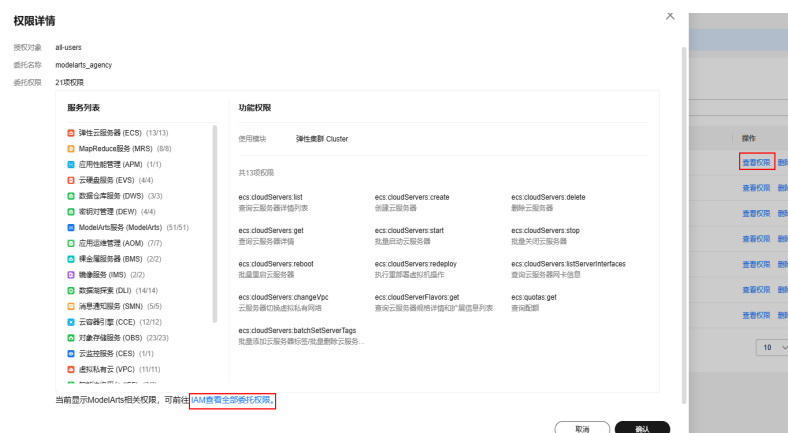
图 2-9 普通模式权限详情示例



修改授权的权限范围

- 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 2-10 去 IAM 修改委托权限



- 进入**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 2-11 手动创建的委托

The screenshot shows a web form for creating a delegation. It has two tabs: '基本信息' (Basic Information) and '授权记录' (Authorization Record). The '基本信息' tab is active. The form contains the following fields:

- 委托名称 (Delegation Name): modelarts_agency
- URN: (empty)
- * 委托类型 (Delegation Type): 云服务 (Cloud Service)
- * 云服务 (Cloud Service): ModelArts
- * 持续时间 (Duration): 永久 (Permanent)
- 描述 (Description): Created by ModelArts service.

At the bottom of the form are two buttons: '确定' (Confirm) and '取消' (Cancel).

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[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MaaS服务的过程中，用户可能会遇到权限配置的问题，如未正确配置权限或缺失权限，MaaS控制台将显示权限缺失的提示。这种情况下，部分功能将无法正常运行，严重影响用户的使用体验。面对权限问题，用户可能会感到困惑，不知道如何高效地解决权限报错提示。为确保MaaS服务的正常运行，建议用户参照本文档提供的指导，及时配置缺失的权限，避免因权限不足而导致的功能异常和系统故障。

前提条件

MaaS控制台出现权限报错相关提示。

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、SW等服务会产生费用，详情请参见[计费概述](#)。

场景一：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，[ModelArts Studio \(MaaS \) 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[配置MaaS委托授权](#)。

子用户：联系管理员进行配置。

图 2-12 获取依赖服务授权提示



场景二：主用户添加缺失的权限

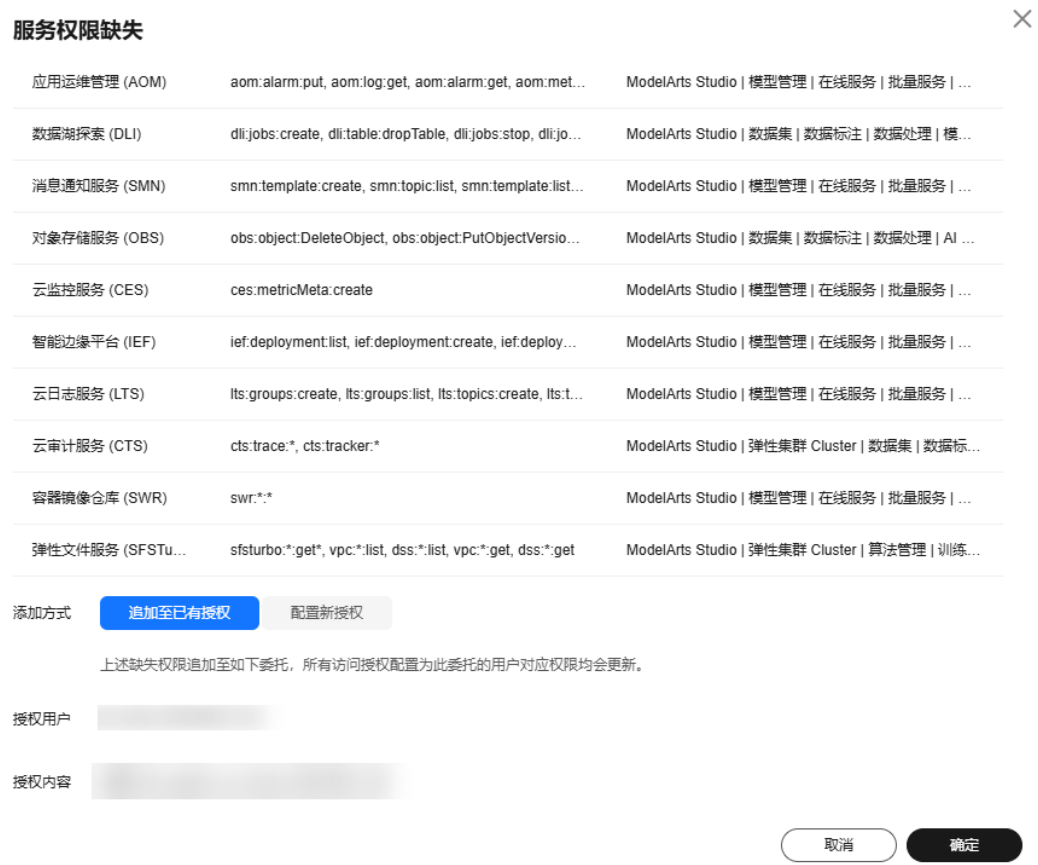
如果您的权限不足，[ModelArts Studio \(MaaS \) 控制台](#)顶部会出现缺失部分服务权限提示。

图 2-13 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 2-14 服务权限缺失



场景三：子用户添加缺失的权限

如果您的权限不足，**ModelArts Studio (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 2-15 访问授权对话框



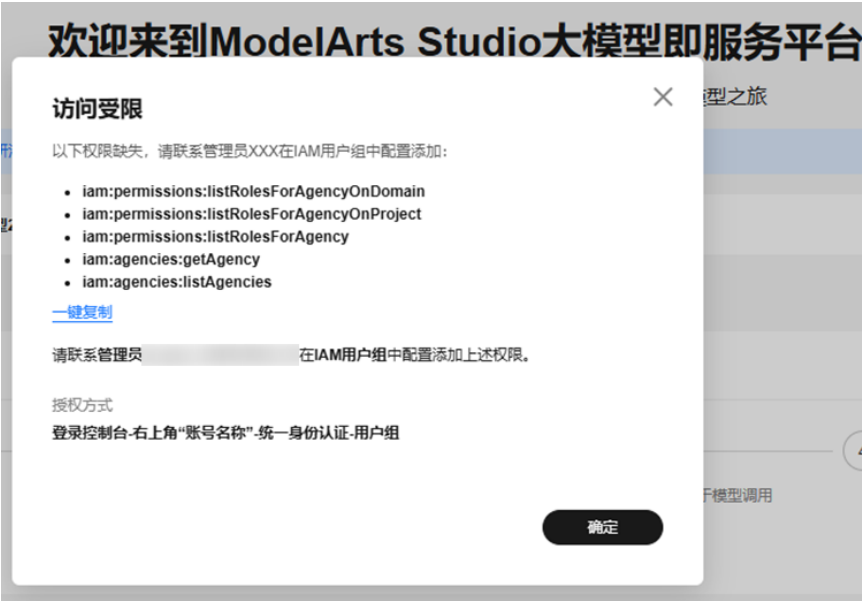
缺失权限的说明请参见[表2-3](#)。更多信息，请参见[授权项](#)。

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。

- 创建自定义策略。
 - 子账号在“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 2-16 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. 管理员登录IAM管理控制台，在左侧导航栏，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 2-17 创建自定义策略

策略名称

polycyn5pk6j

策略配置方式

可视化视图

JSON视图

策略内容

1 {
2 "Version": "1.1",
3 "Statement": [
4 {
5 "Effect": "Allow",
6 "Action": [
7 "iam:agencies:listAgencies",
8 "iam:agencies:getAgency",
9 "iam:permissions:listRolesForAgency",
10 "iam:permissions:listRolesForAgencyOnProject",
11 "iam:permissions:listRolesForAgencyOnDomain"
12]
13 }
14]
15 }]

从已有策略复制 格式化内容

策略描述

请输入策略描述 (可选)

作用范围

全局级服务

确定 取消

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，单击“格式化内容”。	<pre>{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["iam:permissions:listRolesForAgencyOnDomain", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:listRolesForAgency", "iam:agencies:getAgency", "iam:agencies:listAgencies"] }] }</pre>
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. 管理员为用户组添加自定义策略。
- a. 在IAM管理控制台左侧导航栏，选择“用户组”。

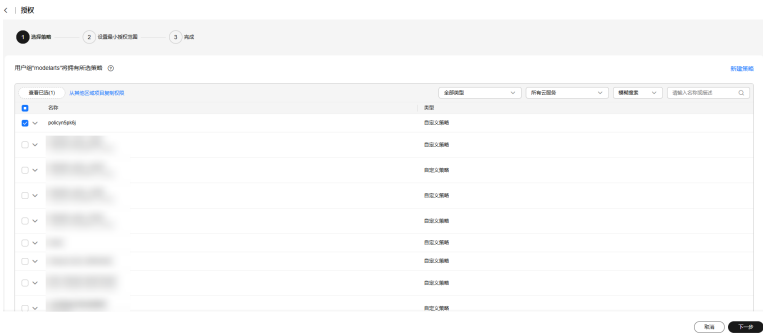
b. 在“用户组”页面，按需搜索目标用户组名称，在操作列单击“授权”。

图 2-18 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 2-19 授权页面

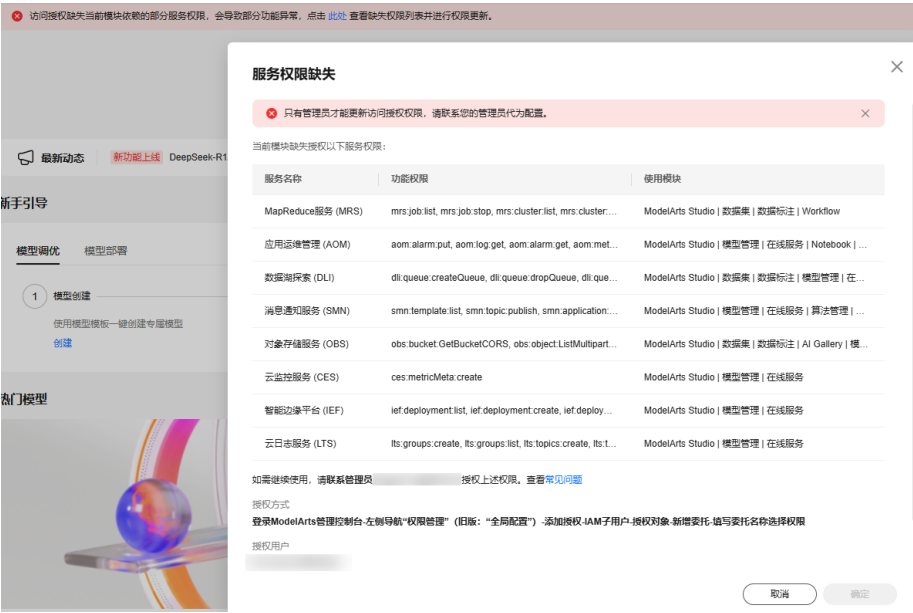


- d. 在“权限生效时间提醒”对话框，仔细阅读相关信息，然后单击“知道了”。

3. 子账号登录ModelArts Studio (MaaS) 控制台控制台，查看“访问受限”对话框是否消失。
- 如果消失，表示权限配置成功，您可以正常使用MaaS。

– 如果出现“服务权限缺失”对话框，请执行下一步。
4. 查看并配置缺失的服务权限。
- a. 子账号登录ModelArts Studio (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 2-20 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[配置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 OBS 桶

在ModelArts Studio大模型即服务平台创建自定义模型时，需要在对象存储服务OBS中创建OBS桶，用于存放模型权重文件、训练数据集或者是存放永久保存的日志。

创建OBS桶和上传文件的操作指导请参见[OBS控制台快速入门](#)。

说明

- 仅“中国-香港”和“中东-利雅得”区域支持使用ModelArts Studio大模型即服务平台（MaaS）。
- OBS桶必须和MaaS服务在同一个Region下，否则无法选择到该OBS路径。

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 \(MaaS \) 控制台](#)，在顶部导航栏选择目标区域。
2.

在左侧导航栏，单击“模型广场”。
3.

在“模型广场”页面的“筛选”区域，按需选择模型类型、上下文长度、高级能力、模型系列、支持作业进行筛选，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
模型类型	支持按照文本生成等模型类型进行筛选。
上下文长度	支持按照64K、32K、16K、≤8K上下文长度进行筛选。
高级能力	支持按照深度思考等多个能力进行筛选。
模型系列	支持按照DeepSeek、通义千问等模型系列进行筛选。
支持作业	支持按照部署等支持作业进行筛选。

4.

在“模型广场”页面的目标模型卡片，按需选择以下操作。

- 鼠标悬浮于模型卡片，可以看到操作按钮，您可以按需单击“模型部署”等。关于部署模型服务的说明，请参见[使用ModelArts Studio \(MaaS \) 部署模型服务](#)。
- 单击模型卡片，进入模型详情页面，可以查看模型的介绍、基本信息和版本信息等。在页面右上角，您可以按需单击“模型部署”等操作，使用模型进行训推。
 - 页面右上角只显示该模型支持的操作。不同模型显示的操作可能不同，请以实际环境为准。
 - 如果模型涉及计费，基本信息页签会显示计费的相关信息。

模型介绍

下表列举了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	中国-香港、中东-利雅得	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	中国-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均衡策略，并设置了多Token预测训练目标以获得更强大的性能。
	DeepSeek-V3.1	文本生成	对话问答	中文、英文	中国-香港、中东-利雅得	DeepSeek-V3.1是一个同时支持思考模式和非思考模式的混合模型，效果与DeepSeek-R1-0528相当，但响应速度更快，且在工具使用方面进行了优化。

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-R1-Distill-Qwen-14B	文本生成	对话问答、文本生成推理	中文、英文	中国-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	中国-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。
Deepseek-Coder		文本生成	对话问答、文本推理	中文、英文	中国-香港	Deepseek Coder由一系列代码语言模型组成，每个模型都从头开始在2T标记上进行训练，其中87%为代码，13%为英文和中文的自然语言。在编码能力方面，DeepSeek Coder在多种编程语言和各种基准测试中均在开源代码模型中取得了较高性能。
通义千问	QwQ	文本生成	对话问答	英文	中国-香港	QwQ是通义千问系列的推理模型。与传统的指令调优模型相比，具有思维和推理能力的QwQ在下游任务（尤其是疑难问题）中可以实现显著的性能提升。
通义千问2.5	Qwen2.5	文本生成	多语言处理、数学推理、对话问答	中文、英文	中国-香港	Qwen2.5是阿里云研发的Qwen系列的新的语言模型。对于Qwen2.5，发布了许多基本语言模型和指令调整的语言模型，参数范围从5亿到720亿。
	Qwen2.5-VL	图像理解	图像理解、对话问答	中文、英文	中国-香港	通义千问2.5-VL是阿里云通义千问团队开源的多模态视觉语言模型，具备强大的视觉和语言理解能力。

模型系列		模型类型	应用场景	支持语言	支持地域	模型介绍
通义千问3	Qwen3	文本生成	对话问答	中文、英文	中国-香港	Qwen3是Qwen团队研发的大语言模型和大型多模态模型系列，在大规模语言和多模态数据上进行预训练，通过高质量的数据进行后期微调。
Kimi	Kimi-K2	文本生成	对话问答	中文、英文	中国-香港	Kimi K2是一款最先进的混合专家（MoE）语言模型，拥有320亿激活参数和1万亿总参数。通过Muon优化器训练，Kimi K2在前沿知识、推理和编程任务上表现出色，同时在智能体能力方面进行了精心优化。

4.2 在 ModelArts Studio (MaaS) 预置服务中开通商用服务

MaaS预置服务的商用服务为企业用户提供高性能、高可用的推理API服务，支持按Token用量计费的模式。该服务适用于需要商用级稳定性、更高调用频次和专业支持的场景。

操作场景

- 企业智能客服：企业希望利用推理API优化客服系统，实现智能问答、意图识别，提升客服效率与客户满意度。
- 内容创作辅助：媒体、广告公司借助理AI进行文案创作、创意生成，提高内容产出的效率与质量。
- 智能数据分析：金融、电商企业通过推理API对海量数据深度分析，挖掘数据价值，辅助决策制定。

约束限制

- 该功能仅支持中国-香港区域。
- 开通商用服务时，将自动开通该服务下所有版本，不支持单独开通某版本。
- 暂不支持关闭商用服务。

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。详细信息，请参见[ModelArts Studio \(MaaS \) 模型推理计费项](#)。

优惠券说明

- 当有优惠折扣时，预置服务的商用服务页签会出现相关提示。开通商用服务时，会默认领取可用的优惠券。在扣费时，会优先抵扣优惠券。
- 不同优惠券活动的适用范围和领取条件各不相同，能否成功领取请以实际活动规则为准。
- 模型服务的优惠折扣的发放和使用情况，请前往“费用中心 > 优惠折扣 > 优惠券”进行查看。

服务调用说明

请求可能会根据实际情况路由到其他区域实例。

开通商用服务

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”。
 3. 在“预置服务 > 商用服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
 4. 在开通预置模型服务对话框，按需勾选商用服务（默认全选），勾选“我已阅读并同意上述说明，及《ModelArts Studio 服务声明》”，单击“一键开通”。
- 勾选目标商用服务后，会自动开通该服务下的所有模型版本。
- 商用服务列表的参数说明请参见下表。

表 4-3 商用服务列表参数说明

参数	说明
服务名称	商用服务的名称。在服务名称左侧单击  图标，可以查看该服务的版本。“model参数”列显示的名称可用于模型调用时使用，即model参数的值。 图 4-1 查看服务版本 
付费状态	• 开通：已开通商用服务。 • 未开通：未开通商用服务。
类型	商用服务的类型。

参数	说明
计费方式	商用服务的计费方式，不同模型的计费方式可能不同，请以实际环境为准。更多信息，请参见 计费项 (ModelArts Studio) 。
推理定价	商用服务的推理定价，不同模型的推理定价可能不同，请以实际环境为准。更多信息，请参见 计费项 (ModelArts Studio) 。
优惠折扣	商用服务已有的优惠折扣，“--”表示没有优惠，请以实际环境为准。
模型限流	当前账号下，访问同一模型下所有服务的总额度。 - TPM：每分钟处理的Tokens数（输入+输出）。 - RPM：每分钟处理的请求数。
调用统计	单击  图标，跳转至“服务调用详情”页面，查看商用服务在指定时间段内的调用数据和监控指标详情。更多信息，请参见在 ModelArts Studio (MaaS) 查看在线推理的调用数据和监控指标 。
操作	商用服务支持的相关操作。 - 关闭服务：该按钮置灰，表示暂不支持关闭服务，未使用服务时不会产生费用。 - 调用说明：单击“调用说明”，选择服务版本，在“调用说明”面板查看调用商用服务的相关信息和操作步骤。更多信息，请参见调用ModelArts Studio (MaaS) 部署的模型服务。

流控规则说明

为了保证用户调用模型的公平性，MaaS设置了基础限流。如果超出限制，API请求将会失败，需等到解除限流条件时再次调用。

- TPM（Tokens Per Minute）：每分钟处理的Tokens数（输入+输出）。
- RPM（Requests Per Minute）：每分钟处理的请求数。

如果模型服务的RPM为300，意味着每秒最多可以处理10个请求（300/30=10）。当用户1秒内发送300个请求会远远超出服务的处理能力，导致请求失败。

建议您均匀地发送API请求，避免短时间内发送大量请求。根据API网关的限流机制，如果1秒内的请求数超过RPM/30*1，超额部分的请求可能会触发API网关的速率限制拦截，导致请求失败并返回错误码429（Too Many Requests）。

注意事项：

即使用户按照RPM/30的速率发送请求，由于网络延迟和请求到达时间的不确定性，仍有可能出现少量失败请求。实际以请求到达服务端的时间为准，而不是以发送请求的时间为准。

欠费说明

当您使用某个模型服务欠费后，会触发欠费冻结，进入宽限期，资源状态变为“已冻结”。此时欠费冻结的资源不可使用，未开通的模型不支持再开通。

宽限期和保留期的时间均为15天。保留期到期后，如果您仍未支付账户欠款，那么您账号名下此模型相关资源和订单记录会被清理，数据无法恢复。如需重新使用模型服务，需要结清欠费后重新开通模型服务。

请您及时关注账户的余额情况，并前往[充值页面](#)及时充值，以保障线上服务的稳定。

图 4-2 按需计费资源生命周期



常见问题

1. 有计费示例吗？
计费项和计费示例请参考[MaaS模型推理计费项](#)。
2. 开通付费服务后，可以关闭吗？
暂不支持关闭付费服务，未使用服务时不会产生费用。
3. 使用商用服务，模型状态显示冻结，如何处理？
此时欠费冻结的资源不可使用，未开通的模型不支持再开通。您可以通过充值进行解冻，被冻结的资源实例将恢复使用，未开通的模型将支持开通。详细操作请参见[账号充值](#)。

4.3 在 ModelArts Studio (MaaS) 创建自定义接入点

MaaS支持用户创建自定义接入点，通过自定义接入点名称进行模型调用（model参数设置），实现不同业务场景或模型版本的分流与精细化管理。

操作场景

在企业 and 开发者的AI应用开发与运营过程中，面临着推理服务调用管理无序、流量控制困难、成本核算模糊等问题。多个业务线共用同一推理服务，导致资源争抢、服务性能不稳定，同时缺乏有效的调用限制手段，难以追溯各业务模块的资源消耗情况。

MaaS支持自定义接入点功能，通过创建独立的调用入口，允许用户设置限流规则，并基于自定义接入点名称实现费用的精准统计，帮助用户高效管理推理服务资源，优化使用成本。

约束限制

- 该功能仅支持中东-利雅得区域。
- 最多可以同时存在10个自定义接入点。
- 同一账户下不允许存在同名的自定义接入点。已删除的接入点名称不允许新建时使用。
- 自定义接入点创建后，不支持修改模型服务。

- 创建的自定义接入点需遵循平台相关的规则和规范，不得进行违规调用。

计费说明

自定义接入点功能本身不收费。调用模型服务或使用资源可能会产生费用。您可以通过接入点名称在费用中心查询服务使用账单。

调用AI开发平台ModelArts-在线服务：由ModelArts计费。关于计费详情，请参见[ModelArts推理部署计费项](#)。

前提条件

已在ModelArts部署模型为在线服务。具体操作，请参见[在ModelArts部署模型为在线服务](#)。

创建自定义接入点

- 登录[ModelArts Studio \(MaaS \) 控制台](#)，在顶部导航栏选择目标区域。
- 在左侧导航栏，选择“在线推理”。
- 单击“自定义接入点”页签，在页面右上角单击“创建自定义接入点”。
- 在“创建自定义接入点”面板，配置相关参数。

表 4-4 创建自定义接入点参数说明

参数	说明
名称	自定义接入点的名称。自定义接入点名称具有 唯一性 ，不能重复，不支持特殊字符。输入长度范围为1~64个字符。
描述	自定义接入点的描述，最多支持256字符。
服务来源	选择“AI开发平台ModelArts-在线服务”。 AI开发平台ModelArts-在线服务：新版在线服务，由ModelArts进行计费。
模型服务	单击“选择模型服务”，在“选择模型服务”对话框，按需选择区域和服务，单击“确定”。
接入点流量控制	勾选“接入点流量控制”，手动设置接入点的RPM和TPM流控。如果该账号下访问同一模型的所有接入点限流总和等于该模型的总限流额度，就能有效避免不同接入点之间争夺流量配额。 - 用户可以针对每个接入点设置不同的RPM和TPM流控，但不能超过账号的模型限流值。 - RPM和TPM流控需为正整数。

- 确认配置信息及计费无误后，单击“立即创建”。
- 创建成功后，“自定义接入点”页签会显示接入点的相关信息，您可以进行调用、在线体验等操作。

在线体验自定义接入点

只有当自定义接入点的“状态”为“使用中”，才能进行在线体验。

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. 单击“自定义接入点”页签，在目标接入点的“操作”列，单击“在线体验”。
关于在线体验的更多信息，请参见[在ModelArts Studio \(MaaS\) 体验文本对话](#)。

调用自定义接入点

只有当自定义接入点的“状态”为“使用中”，才能被成功调用。服务调用产生的内容由AI生成，不代表MaaS观点，平台不保证其合法性、真实性、准确性，不承担相关法律责任。

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. 单击“自定义接入点”页签，在目标接入点的“操作”列，单击“调用说明”。
4. 在“调用说明”页面，按照页面提示获取API Key，复制调用示例并替换接口信息、API Key，进行API调用。
 - 在“自定义接入点”页签的“model参数”列显示的名称，为调用服务时代码的model参数值。用户可以根据不同的model参数进行不同接入点的调用。
 - 关于如何创建API Key，请参见[在ModelArts Studio \(MaaS\) 管理API Key](#)。
 - 关于调用示例的参数说明，请参见[调用ModelArts Studio \(MaaS\) 部署的模型服务](#)。

查看自定义接入点的调用统计

您可以查看自定义接入点在指定时间段内的调用数据和监控指标详情，包括调用次数、调用失败次数、调用总Tokens数等信息，了解服务的使用情况和性能变化，从而更有效地进行模型评估、问题定位、故障排除和性能优化。

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. 单击“自定义接入点”页签，在目标接入点的“调用统计”列，单击图标，跳转至“服务调用详情”页面，查看调用详情。
关于调用统计的更多信息，请参见[在ModelArts Studio \(MaaS\) 查看在线推理的调用数据和监控指标](#)。

编辑自定义接入点

您可以按需修改自定义接入点信息，例如描述、限流等。自定义接入点的模型服务不支持修改。

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. 单击“自定义接入点”页签，在目标接入点的“操作”列，单击“更多 > 编辑”。

4. 在“编辑自定义接入点”面板，按需修改相关参数，单击“更新”。
关于参数说明，请参见[创建自定义接入点参数说明](#)。

停用/启用自定义接入点

当自定义接入点“状态”为“使用中”，可以停用自定义接入点。停用接入点后，该接入点的推理能力将停用，支持重新启用。由于出账存在时延，可能在您停用后仍会收到由该服务产生的账单。

当自定义接入点“状态”为“停用”，可以启用自定义接入点。

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”，在“自定义接入点”页签按需选择以下操作。
 - 停用自定义接入点
 - i. 在目标接入点的“操作”列，单击“更多 > 停用”。
 - ii. 在“停用自定义接入点”对话框，输入**YES**，单击“确定”。
停用后，该接入点的状态会显示为“停用”。
 - 启用自定义接入点
 - i. 在目标接入点的“操作”列，单击“更多 > 启用”。
 - ii. 在“启用”对话框，单击“确定”。
启用后，该接入点的状态会显示为“使用中”。

删除自定义接入点

当自定义接入点不再需要时，您可以进行删除操作。删除后，该接入点的推理能力将停用，全部信息将被删除且无法恢复，请谨慎操作。

由于出账存在时延，可能在您删除后仍会收到由该服务产生的账单。

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. 单击“自定义接入点”页签，在目标接入点的“操作”列，单击“更多 > 删除”。
4. 在“删除自定义接入点”对话框，查看删除提示信息，确认无误后输入**DELETE**，单击“确定”。
删除后，“自定义接入点”页签将不再显示该接入点。

常见问题

- **创建的自定义接入点数量达到上限怎么办？**
您可以删除不再使用的接入点，然后新建接入点。
- **如何确定Tokens的消耗数量？**
您可以通过“调用统计”页面查看模型服务调用的总Tokens数、输入Tokens数、输出Tokens数等信息，详情请参见[查看自定义接入点的调用统计](#)。
- **修改自定义接入点的限流设置后，多久会生效？**
修改保存后，限流设置会立即生效，后续调用将按照新规则执行。

4.4 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。

操作场景

从模型广场或我的模型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在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

- 该功能仅支持中国-香港区域。
- 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。

说明

不同地域支持的模型可能不同，请以实际环境为准。

表 4-5 模型默认最大输入输出长度

模型	默认最大输入输出长度 (token)
DeepSeek-R1-8K DeepSeek-V3-8K DeepSeek-R1-Distill-Qwen-14B-8K DeepSeek-R1-Distill-Qwen-32B-8K	8192
DeepSeek-R1-16K DeepSeek-V3-16K QwQ-32B-16K	16384
DeepSeek-R1-32K DeepSeek-R1-Distill-Qwen-32B-32K DeepSeek-V3-32K Deepseek-Coder-33B QwQ-32B-32K Qwen2.5-VL-7B-32K Qwen3-8B-32K Qwen3-32B-32K	32768

模型	默认最大输入输出长度 (token)
DeepSeek-V3-64K Qwen2.5-32B-64K Qwen3-235B-A22B-64K Kimi-K2	65536
DeepSeek-V3.1	131072
其他模型	4096

前提条件

- 已准备专属资源池，详细请参见[准备ModelArts Studio \(MaaS \) 资源](#)。
- 在“我的模型”页面存在[已创建成功的模型](#)或直接使用[模型广场](#)的模型。

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，单击“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 4-6 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”或“我的模型”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。

参数		说明
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。
	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”页面，然后单击“我的服务”页签。
3. 在“我的服务”页签，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”：可以查看服务资源监控指标相关信息。

表 4-7 资源监控参数说明

参数	说明
时间范围	支持按照近1小时、近3小时、近12小时、近24小时、近7天、自定义时间段统计服务的资源使用情况。 自定义时间支持最多查看30天的数据。
CPU使用率 (%)	服务的CPU使用情况。
内存使用率 (%)	服务的内存使用情况。
NPU算力使用率 (%)	服务的NPU算力使用情况。
NPU显存利用率 (%)	服务的NPU显存使用情况。
磁盘读取速率 (bit/min)	服务的磁盘读取速率。
磁盘写入速率 (bit/min)	服务的磁盘写入速率。
上行速率 (bit/min)	当前服务的出口方向网络流速。
下行速率 (bit/min)	当前服务的入口方向网络流速。

- “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。
4. 在“服务详情”页面上方，您可以按需进行如下操作。
- 查看服务的调用数据：单击“调用统计”，跳转至“服务调用详情”页面查看监控数据和调用失败明细相关信息。详细信息，请参见[在ModelArts Studio \(MaaS \) 查看在线推理的调用数据和监控指标](#)。
 - 停止/启动服务：具体操作，请参见[停止/启动部署服务](#)。
 - 删除服务：具体操作，请参见[删除部署服务](#)。
 - 调用服务：单击“调用说明”，按照页面提示进行调用。详细信息，请参见[调用ModelArts Studio \(MaaS \) 部署的模型服务](#)。

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](#)。

4.5 在 ModelArts Studio (MaaS) 管理我的服务

4.5.1 在 ModelArts Studio (MaaS) 启动/停止/删除服务

约束限制

该功能仅支持中国-香港区域。

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，单击“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，单击“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。

4.5.2 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

该功能仅支持中国-香港区域。

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。

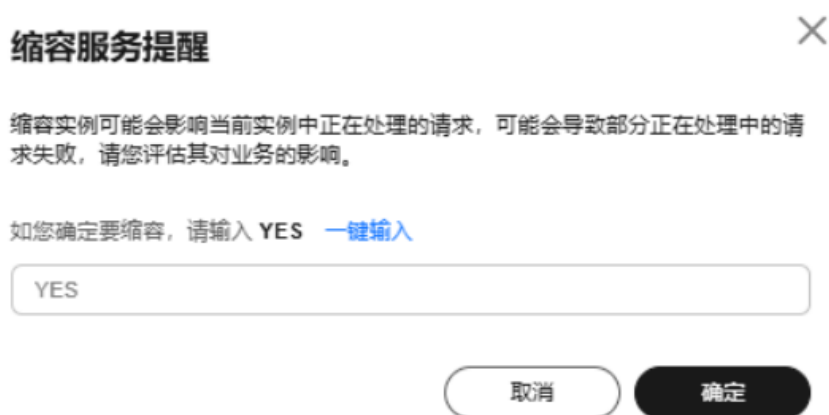
计费说明

扩容模型服务实例数后，在调用MaaS预置服务时，会产生计算资源和存储资源的累计值计费、Token费用详情请见[ModelArts Studio \(MaaS \) 模型推理计费项](#)。

扩缩实例数

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，单击“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，按需选择以下操作。
 - **扩容**：按需增加“变更后实例数”，单击“确定”，在“扩缩容服务”对话框，单击“确定”。
 - **缩容**：按需减少“变更后实例数”，单击“确定”，在“缩容服务提醒”对话框，查看提示信息，确认无误后输入YES，单击“确定”。

图 4-3 缩容服务提醒



修改完后，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。

后续操作

- **调用模型服务**：模型服务扩缩容后，可以在其他业务环境中调用该模型服务进行预测。
- **查看预置服务的调用数据**：MaaS提供调用统计功能，可以查看模型服务在指定时间段内的调用数据详情，监控服务使用情况和资源消耗。

4.5.3 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

该功能仅支持中国-香港区域。

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，单击“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。
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。

4.5.4 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

该功能仅支持中国-香港区域。

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，单击“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”。

后续操作

调用模型服务：模型服务扩缩容后，可以在其他业务环境中调用该模型服务进行预测。

4.6 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-商用服务或自定义接入点。

操作场景

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

- 使用预置服务：在“在线推理”页面的“预置服务”页签，已开通商用服务（付费状态为“开通”）。具体操作，请参见[ModelArts Studio \(MaaS \) 在线推理服务](#)。
- 使用我的服务：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](#)。
- 使用自定义接入点：已创建自定义接入点。具体操作，请参见[在ModelArts Studio \(MaaS \) 创建自定义接入点](#)。

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](#)。

- 登录[ModelArts Studio \(MaaS \) 控制台](#)，在顶部导航栏选择目标区域。
- 在左侧导航栏，单击“API Key管理”。
- 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 4-8 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，单击“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“关闭内容审核服务”对话框，选择是否启用内容审核（默认启用）。

-

启用内容审核，可以阻止在线推理中的输入输出中出现不合规的内容，但可能会对接口性能产生较大影响。

-

关闭内容审核服务，将不会审核在线推理中的输入输出，模型服务可能会有违规风险，请谨慎关闭。

关闭“内容审核”开关，需要在弹窗中确认是否停用内容审核服务，勾选“我已阅读并同意上述说明”后，单击“确定”关闭。
4.

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。

-

Rest API示例代码如下所示：

■ 使用Python调用示例。

```
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cURL调用示例。

```
curl -X POST "https://example.com/v1/chat/completions" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $API_KEY" \
-d '{
  "model": "DeepSeek-R1",
  "messages": [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "你好"}
  ],
  "stream": true,
  "stream_options": { "include_usage": true },
  "temperature": 0.6
}'
```

– 使用OpenAI SDK调用示例。

```
# 安装环境命令。
pip install --upgrade "openai>=1.0"
# OpenAI SDK调用示例。
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4-9](#)仅介绍关键参数，详细参数解释请参见vLLM官网。使用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 4-9 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。
messages.role	是	无	Str	不同的role对应不同的消息类型。 - system：开发人员输入的指令，例如模型应遵循的答复格式、扮演的角色等。 - user：用户输入的消息，包括提示词和上下文信息。 - assistant：模型生成的回复内容。 - tool：模型调用工具返回的信息。
messages.content	是	无	Str	- 当role为system时：给AI模型设定的人设。 <code>{"role": "system", "content": "你是一个乐于助人的AI助手"}</code> - 当role为用户时：用户输入的问题。 <code>{"role": "user", "content": "9.11和9.8哪个大?"}</code> - 当role为assistant时：AI模型输出的答复内容。 <code>{"role": "assistant", "content": "9.11大于9.8"}</code> - 当role为tool时：AI模型调用的工具响应信息。 <code>{"role": "tool", "content": "上海今天天气晴，气温10度"}</code>
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}。

参数	是否必选	默认值	参数类型	描述
max_tokens	否	16	Int	当前任务允许的生成Token数上限，包括模型输出的Tokens和深度思考的Reasoning Tokens。
top_k	否	-1	Int	在生成过程中，候选集大小限定了采样的范围。以取值50为例，这意味着每一步仅会考虑得分排在前50位的Token构成候选集进行随机抽样。增大此值将提高输出的随机性，减小此值会增强输出的确定性。
top_p	否	1.0	Float	模型核采样（ nucleus sampling ）。仅保留累计概率刚好超过阈值p的那一部分词，其余全部屏蔽，最后在这份候选词里重新归一化并采样。 设置值越小，候选词越少，模型输出越集中和保守；设置值越大，候选词越多，模型输出越开放和多样。 通常情况只建议调整temperature或top_p，不要同时修改两个参数。 取值范围：0~1，设置为“1”表示考虑所有Tokens。
temperature	否	0.6	Float	模型采样温度。设置的值越高，模型输出越随机；设置的值越低，输出越确定。 通常情况只建议调整temperature或top_p，不要同时修改两个参数。 temperature取值建议：DeepSeek-R1、DeepSeek-V3、Qwen3系列建议值为0.6，Qwen2.5-VL系列建议值为0.2。
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	为每个输入的消息生成的响应数。 - 不使用beam_search场景下，n取值建议为$1 \leq n \leq 10$。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$1 < n \leq 10$。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

- 普通requests包、OpenAI SDK、curl命令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 4-10 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 4-11 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。

错误码	错误内容	说明
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。

内容审核说明

- 流式请求
 - 如果触发内容审核，则会返回错误：错误码403。您可以通过错误码ModelArts.81011来判断。返回内容如下：

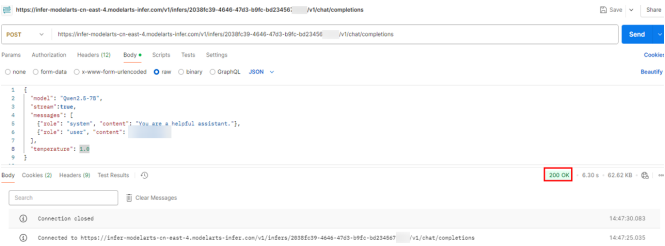
```
{
  "error_code": "ModelArts.81011",
  "error_msg": "May contain sensitive information, please try again."
}
```

图 4-4 报错示例



- 如果未触发内容审核，则使用postman调用返回参考如下，返回码200。

图 4-5 正常返回示例



- 如果输出有敏感信息，则会在输出流后面拼接如下数据：
data: {"id":"chatcmpl-
*****", "object":"chat.completion", "created":1678067605, "model":"*****", "choices":
[{"delta":{"content":"这是流式响应的开始。"}, "index":0}]
data: {"id":"chatcmpl-
*****", "object":"chat.completion", "created":1678067605, "model":"*****", "choices":
[{"delta":{"content":"继续输出结果。"}, "index":0}]
data: {"id":"chatcmpl-
*****", "object":"chat.completion", "created":1678067605, "model":"*****", "choices":
[{"finish_reason":"content_filter", "index":0}]}
data: [DONE]

触发内容审核之后，"finish_reason"是"content_filter"；正常的流式停止是"finish_reason":"stop"。

- 非流式请求

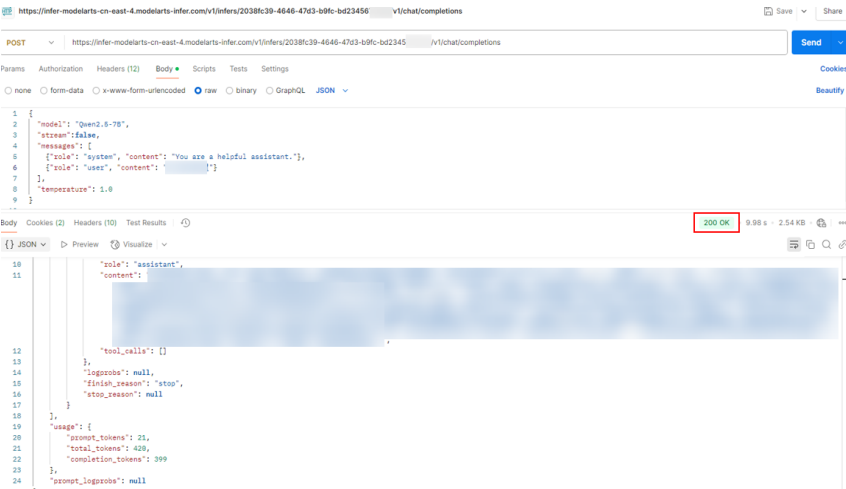
- 如果触发内容审核，则会返回错误：错误码403。您可以通过错误码ModelArts.81011来判断。

返回内容如下：

```
{
  "error_code": "ModelArts.81011",
  "error_msg": "May contain sensitive information, please try again."
}
```

- 如果未触发，则正常返回，示例如下：

图 4-6 正常返回示例



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 4-12 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。

类型/接口	/v1/models	/v1/chat/completions
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。
请求示例	<pre>GET https://example.com/v1/infers/937cabe5-d673-47f1-9e7c-2b4de06*****/v1/models HTTP/1.1 Authorization: Bearer YOUR_API_KEY</pre>	<pre>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 }</pre>
响应示例	<pre>{ "data": [{ "id": "*****", "description": "最新一代大模型" }, { "id": "*****", "description": "性价比较高的替代方案" }] }</pre>	<pre>{ "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 } }</pre>

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](#)
- [在ModelArts Studio \(MaaS \) 查看在线推理的调用数据和监控指标](#)

4.7 ModelArts Studio (MaaS) API 调用规范

4.7.1 对话 Chat/POST

MaaS平台提供功能丰富的在线推理能力，既有免部署可直接调用的预置模型服务，同时也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 4-13 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 。

预置商用服务支持模型列表

模型系列	模型版本	支持地域	model参数值	序列长度	Function Call功能
DeepSeek	DeepSeek-V3.1	中国-香港	deepseek-v3.1	131072	支持
Qwen3	Qwen3-32B-32K	中国-香港	qwen3-32b	32768	不支持

思维链说明

思维链（Chain of Thought，简称CoT）是指模型在解决复杂问题时，能够生成一系列中间推理步骤的能力。这种能力使得模型不仅能够给出最终答案，还能展示出其推理过程，从而提高模型的可解释性和透明度。

仅DeepSeek-V3.1模型支持开启或关闭思维链。

DeepSeek-V3.1模型的约束限制如下：

- Function Call功能和思维链不兼容，不建议同时使用。
- 开启思维链不支持前缀续写。
- 不截断思维链只截断content能力不生效。
- 开启思维链后guided_choice能力不可用，reasoning_content和guided_decoding不兼容。

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 4-14 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4-13 。
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： <pre>"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?" }]</pre> 更多信息，请参见 表4-15 。
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4-16 。
max_tokens	否	无	Int	当前任务允许的生成Token数上限，包括模型输出的Tokens和深度思考的Reasoning Tokens。
top_k	否	-1	Int	在生成过程中，候选集大小限定了采样的范围。以取值50为例，这意味着每一步仅会考虑得分排在前50位的Token构成候选集进行随机抽样。增大此值将提高输出的随机性，减小此值会增强输出的确定性。
top_p	否	1.0	Float	模型核采样（nucleus sampling）。仅保留累计概率刚好超过阈值p的那一部分词，其余全部屏蔽，最后在这份候选词里重新归一化并采样。 设置值越小，候选词越少，模型输出越集中和保守；设置值越大，候选词越多，模型输出越开放和多样。 通常情况只建议调整temperature或top_p，不要同时修改两个参数。 取值范围：0~1，设置为“1”表示考虑所有Tokens。

参数名称	是否必选	默认值	参数类型	说明
temperature	否	1.0	Float	模型采样温度。设置的值越高，模型输出越随机；设置的值越低，输出越确定。 通常情况只建议调整temperature或top_p，不要同时修改两个参数。 temperature取值建议：DeepSeek-R1、DeepSeek-V3、Qwen3系列建议值为0.6，Qwen2.5-VL系列建议值为0.2。
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	Boolean	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	为每个输入的消息生成的响应数。 - 不使用beam_search场景下，n取值建议为$1 \leq n \leq 10$。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$1 < n \leq 10$。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	Boolean	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。
chat_template_kwargs.thinking	否	false	Boolean	默认关闭思维链。 仅支持DeepSeek-V3.1模型 ，约束限制请参见 思维链说明 。 开启思维链示例如下： <pre>{ "model": "deepseek-v3.1", "messages": [{ "role": "system", "content": "You are a helpful assistant." }, { "role": "user", "content": "你好" }], "chat_template_kwargs": { "thinking": true } }</pre>

表 4-15 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	String	不同的role对应不同的消息类型。 - system：开发人员输入的指令，例如模型应遵循的答复格式、扮演的角色等。 - user：用户输入的消息，包括提示词和上下文信息。 - assistant：模型生成的回复内容。 - tool：模型调用工具返回的信息。

参数名称	是否必选	默认值	参数类型	说明
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为用户时：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" } - 当role为tool时：AI模型调用的工具响应信息。 { "role": "tool", "content": "上海今天天气晴，气温10度" }

表 4-16 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	Boolean	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 4-17 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）。

参数名称	类型	说明
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。

DeepSeek-V3 文本生成（非流式）请求示例

- Rest API请求示例：

- Python请求示例：

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/chat/completions" # API地址
    api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key

    # Send request.
    headers = {
        'Content-Type': 'application/json',
        'Authorization': f'Bearer {api_key}'
    }
    data = {
        "model": "deepseek-v3", # 模型名称
        "messages": [
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": "你好"}
        ]
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

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

- cURL请求示例

```
curl -X POST "https://api.modelarts-maas.com/v1/chat/completions" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $MAAS_API_KEY" \
-d '{
  "model": "deepseek-v3",
  "messages": [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "你好"}
  ]
}'
```

- OpenAI SDK请求示例：

```
from openai import OpenAI
```

```
base_url = "https://api.modelarts-maas.com/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY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="deepseek-v3", # 模型名称
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "你好"}
    ]
)

print(response.choices[0].message.content)
```

DeepSeek-V3 文本生成（流式）请求示例

- Python请求示例：

```
from openai import OpenAI

base_url = "https://api.modelarts-maas.com/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY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="deepseek-v3", # 模型名称
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "你好"}
    ],
    stream = True
)

for chunk in response:
    if not chunk.choices:
        continue

    print(chunk.choices[0].delta.content, end="")
```

- cURL请求示例：

```
curl -X POST "https://api.modelarts-maas.com/v1/chat/completions" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $MAAS_API_KEY" \
-d '{
  "model": "deepseek-v3",
  "messages": [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "你好"}
  ],
  "stream": true,
  "stream_options": { "include_usage": true }
}'
```

DeepSeek-V3.1 文本生成（非流式）请求示例

- Rest API请求示例：

- Python请求示例：

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/chat/completions" # API地址
    api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key

    # Send request.
    headers = {
        'Content-Type': 'application/json',
```

```
'Authorization': f'Bearer {api_key}'
}
data = {
    "model": "deepseek-v3.1", # model参数
    "messages": [
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "你好"}
    ],
    "chat_template_kwargs": {
        "thinking": True # 是否开启深度思考模式，默认关闭
    }
}
response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

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

– cURL请求示例

```
curl -X POST "https://api.modelarts-maas.com/v1/chat/completions" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $MAAS_API_KEY" \
-d '{
    "model": "deepseek-v3.1",
    "messages": [
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "你好"}
    ],
    "chat_template_kwargs": {
        "thinking": true
    }
}'
```

• OpenAI SDK请求示例:

```
from openai import OpenAI

base_url = "https://api.modelarts-maas.com/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY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="deepseek-v3.1", # model参数
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "你好"},
    ],
    extra_body={
        "chat_template_kwargs": {
            "thinking": True # 是否开启深度思考模式，默认关闭
        }
    }
)

print(response.choices[0].message.content)
```

Qwen2.5-VL-7B 图像理解（非流式）请求示例

• Rest API请求示例:

– Python请求示例:

```
import requests
import json
import base64

# 图片转Base64编码格式
def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode("utf-8")
```

```
base64_image = encode_image("test.png")

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/chat/completions" # API地址
    api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key

    # Send request.
    headers = {
        'Content-Type': 'application/json',
        'Authorization': f'Bearer {api_key}'
    }
    data = {
        "model": "qwen2.5-vl-7b", # model参数
        "messages": [
            {
                "role": "user",
                "content": [
                    {
                        "type": "text",
                        "text": "描述下图片里的内容"
                    },
                    {
                        "type": "image_url",
                        # 需要注意，Base64，图像格式（即image/{format}）需要与支持的图片列表中的
                        # Content Type保持一致。"f"是字符串格式化的方法。
                        # PNG图像： f"data:image/png;base64,{base64_image}"
                        # JPEG图像： f"data:image/jpeg;base64,{base64_image}"
                        # WEBP图像： f"data:image/webp;base64,{base64_image}"
                        "image_url": {
                            "url": f"data:image/png;base64,{base64_image}"
                        }
                    }
                ]
            }
        ]
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

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

- cURL请求示例

```
curl -X POST "https://api.modelarts-maas.com/v1/chat/completions" \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $MAAS_API_KEY" \
-d '{
  "model": "qwen2.5-vl-72b",
  "messages": [
    {
      "role": "user",
      "content": [
        {
          "type": "text",
          "text": "描述下图片里的内容",
        },
        {
          "type": "image_url",
          "image_url": {
            "url": "data:image/png;base64,$BASE64_IMAGE"
          }
        }
      ]
    }
  ]
}'
```

• OpenAI SDK请求示例：

```
import base64
from openai import OpenAI

base_url = "https://api.modelarts-maas.com/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key

# 图片转Base64编码格式
def encode_image(image_path):
    with open(image_path, "rb") as image_file:
        return base64.b64encode(image_file.read()).decode("utf-8")
```

```
base64_image = encode_image("test.png")

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

response = client.chat.completions.create(
    model = "qwen2.5-vl-72b", # model参数
    messages = [
        {
            "role": "user",
            "content": [
                {"type": "text", "text": "描述下图片里的内容"},
                {
                    "type": "image_url",
                    # 需要注意, Base64, 图像格式 ( 即image/{format} ) 需要与支持的图片列表中的Content
                    # Type保持一致。"f"是字符串格式化的方法。
                    # PNG图像: f"data:image/png;base64,{base64_image}"
                    # JPEG图像: f"data:image/jpeg;base64,{base64_image}"
                    # WEBP图像: f"data:image/webp;base64,{base64_image}"
                    "image_url": {
                        "url": f"data:image/png;base64,{base64_image}"
                    }
                }
            ]
        }
    ]
)

print(response.choices[0].message.content)
```

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。
2. **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, 直接比较对应位数的数字即可)",
        "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, 但实际上并非如此, 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
  ]
}
```

4.7.2 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 4-18 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 4-19 响应参数

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

4.7.3 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 4-20 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。
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 .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 sensitive.....	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。

HTTP 状态码	错误码	错误信息	说明
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。
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。

HTTP 状态码	错误码	错误信息	说明
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。
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。

4.8 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://xxxxxxxxxxxxxxxxx")
# 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) 在线体验

5.1 在 ModelArts Studio (MaaS) 体验文本对话

在ModelArts Studio大模型即服务平台，运行中的模型服务可以在“文本对话”页面在线体验模型服务的推理效果。

约束限制

该功能仅支持中东-利雅得区域。

前提条件

已在“在线推理”页面创建自定义接入点。具体操作，请参见[在ModelArts Studio \(MaaS \) 创建自定义接入点](#)。

操作步骤

1. 登录[ModelArts Studio \(MaaS \) 控制台](#)，在顶部导航栏选择目标区域。
2. 任选以下方式进行模型体验。
 - 方式一
 - i. 在左侧导航栏中，选择“文本对话”。
 - ii. 在“文本对话”页面，单击“请选择模型服务”，在“自定义接入点”页签选择要体验的模型服务，单击“确定”。
 - 方式二
 - i. 在左侧导航栏中，选择“在线推理”。
 - ii. 在“在线推理”页面，单击“自定义接入点”页签，然后单击操作列的“在线体验”，进入“文本对话”页面。
3. 在“文本对话”右上角，单击“参数设置”，按需拖动或直接输入数值配置推理参数。单击“恢复默认”可以将参数值调回默认值。

图 5-1 设置推理参数

参数设置

温度/Temperature ②

0.7

012

核采样/top_p ②

1

0.10.551

top_k ②

20

15001000

恢复默认

表 5-1 参数设置

参数	说明
温度/Temperature	设置推理温度，用于控制生成文本的随机性和创造性，Temperature数值越大随机性越大。 - 数值较低，输出结果更加集中和确定。 - 数值较高，输出结果更加随机，更有创意性。 取值范围：0~2 默认值：不同模型的默认值不同，请以实际环境为准。
核采样/top_p	设置推理核采样，用于调整输出文本的多样性。top_p数值越大，生成文本的多样性就越高。 - 数值较低，输出可选的tokens类型越少，更有确定性。 - 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：0.1~1 默认值：不同模型的默认值不同，请以实际环境为准。 详细解释：top_p可以设置tokens候选列表的大小，将可能性之和刚好超过设定值P的top tokens列入候选名单，然后从候选名单中随机采样，生成一个token。

参数	说明
top_k	用于控制输出tokens的多样性。top_k值越大输出的tokens类型越丰富。选择在模型的输出结果中选择概率最高的前K个结果。 • 数值较低，输出可选的tokens类型越少，更有确定性。 • 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：1~1000 默认值：20 详细解释：top_k可以设置保留概率最高的前K个tokens，从中随机抽取一个token作为最终输出。这种方法可以限制输出序列的长度，并仍然保持样本的一定多样性。

4. 在对话框中输入问题或者使用控制台提供的推荐词，查看返回结果，在线体验模型服务。
- 模型输出内容不代表平台观点，平台不保证其合法性、真实性、准确性，不承担相关责任。

6 ModelArts Studio (MaaS) 模型管理

6.1 在 ModelArts Studio (MaaS) 创建模型

MaaS提供了开源大模型（DeepSeek、通义千问等），您可以使用这些基础模型，结合自定义的模型权重文件（权重类文件、词表类文件和配置类文件），创建个人专属的模型。创建成功的模型可以进行推理等操作。

操作场景

在当今数字化时代，人工智能应用愈发广泛。许多开发者和研究人员期望拥有个性化的大模型，用于各种特定场景，例如开发智能客服提升服务效率、辅助代码写作等。通常情况下，从头训练一个大模型需要大量的时间、计算资源和资金。多数开发者难以承担从头训练大模型的高昂成本，且技术门槛极高，涉及复杂的算法优化、海量数据处理等难题。

即使选择对开源模型进行微调，实际操作过程中仍存在阻碍，例如模型权重文件格式兼容性问题频发、本地训练的PyTorch权重文件与云平台不兼容，导致模型无法加载、不同模型的参数配置差异大等。

MaaS推出预置模型+自定义权重的全流程方案：

- 极简操作，快速适配：支持直接上传Hugging Face标准格式的权重文件，平台自动完成与芯片的算力适配，无需编写额外适配代码。
- 模板化配置，即开即用：内置DeepSeek、通义千问等主流模型的配置模板，用户无需手动调整复杂参数，大幅缩短模型开发周期。
- 弹性算力，高效运行：提供灵活的算力资源按需分配机制，可根据模型规模和业务需求动态调整算力，为业务高效运行提供强大保障。

为什么要创建我的模型

MaaS模型广场提供了丰富的基础模型，您可以直接使用这些模型进行部署模型服务等操作。当基础模型无法满足个性化需求时，您可以基于模型广场的模型创建专属的个性化模型，以实现更优的效果，同时便于版本管理和持续优化。

- 满足个性化需求：MaaS支持结合自定义权重文件，创建个人专属模型。模型广场预置模型是通用的，难以契合所有用户的特定需求，如企业需要将大模型应用于特定业务场景，预置模型因缺乏针对性难以满足需求，自定义模型可以凭借定制化权重文件实现个性化功能。

- 实现更好的效果：在某些复杂场景中，模型广场预置模型的表现可能不尽人意。例如在专业领域的对话问答、代码生成等场景，通过创建个人模型并修改权重配置，能优化模型运行效果，在专业任务处理上比预置模型更具优势。
- 便于版本管理和优化：MaaS提供模型版本管理功能，一个模型最多可支持创建10个版本。创建个人模型后，您可以通过新增版本不断优化模型，提升可追溯性。

计费说明

创建模型本身不收费，但使用过程中涉及的OBS存储、计算等资源会产生费用，详情请参见[ModelArts Studio \(MaaS \) 模型推理计费项](#)。

约束限制

该功能仅支持中国-香港区域。

用于生成专属模型的模型权重文件需要满足Hugging Face上的对应模型的文件格式要求。

- 模型权重文件夹下包括权重类文件、词表类文件和配置类文件。
- 可以使用transformers的from_pretrained方法对模型权重文件夹进行加载。

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](#)。

步骤一：准备权重配置文件

参考[Hugging Face官网](#)，准备好用于生成专属模型的模型权重文件。

说明

如果Hugging Face网站打不开，请在互联网上搜索解决方案。

步骤二：将权重配置文件上传至 OBS 桶

关于如何将权重文件存储到OBS桶，请参见[上传概述](#)。

单次上传本地文件到OBS的总大小不能超过5GB。如果需要上传超过5GB的大对象，可以使用OBS Browser+、obsutil工具上传，或使用OBS SDK及API的多段接口上传，上限为48.8TB，详情请参见[如何上传超过5GB的大对象](#)。

步骤三：创建我的模型

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“我的模型”。
3. 在“我的模型”页面右上角，单击“创建模型”。
4. 在“创建模型”页面，配置相关参数。

表 6-1 创建模型参数说明

参数	说明
来源模型	单击“选择基础模型”，在弹窗中选择模型，单击“确定”。 关于模型系列的详细介绍，请参见 在ModelArts Studio (MaaS) 模型广场查看预置模型 。
模型名称	自定义模型名称。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、下划线 (_)、中划线 (-) 和 (.)。
描述	自定义模型简介。最大支持100字符。
权重设置与词表	默认选择“自定义权重”。权重文件指的是模型的参数集合。
自定义权重存储路径	单击“自定义权重存储路径”右侧的文件图标，选择 步骤二 存放模型权重文件的OBS路径（必须选择到模型文件夹），然后单击“确定”。

5. 参数配置完成后，单击“创建”，创建自定义模型。
- 在模型列表，当模型“状态”变成“创建成功”时，表示模型创建完成。

步骤四：查看我的模型详情

- 模型创建完成后，您可以在“模型详情”页面查看模型的基本信息和版本信息。
- 在“我的模型”页面，单击目标模型名称，进入模型详情页面，查看模型的“基本信息”和“我的版本”。
- 基本信息：可以查看模型名称、模型ID、模型类型、来源模型、创建时间等信息。
 - 我的版本：可以查看已创建的模型版本，单击版本号进入“版本详情”页面，可以查看各个模型版本的详细信息和任务记录。
 - 版本信息：可以查看模型名称、状态、创建时间、基本模型及版本、权重与词表路径等信息。
 - 任务记录：可以查看任务名称、作业类型、状态、创建时间等信息。

（可选）步骤五：新增模型版本

- 为了提升模型的可追溯性和优化效率，MaaS提供了模型版本管理功能。通过此功能，您能够创建模型的新版本。一个模型最多支持创建10个版本。
1. 在[ModelArts Studio \(MaaS \) 控制台](#)左侧导航栏，单击“我的模型”进入模型列表。
 2. 单击目标模型名称，进入模型详情页面。
 3. 在“我的版本”区域，单击“新增版本”。
 4. 在“新增版本”页面，配置模型新版本的参数。

表 6-2 新增模型版本参数说明

参数	说明
新版本号	系统自动编号，不可修改。
版本描述	自定义模型版本简介。最大支持100字符。
选择基础模型版本	选择基础模型的版本。
选择权重路径	单击文件图标，选择 步骤二 存放模型权重文件的OBS路径（必须选择到模型文件夹），然后单击“确定”。

5. 配置完成后，单击“确定”，新增模型版本。
- 在版本列表，当新增版本的“状态”变成“创建成功”时，表示模型新版本创建完成。

（可选）步骤六：删除我的模型

当不需要模型时，可以进行删除操作。**删除操作无法恢复，请谨慎操作。**

1. 在[ModelArts Studio \(MaaS \) 控制台](#)左侧导航栏，选择“我的模型”进入模型列表。
2. 在模型列表，单击目标模型名称，进入“模型详情”页面。
3. 在“我的版本”区域，单击版本号，进入“版本详情”页面。查看该版本的模型“任务记录”是否为空。

– 是，表示模型未被用于训推任务，可以直接删除。则直接执行下一步。

– 否，表示模型已被用于训推任务，需要先删除所有任务，再执行下一步。

删除任务：单击操作列的“删除”，在“删除作业”对话框，输入“DELETE”，单击“确定”。
4. 确认该模型的各个版本的“任务记录”都为空。

当模型存在任务记录会删除失败。
5. 在“模型详情”页面，单击右上角的“删除”，在弹窗中输入“DELETE”，单击“确定”，删除模型。

当模型列表未显示该模型，表示删除成功。

后续操作

您可以将创建成功的模型进行部署，并调用模型服务。具体操作，请参见[调用 ModelArts Studio \(MaaS \) 部署的模型服务](#)。

常见问题

创建模型时，报错“Modelarts.6206:Key fields describing the model structure are missing from config.json, or their values are inconsistent with standard open source”如何处理？

您可以按照以下步骤进行排查：

1. 查看config.json文件是否存在。

2. 查看config.json文件格式是否符合要求。关于格式要求，请参见[Hugging Face官网](#)。

7 ModelArts Studio (MaaS) 管理与统计

7.1 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。

操作场景

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 7-1 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 7-1 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 \(MaaS \) 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，中国-香港区域的API Key必须通过中国-香港控制台创建，且仅能在该区域内调用服务。

相关文档

- 您可以使用已创建的API Key，在调用MaaS部署的模型服务时进行鉴权认证。具体操作，请参见[调用ModelArts Studio \(MaaS \) 部署的模型服务](#)。

- MaaS提供了基于MaaS DeepSeek API和Cherry Studio等第三方平台实现AI相关应用的最佳实践，帮助您快速实现AI应用落地。具体操作，请参见[使用ModelArts Studio \(MaaS \) DeepSeek API搭建AI应用](#)。

7.2 查看 ModelArts Studio (MaaS) 调用数据和监控指标

7.2.1 在 ModelArts Studio (MaaS) 查看在线推理的调用数据和监控指标

MaaS提供调用统计功能，支持查看我的服务、预置服务-商用服务、自定义接入点在指定时间段内的调用数据和监控指标详情，包括总调用次数、总调用失败次数、总调用Tokens数、输入Tokens数、输出Tokens数、端到端时延等信息，并以分钟为最小时间粒度展示数据趋势，帮助您了解服务的使用情况和性能变化，从而更有效地进行模型评估、问题定位、故障排除和性能优化。

操作场景

- 资源消耗监控：跟踪模型服务的Tokens使用量，避免超额使用。
- 成本分析：根据输入/输出Tokens的分布，优化调用策略以降低成本。
- 性能指标：支持查看模型的多种常见性能指标，进行性能优化。
- 服务优化：通过分析调用频率与Tokens消耗的关系，调整服务配置或扩容计划。
- 异常排查：快速定位特定时间段的调用量激增、异常消耗和调用失败问题。

约束限制

- 区域限制：
 - 中国-香港区域支持查看我的服务、预置服务-商用服务的调用数据和监控指标。
 - 中东-利雅得区域支持查看自定义接入点的调用数据和监控指标。
- 统计范围：
 - 仅统计预置服务-商用服务、自定义接入点、我的服务的调用数据。
 - 调用统计数据仅统计通过API接口调用产生的数据。
- 数据更新延迟：调用数据统计可能存在1~2小时的延迟，数据不能实时反映最新调用情况。
- 时间范围限制：
 - 支持预设时间段：今天、昨天、近3天、近7天、近14天。
 - 自定义时间段：最长不超过30天。

计费说明

- 调用统计功能本身不收费。
- 在MaaS进行模型调用时，可能涉及到相关资源收费。具体信息，请参见[模型推理计费项](#)。

前提条件

满足以下任一条件：

- 预置服务-商用服务：已开通预置服务的商用服务并产生调用记录。具体操作，请参见[在ModelArts Studio \(MaaS \) 预置服务中开通商用服务](#)。
- 自定义接入点：已创建自定义接入点并产生调用记录。具体操作，请参见[在ModelArts Studio \(MaaS \) 创建自定义接入点](#)。
- 我的服务：已在“我的服务”页面部署模型服务并产生调用记录。具体操作，请参见[使用ModelArts Studio \(MaaS \) 部署模型服务](#)。

查看服务调用的监控数据

在“调用统计”页面，您可以查看整体服务或单个服务通过API接口调用产生的数据详情。

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“管理与统计 > 调用统计”。
3. 在“调用统计”页面的“在线推理”页签，按需选择“时间范围”、“服务类型”、“调用方式”和“IP地址”。

表 7-2 调用统计筛选参数说明

参数		说明
时间范围		支持按照今天、昨天、近三天、近7天、近14天、自定义时间段统计服务的调用数据。 时间范围与时间精度过滤规则： ● 时间范围≤1天，支持精度：按分钟、按小时、按天。 ● 时间范围2-7天，支持精度：按小时、按天。 ● 时间范围8-30天，支持精度：按天。
服务类型	我的服务	在“我的服务”页面部署的模型服务。更多信息，请参见 使用ModelArts Studio (MaaS) 部署模型服务 。
	预置服务-商用服务	在“预置服务 > 商用服务”页签开通的商用服务。更多信息，请参见 在ModelArts Studio (MaaS) 预置服务中开通商用服务 。
	自定义接入点	在“自定义接入点”页签创建的接入点服务。更多信息，请参见 在ModelArts Studio (MaaS) 创建自定义接入点 。
调用方式		调用MaaS部署的模型服务时，使用API Key进行鉴权认证，默认为“全部API Key”，您也可以按需勾选API Key。更多信息，请参见 调用ModelArts Studio (MaaS) 部署的模型服务 和 在ModelArts Studio (MaaS) 管理API Key 。
IP地址		已产生调用量的客户端源IP地址（公网IP），来源于APIG日志中的http_x_forwarded_for字段值。当该字段包含多个值时，系统将采用第一个值；当字段值为-时，显示为空字符串。 IP地址默认显示为“全部”，您也可以按需勾选IP地址。

4. 在“在线推理”页签，查看整体服务的总调用次数、总调用失败次数、总调用Tokens数等信息。
监控指标默认保留三位小数。

表 7-3 整体服务的参数说明

参数	说明
总调用次数	服务的调用总次数。
总调用失败次数	服务的调用失败总次数，即4xx和5xx错误的总和。
总调用Tokens数	服务的调用总Tokens数。
输入Tokens数	服务的调用输入Tokens数。
输出Tokens数	服务的调用输出Tokens数。

5. 在“在线推理”页签的“服务列表”区域，查看单个服务的调用次数、调用失败次数、调用失败率等信息。
服务列表只显示已开通的预置服务-商用服务、已创建的自定义接入点或已部署成功的我的服务。

表 7-4 服务列表参数说明

参数	说明
服务名称/版本	调用服务的名称或版本。 仅商用服务支持服务版本。您可以单击  图标，查看服务各版本的统计信息。
调用次数	服务的调用次数。
调用失败次数	服务调用失败的次数。
调用失败率（ % ）	调用失败次数占调用总次数的比例。
调用总Tokens数（ 千 tokens ）	服务调用的总Tokens数。
输入Tokens数（ 千 tokens ）	输入的总Tokens数。
输出Tokens数（ 千 tokens ）	输出的总Tokens数。
端到端时延（ ms ）	单位时间内成功请求的端到端时延。
首Token时延（ ms ）	从接收请求到生成第一个输出Token所需的时间。
增量Token时延（ ms ）	生成后续每个输出Token所需的时间间隔。
平均生成时长（ s ）	平均生成每图片/视频实际花费的时间。

 说明

如果指标显示为“-”，表示服务不涉及该指标。“服务调用详情”的“监控”页签，仅显示服务涉及的指标。

6. 在“在线推理”页签的“服务列表”区域，单击目标服务右侧的“查看监控”，在“服务调用详情”页面的“监控”或“调用失败明细”页签查看调用相关信息。
- 在页面上方，您可以单击服务名称进行切换，也可以按需选择服务的版本（仅商用服务支持服务版本）。服务切换只显示已开通的预置服务-商用服务、已创建的自定义接入点或已部署成功的我的服务。
- “监控”页签：查看该服务的调用次数、调用失败率、输入Tokens大小、输出Tokens大小、端到端时延等变化趋势。

表 7-5 监控参数说明

参数		说明
筛选项	时间范围	默认为在“在线推理”页签选择的时间范围，您也可以按需修改。
	时间精度	时间精度与选择的时间范围有关，过滤规则如下： - 时间范围≤1天：支持按分钟、小时、天进行统计。 - 时间范围为2~7天：支持按小时、天进行统计。 - 时间范围为8~30天：支持按天进行统计。
	调用方式	默认为在“在线推理”页签选择的调用方式，您也可以按需修改。
	IP地址	默认为在“在线推理”页签选择的IP地址，您也可以按需修改。
监控指标	调用次数（次）	服务调用、成功、失败的次数。
	调用tokens量（千tokens）	单位时间内服务的调用总tokens数。

参数		说明
	首Token时延 (ms)	从接收请求到生成第一个输出Token所需的时间，仅统计流式响应。受限于模型版本约束，部分模型版本在非流式场景下不支持该指标展示，请将该服务的模型升级至最新版本后查看。关于升级模型服务的操作，请参见在ModelArts Studio (MaaS) 升级模型服务。 • AVG：首Token时延的平均值。 • MAX：首Token时延的最大值。 • P50：50%的首Token时延低于该值。 • P80：80%的首Token时延低于该值。 • P90：90%的首Token时延低于该值。 • P99：99%的首Token时延低于该值。
	输入Tokens大小 (千 tokens)	输入Token长度。 • AVG：输入Token长度的平均值。 • MAX：输入Token长度的最大值。 • P50：50%的输入Token长度低于该值。 • P80：80%的输入Token长度低于该值。 • P90：90%的输入Token长度低于该值。 • P99：99%的输入Token长度低于该值。
	RPM (次/分钟)	每分钟处理的请求数。
	调用失败率 (%)	调用失败次数占调用总次数的比例。
	错误发生次数	各错误码的发生次数。
	端到端时延 (ms)	单位时间内成功请求的端到端时延。 • AVG：端到端时延的平均值。 • MAX：端到端时延的最大值。 • P50：50%的端到端时延低于该值。 • P80：80%的端到端时延低于该值。 • P90：90%的端到端时延低于该值。 • P99：99%的端到端时延低于该值。

参数		说明
	增量Token时延 (ms)	生成后续每个输出Token所需的时间间隔， 仅统计流式响应 。受限于模型版本约束，部分模型版本在非流式场景下不支持该指标展示，请将该服务的模型升级至最新版本后查看。关于升级模型服务的操作，请参见 在ModelArts Studio (MaaS) 升级模型服务 。 • AVG：增量Token时延的平均值。 • MAX：增量Token时延的最大值。 • P50：50%的增量Token时延低于该值。 • P80：80%的增量Token时延低于该值。 • P90：90%的增量Token时延低于该值。 • P99：99%的增量Token时延低于该值。
	输出Tokens大小 (千 tokens)	输出Token长度。 • AVG：输出Token长度的平均值。 • MAX：输出Token长度的最大值。 • P50：50%的输出Token长度低于该值。 • P80：80%的输出Token长度低于该值。 • P90：90%的输出Token长度低于该值。 • P99：99%的输出Token长度低于该值。
	TPM (千 tokens/分钟)	每分钟处理的Tokens数 (输入+输出)。
	平均生成时长 (s)	平均生成每图片/视频实际花费的时间。

- “调用失败明细” 页签：查看调用失败的相关信息，如错误码、发生次数、错误信息等，进行问题定位和修复等。

表 7-6 调用失败明细参数说明

参数		说明
筛选项	时间范围	默认为在“在线推理”页签选择的时间范围，您也可以按需修改。
	调用方式	默认为在“在线推理”页签选择的调用方式，您也可以按需修改。
	IP地址	默认为在“在线推理”页签选择的IP地址，您也可以按需修改。
错误信息	错误码	报错的错误码，包含4xx和5xx。单击4xx或5xx前的  图标，可查看详细的错误码、发生次数、占比和错误信息。

参数		说明
	发生次数	4xx和5xx错误发生的次数。
	占比 (%)	该错误码发生次数占全部错误次数的比例。
	错误信息	4xx和5xx错误的描述信息。

导出服务调用的监控数据

- “服务调用详情”页面提供监控数据导出功能，支持导出所有或指定监控指标折线图对应的数据。
- 在“调用统计”页面的“在线推理”页签，在“服务列表”区域单击目标服务右侧的“查看监控”。
 - 在“服务调用详情”页面，按需选择“时间范围”、“服务类型”、“调用方式”和“IP地址”。
- 关于参数的说明，请参见[表7-5](#)。
- 在页面右上角，单击“导出”。
 - 在导出监控数据对话框，按需选择监控指标（默认为全选），然后单击“确定”。
- 导出的文件为.XLSX格式，每个页签对应一个监控指标折线图数据，由时间列和对应折线图的指标列组成。

常见问题

- 为什么调用了模型，但是查不到消耗Tokens数等信息？
由于数据更新存在延迟，消耗Tokens数等统计数据的更新延迟为小时级别，请耐心等待后再查询。
- 输入和输出Tokens的统计逻辑是什么？
 - 输入Tokens：用户请求中的文本经过分词后的Token总数。
 - 输出Tokens：模型响应结果的Token总数，包含终止符。

8 ModelArts Studio (MaaS) 业务最佳实践

8.1 使用 ModelArts Studio (MaaS) DeepSeek API 搭建 AI 应用

您可以使用MaaS DeepSeek API搭配Cherry Studio、Cursor等实现AI相关应用。

- Cherry Studio: 使用Cherry Studio调用部署在ModelArts Studio上的DeepSeek模型，构建个人AI助手。具体操作，请参见[基于ModelArts Studio \(MaaS \) DeepSeek API和Cherry Studio快速构建个人AI智能助手](#)。
- Cursor: 使用Cursor调用部署在ModelArts Studio上的DeepSeek模型，构建代码编辑器。具体操作，请参见[基于ModelArts Studio \(MaaS \) DeepSeek API和Cursor快速构建代码编辑器](#)。
- Cline: 使用Cline调用部署在ModelArts Studio上的DeepSeek模型，构建AI编程助手。具体操作，请参见[基于ModelArts Studio \(MaaS \) DeepSeek API和Cline快速构建AI编程助手](#)。