ModelArts Studio（MaaS）

用户指南

文档版本	01
发布日期	2025-10-27

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

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，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Llama、Baichuan、Yi、Qwen、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进行微调或直接应用，减轻模型集成的负担。

- **零代码、免配置、免调优模型开发**

  平台结合与100+客户适配、调优开源大模型的行业实践经验，沉淀了大量适配昇腾和调优推理参数的最佳实践。通过为客户提供一键式训练、自动超参调优等能力，和高度自动化的参数配置机制，使得模型优化过程不再依赖于手动尝试，显

著缩短了从模型开发到部署的周期，确保了模型在各类应用场景下的高性能表现，让客户能够更加聚焦于业务逻辑与创新应用的设计。

- **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，提供MCP服务，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“华东二”、“西南-贵阳一”和“华北-乌兰察布一”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
	体验免费服务	ModelArts Studio大模型即服务平台给用户提供了免费服务，无需部署即可一键体验预置模型服务。	在ModelArts Studio（MaaS）预置服务中体验免费服务

模块	操作	说明	相关文档
	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或者自定义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
在线体验	模型在线体验	您可以在“模型体验”页面，使用预置服务的商用服务、预置服务的免费服务或者自部署的模型服务进行功能体验。	ModelArts Studio（MaaS）在线体验
API调用	调用模型服务	您可以对预置服务的商用服务、预置服务的免费服务或者自部署的模型服务进行API调用。	调用ModelArts Studio（MaaS）部署的模型服务
模型管理	创建模型	ModelArts Studio提供了基于昇腾云算力适配的开源大模型，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调优、压缩、推理等操作。	在ModelArts Studio（MaaS）创建模型
	模型压缩	在ModelArts Studio大模型即服务平台支持对模型广场的预置模型或者自定义模型进行压缩，以此提升推理服务性能、降低部署成本。	使用ModelArts Studio（MaaS）压缩模型
模型训练	模型调优	完成数据集的准备后，可以在ModelArts Studio大模型即服务平台对模型广场的预置模型或者自定义模型进行调优。模型调优，即使用训练数据集和验证数据集训练模型。	使用ModelArts Studio（MaaS）调优模型

模块	操作	说明	相关文档
应用中心	管理应用	MaaS应用广场提供了多种AI原型应用，帮助您“一键复制”完成基础应用搭建。	• ModelArts Studio（MaaS）应用广场概述 • 在ModelArts Studio（MaaS）应用广场一键复制应用 • 在ModelArts Studio（MaaS）应用管理创建应用
	管理MCP服务	MaaS支持本地部署和云端部署MCP服务。 • 本地部署：不可以直接开通使用，仅提供元数据。您可以在“MCP广场”页面查看支持本地部署的MCP服务和JSON配置文件，然后在“MCP管理”页面通过NPX、UVX等方式进行部署。 • 云端部署：可以直接在“MCP广场”页面开通使用，包括MCP官方、三方平台以及MaaS云端部署的MCP服务，提供SSE访问方式。	• ModelArts Studio（MaaS）MCP概述 • 在ModelArts Studio（MaaS）MCP广场开通预置MCP服务 • 在ModelArts Studio（MaaS）创建自定义MCP服务
	应用体验	ModelArts Studio大模型即服务平台提供了MaaS应用体验中心，为具体的应用场景提供一整套解决方案。	在ModelArts Studio（MaaS）应用体验中心查看应用解决方案
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、预置服务的免费服务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。

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	策略类型	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	系统策略	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	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”。

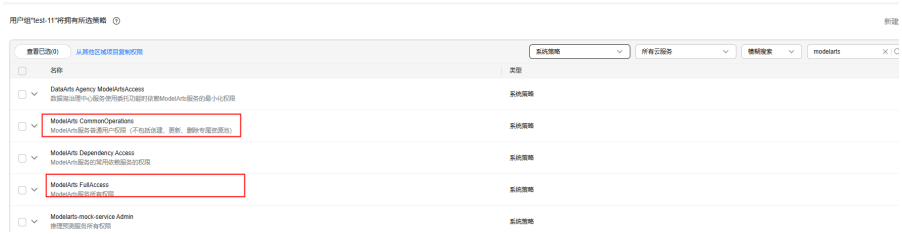
步骤2 配置用户组权限。

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. 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、调优或压缩模型时，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

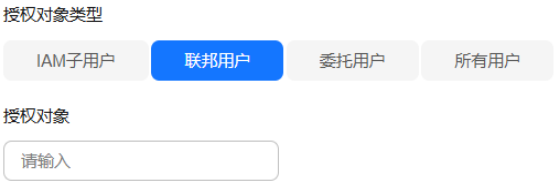
已注册华为账号并开通华为云，详情请见**注册华为账号并开通华为云**。

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  • 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  • 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  	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 	-

3. 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。

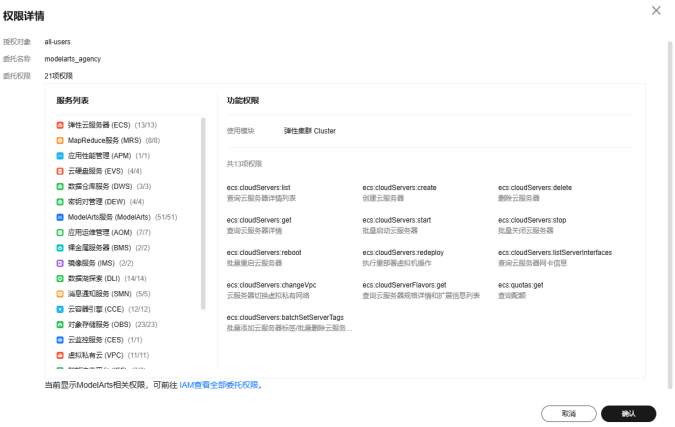
   也可前往“权限管理”页面**查看授权的权限列表**，查看已配置权限的权限详情。

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

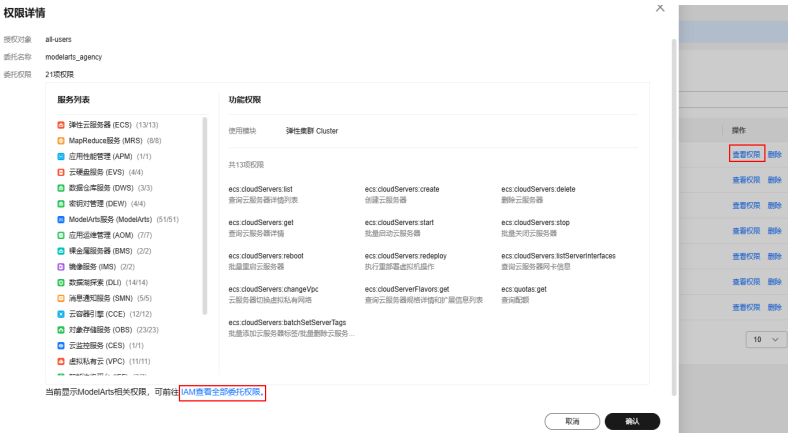
图 2-9 普通模式权限详情示例



修改授权的权限范围

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 2-10 去 IAM 修改委托权限



2. 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

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

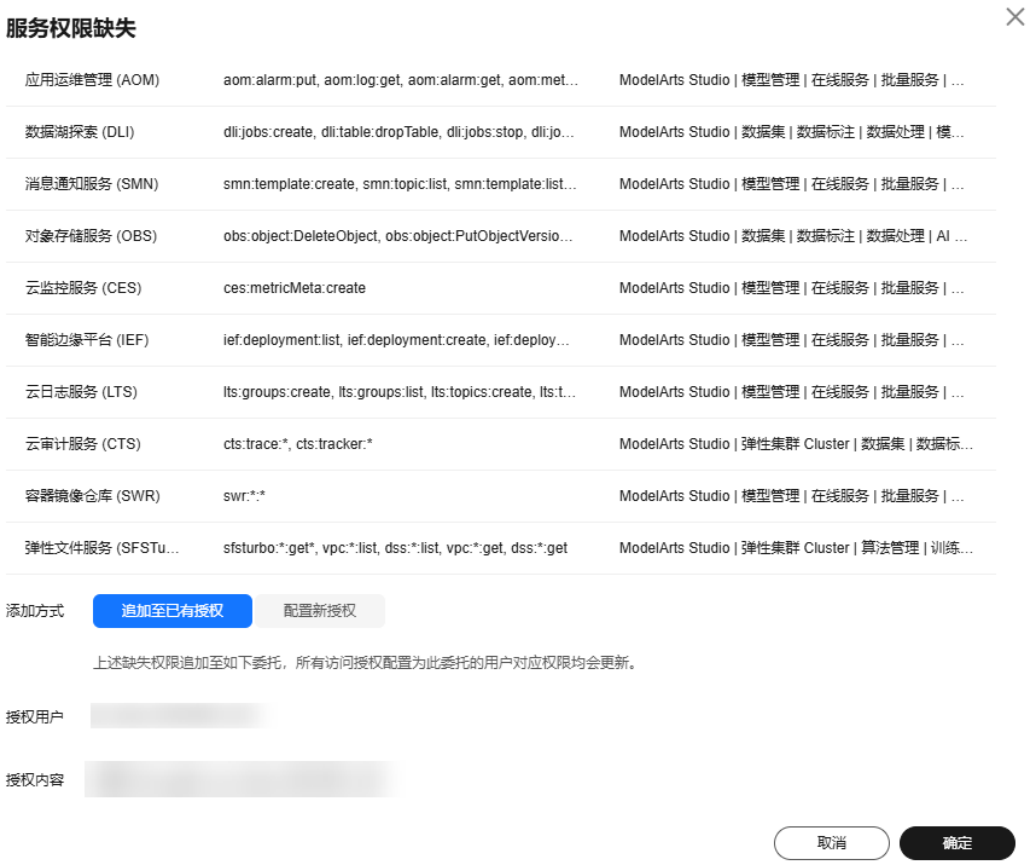
场景二：主用户添加缺失的权限

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

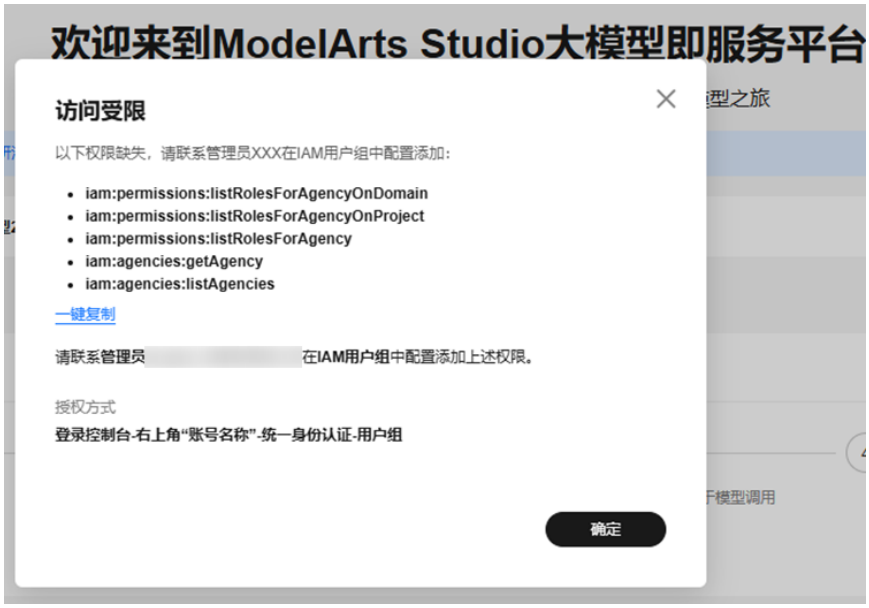
缺失权限的说明请参见**表2-3**。更多信息，请参见**授权项**。

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。

1. 创建自定义策略。
   a. 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

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

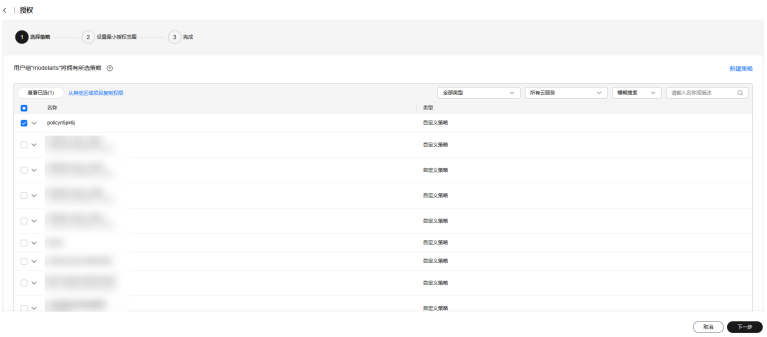
参数	说明	配置示例
策略内容	在Statement参数的[]中粘贴**步骤1.a**保存的权限策略，单击“格式化内容”。	{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["iam:permissions:listRolesForAgencyOnDomain", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:listRolesForAgency", "iam:agencies:getAgency", "iam:agencies:listAgencies"]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

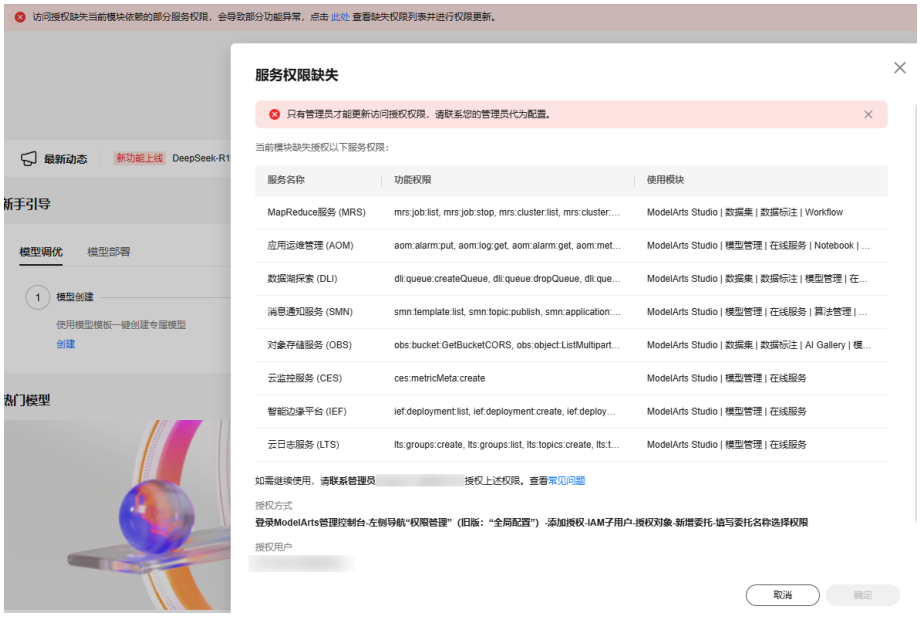
图 2-19 授权页面



   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OBS桶、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、调优或压缩模型时，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“华东二”、“西南-贵阳一”和“华北-乌兰察布一”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调优、压缩或部署时，需要选择资源池。MaaS服务支持专属资源池和公共资源池。

- 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**。

  须知

  MaaS服务只支持使用驱动版本是23.0.5的专属资源池，其他版本会导致任务失败。当专属资源池的驱动版本不适配时，可以参考**升级Standard专属资源池驱动**升级驱动。

- 公共资源池：公共资源池提供公共的大规模计算集群，根据用户作业参数分配使用，资源按作业隔离。MaaS服务可以使用ModelArts Standard形态下提供的公共资源池完成模型训推，按照使用量计费，方便快捷。

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、模型系列和支持作业进行筛选，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、图像理解、重排序和向量模型等进行筛选。 如果您同时选择了多个模型类型，页面会显示所选模型类型的合集。
上下文长度	支持按照64K、32K、16K、≤8K等进行筛选。 如果您同时选择了多个上下文长度，页面会显示所选上下文长度的模型合集。
高级能力	支持按照深度思考等多个能力进行筛选。

筛选项	说明
模型系列	支持按照DeepSeek、通义千问2、通义千问2.5、ChatGLM、Deepseek Coder等进行筛选。不同地域支持的模型系列不同，详情请参见**模型介绍**。 如果您同时选择了多个模型，页面会显示所选模型系列的合集。
支持作业	支持按照部署、调优进行筛选。 如果您同时选择了多个支持作业，页面会显示所选支持作业的模型交集，例如选择部署和调优，页面会显示同时支持部署与调优的模型。

4. 在“模型广场”页面的目标模型卡片，按需选择以下操作。

   模型卡片显示了模型的简要信息，例如模型介绍、模型类型、支持的能力、上下文长度、更新时间等信息。

   图 4-1 模型卡片示例

   - 鼠标悬浮于模型卡片，可以看到操作按钮，您可以按需单击“模型调优”、“模型部署”等。

     模型卡片上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。

     - 在线体验：
       - 未开通模型服务：单击“在线体验”，会弹出“开通模型服务”对话框，请仔细查看相关信息，勾选“我已阅读并同意上述说明，及《ModelArts Studio 服务声明》”，单击“确认开通”，跳转至“文本对话”页面进行在线体验。更多信息，请参见**在ModelArts Studio（MaaS）体验文本对话**。
       - 已开通模型服务：单击“在线体验”，会跳转至“文本对话”页面进行在线体验。
     - 模型部署：单击“模型部署”，会跳转至“部署模型服务”页面。具体操作，请参见**使用ModelArts Studio（MaaS）部署模型服务**。
     - 推理调用：
       - 未开通模型服务：单击“推理调用”，会弹出“调用说明”面板，在“开通模型服务”区域，仔细查看相关信息，勾选“我已阅读并同意上述说明，及《ModelArts Studio 服务声明》”，单击“立即

开通”，参照“调用模型区域”信息，调用模型服务。更多信息，请参见**调用ModelArts Studio（MaaS）部署的模型服务**。

- 已开通模型服务：单击“推理调用”，会弹出“调用说明”面板，参照“调用模型区域”信息，调用模型服务。

- 模型调优：单击“模型调优”，会跳转至“创建调优作业”页面。具体操作，请参见**使用ModelArts Studio（MaaS）调优模型**。

– 单击模型卡片，进入模型详情页面，可以查看模型的介绍、支持的版本、版本功能信息、备案信息等。不同的模型版本能力和操作可能不同，请以实际环境为准。

- 在页面右上角，您可以按需单击“模型部署”、“推理调用”等操作（部分操作支持选择版本），使用模型进行训推。

- 在版本卡片右侧，您可以按需单击“部署”、“推理调用”等操作，使用模型进行训推。

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	西南-贵阳一、华北-乌兰察布一、华东二	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	西南-贵阳一、华北-乌兰察布一	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-V3.1	文本生成	对话问答	中文、英文	西南-贵阳一	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-V3.2-Exp	文本生成	对话问答	中文、英文	西南-贵阳一	V3.2-Exp版本在V3.1-Terminus的基础上引入了DeepSeek稀疏注意力机制，探索并验证了针对长文本训练和推理效率的优化方法。
	DeepSeek-R1-Distill-Qwen-14B	文本生成	对话问答、文本生成推理	中文、英文	西南-贵阳一、华北-乌兰察布一	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	西南-贵阳一、华北-乌兰察布一	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。
ChatGLM	GLM-4	文本生成	对话问答、长文本推理、代码生成	中文、英文	西南-贵阳一、华东二	GLM-4-9B是智谱AI推出的最新一代预训练模型GLM-4系列中的开源版本。在语义、数学、推理、代码和知识等多方面的数据集测评中，GLM-4-9B及其人类偏好对齐的版本GLM-4-9B-Chat均表现出较高的性能。
	ChatGLM3	文本生成	对话问答、数学推理、代码生成	中文、英文	西南-贵阳一、华东二	ChatGLM3-6B是ChatGLM系列最新一代的开源模型，在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，ChatGLM3-6B引入了更强大的基础模型和更完整的功能支持。

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-Coder		文本生成	对话问答、文本推理	中文、英文	西南-贵阳一、华东二	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。
Yi		文本生成	代码生成、数学推理、对话问答	中文、英文	西南-贵阳一、华东二	Yi系列模型是01.AI从零训练的下一代开源大语言模型。Yi系列模型是一个双语的语言模型，在3T多语言语料库上训练而成，是全球最强大的大语言模型之一。Yi系列模型在语言认知、常识推理、阅读理解等方面表现优异。
通义千问	Qwen	文本生成	对话问答、智能创作、文本摘要、翻译、代码生成、数学推理	中文、英文	西南-贵阳一、华东二	通义千问-14B（Qwen-14B）是阿里云研发的通义千问大模型系列的140亿参数规模的模型。通义千问-72B（Qwen-72B）是阿里云研发的通义千问大模型系列的720亿参数规模的模型。通义千问-7B（Qwen-7B）是阿里云研发的通义千问大模型系列的70亿参数规模的模型。
	Qwen Image	图像生成	文生图	中文、英文	西南-贵阳一	Qwen-Image的图像生成与编辑通用能力强劲，在文本渲染的场景下表现出色。
	Qwen-Image-Edit	图像生成	文生图、图像编辑	中文、英文	西南-贵阳一	该模型是Qwen-Image的图像编辑版本，合入了其文本渲染能力，支持精准的图中文字修改。

模型系列		模型类型	应用场景	支持语言	支持地域	模型介绍
	QwQ	文本生成	对话问答	英文	西南-贵阳一	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1.5		文本生成	代码生成、数学推理、对话问答	中文、英文	西南-贵阳一、华东二	Qwen1.5是阿里云研发的通义千问大语言模型系列，包括不同模型大小的基础语言模型和对话聊天模型，可适应多种自然语言和代码。Qwen1.5版本开源了包括0.5B、1.8B、4B、7B、14B和72B在内的六种大小的基础和聊天模型，同时，也开源了量化模型。不仅提供了Int4和Int8的GPTQ模型，还有AWQ模型，以及GGUF量化模型。
通义千问2	Qwen2	文本生成	多语言处理、数学推理、对话问答	中文、英文	西南-贵阳一、华东二	Qwen2是阿里云研发的Qwen系列的新的大型语言模型。对于Qwen2，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，包括专家混合模型，并在一系列针对语言理解，语言生成，多语言能力，编码，数学，推理等的基准测试中表现出对专有模型的竞争力。
	Qwen2-VL	图像理解	图像理解、对话问答	中文、英文	西南-贵阳一、华东二	Qwen2-VL是阿里云推出的具有70亿参数的大型视觉语言模型，专注于图像和文本的多模态理解和生成任务。
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	西南-贵阳一、华东二	Qwen2.5是阿里云研发的Qwen系列的新的大型语言模型。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。

模型系列		模型类型	应用场景	支持语言	支持地域	模型介绍
	Qwen2.5-VL	图像理解	图像理解、对话问答	中文、英文	西南-贵阳一	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。
通义千问3	Qwen3	文本生成	对话问答	中文、英文	西南-贵阳一	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。
通义万相	Wan2.1-T2V	视频生成	文字生成视频	中文、英文	西南-贵阳一	Wan2.1-T2V系列模型在开源和闭源模型中建立了新的SOTA性能基准。在生成高质量且具有显著动态效果的视觉内容方面表现出色，支持中文和英文文本输入并支持480P和720P分辨率的视频生成。
	Wan2.1-I2V	视频生成	图片生成视频	中文、英文	西南-贵阳一	Wan2.1模型在生成高质量、有显著动态效果的视频方面表现优异，支持中文和英文。
	Wan2.2-T2V	视频生成	文字生成视频	中文、英文	西南-贵阳一	该文生视频模型采用混合专家（MoE）架构，视频合成更加稳定，支持了更多样化的风格场景。
	Wan2.2-I2V	视频生成	图片生成视频	中文、英文	西南-贵阳一	Wan2.2内置丰富的美学数据集，可轻松定制个性化的电影级画面。
Kimi	Kimi-K2	文本生成	对话问答	中文、英文	西南-贵阳一	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。

模型系列		模型类型	应用场景	支持语言	支持地域	模型介绍
BGE	bge-m3	向量模型	文本向量化	中文、英文	西南-贵阳一	BGE-M3以其在多语言、多功能和多粒度方面的灵活性而著称。它为超过100种工作语言的语义检索提供了统一的支持，可以同时完成三种常见的检索功能：密集检索、多向量检索和稀疏检索。此外，它还能够处理不同粒度的输入，从短句子到长达8192个token的长文档。
	bge-reranker-v2-m3	重排序	检索结果再排序	中文、英文	西南-贵阳一	一个轻量级的交叉编码器模型，基于BGE-M3模型开发，具有强大的多语言能力，易于部署，具有快速的推理能力。

模型分为量化模型和非量化模型，其中，量化模型又包括SmoothQuant-W8A8和AWQ-W4A16两种。

- AWQ-W4A16量化模型可以由非量化模型压缩后生成，也可以直接使用开源AWQ权重。
- SmoothQuant-W8A8量化模型只能由非量化模型压缩生成。

ModelArts Studio大模型即服务平台已预置非量化模型与AWQ-W4A16量化模型的模型模板。

- 非量化模型可以支持调优、压缩、部署等操作。
- 量化模型仅支持部署操作。当需要获取SmoothQuant-W8A8量化模型时，则可以通过对非量化模型进行模型压缩获得。

4.2 在 ModelArts Studio（MaaS）预置服务中体验免费服务

MaaS预置服务即ModelArts Studio平台已部署好的服务，无需部署即可一键体验预置模型服务。

免费服务说明

免费服务仅提供基础体验能力，且存在严格的速率限制。平台可能会不定期调整其适用模型、免费额度、有效期等内容。请以实际环境为准。

免费服务支持以下两种模式：

- 模式一：免费服务的额度无限制，无需领取即可使用，有失效时间，在失效时间之前可免费使用。失效时间可以在“预置服务 > 免费服务”页签的“失效时间”列查看。
- 模式二：免费服务的额度有限制，需要领取后使用。

使用免费服务进行推理

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“预置服务”页签，单击“免费服务”页签。
3. 在“免费服务”页签，任选以下方式免费使用预置服务。
   - 方式一：在业务环境中调用模型服务的API进行推理。

     在目标服务右侧，单击操作列的“调用说明”，在“调用说明”页面获取调用示例，在业务环境中调用API进行体验。操作指导请参见**调用ModelArts Studio（MaaS）部署的模型服务**。

     - 预置服务默认启用内容审核，且“调用说明”页面不显示该参数。
     - 当您调用模型服务的API，返回状态码“429 Too Many Requests”时，表示请求超过流控，请稍后重新调用。
   - 方式二：在“文本对话”页面进行推理。

     在目标服务右侧，单击“操作”列的“在线体验”，跳转到“文本对话”页面，开始问答体验。操作指导请参见**在ModelArts Studio（MaaS）体验文本对话**。

图 4-2 修改参数

您可以在右上角单击“参数设置”，按需修改相关参数，以获取更好的推理效果。

表 4-3 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。

参数	说明
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。
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。

后续操作

- 查看免费服务调用数据：在“免费服务”页签，单击“调用统计”列的图标，可以查看目标服务的调用次数、Tokens数、首Token时延等指标信息。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**。
- 当免费Token额度用完或者免费服务失效，您可以部署为我的服务付费使用，或开通商用服务付费使用。
  - 部署为我的服务付费使用：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”，进行相关配置。操作指导请参见**使用ModelArts Studio（MaaS）部署模型服务**。

    模型服务部署成功后，可以使用我的服务进行体验或调用等操作。具体操作，请参见**在MaaS体验模型服务**和**调用MaaS部署的模型服务**。
  - 开通商用服务付费使用：在“预置服务 > 商用服务”页签，开通商用服务。操作指导请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。

    开通商用服务后，可以使用预置服务进行体验或调用等操作。具体操作，请参见**在MaaS体验模型服务**和**调用MaaS部署的模型服务**。

4.3 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、广告公司借助推理API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。

免费服务与商用服务的区别

- 免费服务：仅提供基础体验能力，且存在严格的速率限制。平台可能会不定期调整其适用模型、免费额度、有效期等内容，请以实际环境为准。免费服务仅适用于体验模型。
- 商用服务：提供商用级别的API推理服务，开通后您可以获取付费API服务。商用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。

约束限制

- 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**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”。

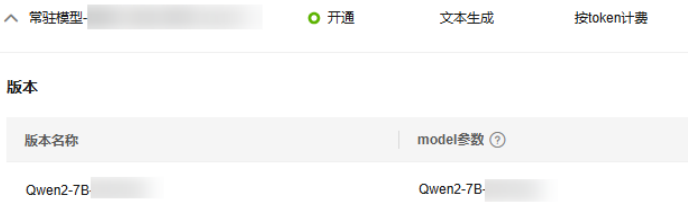
   勾选目标商用服务后，会自动开通该服务下的所有模型版本。

图 4-3 开通预置模型服务

已开通的商用服务示例如下。商用服务列表的参数说明请参见下表。

图 4-4 已开通商用服务

表 4-4 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 4-5** 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。
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**。 • 在线体验：单击“在线体验”，选择服务版本，跳转至模型对应的体验页面，进行在线体验。更多信息，请参见**ModelArts Studio（MaaS）在线体验**。

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

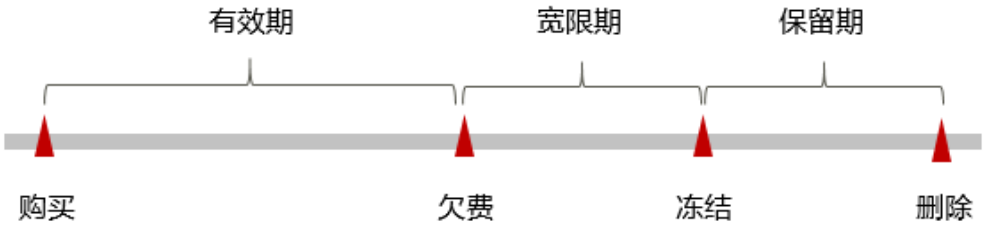
当您使用某个模型服务欠费后，对应资源实例不会立即停止服务，资源进入宽限期。您需支付按需资源在宽限期内产生的费用，相关费用可在**管理控制台** > 费用中心 > 总览“欠费金额”查看，华为云将在您充值时自动扣取欠费金额。此时不会冻结资源， 只会影响用户开通新资源、开通新服务。 已有资源可正常使用。

如果您在宽限期内仍未支付欠款，特定资源会触发欠费冻结，进入保留期，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。

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。对应模型的付费状态变为未开通。

欠费后请您及时充值，详细操作请参见**账户充值**。

图 4-6 按需计费资源生命周期



说明

华为云根据**客户等级**定义了不同客户的宽限期和保留期时长。

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**账户充值**。

4.4 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

在企业和开发者的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。

商用服务、免费服务与自定义接入点的区别

- 免费服务：仅提供基础体验能力，且存在严格的速率限制。平台可能会不定期调整其适用模型、免费额度、有效期等内容，请以实际环境为准。免费服务仅适用于体验模型。
- 商用服务：提供商用级别的API推理服务，开通后您可以获取付费API服务。商用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。
- 自定义接入点：可根据业务需求创建自定义接入点。自定义接入点支持独立设置流控、独立出账及独立监控能力。

约束限制

- 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。

调用在线推理-预置服务中的商用服务：按Token计费，计费模式与所选基础模型的计费模式一致。关于计费详情，请参见**ModelArts Studio（MaaS）模型推理计费项**。

前提条件

已在MaaS开通预置服务中的商用服务。具体操作，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。

创建自定义接入点

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”。
3. 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”。
4. 在“创建自定义接入点”面板，配置相关参数。

表 4-5 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“商用服务”。 商用服务：在线推理-预置服务中的商用服务。
模型服务	“服务来源”为“商用服务”：单击“选择模型服务”，在“选择模型服务”对话框，按需选择模型服务的版本，单击“确定”。 默认支持预置服务中的全部商用服务（开通和未开通），免费服务不支持。
模型限流	仅“模型来源”选择商用模型，显示该参数。 选择商用模型后，会显示当前账号下访问该模型服务的总限流。 • RPM（Requests Per Minute）：每分钟处理的请求数。 • TPM（Tokens Per Minute）：每分钟处理的Tokens数（输入+输出）。
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。

5. 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

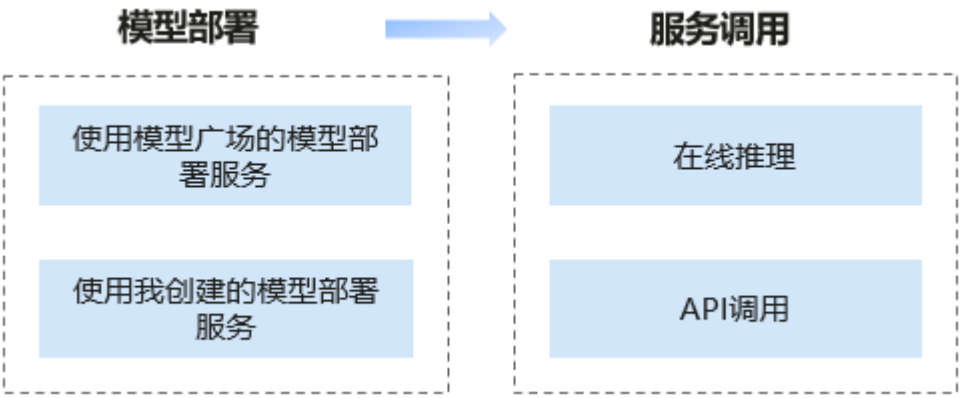
- **创建的自定义接入点数量达到上限怎么办?**
  您可以删除不再使用的接入点，然后新建接入点。
- **如何确定Tokens的消耗数量?**
  您可以通过以下两种方式查看Tokens的消耗数量。
  - 通过“调用统计”页面查看模型服务调用的总Tokens数、输入Tokens数、输出Tokens数等信息，详情请参见**查看自定义接入点的调用统计**。
  - 在费用中心通过自定义接入点名称查询账单详情（该方式仅支持商用模型接入点）。账单中会显示接入点的输入Tokens数、输出Tokens数等信息。
- **修改自定义接入点的限流设置后，多久会生效?**
  修改保存后，限流设置会立即生效，后续调用将按照新规则执行。

4.5 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“模型体验”或其他业务环境中可以调用。

当模型广场的模型无法满足您的个性化需求时，您可以基于模型广场的模型，创建个人专属模型，部署为我的服务。

图 4-7 部署模型服务使用流程



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

- ModelArts Studio大模型即服务平台的模型推理的最大输入输出长度如下表所示。

  说明

  不同地域支持的模型可能不同，请以实际环境为准。

  表 4-6 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）
Qwen Image Qwen-Image-Edit	输入：2000（输入超过800 tokens会被截断）

模型	默认最大输入输出长度（token）
Wan2.1-T2V-14B Wan2.1-T2V-1.3B	输入：4096
Wan2.2-I2V-A14B Wan2.2-T2V-A14B	输入：1000
Qwen-14B	2048
Qwen2.5-72B-8K DeepSeek-V3-8K DeepSeek-R1-Distill-Qwen-14B-8K DeepSeek-R1-Distill-Qwen-32B-8K BGE-M3 BGE-Reranker-V2-M3	8192
DeepSeek-R1-16K DeepSeek-V3-16K QwQ-32B-16K	16384
QwQ-32B-32K Qwen2-72B-32K Qwen2.5-7B-32K Qwen2.5-32B-32K Qwen2.5-72B-32K Qwen2.5-VL-7B-32K Qwen2.5-VL-72B-32K Qwen3-4B-32K Qwen3-8B-32K Qwen3-14B-32K Qwen3-32B-32K Qwen3-235B-A22B-32K DeepSeek-R1-32K DeepSeek-R1-32K-0528 DeepSeek-R1-Distill-Qwen-32B-32K DeepSeek-V3-32K Deepseek-Coder-33B-32K	32768

模型	默认最大输入输出长度（token）
Qwen3-235B-A22B-64K Qwen3-32B-64K DeepSeek-V3-64K Deepseek-V3.1-64K DeepSeek-V3.2-Exp DeepSeek-R1-64K DeepSeek-R1-64K-0528 Kimi-K2	65536
DeepSeek-V3.1-128K Qwen3-30B-A3B-128K Qwen3-Coder-480B-A35B	131072
其他模型	4096

- 如果不支持公共资源池，“公共资源池”按钮会置灰，鼠标悬停时，会提示：该模型版本暂不支持公共资源池部署；如果专属资源池不匹配，勾选按钮会置灰，鼠标悬停时，会出现相关提示，请按照提示进行相关操作。

  使用历史模型在专属资源池部署时，驱动版本需为23.0.5或23.0.6；使用DeepSeek模型新版本时，驱动版本需为24.0.1。如果驱动版本不正确会导致部署任务创建失败。

前提条件

- 已准备公共资源池或专属资源池，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 4-7 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。

参数		说明
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	资源池分为公共资源池与专属资源池。 ● 公共资源池供由所有租户共享使用。 – 如果支持公共资源池，但是没开白名单，“资源池类型”选择“公共资源池”时，下方会出现提示：公共资源池暂未完全公开，如需申请使用，请联系与您对接的销售人员或拨打4000-955-988获得支持，您也可以在线提交**售前咨询**。 – 如果不支持公共资源池，“公共资源池”按钮会置灰，鼠标悬停时，会提示：该模型版本暂不支持公共资源池部署；如果专属资源池不匹配，勾选按钮会置灰，鼠标悬停时，会出现相关提示，请按照提示进行相关操作。 ● 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	内容审核	选择是否打开内容审核，默认启用。 ● 开关打开（默认打开），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。 ● 开关关闭，停用内容审核服务，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，勾选后，单击“确定”关闭。

参数		说明
	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。 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**。 • “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** 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	设定服务在运行指定时间后自动停止。 • 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”。

   “资源池类型”选择“公共资源池”时，会出现“计费提醒”对话框，请您仔细阅读预估费用信息，然后单击“确定”，创建部署任务。模型部署会基于资源占用时长进行计费。服务状态为运行中时会产生费用，最终实际费用以账单为准。

   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“公共资源池”时，模型部署会基于资源占用时长进行计费。

   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，可以进行在线体验或API调用。具体操作，请参见**在ModelArts Studio（MaaS）体验文本对话**或**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 4-8 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**。
   - 在线体验：单击“在线体验”，进行在线文本对话。详细信息，请参见**在ModelArts Studio（MaaS）体验文本对话**。

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**。
- 在线体验模型请参考**在ModelArts Studio（MaaS）体验文本对话**。
- API调用请参考**调用ModelArts Studio（MaaS）部署的模型服务**。
- 如果模型服务部署失败，您可以参考**ModelArts Studio（MaaS）模型服务部署失败，报错：jod failed: real time create service failed**进行定位。

4.6 在 ModelArts Studio（MaaS）管理我的服务

4.6.1 在 ModelArts Studio（MaaS）启动/停止/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，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。

定时启停部署服务

华东二和华北-乌兰察布一支持通过**FunctionGraph控制台**实现定时启停，西南-贵阳一支持调用接口实现启停，请您按需选择以下步骤。

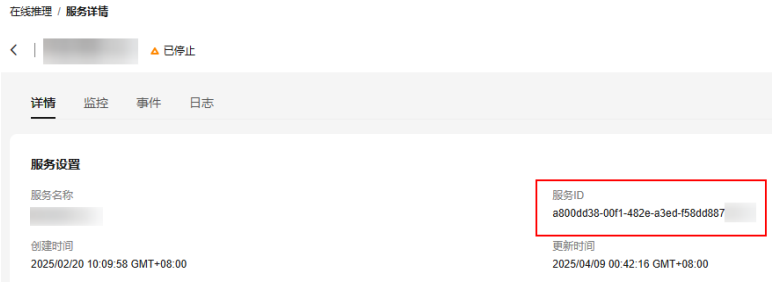
华东二和华北-乌兰察布一：

华为云函数工作流FunctionGraph提供定时触发器，可以帮助用户实现ModelArts Studio定时批量启停的计划，适用于需要通过停止不使用的实例并在需要使用实例时自动启动实例，来帮助降低运营成本的场景。更多信息，请参见**定时开关机解决方案概述**和**资源和成本规划**。

1. 创建rf_admin_trust委托和IAM Agency Management FullAcces策略，为rf_admin_trust委托添加IAM Agency Management FullAcces策略。具体操作，请参见**准备工作**。
2. 获取ModelArts Studio模型服务ID。
   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，然后单击目标服务名称。
d. 在“服务详情”页面，获取服务ID。

图 4-8 获取服务 ID



3. 登录**华为云解决方案实践**，在文本框搜索**定时开关机**，在“定时开关机”卡片，单击“一键部署”，跳转至“立即创建资源栈”页面，在顶部导航栏选择目标区域，进行定时开关机相关配置。具体操作，请参见**快速部署**。

   关于如何查看函数、编辑环境变量、查看执行日志的具体操作，请参见**开始使用**。

图 4-9 一键部署定时开关机

4. 消息通知服务SMN会自动发送受邀订阅主题链接的短信，您可以单击访问链接，使用浏览器打开即可确认订阅。

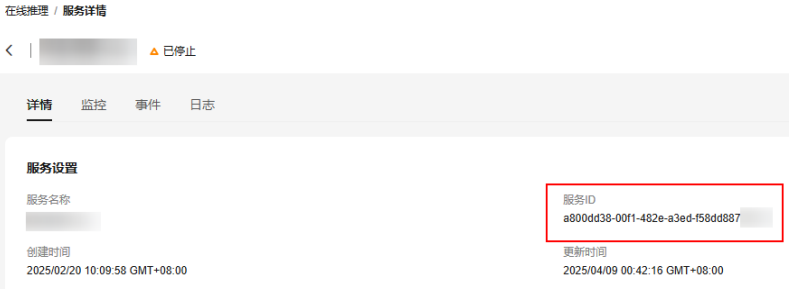
5. （可选）如果不需要定时启停功能，可以进行卸载。具体操作，请参见**快速卸载**。

西南-贵阳一：

MaaS支持调用接口实现启停功能，适用于需要通过停止不使用的实例并在需要使用实例时自动启动实例，来帮助降低运营成本的场景。您可以参考以下示例代码，按需修改相关参数，创建自己的启停任务。

1. 通过IAM获取token和project_id。
   - 关于如何获取token，请参见**获取IAM用户Token（使用密码）**。
   - 关于如何获取project_id，请参见**获取账号、IAM用户、项目、用户组、区域、委托的名称和ID**。
2. 获取ModelArts Studio模型服务ID（maas_server_id）。
   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，然后单击目标服务名称。
   d. 在“服务详情”页面，获取服务ID。

   图 4-10 获取服务 ID

3. 使用以下代码示例，调用get_server和post_server方法。
   - get_server方法：用于确认ModelArts服务状态。通过调用该方法，可以获取指定服务的当前状态。
   - post_server方法：用于控制ModelArts服务状态。通过指定操作类型（action_type），可以对服务进行启动或停止操作。action_type参数说明如下：
     - restart：重启服务。
     - terminate：停止服务。

   代码示例如下，请您根据实际情况进行修改。{token}、{project_id}、{maas_server_id}请替换为前两个步骤获取的值。

```
import http.client
import json
import traceback


def get_server(project_id, maas_server_id, token):    //用于确认ModelArts服务状态。通过调用该方法，可以获取指定服务的当前状态。
    try:
        conn = http.client.HTTPSConnection(f"modelarts.cn-southwest-2.myhuaweicloud.com")
        headers = {'X-Auth-Token': f"{token}"}

        conn.request("GET", f"/v1/{project_id}/maas/services/{maas_server_id}", None, headers)
```



```
        res = conn.getresponse()
        status = res.read().decode("utf-8")
        print("status", status)
    except Exception:
        print("Failed to create service,"
              f"exception: {traceback.format_exc()}")


def post_server(project_id, maas_server_id, token, action_type):    //用于控制ModelArts服务状态。通过指定操作类型（action_type），可以对服务进行启动或停止操作。
    try:
        conn = http.client.HTTPSConnection("modelarts.cn-southwest-2.myhuaweicloud.com")
        headers = {'X-Auth-Token': f"{token}", 'Content-Type': 'application/json'}

        body = json.dumps({
            "action_type": action_type      //restart：重启服务；terminate：停止服务。
        })

        conn.request("POST", f"/v1/{project_id}/maas/services/{maas_server_id}", body, headers)
        res = conn.getresponse()
        status = res.read().decode("utf-8")
        print("status", status)
    except Exception:
        print("Failed to create service,"
              f"exception: {traceback.format_exc()}")
```

在get_server方法中，通过GET请求获取服务状态。状态信息存储在status变量中，并通过print("服务状态:", status)输出。服务状态反映了服务的当前运行情况，以下是所有可能的状态：

- Creating：创建中
- Initing：初始化中
- Pending：等待中
- Waiting：等待中
- Deploying：部署中
- Running：运行中
- Concerning：关注中
- Failed：失败
- Completed：已完成
- Terminating：终止中
- Terminated：已终止
- Deleting：删除中
- Deleted：已删除
- Unknown：未知
- Abnormal：异常
- Restarting：重启中
- Upgrade：升级
- Scale：扩缩容

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。

4.6.2 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。

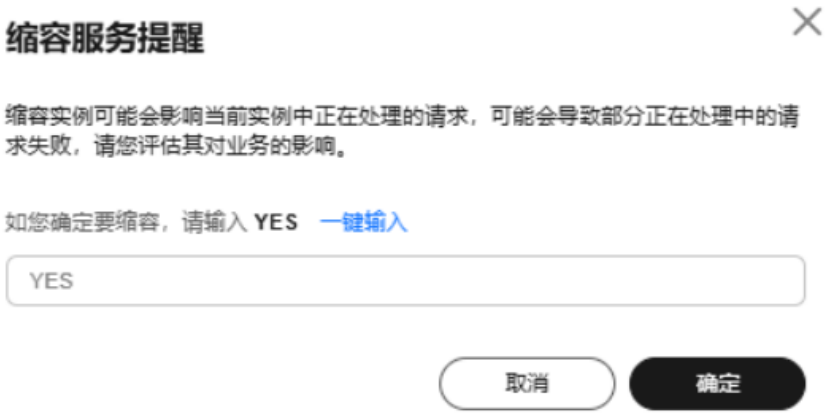
计费说明

扩容模型服务实例数后，在调用MaaS预置服务时，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

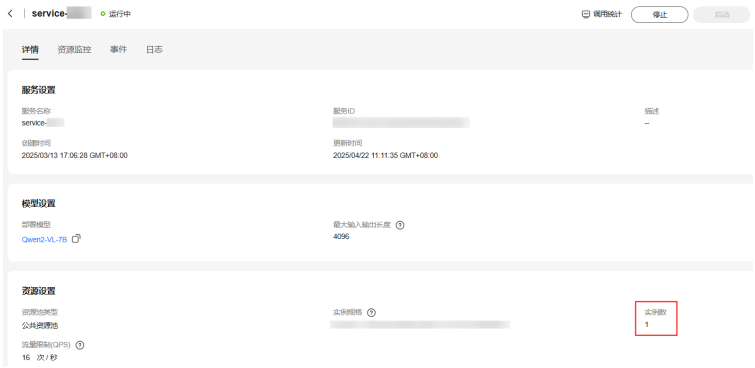
1. 登录**ModelArts Studio（MaaS）控制台**，在顶部导航栏选择目标区域。
2. 在左侧导航栏，选择“在线推理”。
3. 在“在线推理”页面，单击“我的服务”页签，在目标模型服务右侧，单击操作列的“更多 > 扩缩容”，进入扩缩容页面。
4. 在“扩缩容”页面，按需选择以下操作。
   - **扩容**：按需增加“变更后实例数”，单击“确定”，在“扩缩容服务”对话框，单击“确定”。
   - **缩容**：按需减少“变更后实例数”，单击“确定”，在“缩容服务提醒”对话框，查看提示信息，确认无误后输入YES，单击“确定”。

图 4-11 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。

图 4-12 查看实例数



后续操作

- **模型体验**：模型服务扩缩容后，可以在“模型体验”调用该模型服务进行功能体验。
- **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。

4.6.3 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。

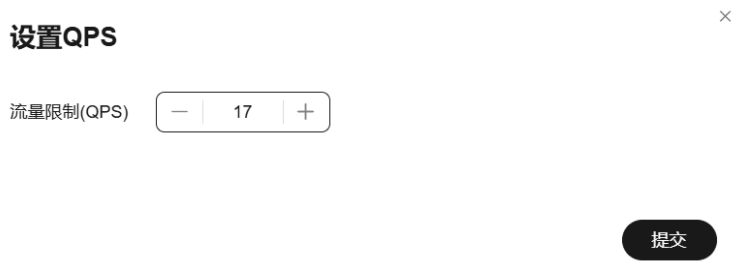
约束限制

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，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。

图 4-13 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。

4.6.4 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，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”。

图 4-14 服务升级

后续操作

- **模型体验**：模型服务扩缩容后，可以在“模型体验”调用该模型服务进行功能体验。
- **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。

4.7 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**。

约束限制

对于支持图片上传的模型，单个图片文件的大小不超过10MB。如果以Base64编码形式上传图片，需确保编码后的图片小于10MB。

前提条件

- 使用预置服务：在“在线推理”页面的“预置服务”页签，使用有效期内的免费服务或者已开通商用服务（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**。

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 4-9 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4-10**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 4-10 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。

参数	是否必选	默认值	参数类型	描述
messages.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"} • 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"}
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	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	是否必选	默认值	参数类型	描述
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。
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≤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]。

参数	是否必选	默认值	参数类型	描述
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 4-11 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。

参数	参数类型	描述
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 4-12 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。

内容审核说明

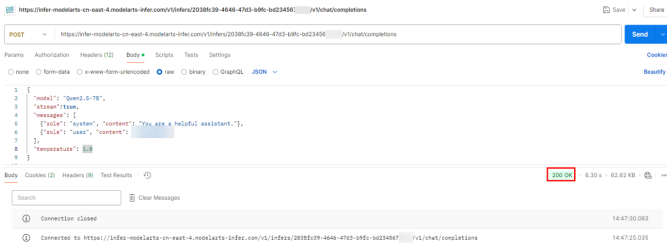
- **流式请求**
  - 如果触发内容审核，则会返回错误：错误码403。您可以通过错误码ModelArts.81011来判断。返回内容如下：

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

图 4-15 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 4-16 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"。

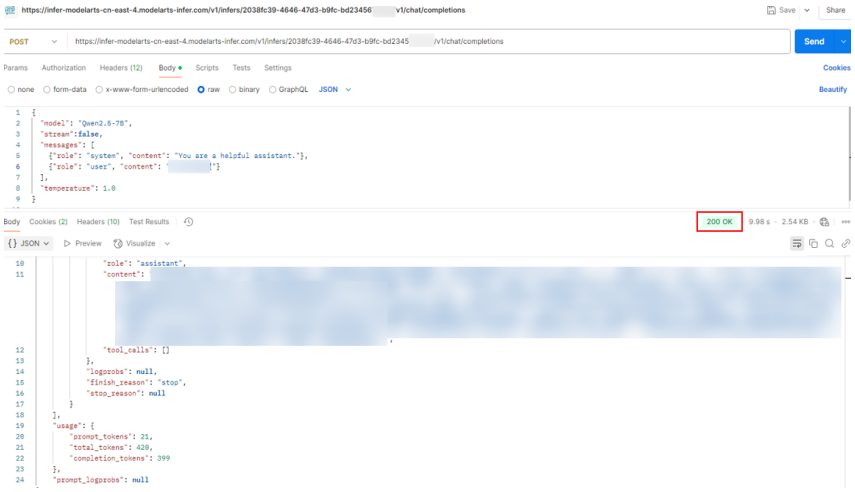
- **非流式请求**
  - 如果触发内容审核，则会返回错误：错误码403。您可以通过错误码ModelArts.81011来判断。

    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 4-17 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 4-13 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，例如"GLM-4-9B"。 • 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	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
- 在ModelArts Studio（MaaS）查看在线推理的调用数据和监控指标

4.8 ModelArts Studio（MaaS）API 调用规范

4.8.1 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。

约束限制

对于支持图片上传的模型，单个图片文件的大小不超过10MB。如果以Base64编码形式上传图片，需确保编码后的图片小于10MB。

接口信息

表 4-14 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-V3-64K	西南-贵阳一	DeepSeek-V3	65536	支持
	DeepSeek-R1-64K	西南-贵阳一	DeepSeek-R1	65536	支持

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-R1-64K-0528	西南-贵阳一	deepseek-r1-250528	65536	支持
	DeepSeek-V3.1	西南-贵阳一	deepseek-v3.1	131072	支持
	DeepSeek-V3.2-Exp	西南-贵阳一	deepseek-v3.2-exp	65536	支持
Qwen2.5	Qwen2.5-VL-7B-32K	西南-贵阳一	qwen2.5-vl-7b	32768	不支持
	Qwen2.5-VL-72B-32K	西南-贵阳一	qwen2.5-vl-72b	32768	不支持
Qwen3	Qwen3-32B-32K	西南-贵阳一	qwen3-32b	32768	不支持
	Qwen3-235B-A22B-32K	西南-贵阳一	qwen3-235b-a22b	32768	不支持
	Qwen3-30B-A3B	西南-贵阳一	qwen3-30b-a3b	131072	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、DeepSeek-V3.2-Exp和Qwen3-30B-A3B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。

DeepSeek-V3.2-Exp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。
- 不截断思维链只截断content能力不生效。
- 不支持前缀续写、guided_choice能力。

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 4-15 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4-14。
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4-16。
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4-17。
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	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	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≤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。

chat_template_kwargs.thinking	否	false	Bool	默认关闭思维链。**仅支持DeepSeek-V3.1和DeepSeek-V3.2-Exp模型**，约束限制请参见**思维链说明**。
				开启思维链示例如下：
				{
				"model": "deepseek-v3.1",
				"messages": [{
				"role": "system",
				"content": "You are a helpful assistant."
				}, {
				"role": "user",
				"content": "你好"
				}],
				"chat_template_kwargs": {
				"thinking": true
				}
				}

chat_template_kwargs.enable_thinking	否	true	Bool	默认开启思维链。**仅支持Qwen3-30B-A3B模型。**
				关闭思维链示例如下：
				{
				"model": "qwen3-30b-a3b",
				"messages": [{
				"role": "system",
				"content": "You are a helpful assistant."
				}, {
				"role": "user",
				"content": "你好"
				}],
				"chat_template_kwargs": {
				"enable_thinking": false
				}
				}

表 4-16 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	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。
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"} ● 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 4-17 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 4-18 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。

参数名称	类型	说明
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。

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)
```

Qwen3-30B-A3B 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": "qwen3-30b-a3b",  # model参数
        "messages": [
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": "你好"}
        ],
        "chat_template_kwargs": {
            "enable_thinking": False # 是否开启深度思考模式，默认开启
        }
    }
    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": "qwen3-30b-a3b",
    "messages": [
      {"role": "system", "content": "You are a helpful assistant."},
      {"role": "user", "content": "你好"}
    ],
     "chat_template_kwargs": {
       "enable_thinking": false
     }
  }'
```

- 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="qwen3-30b-a3b",  # model参数
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "你好"},
    ],
    extra_body={
        "chat_template_kwargs": {
            "enable_thinking": False  # 是否开启深度思考模式，默认开启
        }
    }
)

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"{base64_image}"
                  # JPEG图像：  f"{base64_image}"
                  # WEBP图像：  f"{base64_image}"
                  "image_url": {
                    "url": f"{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 $MAAS_API_KEY" \
  -d '{
    "model": "qwen2.5-vl-72b",
    "messages": [
      {
        "role": "user",
        "content": [
          {"type": "text", "text": "描述下图片里的内容"},
          {"type": "image_url", "image_url": {"url": "$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"{base64_image}"
                    # JPEG图像：  f"{base64_image}"
                    # WEBP图像：  f"{base64_image}"
                    "image_url": {
                        "url": f"{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。\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
}
```

4.8.2 图片生成

图片生成API用于根据给定的文本提示词同步生成图像。其业务逻辑为接收包含模型名称、文本提示词以及图片生成参数（如图像尺寸、随机数种子等）的请求，调用相应模型进行图片生成，并返回生成结果的URL以及相关状态和使用信息。

前提条件

- 预置服务：已在“在线推理 > 预置服务”页签开通Qwen_Image或Qwen-Image-Edit模型的商用服务。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：已在“在线推理 > 自定义接入点”页签为Qwen_Image或Qwen-Image-Edit模型创建了自定义接入点。具体操作，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

约束限制

对于支持图片上传的模型，单个图片文件的大小不超过10MB。如果以Base64编码形式上传图片，需确保编码后的图片小于10MB。

接口信息

表 4-19 接口信息

名称	说明	取值
API地址	调用图片生成的API地址。	https://api.modelarts-maas.com/v1/images/generations
model参数	model参数调用名称。	您可以通过任选以下方式获取model参数值。 • 从**表4-20**的“**model参数值**”列获取。 • 在“预置服务 > 商用服务”页签的服务名称左侧，单击图标，在“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。 • 在“自定义接入点”页签的“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

支持模型列表

表 4-20 支持模型列表

模型	模型版本	支持地域	model参数值	应用场景
Qwen_Image	qwen-image	西南-贵阳一	qwen-image	文字生成图像
Qwen-Image-Edit	qwen_image_edit	西南-贵阳一	qwen_image_edit	图像编辑

请求参数说明

- **Qwen-Image模型**

表 4-21 请求 body 参数（body 体需要小于 8M）

参数名称	参数类型	是否必填	默认值	说明	示例值
model	string	是	无	模型名称，具体请参见支持模型列表的“**model参数值**”列。	qwen-image
prompt	string	是	无	文本提示词，用于引导模型生成图像，支持中英文。长度支持2000 tokens以下（每个单词和标点都算一个token）。如果长度超过800，会自动截断为800 tokens。	A running cat

参数名称	参数类型	是否必填	默认值	说明	示例值
size	string	是	无	生成图像的尺寸要求： ● 宽度和高度范围：[512,3072]像素。 ● 推荐尺寸：2048x2048、1536x1536、1024x1024。 ● 尺寸要求：宽度和高度必须是16的倍数，否则系统将自动向下调整至最近的16的倍数。 ● 分辨率支持：面积 = 宽度 x 高度。 ● 最小分辨率：512x512。 ● 最大分辨率：3072x3072（例如，1024x4096是被允许的）。 ● 宽高比限制：宽高比需在1:12至12:1之间，超出此范围将导致错误。 ● 处理超出尺寸范围的输入：系统将按比例缩放输入尺寸，以使总面积接近允许的最小或最大值。	1024x1024
response_format	string	否	b64_json	返回格式，可取值为[url, b64_json]，目前仅支持b64_json。	b64_json
seed	int	否	范围内随机数	随机种子，取值范围为[0, 2147483648]，您可以按需配置，不配置则在范围内随机。	333
watermark	bool	否	无	是否对图片进行水印处理。 ● true：对图片进行水印处理。 ● false：对图片不进行水印处理。	true

- **Qwen-Image-Edit模型**

表 4-22 请求 body 参数（body 体需要小于 8M）

参数名称	参数类型	是否必填	默认值	说明	示例值
model	string	是	无	模型名称，具体请参见**支持模型列表**的“**model参数值**”列。	qwen_image_edit
prompt	string	是	无	文本提示词，用于引导模型生成图像，支持中英文。长度支持2000 tokens以下（每个单词和标点都算一个token）。如果长度超过800，会自动截断为800 tokens。	A running cat
size	string	是	无	生成图像尺寸，需要介于[512x512, 2048x2048]之间，height和width需要被16整除，否则会向下兼容。推荐值如下： • 2048x2048 • 1024x1024 • 1328x1328 • 1472x1104 • 1584x1056 • 1664x936 • 2016x864	1024x1024
image	string	是	无	返回格式，仅支持base64。图片尺寸最大支持1024x1024，最小支持512x512，base64编码最大长度限制3145728（即1024x1024x3）。	base64
seed	int	否	随机值	随机种子，取值范围为[0, 2147483648]，默认值为1。	333
watermark	bool	否	无	是否对图片进行水印处理。 • true：对图片进行水印处理。 • false：对图片不进行水印处理。	true

响应参数说明

状态码：200

参数	参数类型	说明
model	string	本次请求使用的模型。
created	int	任务创建时间的Unix时间戳（毫秒）。
data	list[dict]	图像数据列表，与输入image的格式一致。
error	error结构	固定返回null。

usage	usage结构	结构内容为json结构体，KV值可自定义，例如：
		{
		"model_latency"：6000，
		"prompt_tokens": 0,
		"completion_tokens": 0,
		"total_tokens": 0
		}

Qwen-Image-Edit 请求示例

- Rest API请求示例：
  - Python示例：

```
import requests
import json
import base64


#  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.jpg")

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/images/generations"  # 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": "qwen_image_edit",  # model参数
        "prompt": "将湖面颜色修改为蓝色",  # 支持中英文
        "size": "1024x1024",
        # 生成图像尺寸qwen_image_edit要求介于[512x512,2048x2048]。
        # 推荐：2048x2048,1536x1536,1024x1024, 512x512，其中height和width需要被16整除，否则会向下兼容。
        "image": f"{base64_image}",  # 支持图片格式，仅支持b64_json。
        "seed": 44  # 取值范围在[0, 2147483648]， 随机种子。
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)
    # Print result.
    print(response.status_code)
    print(response.text)
```

  - cURL示例

```
curl -X POST https://api.modelarts-maas.com/v1/images/generations \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
```

```
    "model": "qwen_image_edit",
    "prompt": "将湖面颜色修改为蓝色",
    "size": "1024x1024",
    "image": f"$BASE64_IMAGE",
    "seed": 44
}'
```

- OpenAI SDK请求示例：

```
import base64
from openai import OpenAI

#  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.jpg")
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.images.generate(
    model="qwen_image_edit",
    prompt="将湖面颜色修改为蓝色",
    size="1024x1024",
    extra_body={
        "image": f"{base64_image}",
        "seed": 44
    }
)

print(response.data[0].b64_json)
```

Qwen-Image-Edit 响应示例

```
{
    "model": "qwen_image_edit",
    "created": 1757948064800,
    "data":
    [
        {
            "url": null,
            "b64_json": "....JtSWh//2Q=="
        }
    ],
    "usage":
    {
        "model_latency": 17509,
        "prompt_tokens": 0,
        "completion_tokens": 0,
        "total_tokens": 0
    },
    "error": null
}
```

4.8.3 视频生成

4.8.3.1 创建视频生成任务

创建视频生成任务API用于根据给定的输入信息，如文本提示词、图片（仅I2V模式）等，结合指定的模型及视频处理参数，生成相应的视频。其业务逻辑是将用户输入的各种参数传递给后端模型进行处理，最终输出生成的视频链接。

约束限制

对于支持图片上传的模型，单个图片文件的大小不超过10MB。如果以Base64编码形式上传图片，需确保编码后的图片小于10MB。

接口信息

表 4-23 接口信息

名称	说明	取值
API地址	调用创建视频生成任务的API地址。	https://api.modelarts-maas.com/v1/video/generations
model参数	model参数调用名称。	您可以通过任选以下方式获取model参数值。 • 从**表4-24**的“**model参数值**”列获取。 • 在“预置服务 > 商用服务”页签的服务名称左侧，单击图标，在“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。 • 在“自定义接入点”页签的“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

支持模型列表

表 4-24 支持模型列表

模型系列	模型版本	支持地域	model参数值	应用场景
通义万相	Wan2.1-T2V-1.3B	西南-贵阳一	wan2.1-t2v-1.3b	文字生成视频
	Wan2.1-T2V-14B	西南-贵阳一	wan2.1-t2v-14b	文字生成视频
	Wan2.1-I2V-14B-480P	西南-贵阳一	wanx2.1-i2v-14b-480p	图片生成视频（首帧）
	Wan2.1-I2V-14B-720P	西南-贵阳一	wanx2.1-i2v-14b-720p	图片生成视频（首帧）
	Wan2.2-I2V-A14B	西南-贵阳一	Wan2.2-I2V-A14B	图片生成视频
	Wan2.2-T2V-A14B	西南-贵阳一	Wan2.2-T2V-A14B	文字生成视频

请求参数说明

表 4-25 请求 body 参数（body 体需要小于 8M）

参数名称	参数类型	是否必填	说明
model	string	是	模型名称，具体请参见表4-24的“**model参数值**”列。
input	object	是	输入的基本信息，如提示词、图片。关于子参数的说明，请参见表4-26。
parameters	object	否	视频内容生成参数。关于子参数的说明，请参见表4-27。

表 4-26 input 子参数说明

参数名称	参数类型	是否必填	默认值	说明
prompt	string	• 文生视频模型必填 • 图生视频模型选填	无	文本提示词，支持中英文，不超过1000字符。
img_url	string	是	无	**说明** 仅支持图生视频模型。 输入给模型的图片内容，填写图片的Base64编码内容。 目前仅支持JPEG（JPG）格式的图片，请按照如下格式输入： • 示例一： "...." • 示例二： "...."

表 4-27 parameters 子参数说明

参数名称	参数类型	是否必填	默认值	说明
size	string	否	1280*720	直接设置为目标分辨率的具体数值。 Wan2.1格式为宽*高，例如720*1280。支持如下档位分辨率： ● 480P档位：不同视频宽高比对应的分辨率如下： – 16:9：832*480 – 9:16：480*832 ● 720P档位：不同视频宽高比对应的分辨率如下： – 16:9：1280*720 – 9:16：720*1280 Wan2.2格式为宽x高，例如720x1280。支持如下档位分辨率： ● 480P档位：不同视频宽高比对应的分辨率如下： – 16:9：832x480 – 9:16：480x832 ● 720P档位：不同视频宽高比对应的分辨率如下： – 16:9：1280x720 – 9:16：720x1280 **说明** Wan2.1-I2V-14B模型分为480P和720P两个版本，调用时只支持模型版本对应的分辨率。
fps	integer	否	16	生成视频每秒的帧数，当前只支持8和16。
duration	integer	否	5	生成视频时长，单位为秒，当前只支持3s和5s。
seed	integer	否	0	随机种子，用于控制生成内容的随机性。取值范围为[0，2147483648]。

响应参数说明

参数名称	参数类型	说明
task_id	string	任务ID。

Wan2.1-T2V-14B 文生视频请求示例

Rest API的示例代码如下。

- 使用Python调用示例：

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/video/generations"  # 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": "wan2.1-t2v-14b",  # model参数
        "input": {
            "prompt": "小猫在散步",
        },
        "parameters": {
            "size": "720*1280",    # 根据调用模型，填写相应的分辨率
            "fps": 16,
            "duration": 5,
            "seed": 0
        }
    }
    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://api.modelarts-maas.com/v1/video/generations" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
    "model": "wan2.1-t2v-14b",
    "input": {
        "prompt": "小猫在散步"
    },
    "parameters": {
        "size": "720*1280",
        "fps": 16,
        "duration": 5,
        "seed": 0
    }
}'
```

Wan2.1-I2V-14B-720P 图生视频请求示例

- 使用Python调用示例。

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/video/generations"  # 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": "wan2.1-i2v-14b-720p",  # model参数
```

```
    "input": {
        "prompt": "小猫在散步",
        "img_url": "...."  # jpg图片base64编码。
    },
    "parameters": {
        "size": "720*1280",
        "fps": 16,
        "duration": 5,
        "seed": 0
    }
}
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://api.modelarts-maas.com/v1/video/generations" \
 -H "Content-Type: application/json" \
 -H "Authorization: Bearer $MAAS_API_KEY" \
 -d '{
  "model": "wan2.1-i2v-14b-720p",
  "input": {
      "prompt": "小猫在散步",
      "img_url": "...."
  },
  "parameters": {
      "size": "720*1280",
      "fps": 16,
      "duration": 5,
      "seed": 0
  }
}'
```

Wan2.2-T2V-A14B 文生视频请求示例

Rest API的示例代码如下。

- Python示例

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/video/generations"  # 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": "Wan2.2-T2V-A14B",  # model参数
        "input": {
            "prompt": "小猫在散步",
        },
        "parameters": {
            "size": "720x1280",  # 根据调用模型，填写相应的分辨率，支持"1280x720"、 "720x1280"、
"480x832"或"832x480"。
            "fps": 16,
            "duration": 5,
            "seed": 0
        }
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

    # Print result.
```

```
    print(response.status_code)
    print(response.text)
```

- cURL示例

```
curl -X POST "https://api.modelarts-maas.com/v1/video/generations" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
    "model": "Wan2.2-T2V-A14B",
    "input": {
        "prompt": "小猫在散步"
    },
    "parameters": {
        "size": "720x1280",
        "fps": 16,
        "duration": 5,
        "seed": 0
    }
}'
```

Wan2.2-I2V-A14B 图生视频请求示例

Rest API的示例代码如下。

- Python示例

```
import base64
import requests
import json


#  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(r"D:\Pictures\image.jpg") # jpg图片base64编码。

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/video/generations"  # 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": "Wan2.2-I2V-A14B",  # model参数
        "input": {
            "prompt": "小猫在散步",
            "img_url": f"{base64_image}"
        },
        "parameters": {
            "size": "720x1280",# 根据调用模型，填写相应的分辨率，支持"1280x720"、 "720x1280"、
"480x832"或"832x480"。
            "fps": 16,
            "duration": 5,
            "seed": 0
        }
    }
    response = requests.post(url, headers=headers, data=json.dumps(data), verify=False)

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

- cURL示例

  请将示例中的“$BASE64_IMAGE”替换为实际的base64。

```
curl -X POST "https://api.modelarts-maas.com/v1/video/generations" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
    "model": "Wan2.2-I2V-A14B",
    "input": {
        "prompt": "小猫在散步",
        "img_url": "$BASE64_IMAGE"
    },
    "parameters": {
        "size": "720x1280",
        "fps": 16,
        "duration": 5,
        "seed": 0
    }
}'
```

响应示例

```
{
  "task_id": "e0cc914f-66bb-402a-912b-990fa1e4ab42",
}
```

4.8.3.2 查询视频生成任务

查询视频生成任务API用于根据任务ID查询视频生成任务的状态和结果。

接口信息

表 4-28 接口信息

名称	说明	取值
API地址	查询视频生成任务的API地址，需要在链接末尾拼接生成任务的task_id。 task_id可以通过**创建视频生成任务**API获取。	https://api.modelarts-maas.com/v1/video/generations/task_id

请求参数说明

参数名称	参数类型	是否必填	说明
task_id	string	是	要查询的视频生成任务ID，需拼接在查询API末尾处，可以通过**创建视频生成任务**API获取。

响应参数说明

表 4-29 响应参数

参数名称	参数类型	说明
task_id	string	任务ID。

参数名称	参数类型	说明
status	string	任务状态。取值如下： • queued：排队中 • running：运行中 • succeeded：成功 • failed：失败 • timeout：超时
error	object	错误提示信息。关于子参数的说明，请参见表4-30。
content	object	生成的视频内容信息。关于子参数的说明，请参见表4-31。
usage	object	任务的Token用量。关于子参数的说明，请参见表4-32。
created_at	integer	任务创建时间的Unix时间戳（秒）。
updated_at	integer	任务状态更新时间的Unix时间戳（秒）。

表 4-30 error 子参数

参数名称	参数类型	说明
code	integer	错误码。任务状态为成功时，返回0。
message	string	报错信息。

表 4-31 content 子参数

参数名称	参数类型	说明
result_url	string	图片/文本生成视频的URL。该URL有效期为24小时，请注意及时下载转存。

表 4-32 usage 子参数

参数名称	参数类型	说明
model_latency	integer	从模型收到请求到返回结果的端到端时延（毫秒）。
completion_tokens	integer	模型生成内容的消耗Token数。

参数名称	参数类型	说明
prompt_tokens	integer	用户输入Token数。
total_tokens	integer	总消耗Token数。

请求示例

- Python示例：

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/video/generations/task_id" # API地址。请将task_id替换为实际的ID，您可以通过创建视频生成任务API获取。
    api_key = "MAAS_API_KEY"  # 把yourApiKey替换成已获取的API Key

    # Send request.
    headers = {
        'Content-Type': 'application/json',
        'Authorization': f'Bearer {api_key}'
    }
    response = requests.get(url, headers=headers, verify=False)

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

- cURL示例：

```
curl -X GET 'https://api.modelarts-maas.com/v1/video/generations/task_id' \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY"
```

响应示例

```
{
    "task_id": "330b74a382a6d42044f146f389cd698e",
    "status": "succeeded",
    "error": {
        "code": 0,
        "message": ""
    },
    "content": {
        "result_url": "https://modelarts.obs.com/example.mp4"
    },
    "usage": {
        "model_latency": 43564,
        "completion_tokens": 124800,
        "prompt_tokens": 0,
        "total_tokens": 124800
    },
    "created_at": 1751894112234,
    "updated_at": 1751894156753
}
```

4.8.4 创建文本向量化

创建文本向量化API用于将文本数据转换为数值向量，以便于在机器学习和自然语言处理任务中使用。这些向量可以捕捉文本的语义信息，使得机器学习模型能够理解和处理文本数据。

前提条件

- 预置服务：已在“在线推理 > 预置服务”页签开通BGE-M3模型的商用服务。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：已在“在线推理 > 自定义接入点”页签为BGE-M3模型创建了自定义接入点。具体操作，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

接口信息

表 4-33 接口信息

名称	说明	取值
API地址	调用图片生成的API地址。	https://api.modelarts-maas.com/v1/embeddings
model参数	model参数调用名称。	您可以通过任选以下方式获取model参数值。 • 从**表4-34**的“**model参数值**”列获取。 • 在“预置服务 > 商用服务”页签的服务名称左侧，单击图标，在“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。 • 在“自定义接入点”页签的“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

支持模型列表

表 4-34 支持模型列表

模型	模型版本	支持地域	model参数值	应用场景
BGE-M3	bge-m3	西南-贵阳一	bge-m3	文本向量化

请求参数说明

表 4-35 请求 body 参数

参数	是否必填	默认值	参数类型	描述
model	是	无	string	模型名称，具体请参见**表4-34**的“**model参数值**”列。

参数	是否必填	默认值	参数类型	描述
input	是	无	string	支持字符串或字符串列表，总输入长度不超过8K。
encoding_format	否	float	string	指定文本向量化结果的输出格式。取值为float或base64。

响应参数说明

状态码：200

参数	参数类型	说明
id	string	请求ID。
object	string	对象类型，始终为 "list"。
created	integer	时间戳。
model	string	模型名称。
data	object[]	模型生成结果数据集。
data.index	integer	序号。
data.object	enum<string>	对象类型。
data.embedding	number[]	模型生成的嵌入向量列表。
usage	object	请求的使用信息。
usage.prompt_tokens	integer	提示词Token计数。
usage.total_tokens	integer	请求使用的Token总数。
usage.completion_tokens	integer	推理Token计数。
usage.prompt_tokens_details	object	输入Prompt使用情况详情。

请求示例

- Rest API的示例代码如下。
  - 使用Python调用示例。

```
import requests
import json
```

```
if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/embeddings"  # 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}'
    }
    texts = ["这是一只小猫", "这是一只小狗"]
    data = {
        "model": "bge-m3",  # 模型名称
        "input": texts,  # input类型可为string or string[]
        "encoding_format": "float"  # 取值范围：  "float","base64"
    }
    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://api.modelarts-maas.com/v1/embeddings" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
    "model": "bge-m3",
    "input": [
      "这是一只小猫",
      "这是一只小狗"
    ],
    "encoding_format": "float"
  }'
```

- 使用OpenAI SDK调用示例。

```
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
texts = ["这是一只小猫", "这是一只小狗"]

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

response = client.embeddings.create(
    model="bge-m3",  # model参数
    input=texts,  # input类型可为string or string[]
    encoding_format="float"  # 取值范围：  "float","base64"
)

# Print result.
print(response.data)
```

响应示例

```
{
  "id": "embd-d848df392a67d662f5a76eaa9e33974f",
  "object": "list",
  "created": 1758023320,
  "model": "bge-m3",
  "data": [{
      "index": 0,
      "object": "embedding",
      "embedding": [-0.021697998046875, 0.0322265625, ...]
  }],
  "usage": {
      "prompt_tokens": 7,
      "total_tokens": 7,
      "completion_tokens": 0,
      "prompt_tokens_details": null
```

```
  }
}
```

4.8.5 创建重排序

创建重排序API用于提供灵活的数据项排序功能，以提升用户体验和应用的可定制性。

前提条件

- 预置服务：已在“在线推理 > 预置服务”页签开通bge-reranker-v2-m3模型的商用服务。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：已在“在线推理 > 自定义接入点”页签为bge-reranker-v2-m3模型创建了自定义接入点。具体操作，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

接口信息

表 4-36 接口信息

名称	说明	取值
API地址	调用图片生成的API地址。	https://api.modelarts-maas.com/v1/rerank
model参数	model参数调用名称。	您可以通过任选以下方式获取model参数值。 • 从**表4-37**的“**model参数值**”列获取。 • 在“预置服务 > 商用服务”页签的服务名称左侧，单击图标，在“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。 • 在“自定义接入点”页签的“model参数”列查看取值。更多信息，请参见**在ModelArts Studio（MaaS）创建自定义接入点**。

支持模型列表

表 4-37 支持模型列表

模型	模型版本	支持地域	model参数值	应用场景
bge-reranker-v2-m3	bge-reranker-v2-m3	西南-贵阳一	bge-reranker-v2-m3	检索结果再排序

请求参数说明

表 4-38 请求 body 参数

参数	是否必填	默认值	参数类型	描述
model	是	无	string	模型名称，具体请参见表4-37的“**model参数值**”列。
query	是	无	string	用户查询文本。总输入长度不超过8K。
documents	是	无	string	待排序文档列表（通常为Embedding召回的Top-K结果）。总输入长度不超过8K。

响应参数说明

状态码：200

参数	参数类型	说明
id	string	请求ID。
model	string	模型名称。
usage	object	请求的使用信息。
usage.total_tokens	integer	请求使用的Token总数。
results	object[]	重排序结果集。
results[i].index	integer	序号。
results[i].document	object	原始文档内容信息。
results[i].document.text	number[]	具体文档内容。
results[i].relevance_score	double	相似度得分。

请求示例

Rest API的示例代码如下。

- 使用Python调用示例。

```
import requests
import json

if __name__ == '__main__':
    url = "https://api.modelarts-maas.com/v1/rerank"  # 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": "bge-reranker-v2-m3",
        "query": "牛是一种动物如何冲泡一杯好喝的咖啡？",  # input类型可为string或string[]。
        "documents": [
            "咖啡豆的产地主要分布在赤道附近，被称为‘咖啡带’。",
            "法压壶的步骤：1. 研磨咖啡豆。2. 加入热水。3. 压下压杆。4. 倒入杯中。",
            "意式浓缩咖啡需要一台高压机器，在9个大气压下快速萃取。",
            "挑选咖啡豆时，要注意其烘焙日期，新鲜的豆子风味更佳。",
            "手冲咖啡的技巧：控制水流速度、均匀注水和合适的水温（90-96°C）是关键。"
        ]
    }

    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://api.modelarts-maas.com/v1/rerank" \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MAAS_API_KEY" \
  -d '{
    "model": "bge-reranker-v2-m3",
    "input": "牛是一种动物如何冲泡一杯好喝的咖啡？",
    "documents": [
      "英国首都是伦敦咖啡豆的产地主要分布在赤道附近，被称为‘咖啡带’。",
      "法国首都是巴黎法压壶的步骤：1. 研磨咖啡豆。2. 加入热水。3. 压下压杆。4. 倒入杯中。",
      "意式浓缩咖啡需要一台高压机器，在9个大气压下快速萃取。",
      "猫和狗都是动物挑选咖啡豆时，要注意其烘焙日期，新鲜的豆子风味更佳。"
      "手冲咖啡的技巧：控制水流速度、均匀注水和合适的水温（90-96°C）是关键。"
    ]
  }'
```

响应示例

```
{
  "id": "rerank-dc9e3495b71134e82c50651c32cde9f6",
  "model": "bge-reranker-v2-m3",
  "usage": {
      "total_tokens": 211
  },
  "results": [{
      "index": 4,
      "document": {
          "text": "手冲咖啡的技巧：控制水流速度、均匀注水和合适的水温（90-96°C）是关键。"
      },
      "relevance_score": 0.01898193359375
  }, {
      "index": 1,
      "document": {
          "text": "法压壶的步骤：1. 研磨咖啡豆。2. 加入热水。3. 压下压杆。4. 倒入杯中。"
      },
      "relevance_score": 0.007404327392578125
  }, {
      "index": 2,
      "document": {
          "text": "意式浓缩咖啡需要一台高压机器，在9个大气压下快速萃取。"
```

```
        },
        "relevance_score": 0.0003418922424316406
    }, {
        "index": 3,
        "document": {
            "text": "挑选咖啡豆时，要注意其烘焙日期，新鲜的豆子风味更佳。"
        },
        "relevance_score": 6.014108657836914e-05
    }, {
        "index": 0,
        "document": {
            "text": "咖啡豆的产地主要分布在赤道附近，被称为‘咖啡带’。"
        },
        "relevance_score": 3.349781036376953e-05
    }]
}
```

4.8.6 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 4-39 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 4-40 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": ""
    }
  ]
}
```

4.8.7 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 4-41 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。
400	ModelArts.81013	Content moderation failed when detecting language. The prompt can only contain Chinese.	使用非中文请求，内容审核失败。
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。

HTTP状态码	错误码	错误信息	说明
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。
403	ModelArts.81014	The free service has expired. You can subscribe commercial service.	免费服务已到期。
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。
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。

HTTP状态码	错误码	错误信息	说明
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。
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。

HTTP状态码	错误码	错误信息	说明
404	"object": "error"	"object": "error", "message": "The model `Qwen2.5-72B-32K` 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。

4.9 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://xxxxxxxxxxxxxxx")
# 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。

操作场景

ModelArts Studio提供了“预置服务”和“我的服务”的文本对话功能，帮助您快速体验模型的文本对话效果。

前提条件

- 使用预置服务：在“在线推理 > 预置服务”页签，使用有效期内的免费服务或者已开通商用服务。具体操作，请参见**免费体验MaaS预置服务**或**在ModelArts Studio（MaaS）预置服务中开通商用服务**。
- 使用我的服务：在“在线推理 > 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：在“在线推理 > 自定义接入点”页签，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**。

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“预置服务”、“我的服务”或“自定义接入点”页签，选择要体验的模型服务，单击“确定”。
     - “预置服务”页签：按需单击“商用服务”或“免费服务”页签，选择目标服务进行体验。商用服务支持按需选择版本。

- “我的服务”页签：单击已部署的模型服务进行体验。
- “自定义接入点”页签：单击使用中的自定义接入点。

– 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，任选以下方式进入“文本对话”页面。

- 在“预置服务”页签，按需单击“商用服务”或“免费服务”页签，单击操作列的“在线体验”，进入“文本对话”页面。商用服务支持按需选择版本。
- 在“我的服务”页签，单击操作列的“更多 > 在线体验”，进入“文本对话”页面。
- 在“自定义接入点”页签：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

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。

参数	说明
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。
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。**输入和输出内容已默认开启内容审核。**

图 5-2 体验模型服务

5.2 在移动端体验 ModelArts Studio（MaaS）文本对话

您可以在移动端华为云App上便捷体验ModelArts Studio模型服务的文本对话效果。

操作场景

ModelArts Studio提供了免费版和商用版的文本对话功能，帮助您快速体验DeepSeek-R1-32K和DeepSeek-V3-32K模型的文本对话效果。

计费说明

- 免费模型服务：体验免费模型服务将消耗已领取的免费Token额度，不涉及计费。
- 商用模型服务：体验商用模型服务将消耗已开通的商用服务Token，费用请以实际发生为准。关于如何计费，请参见**计费说明**。

前提条件

已注册华为云账号，并进行实名认证。具体操作，请参见**注册华为账号并开通华为云**和**实名认证**。

场景一：新用户使用网址体验 ModelArts Studio

1. 打开**ModelArts Studio**，跳转至模型服务体验页面。
2. 在模型服务体验页面，将出现“ModelArts Studio 服务声明”对话框，请查看服务声明内容并单击“同意”。
3. 在模型服务体验页面顶部，单击“选择模型服务”，按需领取免费服务额度或开通商用服务。
   - 领取免费服务额度：

     单击“免费服务”页签，在模型服务右侧单击“领取”。免费服务仅提供基础体验能力，且存在严格的速率限制。
   - 开通商用服务：

i. 单击“商用服务”页签，在模型服务右侧单击“开通”。

ii. 在“开通商用服务”页面，选择需要开通的商用服务，单击“立即开通”。

开通商用服务后，该服务会显示“已开通”。商用服务将为您提供商用级别的API推理服务。暂不支持关闭商用服务，未使用商用服务时不会产生费用。关于如何计费，请参见**计费说明**。

4. 在“免费服务”页签或“商用服务”页签，单击已领取额度的免费服务或已开通的商用服务，进行文本对话。

您可以在右上角单击 图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。

表 5-2 参数设置

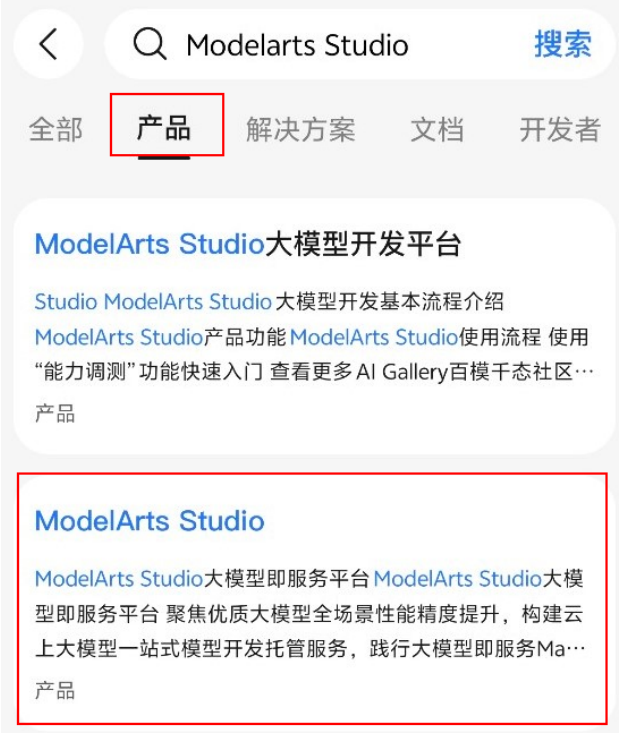
参数	说明
温度/Temperature	设置推理温度。 ● 数值较高，输出结果更加随机。 ● 数值较低，输出结果更加集中和确定。 取值范围：0~2 默认值：0.7
核采样/top_p	设置推理核采样。调整输出文本的多样性，数值越大，生成文本的多样性就越高。 取值范围：0.1~1 默认值：1
top_k	选择在模型的输出结果中选择概率最高的前K个结果。 取值范围：1~1000 默认值：20

5. （可选）免费服务额度使用完后，会显示“额度已用完”。您可以按需开通商用服务付费使用，或前往PC端**ModelArts Studio（MaaS）控制台**，在“在线推理 > 我的服务”页面部署为我的服务进行使用。具体操作，请参见**开通商用服务**或**使用ModelArts Studio（MaaS）部署模型服务**。

场景二：新用户使用华为云 App 体验 ModelArts Studio

1. 下载华为云App。具体信息，请参见**华为云App介绍**。
2. 打开华为云App，在底部导航栏单击“华为云”，在搜索框输入**ModelArts Studio**并单击“搜索”，单击“产品”页签，单击“ModelArts Studio”卡片。

图 5-3 打开 ModelArts Studio 产品



3. 在ModelArts Studio大模型即服务平台，单击“ModelArts Studio控制台”，跳转至模型服务体验页面。
4. 在模型服务体验页面，将出现“ModelArts Studio 服务声明”对话框，请查看服务声明内容并单击“同意”。
5. 在模型服务体验页面顶部，单击“选择模型服务”，按需领取免费服务额度或开通商用服务。
   - 领取免费服务额度：

     单击“免费服务”页签，在模型服务右侧单击“领取”。免费服务仅提供基础体验能力，且存在严格的速率限制。
   - 开通商用服务：

     i. 单击“商用服务”页签，在模型服务右侧单击“开通”。

     ii. 在“开通商用服务”页面，选择需要开通的商用服务，单击“立即开通”。

     开通商用服务后，将为您提供商用级别的API推理服务。暂不支持关闭商用服务，未使用商用服务时不会产生费用。关于如何计费，请参见**计费说明**。
6. 在“免费服务”页签或“商用服务”页签，单击已领取额度的免费服务或已开通的商用服务，进行文本对话。

   您可以在右上角单击参数设置图标，按需拖动调整推理参数。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见**表5-2**。
7. （可选）免费服务额度使用完后，会显示“额度已用完”。您可以按需开通商用服务付费使用，或前往PC端**ModelArts Studio（MaaS）控制台**，在“在线推理 > 我的服务”页面部署为我的服务进行使用。具体操作，请参见**开通商用服务**或**使用ModelArts Studio（MaaS）部署模型服务**。

场景三：老用户使用网址体验 ModelArts Studio

1. 打开**ModelArts Studio**，跳转至模型服务体验页面。

   模型服务体验页面会默认显示您已领取免费额度的免费服务或已开通的商用服务。

2. 在模型服务体验页面，直接进行文本对话，或者在页面上方单击模型服务名称，按需切换免费或商用模型服务进行文本对话。

   关于如何领取免费服务额度或开通商用服务，请参见**领取免费服务额度或开通商用服务**。

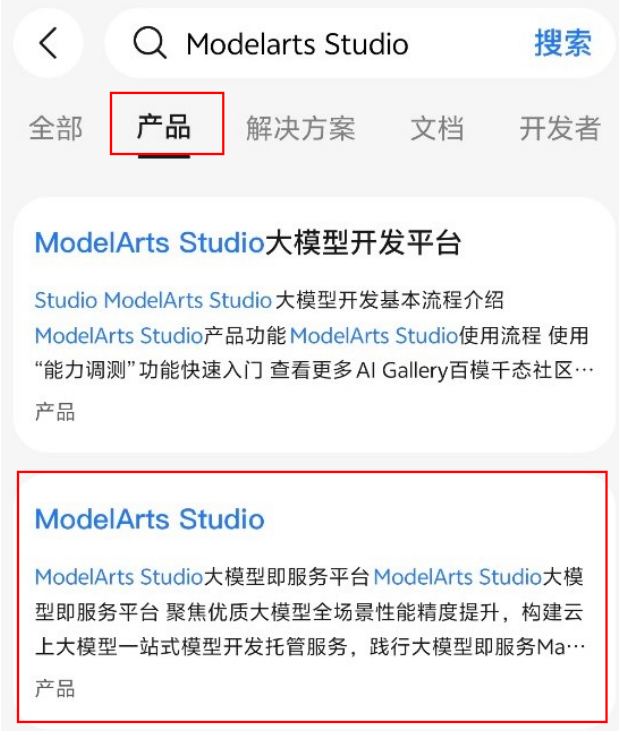
   您可以在右上角单击 图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见**表5-2**。

3. （可选）免费服务额度使用完后，会显示“额度已用完”。您可以按需开通商用服务付费使用，或前往PC端**ModelArts Studio（MaaS）控制台**，在“在线推理 > 我的服务”页面部署为我的服务进行使用。具体操作，请参见**开通商用服务**或**使用ModelArts Studio（MaaS）部署模型服务**。

场景四：老用户使用华为云 App 体验 ModelArts Studio

1. 打开华为云App，在底部导航栏单击“华为云”，在搜索框输入**ModelArts Studio**并单击“搜索”，单击“产品”页签，单击“ModelArts Studio”卡片。

   图 5-4 打开 ModelArts Studio 产品



2. 在ModelArts Studio大模型即服务平台，单击“ModelArts Studio控制台”，跳转至模型服务体验页面。

   模型服务体验页面会默认显示您已领取免费额度的免费服务或已开通的商用服务。

3. 在模型服务体验页面，直接进行文本对话，或者在页面上方单击模型服务名称，按需切换免费或商用模型服务进行文本对话。

关于如何领取免费服务额度或开通商用服务，请参见**领取免费服务额度或开通商用服务**。

您可以在右上角单击 图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见**表5-2**。

4. （可选）免费服务额度使用完后，会显示“额度已用完”。您可以按需开通商用服务付费使用，或前往PC端**ModelArts Studio（MaaS）控制台**，在“在线推理 > 我的服务”页面部署为我的服务进行使用。具体操作，请参见**开通商用服务**或**使用ModelArts Studio（MaaS）部署模型服务**。

6 ModelArts Studio（MaaS）模型管理

6.1 在 ModelArts Studio（MaaS）创建模型

MaaS提供了基于昇腾云算力适配的开源大模型（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调优、压缩、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基于昇腾云算力适配开源大模型，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，平台自动完成与昇腾芯片的算力适配，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、百川、ChatGLM、Llama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在线体验、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，基于昇腾云适配的开源模型创建个人专属模型。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

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。

（可选）步骤二：修改权重配置文件

当选择ChatGLM3-6B、GLM-4-9B、Qwen-7B、Qwen-14B、Qwen-72B、Baichuan2-7B、Baichuan2-13B、Llama2-7B、Llama2-13B和Llama2-80B基础模型（名字必须一致）创建模型时，建议对权重配置参数进行优化调整，以提升模型的预测精度和输出质量，从而构建更贴合具体业务需求的定制化模型。修改后的权重文件要更新至OBS桶中。

Qwen2.5系列模型无需修改权重即可部署。关于如何部署模型服务，请参见**使用ModelArts Studio（MaaS）部署模型服务**。

- **ChatGLM3-6B、GLM-4-9B**

  修改文件“tokenization_chatglm.py”。

  - 第一处

    原内容

    ```
    # Load from model defaults assert self.padding_side == "left"
    ```

    修改为

    ```
    # Load from model defaults # assert self.padding_side == "left"
    ```

  - 第二处

    原内容

    ```
    if needs_to_be_padded:
    difference = max_length - len(required_input)
    ```

```
if "attention_mask" in encoded_inputs:
encoded_inputs["attention_mask"] = [0] * difference + encoded_inputs["attention_mask"]
if "position_ids" in encoded_inputs:
encoded_inputs["position_ids"] = [0] * difference + encoded_inputs["position_ids"]
encoded_inputs[self.model_input_names[0]] = [self.pad_token_id] * difference + required_input
```

修改为

```
if needs_to_be_padded:
difference = max_length - len(required_input)
if "attention_mask" in encoded_inputs:
encoded_inputs["attention_mask"] = encoded_inputs["attention_mask"] + [0] * difference
if "position_ids" in encoded_inputs:
encoded_inputs["position_ids"] = encoded_inputs["position_ids"] + [0] * difference
encoded_inputs[self.model_input_names[0]] =  required_input + [self.pad_token_id] * difference
```

- **Qwen-7B、Qwen-14B和Qwen-72B**
  - 第一处，修改文件“modeling_qwen.py”。

    原内容

    ```
    SUPPORT_BF16 = SUPPORT_CUDA and torch.cuda.is_bf16_supported() SUPPORT_FP16 = SUPPORT_CUDA and torch.cuda.get_device_capability(0)[0] >= 7
    ```

    修改为

    ```
    SUPPORT_BF16 = SUPPORT_CUDA and True SUPPORT_FP16 = SUPPORT_CUDA and True
    ```

  - 第二处，修改文件“tokenizer_config.json”。

    在文件中增加内容

    ```
    chat_template = {% for message in messages %}{{'<|im_start|>' + message['role'] + '\n' + message['content']}}{% if (loop.last and add_generation_prompt) or not loop.last %}{{ '<|im_end|>' + '\n'}}{% endif %}{% endfor %}{% if add_generation_prompt and messages[-1]['role'] != 'assistant' %}{{ '<|im_start|>assistant\n' }}{% endif %}
    ```

- **Baichuan2-7B和Baichuan2-13B**

  在文件“tokenizer_config.json”中增加如下内容。

  ```
  chat_template = {% for message in messages %}{{'<|im_start|>' + message['role'] + '\n' + message['content']}}{% if (loop.last and add_generation_prompt) or not loop.last %}{{ '<|im_end|>' + '\n'}}{% endif %}{% endfor %}{% if add_generation_prompt and messages[-1]['role'] != 'assistant' %}{{ '<|im_start|>assistant\n' }}{% endif %}
  ```

- **Llama2-7B、Llama2-13B和Llama2-80B**

  在文件“tokenizer_config.json”中增加如下内容。

  ```
  chat_template = {% if messages[0]['role'] == 'system' %}{% set loop_messages = messages[1:] %}{% set system_message = messages[0]['content'] %}{% else %}{% set loop_messages = messages %}{% set system_message = false %}{% endif %}{% for message in loop_messages %}{% if (message['role'] == 'user') != (loop.index0 % 2 == 0) %}{{ raise_exception('Conversation roles must alternate user/assistant/user/assistant/...') }}{% endif %}{% if loop.index0 == 0 and system_message != false %}{% set content = '<<SYS>>\\n' + system_message + '\\n<</SYS>>\\n\\n' + message['content'] %}{% else %}{% set content = message['content'] %}{% endif %}{% if message['role'] == 'user' %}{{ bos_token + '[INST] ' + content.strip() + ' [/INST]' }}{% elif message['role'] == 'assistant' %}{{ ' '  + content.strip() + ' ' + eos_token }}{% endif %}{% endfor %}
  ```

步骤三：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**。

步骤四：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	MaaS提供基于昇腾云算力适配的开源大模型供您使用。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**步骤三**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”。

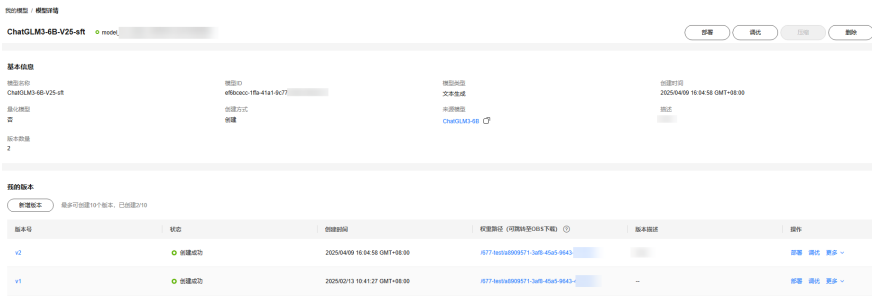
5. 参数配置完成后，单击“创建”，创建自定义模型。
   在模型列表，当模型“状态”变成“创建成功”时，表示模型创建完成。

步骤五：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”。

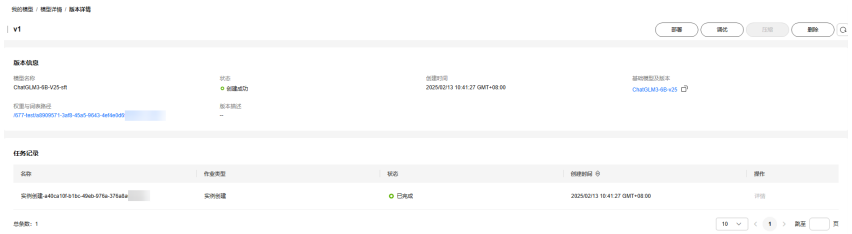
图 6-1 模型详情



- 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。

图 6-2 版本详情



  - 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）步骤六：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**步骤三**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）步骤七：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

- 当模型创建成功后，您可以对模型进行调优或压缩，获得更合适的模型。具体操作，请参见**使用ModelArts Studio（MaaS）调优模型**或**使用ModelArts Studio（MaaS）压缩模型**。
- 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**。

6.2 使用 ModelArts Studio（MaaS）压缩模型

ModelArts Studio大模型即服务平台支持对模型广场或用户自建的模型进行压缩，通过SmoothQuant-W8A8或AWQ-W4A16压缩策略优化模型，从而缓解资源占用问题。

操作场景

模型压缩是优化深度学习模型的技术，旨在减少模型的体积、计算量或内存占用，同时尽可能保持其性能（如准确率）。它是解决大型模型在资源受限场景中部署问题的关键手段。

模型压缩适用于追求更高的推理服务性能、低成本部署以及可接受一定精度损失的场景。

模型压缩的原理如下：

- 参数修剪：删除模型中对性能影响较小的参数，如权重矩阵中绝对值较小的元素，从而减少模型的存储和计算量。

- 量化：将模型参数的数据类型从高精度（如32位浮点数）转换为低精度（如8位整数），在不损失太多精度的情况下减少模型的存储和计算需求。
- 知识蒸馏：将复杂的大模型（教师模型）的知识传递给一个较小的模型（学生模型），使学生模型在较小的规模下仍能保持较好的性能。

  ModelArts Studio（MaaS）大模型即服务平台当前支持的模型压缩策略主要是SmoothQuant-W8A8和AWQ-W4A16两种量化压缩策略。

表 6-3 模型压缩策略介绍

压缩策略	说明	适用场景
SmoothQuant-W8A8	SmoothQuant是一种同时确保准确率与推理高效的训练后量化（PTQ）方法，W8A8可实现8-bit权重、8-bit激活（W8A8）量化，引入平滑因子来平滑激活异常值，将量化难度从较难量化的激活转移到容易量化的权重上。 W8表示将权重（Weight）量化为8位整数（INT8）；A8表示将激活（Activation）量化为8位整数。	• 长序列的场景 • 大并发量的场景
AWQ-W4A16	AWQ是一种大模型低比特权重的训练后量化（PTQ）方法，W4A16可实现4-bit权重、16-bit激活（W4A16）量化，通过激活值来选择并放大显著权重，以提高推理效率。 W4表示将大部分权重量化至4位整数（INT4），大幅减少存储占用。A16表示保持激活值为16位浮点数（FP16/BF16），避免因激活量化引入额外误差。	• 小并发量的低时延场景 • 更少推理卡数部署的场景

约束限制

表6-4列举了在MaaS平台上支持模型压缩的模型，不在表格里的模型不支持在MaaS平台上使用模型压缩功能。

表 6-4 支持模型压缩的模型

模型名称	SmoothQuant-W8A8	AWQ-W4A16
Llama2-13B	√	√
Llama2-70B	√	√
Llama2-7B	√	√
Llama3-70B	√	√
Llama3-8B	√	√
Qwen1.5-14B	√	√

模型名称	SmoothQuant-W8A8	AWQ-W4A16
Qwen1.5-72B	√	√
Qwen1.5-7B	√	√
Qwen2-72B	√	x
Qwen2-72B-1K	√	x
Qwen2.5-72B	√	x
Qwen2.5-32B	√	√

前提条件

- 已准备好用于存放压缩后模型权重文件的OBS桶，OBS桶必须和MaaS服务在同一个Region下。关于如何创建OBS桶和上传文件，请参见**OBS控制台快速入门**。
- 如果需要对已创建的模型进行压缩，则该模型需支持压缩，且在“模型管理 > 我的模型”页面中，模型的“状态”为“创建成功”。
  - 关于是否支持压缩，请参见**约束限制**。
  - 关于如何创建模型，请参见**在ModelArts Studio（MaaS）创建模型**。

创建压缩作业

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“模型压缩”进入作业列表。
3. 单击“创建压缩作业”进入创建页面，完成创建配置。

表 6-5 创建压缩作业参数说明

参数		说明
作业设置	作业名称	自定义压缩作业名称。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	自定义压缩任务简介。最大支持1000字符。
模型设置	来源模型	单击“选择模型”，选择“模型广场”或“我的模型”下面的模型。

参数		说明
	压缩策略	支持SmoothQuant-W8A8和AWQ-W4A16两种压缩策略。 ● **SmoothQuant-W8A8：**SmoothQuant是一种同时确保准确率与推理高效的训练后量化（PTQ）方法，W8A8可实现8-bit权重、8-bit激活（W8A8）量化，引入平滑因子来平滑激活异常值，将量化难度从较难量化的激活转移到容易量化的权重上。 ● **AWQ-W4A16：**AWQ是一种大模型低比特权重的训练后量化（PTQ）方法，W4A16可实现4-bit权重、16-bit激活（W4A16）量化，通过激活值来选择并放大显著权重，以提高推理效率。
	压缩后模型名称	设置压缩后产生的新模型的名称。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、下划线（_）、中划线（-）和英文半角句号（.）。
参数设置	平滑系数/Migration Strength	设置SmoothQuant量化的迁移系数，仅“压缩策略”选择“SmoothQuant-W8A8”时才需要配置。建议使用默认值。 取值范围：0~1 默认值：0.5
	压缩后模型权重保存路径	选择压缩后模型权重文件存放的OBS路径。
资源设置	资源池类型	资源池分为公共资源池与专属资源池。 ● 公共资源池由所有租户共享使用。 ● 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。
更多选项	永久保存日志	选择是否打开“永久保存日志”开关。 ● 开关关闭（默认关闭）：表示不永久保存日志，则任务日志会在30天后会被清理。可以在任务详情页下载全部日志至本地。 ● 开关打开：表示永久保存日志，此时必须配置“日志路径”，系统会将任务日志永久保存至指定的OBS路径。

参数		说明
	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“创建中”、“已完成”、“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。 • 开关关闭（默认关闭）：表示任务将一直运行直至完成。 • 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任务，准备排队等状态不扣除运行时长。

4. 参数配置完成后，单击“提交”。

   “资源池类型”选择“公共资源池”时，会出现“计费提醒”对话框，请您仔细阅读预估压缩时长和费用信息，然后单击“确定”，创建压缩任务。模型压缩运行时会产生费用，压缩时长与选取模型及压缩方式有关。该预估费用不包含OBS存储费用。预估费用基于目录价和预估时长计算，估算存在波动性，最终以实际发生为准。

   在“模型压缩”列表中，当压缩作业的“状态”变成“已完成”时，表示模型压缩完成。

模型压缩时长估算

表 6-6 模型压缩时长估算

模型名称	SmoothQuant-W8A8	AWQ-W4A16
Llama2-13B	5~10分钟	60分钟
Llama2-70B	20~30分钟	3小时
Llama2-7B	5~10分钟	40分钟

模型名称	SmoothQuant-W8A8	AWQ-W4A16
Llama3-70B	20~30分钟	3小时
Llama3-8B	5~10分钟	40分钟
Qwen1.5-14B	5~10分钟	60分钟
Qwen1.5-72B	20~30分钟	3小时
Qwen1.5-7B	5~10分钟	40分钟
Qwen2-72B	20~30分钟	-
Qwen2-72B-1K	20~30分钟	-
Qwen2.5-72B	40分钟	-
Qwen2.5-32B	20~30分钟	2小时

查看压缩作业信息

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“模型压缩”进入作业列表。
3. 单击作业名称，进入压缩作业详情页面，可以查看作业详情和日志。
   - “详情”：可以查看作业的基本信息，包括作业、模型、资源等设置信息。
   - “日志”：可以搜索、查看和下载作业日志。

删除压缩作业

说明

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“模型压缩”进入列表。
3. 选择压缩作业，单击操作列的“删除”，在弹窗中输入“DELETE”，单击“确定”，删除作业。

后续操作

模型压缩后，您可以将其部署为我的服务，进行在线体验或API调用。具体操作，请参见**使用ModelArts Studio（MaaS）部署模型服务**、**使用ModelArts Studio（MaaS）部署模型服务**和**调用ModelArts Studio（MaaS）部署的模型服务**。

7 ModelArts Studio（MaaS）模型训练

7.1 使用 ModelArts Studio（MaaS）调优模型

在ModelArts Studio大模型即服务平台支持对模型广场的预置模型或用户自建的模型进行调优，通过多种训练方法（如全参微调、增量预训练等）优化模型性能，从而获得更符合业务需求的模型。

操作场景

从模型广场或“我的模型”中选择一个模型进行调优，当模型完成调优作业后会产生一个新的模型，呈现在“我的模型”列表中。

约束限制

训练数据集的名称不能含有中文。

前提条件

- 已准备好训练数据集，并存放于OBS桶中，OBS桶必须和MaaS服务在同一个Region下。关于如何创建OBS桶和上传文件，请参见**OBS控制台快速入门**。
- 当需要永久保存日志时，需要准备好存放日志的OBS路径，OBS桶必须和MaaS服务在同一个Region下。
- 如果需要对已创建的模型进行调优，则该模型需支持调优，且在“模型管理 > 我的模型”页面中，模型的“状态”为“创建成功”。
  - 关于是否支持调优，请参见**支持模型微调的模型**。
  - 关于如何创建模型，请参见**在ModelArts Studio（MaaS）创建模型**。
- 已准备公共资源池或专属资源池，详情请参见**准备ModelArts Studio（MaaS）资源**。

创建调优作业

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“模型调优”。
3. 在“模型调优”页面右上角，单击“创建调优作业”，完成创建配置。

表 7-1 创建调优作业参数说明

参数		说明
作业设置	作业名称	自定义调优作业名称。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	自定义调优作业简介。支持1000字符。
模型设置	来源模型	单击“选择模型”，在“选择模型”对话框中选择“模型广场”或“我的模型”下面的模型。关于模型的介绍，请参见**在ModelArts Studio（MaaS）模型广场查看预置模型**和**在ModelArts Studio（MaaS）创建模型**。
	调优类型	MaaS支持全参微调、LoRA微调和增量预训练三种调优类型。不同模型支持的调优类型不同，不支持的调优类型会置灰，详情请参见**表7-3**。 • **全参微调：**直接在模型上训练，影响模型全量参数的微调训练，效果较好，收敛速度较慢，训练时间较长。 • **LoRA微调：**冻结原模型，通过往模型中加入额外的网络层，并只训练这些新增的网络层参数，效果接近或略差于全参训练，收敛速度快，训练时间短。 • **增量预训练**：在现有预训练模型基础上，利用新数据或特定领域的数据增强模型的能力和性能。允许模型逐步适应新的作业和数据，避免过拟合和欠拟合问题，进一步提高模型的泛化能力。
	调优后模型名称	设置调优后产生的新模型的名称。 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、下划线（_）、中划线（-）和半角句号（.）。
	调优后模型权重存放路径	选择调优后模型权重文件的OBS存放路径。训练后将在指定路径下自动创建以作业ID命名的新文件夹进行权重存储。
数据设置	选择数据集格式	支持选择MOSS、Alpaca和ShareGPT。训练数据需要按照对应格式，上传符合规范的数据集，以更好完成训练作业。关于数据集示例，请参见**支持的数据集格式**。 **说明** 如果数据集选择错误，您可以通过以下方式查看日志详情。更多信息，请参见**调优数据集异常日志说明**。 • 登录**ModelArts Studio（MaaS）控制台**，在“模型调优”页面单击目标作业，在作业详情的日志页签查看详情。 • 登录**ModelArts管理控制台**，在“模型训练 > 训练作业”页面单击目标作业，在日志页签查看详情。

参数		说明
	添加数据集	选择存放训练数据集的OBS路径，必须选择到文件。单次上传本地文件到OBS的总大小不能超过5GB，详情请参见**如何上传超过5GB的大对象**。 **说明** 数据集必须满足要求（请参见**支持模型微调的模型**和**支持的数据集格式**），否则调优会失败。
超参设置	数据条数	输入数据集中的总数据条数。
	迭代轮次/Epoch	训练过程中模型遍历整个数据集的次数。不同量级数据集的建议值：百量集4~8；千量集2~4；更大数量级1~2。
	迭代步数/Iterations	计算得出的模型参数/权重更新的次数。在调优过程中，Qwen2-72B-1K模型的每一个Iterations会消耗512条训练数据，其他模型的每一个Iterations会消耗32条训练数据。 当数据集是数百量级，则建议迭代4~8个epoch（epoch表示整个数据集被完整地用于一次训练的次数）；当数据集是数千量级，则建议迭代2~4个epoch；当数据集是更大数量，则建议迭代1~2个epoch。 总Iterations = 整个数据集完整训练需要的Iterations * epoch。例如，当一个数据集有3200条数据，完整训练一个数据集的Iterations为100，迭代2个epoch，总Iterations就是200。 取值范围：1~100000
	学习率/learning_rate	设置每个迭代步数（iteration）模型参数/权重更新的速率。学习率设置的过高会导致模型难以收敛，过低则会导致模型收敛速度过慢。 取值范围：[0, 0.1] 默认值：0.00002 建议微调场景的学习率设置在10^{-5}这个量级。
	Checkpoint保存个数	训练过程中保存Checkpoint的个数。最小值为1，最大值为“迭代步数/Iterations”的参数值，不超过10。Checkpoint会自动存储到“调优后模型权重保存路径”的OBS路径下。
资源设置	资源池类型	资源池分为公共资源池与专属资源池。 • 公共资源池由所有租户共享使用。 • 专属资源池需单独创建，不与其他租户共享。
	规格	选择规格，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。
	计算节点个数	当计算节点个数大于1，将启动多节点分布式训练。详细信息，请参见**分布式训练功能介绍**。

参数		说明
更多选项	永久保存日志	选择是否打开“永久保存日志”开关。 ● 开关关闭（默认关闭）：表示不永久保存日志，则作业日志会在30天后会被清理。可以在作业详情页下载全部日志至本地。 ● 开关打开：表示永久保存日志，此时必须配置“日志路径”，系统会将作业日志永久保存至指定的OBS路径。
	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，当作业发生特定事件（如作业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“创建中”、“已完成”、“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。 ● 开关关闭（默认关闭）：表示作业将一直运行直至完成。 ● 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作业，准备排队等状态不扣除运行时长。

参数		说明
	自动重启	选择是否打开“自动重启”开关。 • 开关关闭（默认关闭）：表示不启用自动重启。 • 开关打开：表示当由于环境问题导致训练作业异常时，系统将自动修复异常或隔离节点，并重启训练作业，提高训练成功率。 打开开关后，可以设置“最大重启次数”和是否启用“无条件自动重启”。 – 重启次数的取值范围是1~128，缺省值为3。创建调优作业后不支持修改重启次数，请合理设置次数。 – 开启无条件自动重启后，只要系统检测到训练异常，就无条件重启训练作业。为了避免无效重启浪费算力资源，系统最多只支持连续无条件重启3次。 如果训练过程中触发了自动重启，则平台会自动获取最新的Checkpoint，并从该点重启作业。

4. 参数配置完成后，单击“提交”。

   “资源池类型”选择“公共资源池”时，会出现“计费提醒”对话框，请您仔细阅读预计调优运行时间和预计消耗费用信息，然后单击“确定”，创建调优作业。该预估费用不包含OBS存储费用。预估费用基于目录价和预估时长计算，估算存在波动性，最终以实际发生为准。

   在“模型调优”列表中，当模型调优作业的“状态”变成“已完成”时，表示模型调优完成。

支持模型微调的模型

表7-2列举了支持模型调优的模型，不在表格里的模型不支持使用MaaS调优模型。

表 7-2 支持模型微调的模型

模型名称	全参微调	LoRA微调	增量预训练
Baichuan2-13B	√	√	x
ChatGLM3-6B	√	√	x
GLM-4-9B	√	√	x
Llama2-13B	√	√	x
Llama2-70B	√	√	x
Llama2-7B	√	√	x
Llama3-70B	√	√	x
Llama3-8B	√	√	x
Qwen-14B	√	√	x

模型名称	全参微调	LoRA微调	增量预训练
Qwen-72B	√	√	x
Qwen-7B	√	√	x
Qwen1.5-14B	√	√	x
Qwen1.5-32B	√	√	x
Qwen1.5-72B	√	√	x
Qwen1.5-7B	√	√	x
Qwen2-72B	√	√	x
Qwen2-72B-1K	√	√	x
Qwen2-7B	√	√	x
Qwen2-1.5B	√	√	x
Qwen2-0.5B	√	√	x
Qwen2.5-72B	√	√	x
Qwen2.5-32B	√	√	x
Qwen2.5-14B	√	√	√
Qwen2.5-7B	√	√	x
Qwen2.5-72B-1K	√	√	x
Qwen2-VL-7B	√	√	x
Qwen2.5-72B-8K	√	√	x

支持的数据集格式

创建模型调优作业时，支持选择MOSS、Alpaca和ShareGPT这三种数据集格式。

- MOSS：用于存储和交换机器学习模型数据的数据集格式，文件类型为JSONL。
- Alpaca：用于训练语言模型的数据集格式，文件类型为JSONL。
- ShareGPT：用于分享GPT模型对话结果的数据集格式，文件类型为JSONL。

说明

- 请按数据集格式要求准备数据，否则会导致调优作业失败。
- 对于csv、xlsx文件，平台会将其转为Alpaca格式或MOSS格式，具体请参见**表7-3**。

表 7-3 模型与数据集格式说明

模型	调优类型	数据集格式（JSONL）	数据集格式（xlsx和csv）
Qwen2.5-72B及其余模型系列（权重格式为Megatron的模型，具体请参见表7-7）	全参微调、LoRA微调	MOSS、Alpaca、ShareGPT	MOSS
Qwen2.5-7B、Qwen2.5-14B、Qwen2.5-32B、Qwen2.5-72B-1K	全参微调、LoRA微调	Alpaca、ShareGPT	Alpaca
Qwen2.5-14B	增量预训练	Alpaca	不支持

数据集格式示例如下：

1. MOSS数据集格式：JSONL格式

   说明

   MOSS数据集格式仅支持微调。

   JSONL的一行数据就是数据集中的一条样本，建议总的数据样本不少于2000条。数据集示例如下，单轮对话也可以复用此格式。您可以单击下载，获取示例数据集“simple_moss.jsonl”，该数据集可以用于文本生成类型的模型调优。

   ```
   {"conversation_id": 1, "chat": {"turn_1": {"Human":"text","MOSS":"text"},"turn_2": {"Human":"text","MOSS":"text"}}}
   ```

   - "conversation_id"：样本编号。
   - "chat"：多轮对话的内容。
   - "turn_n"：表示是第n次对话，每次对话都有输入（对应Human角色）和输出（对应MOSS角色）。其中Human和MOSS仅用于角色区分，模型训练的内容只有text指代的文本。

2. Alpaca数据集格式

   a. 微调：jsonl格式

   ```
   {
     "instruction": "计算这些物品的数量之和。 ",
     "input": "输入：汽车 - 3，衣服 - 100，书 - 20。",
     "output": "汽车、衣服和书的数量之和为 3 + 100 + 20 = 123。"
   }
   ```

   b. 增量预训练：

   ```
   [
     {"text": "document"},
     {"text": "document"}
   ]
   ```

3. ShareGPT数据集格式

   说明

   - ShareGPT数据集格式仅支持微调。
   - ShareGPT格式支持更多的角色种类，例如human、gpt、observation、function等。它们构成一个对象列表呈现在conversations列中。

     注意：其中human和observation必须出现在奇数位置，gpt和function必须出现在偶数位置。

示例如下：

```
[
 {
  "conversations": [
   {
    "from": "human",
    "value": "human instruction"
   },
   {
    "from": "function_call",
    "value": "tool arguments"
   },
   {
    "from": "observation",
    "value": "tool result"
   },
   {
    "from": "gpt",
    "value": "model response"
   }
  ],
  "system": "system prompt (optional)",
  "tools": "tool description (optional)"
 }
]
```

4. csv、xlsx

   说明

   csv和xlsx格式数据集仅支持微调。

   表格里的一行数据就是一条样本。表格中仅有3个字段：conversation_id、human和assistant。

   - conversation_id：对话ID，可以重复，但必须是正整数。如果有多组Human-assiant对话使用同一个ID，则会按照文件中的顺序，将这几组对话编排成一个多轮对话。
   - human：对话输入，内容不能为空。
   - assistant：对话输出，内容不能为空。

   表 7-4 表格示例

conversation_id	human	assistant
1	text	text

模型调优时长估算

调优时长表示调优作业的“状态”处于“运行中”的耗时。由于训练吞吐有上下限，因此计算出的调优时长是个区间。

- 计算公式：调优时长 = 经验系数 x Iterations ÷（卡数 x 实例数 x 吞吐）+ 前后处理时间，取值请参见下表。
- 单位：小时

表 7-5 参数说明

参数	参数来源	说明
经验系数	非控制台参数	经验系数与模型训练迭代过程中处理的序列长度和批次大小有关。经验系数取值如下： ● Qwen2-72B-1K模型的经验系数 = 512 x 1024 ÷ 3600 ≈ 146 ● 其他模型的经验系数 = 32 x 4096 ÷ 3600 ≈ 36
Iterations	控制台参数	创建调优作业时计算得出的“迭代步数/Iterations”超参值。 **图 7-1** 迭代步数/Iterations 
卡数	控制台参数	和创建调优作业时选择的“规格”相关，例如，“规格”选择的是“Ascend: 4*Ascend-snt9b2(64GB) ”，*号前面的数字是4，则卡数就是4。 **图 7-2** 规格 

参数	参数来源	说明
实例数	控制台参数	创建调优作业时设置的“计算节点个数”。 **图 7-3** 计算节点个数 
吞吐	非控制台参数	吞吐表示每秒每卡处理的Tokens数量，吞吐值的上下限可以参考**表7-6**获取。 单位：tokens/s/p
前后处理时间	非控制台参数	调优时长还包括训练前的数据预处理、训练后格式转换等调优之外的耗时，统称为前后处理时间。 调优作业的前后处理时间和模型参数量相关，不同参数量的时间估值如下： • 72B： 1.35小时 • 32B：0.58小时 • 14B：0.25小时 • 7B：0.15小时

表 7-6 各模型的吞吐数据参考

模型名称	调优类型	吞吐下限取整	吞吐上限取整
Baichuan2-13B	全参微调	1200	1600
	LoRA微调	1300	1800
ChatGLM3-6B	全参微调	2000	2700
	LoRA微调	2300	3100
GLM-4-9B	全参微调	1800	2100
	LoRA微调	2400	2800

模型名称	调优类型	吞吐下限取整	吞吐上限取整
Llama2-13B	全参微调	1300	1800
	LoRA微调	1400	1900
Llama2-70B	全参微调	300	400
	LoRA微调	400	500
Llama2-7B	全参微调	3100	4200
	LoRA微调	3500	4700
Llama3-70B	全参微调	300	400
	LoRA微调	300	500
Llama3-8B	全参微调	2100	2800
	LoRA微调	2300	3100
Qwen-14B	全参微调	1200	1600
	LoRA微调	1400	1900
Qwen-72B	全参微调	300	400
	LoRA微调	300	500
Qwen-7B	全参微调	2100	2900
	LoRA微调	2200	3000
Qwen1.5-14B	全参微调	1300	1700
	LoRA微调	1400	1800
Qwen1.5-32B	全参微调	600	800
	LoRA微调	700	900
Qwen1.5-72B	全参微调	300	400
	LoRA微调	300	500
Qwen1.5-7B	全参微调	2200	3000
	LoRA微调	2600	3600
Qwen2-0.5B	全参微调	12800	17300
	LoRA微调	12800	17300
Qwen2-1.5B	全参微调	7300	9800
	LoRA微调	7300	9900
Qwen2-72B	全参微调	300	300
	LoRA微调	300	400

模型名称	调优类型	吞吐下限取整	吞吐上限取整
Qwen2-72B-1K	全参微调	300	300
	LoRA微调	300	400
Qwen2-7B	全参微调	2300	3200
	LoRA微调	2600	3500
Qwen2.5-72B	全参微调	100	120
	LoRA微调	280	330
Qwen2.5-32B	全参微调	340	410
	LoRA微调	480	570
Qwen2.5-14B	全参微调	1120	1320
	LoRA微调	1410	1660
	增量预训练	1120	1320
Qwen2.5-7B	全参微调	2459	2890
	LoRA微调	3180	3750
Qwen2.5-72B-1K	全参微调	250	300
	LoRA微调	340	400
Qwen2-VL-7B	全参微调	1500	1770
	LoRA微调	2100	2480
Qwen2.5-72B-8K	全参微调	320	400
	LoRA微调	380	480

查看调优作业详情

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“模型调优”进入作业列表。
3. 单击作业名称，进入调优作业详情页面，可以查看作业详情和日志。
   - “详情”：可以查看作业的基本信息，包括作业、模型、数据等设置信息。
   - “日志”：可以搜索、查看和下载作业日志。
     - 查看loss：当作业进入训练流程之后，会按照Step进行loss打印，因此在日志中搜索关键字段“lm loss”即可查看loss。
     - 获取训练吞吐数据：在打印的loss日志中搜索关键字段“elapsed time per iteration”获取每步迭代耗时，总的Token数可以用日志中的“global batch size”和“SEQ_LEN”相乘获得，训练的每卡每秒的吞吐=总Token数÷每步迭代耗时÷总卡数。

停止/继续调优作业

只有作业“状态”处于“运行中”、“等待中”、“告警”和“创建中”，才支持停止调优作业。

- 停止调优作业
  a. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入作业列表。
  b. 选择调优作业，单击操作列的“停止”，在弹窗中单击“确定”，暂停调优作业，作业“状态”变成“已停止”。
- 继续调优作业
  a. 当调优作业处于“已停止”状态时，单击右侧操作列的“继续”。
  b. 在“继续作业”对话框，仔细阅读提示信息，单击“确定”，即可从最新的Checkpoint启动作业，作业“状态”变成“启动中”。
  重新启动的作业将基于调优作业运行时长计费。

删除调优作业

说明

删除操作无法恢复，请谨慎操作。

1. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入作业列表。
2. 选择调优作业，单击操作列的“更多 > 删除”，在弹窗中输入“DELETE”，单击“确定”，删除作业。

查看 Checkpoint 与权重格式转换

- 权重格式有Huggingface和Megatron。
  - Huggingface格式可直接创建为我的模型或者添加为当前调优模型新版本。
  - Megatron格式需要将权重转换为Huggingface格式之后，才能创建为新的模型或者创建已有模型的新版本。
  - 转换权重格式时，会有预估费用提示框。预估费用仅作为参考，与实际收取费用可能存在偏差，请您以实际收取费用为准。
- 只有调优作业为已完成时，才可以添加为当前调优模型版本。

1. 查看Checkpoint。
   a. 在ModelArts Studio左侧导航栏中，选择“模型调优”进入作业列表。
   b. 在调优作业右侧，单击操作列的“更多 > 查看Checkpoint”。
   c. 在“Checkpoint列表”页面，可以查看Checkpoint的格式、保存路径等信息。

表 7-7 支持 Checkpoint 查看的模型

模型系列	模型名称	权重格式	说明
百川2	Baichuan2-13B	Megatron	中间产物需要做权重转换后使用。
ChatGLM3	ChatGLM3-6B		

模型系列	模型名称	权重格式	说明
GLM-4	GLM-4-9B		
Llama 2	Llama2-7B、Llama2-13B、Llama2-70B		
Llama 3	Llama3-8B、Llama3-70B		
通义千问	Qwen-7B、Qwen-14B、Qwen-72B		
通义千问1.5	Qwen1.5-7B、Qwen1.5-14B、Qwen1.5-32B、Qwen1.5-72B		
通义千问2	Qwen2-0.5B、Qwen2-1.5B、Qwen2-7B、Qwen2-72B、Qwen2-72B-1K		
通义千问2.5	Qwen2.5-72B		
	Qwen2.5-72B-8K		Qwen2.5-72B-8K暂不支持权重转换。
	Qwen2.5-14B、Qwen2.5-32B、Qwen2.5-72B、Qwen2.5-72B-1K	Huggingface	全参微调（sft）中间产物可以直接使用，LoRA微调中间产物不可使用。

2. 权重格式转换。下文以Megatron格式为例。
   - 场景一：将Checkpoint创建为我的模型。
     i. 在“Checkpoint列表”页面的“操作”列，单击“创建为我的模型”。
     ii. 在“创建为我的模型”页面，配置相关信息，然后单击“创建”。
     iii. 在“费用提醒”对话框，仔细阅读预估转换时长和费用信息，单击“确定”，跳转至“我的模型”页面创建模型。
     iv. 模型创建成功后，单击模型名称，在“我的版本”区域，单击版本号。
     v. 在“任务记录”区域，可以看到“作业类型”为“权重格式转换”的任务。
   - 场景二：将Checkpoint添加为调优后模型版本。
     i. 在“Checkpoint列表”页面的“操作”列，单击“添加为调优后模型版本”。
     ii. 在“添加为调优后模型版本”页面，配置相关信息，然后单击“创建”。
     iii. 在“费用提醒”对话框，仔细阅读预估转换时长和费用信息，单击“确定”，跳转至“模型详情”页面创建版本。
     iv. 版本创建成功后，单击版本号名称，在“任务记录”区域，可以看到“作业类型”为“权重格式转换”的任务。

后续操作

- 您可以将调优后的模型压缩，缓解资源占用问题。具体操作，请参见**使用ModelArts Studio（MaaS）压缩模型**。
- 您可以将调优后的模型部署为我的服务，并进行在线体验或API调用。具体操作，请参见**使用ModelArts Studio（MaaS）部署模型服务**、**在ModelArts Studio（MaaS）体验文本对话**和**调用ModelArts Studio（MaaS）部署的模型服务**。

相关文档

如果调优任务创建/运行失败，您可以参考以下文档进行定位：

- **ModelArts Studio（MaaS）调优数据集异常日志说明**
- **ModelArts Studio（MaaS）模型调优作业运行失败，报错：Modelarts.6001**
- **在ModelArts Studio（MaaS）创建Qwen2-0.5B或Qwen2-1.5B模型的LoRA微调类型的调优任务，显示创建失败**
- **在ModelArts Studio（MaaS）创建训练任务，显示创建失败**

8 ModelArts Studio（MaaS）应用中心

8.1 ModelArts Studio（MaaS）应用管理

8.1.1 ModelArts Studio（MaaS）应用广场概述

背景信息

随着大模型技术的快速发展，其性能持续提升，应用效果显著增强，同时MCP生态体系日益完善。在此背景下，基于MCP构建应用已成为企业提升核心竞争力的关键路径。然而，企业在实际构建应用时面临多重挑战：一方面，由于场景理解深度不足，难以充分发挥模型与MCP的协同能力；另一方面，不同场景下的应用管理缺乏统一的标准化参考体系，导致应用开发和部署的门槛较高。

为有效解决上述痛点，MaaS提供了一个标准化的应用模板广场，通过汇聚一系列预置的应用模板（如联网搜索MCP+大模型等典型场景模板），为企业提供开箱即用的场景化解决方案，提升应用构建效率和效果。

应用广场介绍

MaaS应用广场提供了多个预置应用模板，帮助您“一键复制”完成基础应用搭建。应用广场中的应用为卡片式布局，每个应用卡片会显示应用名称、简介、标签特性等关键信息。应用广场提供MaaS平台的预置应用和三方平台的应用。

- MaaS平台的预置应用（精选应用）：包括可实现高效信息检索的DeepSeek V3联网搜索、辅助市场调研工作的企业调研助理等。单击应用卡片可以查看应用详情，包括应用的基本信息、关联服务及云产品、MCP和相关的计费信息。支持一键复制应用至“应用管理”页面进行创建。
- 三方平台的应用（热门大模型应用解决方案）：包括可用于智能客户场景的数字人交互智能问答解决方案、助力数据可视化分析的快速体验智能问数等。单击应用卡片后，将跳转至相关页面。您可以按照页面提示创建应用。

计费说明

在应用广场一键复制应用、在应用管理创建应用这两个操作本身不收费，仅调试与预览、调用操作涉及计费。计费项如下：

- 调试和预览计费
  - 使用模型服务进行调试和预览：将消耗Tokens，按照推理输入输出价格计费。费用请以实际账单为准。计费详情，请参见**ModelArts Studio（MaaS）模型推理计费项**。
  - MCP：
    - MCP服务返回的内容将被计入输入Token消耗。
    - 调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。
- 调用计费

  应用发布后，将托管至函数工作流FunctionGraph服务中。实际计费请以FunctionGraph计费为准，详情请参见**FunctionGraph函数工作流计费规则**。您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。
  - 调用应用按调用请求次数、资源用量（函数执行时间）收费。
  - MCP服务返回的内容会输入模型，用于模型的分析理解和总结，将被计入输入Token的消耗。调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

8.1.2 在 ModelArts Studio（MaaS）应用广场一键复制应用

MaaS提供了一个标准化的应用模板广场，通过汇聚一系列预置的应用模板和三方平台的应用模板（如DeepSeek V3联网搜索、企业调研助理等），为企业提供开箱即用的场景化解决方案，显著提升应用构建效率和效果。

计费说明

在应用广场一键复制应用、在应用管理创建应用这两个操作本身不收费，仅调试与预览、调用操作涉及计费。计费项如下：

- 调试和预览计费
  - 使用模型服务进行调试和预览：将消耗Tokens，按照推理输入输出价格计费。费用请以实际账单为准。计费详情，请参见**ModelArts Studio（MaaS）模型推理计费项**。
  - MCP：
    - MCP服务返回的内容将被计入输入Token消耗。
    - 调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。
- 调用计费

  应用发布后，将托管至函数工作流FunctionGraph服务中。实际计费请以FunctionGraph计费为准，详情请参见**FunctionGraph函数工作流计费规则**。您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。
  - 调用应用按调用请求次数、资源用量（函数执行时间）收费。
  - MCP服务返回的内容将被计入输入Token消耗。调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见**注册华为账号并开通华为云**和**实名认证**。
- 已完成ModelArts委托授权。具体操作，请参见**配置ModelArts Studio（MaaS）访问授权**。

一键复制应用并创建

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“应用广场”。
3. 在“应用广场”页面，单击“可复制”标签的应用卡片，查看应用详情。

   应用详情页面会显示应用的基本信息、关联服务及云产品、MCP和计费信息。
4. 在“应用广场”页面右上角，单击“一键复制应用”。
5. 在“应用模板使用确认”面板，按需选择以下操作。
   - 未授权云产品：勾选“一键开通和授权上述所有构建应用所依赖的相关服务”和“我已阅读并同意上述所有协议和计费规则”，然后单击“确定”。

     部分第三方MCP服务需要输入第三方API Key（请以实际环境为准），您可以参照第三方MCP服务的对应文档，获取API Key。
   - 已授权云产品：勾选“我已阅读并同意上述所有协议和计费规则”，然后单击“确定”。
6. 跳转至“创建应用”页面，按需修改应用名称、模型服务及其参数、提示词和MCP服务，在右上角单击“保存草稿”进行预览与调试，或者单击“发布”将应用进行发布。

   关于创建应用的具体操作，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。

后续操作

- 您可以对草稿应用或已发布的应用进行编辑。具体操作，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。
- 您可以对已发布的应用进行调用。具体操作，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。
- 您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。

  数据可能存在1~2小时的时延。

8.1.3 在 ModelArts Studio（MaaS）应用管理创建应用

MaaS提供应用管理功能，支持用户通过可视化操作界面，一键创建AI应用。用户可灵活选择模型服务、设置系统提示词、添加MCP等，并进行应用效果预览与调试。将应用发布后进行调用。

操作场景

在当今快速变化为数字化时代，企业与开发者面临着AI应用开发周期长、成本高、定制化难等挑战。传统开发方式需投入大量时间与技术资源，难以快速响应市场需求。MaaS平台支持用户通过可视化界面一键创建个性化AI应用，灵活配置应用的相关信息，快速完成应用的开发与调试，实现按需调用、按量计费，有效解决开发效率与成本控制问题，帮助企业与开发者低成本、高效率地实现AI能力落地。

约束限制

- 模型服务限制：仅支持在线推理-预置服务或在线推理-我的服务。更多信息，请参见**ModelArts Studio（MaaS）在线推理服务**。
- 模型类型限制：仅支持文本生成类型。
- MCP限制：仅支持MCP广场开通的服务或自定义MCP已部署的服务，且最多可添加5个MCP服务。更多信息，请参见**ModelArts Studio（MaaS）MCP管理**。
- 调用限制：仅已发布的应用可进行调用。更多信息，请参见**ModelArts Studio（MaaS）应用管理**。

计费说明

在应用广场一键复制应用、在应用管理创建应用这两个操作本身不收费，仅调试与预览、调用操作涉及计费。计费项如下：

- 调试和预览计费
  - 使用模型服务进行调试和预览：将消耗Tokens，按照推理输入输出价格计费。费用请以实际账单为准。计费详情，请参见**ModelArts Studio（MaaS）模型推理计费项**。
  - MCP：
    - MCP服务返回的内容将被计入输入Token消耗。
    - 调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。
- 调用计费

  应用发布后，将托管至函数工作流FunctionGraph服务中。实际计费请以FunctionGraph计费为准，详情请参见**FunctionGraph函数工作流计费规则**。您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。
  - 调用应用按调用请求次数、资源用量（函数执行时间）收费。
  - MCP服务返回的内容将被计入输入Token消耗。调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

前提条件

- 预置服务、我的服务或自定义接入点满足以下条件：
  - 预置服务中商用服务：已开通预置服务中的商用服务。具体操作，请参见**在ModelArts Studio（MaaS）预置服务中开通商用服务**。
  - 预置服务中免费服务：使用有效期内的免费服务。具体操作，请参见**在ModelArts Studio（MaaS）预置服务中体验免费服务**。
  - 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**。
  - 我的服务：已在“我的服务”页面部署模型服务且状态为运行中。具体操作，请参见**使用ModelArts Studio（MaaS）部署模型服务**。
- 已在MCP广场开通MCP服务或创建自定义MCP服务。具体操作，请参见**ModelArts Studio（MaaS）MCP管理**或**在ModelArts Studio（MaaS）创建自定义MCP服务**。

使用主账号创建应用

MaaS支持使用主账号或子账号创建应用。主账号在首次创建应用时会弹出FunctionGraph授权提示，按提示信息操作即可一键授权；子账号需要主账号为其授予FunctionGraph Administrator权限，否则无法创建应用，详情请参见**使用子账号创建应用**。

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“应用中心 > 应用管理”。
3. 在“应用管理”页面右上角，单击“创建应用”。
4. 在“创建应用”页面，配置相关信息。

   首次创建应用时，会提示“发布应用需要授权ModelArts Studio获取函数工作流（FunctionGraph）服务中的函数基础信息”，请您“点击授权”，在“授权提醒”对话框，单击“同意授权”，完成FunctionGraph授权避免影响应用发布。

图 8-1 授权提醒

表 8-1 参数说明

参数	说明
应用名称	系统自动创建应用名称，您可以按需修改。支持1~100位只包含中英文、数字、下划线（_）、中划线（-）和英文半角句号（.）的名称。
模型服务	单击“选择模型服务”，按需选择预置服务或我的服务，然后单击“确定”。预置服务中的商用服务支持选择主服务或其子服务版本。 单击⇌图标，可以修改模型服务；单击⚙图标，可以设置模型服务的参数，以获得更好的推理效果。模型参数说明如下： ● 温度/Temperature：该参数用于控制生成文本的随机性或创造性，数值越高，生成内容更具备多样性和创新性，但也更有可能包含错误或不连贯的内容。 ● 核采样/top_p：调整输出文本的多样性，数值越大，生成文本的多样性就越高。 ● top_k：通常是指在模型的输出中选择概率最高的前K个结果。
提示词	自定义应用的角色和任务背景，长度范围为1~4096个字符。

参数	说明
MCP	应用可以通过模型上下文协议（Model Context Protocol）实现各类外部工具的API调用。单击“添加MCP”，按需选择“MCP广场”或“自定义MCP”页签下的MCP，单击“确定”。最多可添加5个MCP。 MCP广场的MCP需要开通才能选中。如果未开通，您可以在目标MCP卡片，单击“立即开通”，部分服务需要输入API Key，然后单击“确认开通”。 使用部分官方预置MCP服务时，您需要填写相应的API Key，请参考对应产品的文档进行操作。

5. 配置完成后，您可以在右上角单击“保存草稿”进行预览与调试，或者单击“发布”将应用进行发布。
   - 保存草稿：保存为草稿后，您可以在“预览”区域，使用模型服务进行调试和预览。您也可以修改模型服务及其参数、提示词等获取更好的效果。

     关于调试和预览的计费，请参见**计费说明**。
   - 发布：单击“发布”后，在“计费提醒”对话框，仔细阅读计费相关信息，然后单击“确定”。

     关于调用的计费，请参见**计费说明**。在“应用管理”页面的目标应用卡片中，当应用状态显示为已发布，表示该应用可以调用。

     发布应用后，您可以在应用卡片单击“编辑”，在“编辑应用”页面按需修改应用名称、模型服务、提示词、添加或删除MCP等，也可以在应用卡片单击“调用”，按照页面信息调用应用。
6. 应用创建成功后，会在“应用管理”页面显示。

   您可以在应用卡片中查看应用的状态、模型服务、应用ID和发布/保存时间，还可以对应用进行编辑、调用或删除等操作。功能按钮置灰，表示应用暂不支持该操作，您可以按照提示处理。

使用子账号创建应用

子账号必须拥有FunctionGraph Administrator权限，才能成功创建应用。

1. 使用主账号为子账号所在的用户组授予FunctionGraph Administrator权限。更多信息，请参见**创建用户组并授权**。

   图 8-2 为用户组授予 FunctionGraph Administrator 权限

2. 使用子账号创建应用。
   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“应用中心 > 应用管理”。
   c. 在“应用管理”页面右上角，单击“创建应用”。
   d. 在“创建应用”页面，配置相关信息。

表 8-2 参数说明

参数	说明
应用名称	系统自动创建应用名称，您可以按需修改。支持1~100位只包含中英文、数字、下划线（_）、中划线（-）和英文半角句号（.）的名称。
模型服务	单击“选择模型服务”，按需选择预置服务或我的服务，然后单击“确定”。预置服务中的商用服务支持选择主服务或其子服务版本。 单击⇄图标，可以修改模型服务；单击⚙图标，可以设置模型服务的参数，以获得更好的推理效果。模型参数说明如下： ● 温度/Temperature：该参数用于控制生成文本的随机性或创造性，数值越高，生成内容更具备多样性和创新性，但也更有可能包含错误或不连贯的内容。 ● 核采样/top_p：调整输出文本的多样性，数值越大，生成文本的多样性就越高。 ● top_k：通常是指在模型的输出中选择概率最高的前K个结果。
提示词	自定义应用的角色和任务背景，长度范围为1~4096个字符。
MCP	应用可以通过模型上下文协议（Model Context Protocol）实现各类外部工具的API调用。单击“添加MCP”，按需选择“MCP广场”或“自定义MCP”页签下的MCP，单击“确定”。最多可添加5个MCP。 MCP广场的MCP需要开通才能选中。如果未开通，您可以在目标MCP卡片，单击“立即开通”，部分服务需要输入API Key，然后单击“确认开通”。 使用部分官方预置MCP服务时，您需要填写相应的API Key，请参考对应产品的文档进行操作。

e. 配置完成后，您可以在右上角单击“保存草稿”进行预览与调试，或者单击“发布”将应用进行发布。

- 保存草稿：保存为草稿后，您可以在“预览”区域，使用模型服务进行调试和预览。您也可以修改模型服务及其参数、提示词等获取更好的效果。

  关于调试和预览的计费，请参见**计费说明**。

- 发布：单击“发布”后，在“计费提醒”对话框，仔细阅读计费相关信息，然后单击“确定”。

  关于调用的计费，请参见**计费说明**。在“应用管理”页面的目标应用卡片中，当应用状态显示为已发布，表示该应用可以调用。

  发布应用后，您可以在应用卡片单击“编辑”，在“编辑应用”页面按需修改应用名称、模型服务、提示词、添加或删除MCP等，也可以在应用卡片单击“调用”，按照页面信息调用应用。

f. 应用创建成功后，会在“应用管理”页面显示。

您可以在应用卡片中查看应用的状态、模型服务、应用ID和发布/保存时间，还可以对应用进行编辑、调用或删除等操作。功能按钮置灰，表示应用暂不支持该操作，您可以按照提示处理。

编辑和发布草稿应用

您可以在“编辑应用”页面按需修改草稿应用的名称、模型服务、提示词、添加或删除MCP等。

1. 在“应用管理”页面，在状态为草稿的目标应用卡片单击“编辑”。
2. 在“编辑应用”页面，按需修改应用名称、模型服务、提示词、添加或删除MCP等，在右上角单击“保存草稿”或“发布”。

   关于参数的说明，请参见**表8-1**。单击“发布”后，会出现“计费提醒”对话框，请仔细阅读计费相关信息，然后单击“确定”。关于如何计费，请参见**计费说明**。

   在“应用管理”页面的目标应用卡片中，当应用状态显示为已发布，表示该应用可以调用。

编辑已发布的应用

您可以在“编辑应用”页面按需修改已发布应用的名称、模型服务、提示词、添加或删除MCP等。

1. 在“应用管理”页面，在状态为已发布的目标应用卡片单击“编辑”。
2. 在“编辑应用”页面，按需修改应用名称、模型服务、提示词、添加或删除MCP等，在右上角单击“更新”。

调用已发布的应用

仅发布的应用支持调用。如果应用的状态为草稿，需要先进行发布才能调用。具体操作，请参见**编辑和发布草稿应用**。

1. 在“应用管理”页面，在状态为已发布目标应用卡片，单击“调用”，在“调用说明”页面按照页面信息进行调用。

   关于创建API Key的操作步骤，请参见**在ModelArts Studio（MaaS）管理API Key**。
2. 在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。

   数据可能存在1~2小时的时延。

删除应用

对于不需要的应用，您可以进行删除操作。**删除后，该应用不可恢复，请谨慎操作。**

1. 在“应用管理”页面，在目标应用卡片单击“删除”。
2. 在“删除应用”对话框，单击“确定”。

常见问题

- 调用应用后，效果不理想怎么办?

  您可以尝试优化系统提示词，明确指令逻辑，调整模型服务的参数等。

- 应用发布后能否修改配置?

  应用发布后，您可以在应用卡片中单击“编辑”修改配置。具体操作，请参见**编辑已发布的应用**。

8.2 ModelArts Studio（MaaS）MCP 管理

8.2.1 ModelArts Studio（MaaS）MCP 概述

什么是 MCP

模型上下文协议MCP（Model Context Protocol）旨在搭建大模型和外部工具之间的信息传递通道。MCP服务是MaaS提供的标准化中间件能力模块，通过预集成第三方服务或平台自研功能，帮助开发者快速扩展AI应用的专业能力。其核心价值在于将复杂的外部服务（如网页搜索、出行服务、开发工具等）封装为即插即用的功能组件，用户无需开发底层接口即可直接调用，显著降低多能力协同的开发成本。关于MCP的更多信息，请参见**MCP官网**。

MCP 技术架构

MCP采用模块化的客户端-服务器架构，解耦前端应用和后端服务。其典型部署结构由三大核心组件构成：

- 主机进程（Host）：作为大模型的运行载体，承载各类终端应用（如IDE、Cursor等），为用户提供直接的AI服务接口。
- MCP客户端（Client）：充当桥梁角色，作为主机进程与服务器间的抽象接口层，通过标准化流程统一通信规范，高效处理协议转换任务，确保数据交互的一致性与稳定性。
- MCP服务器（Server）：通过标准化的MCP协议向客户端提供服务能力，涵盖智能问答、内容生成等核心功能。

在通信机制方面，MCP客户端采用JSON-RPC接口实现与服务器的交互，支持标准化输入/输出（stdio）和流式传输HTTP（Streamable HTTP）等多种传输模式，既保障数据传输的高效性，又能灵活适配不同的应用场景需求，实现低延迟、高吞吐量的通信体验。

MCP 工作流程

当用户输入自然语言请求时，MCP架构通过五层协同处理机制实现高效响应，具体流程如下：

1. 请求解析与转发：主机进程解析用户自然语言请求的意图，并通过标准化接口将结构化指令传递给MCP客户端。
2. 服务发现与路由：客户端基于请求类型，动态查询注册的MCP服务器能力清单，路由至匹配的后端服务资源（例如数据库查询接口、API工具等）。
3. 任务执行：目标MCP服务器接收指令，执行原子化操作（例如数据库检索、第三方API调用、本地计算任务等）。
4. 结果回传：服务器将结构化执行结果（JSON/Protobuf格式）通过双向通道返回客户端，客户端完成协议适配后递交给主机进程。
5. 响应生成与呈现：主机进程融合当前会话上下文与返回数据，驱动大模型生成自然语言响应，最终通过前端界面反馈给用户。

整个处理链路采用全自动化闭环设计，用户无需感知底层复杂逻辑，从而实现毫秒级响应、高精准输出的交互体验。

如何使用 MCP

MaaS支持接入两种MCP服务。

- 预置MCP服务：MaaS提供丰富的MCP Server资源，涵盖地理位置（高德地图、百度地图）、图像编辑（美图影像）、Web搜索（联网增强MCP）等多种优质服务，方便您快速开通并接入应用。具体操作，请参见**在ModelArts Studio（MaaS）MCP广场开通预置MCP服务**。
- 自定义MCP服务：MaaS支持部署开源社区和自行开发的MCP服务。自定义MCP服务会被部署到函数工作流FunctionGraph中，无需配置和管理服务器等基础设施，函数以弹性、免运维、高可靠的方式运行。具体操作，请参见**在ModelArts Studio（MaaS）创建自定义MCP服务**。

MCP 部署方式

MaaS支持本地部署和云端部署MCP服务。

- 本地部署：不可以直接开通使用，仅提供元数据。您可以在“MCP广场”页面查看支持本地部署的MCP服务和JSON配置文件，然后在“MCP管理”页面通过NPX、UVX等方式进行部署。
- 云端部署：可以直接在“MCP广场”页面开通使用，包括MCP官方、三方平台以及MaaS云端部署的MCP服务，提供SSE访问方式。

表 8-3 部署方式说明

维度	本地部署	云端部署
定义	将MCP平台直接部署在企业本地数据中心或自有服务器上，系统完全由企业自主管理。	将MCP平台部署在公有云或第三方托管服务上，通过网络远程管理集群。
通信方式	Stdio（标准输出的本地通信方式）	SSE（远程通信）
部署方式	NPX、UVX	SSE（Remote URL）
优势	安全、自主可控，适合有敏感数据的场景。	• 弹性扩缩容，动态负载均衡。 • 无需企业运维。 • SSE的访问方式，更方便构建标准化的应用。
适用场景	• 重视数据安全，且有成熟的运维团队。 • C端用户结合工具本地部署，部署后主要本机用户访问。	• 访问量较大，需要弹性扩缩容。 • B端用户，部署后，可提供多用户远程访问。
相关文档	**在ModelArts Studio（MaaS）创建自定义MCP服务**	**在ModelArts Studio（MaaS）MCP广场开通预置MCP服务**

计费说明

- 预置MCP服务：在MCP广场开通MCP服务不涉及收费。调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。
- 自定义MCP服务：创建自定义MCP服务不涉及计费。调用自定义MCP服务时，实际计费请以FunctionGraph计费为准，详情请参见**FunctionGraph函数工作流计费规则**。您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。

8.2.2 在 ModelArts Studio（MaaS）MCP 广场开通预置 MCP 服务

MaaS服务提供丰富的MCP Server资源，涵盖地理位置（高德地图、百度地图）、图像编辑（美图影像）、Web搜索（联网增强MCP）等多种优质服务，帮助您快速扩展智能应用能力。您可以在MCP广场选择开通所需的MCP服务，然后将已开通的MCP服务添加到应用中，完成发布后即可实现调用，大幅降低AI开发门槛。

操作场景

随着AI技术的快速发展，应用需要无缝接入企业数据库、代码仓库、API服务等多样化数据源，以提升决策质量。然而，传统数据集成方式面临三大核心挑战：异构数据源接口的碎片化导致开发与维护成本居高不下；跨系统间缺乏有效的上下文感知能力；安全管控机制的不统一。

为了解决这些难题，Anthropic推出的Model Context Protocol（MCP）作为一种开放协议标准，通过定义统一的上下文交互接口，实现了AI系统与数据源之间的即插即用式连接。通过在应用平台中集成MCP协议，可以构建标准化的数据访问层，从而有效支撑应用在复杂业务场景下的上下文感知能力，为AI决策提供更高质量的支持。借助MCP协议，开发者无需为每个外部工具编写复杂的接口，MaaS应用也能够接入海量第三方工具。

关于MCP的详细信息，请参见**ModelArts Studio（MaaS）MCP概述**。

MCP 广场介绍

MCP广场汇聚了优质的MCP服务，可以扩展应用能力。您可以在MCP广场左侧筛选区域，按需选择部署方式和类型。

- 部署方式：MaaS支持本地部署和云端部署MCP服务。关于部署方式的具体说明，请参见**MCP部署方式**。
- 类型：包括开发工具、浏览器自动化、网页搜索、金融服务等多种类型，您可以按需选择。

您可以在MCP广场单击目标服务，查看MCP服务的详细介绍、使用场景等信息。

图 8-3 MCP 广场

预置 MCP 服务的使用流程

1. **开通MCP服务**：在“MCP广场”页面开通所需服务。
2. **创建应用**：在“应用管理”页面创建应用，添加已开通的MCP服务，进行在线调试与预览或者直接发布。
3. **调用已发布的应用**：在“应用管理”页面，调用已发布的应用API。
4. 查看应用调用情况：在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。

约束限制

- 服务开通：部分MCP服务需要在对应第三方平台进行实名认证或额外的资质审核。
- 数据合规：使用第三方MCP服务时，需遵守对应平台的数据使用协议。

计费说明

在MCP广场开通MCP服务不涉及收费。调用MCP服务时，可能会涉及到第三方平台服务的使用费用，请以第三方平台的计费规则为准。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见**注册华为账号并开通华为云**和**实名认证**。
- 已完成ModelArts委托授权。具体操作，请参见**配置ModelArts Studio（MaaS）访问授权**。

开通 MCP 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，单击“MCP广场”。
3. 在“MCP广场”页面，在左侧筛选区域选择“云端部署”，按需选择MCP类型，也可以在搜索框输入MCP名称进行搜索。
4. 在“MCP广场”页面，单击目标服务卡片，查看服务的详情，然后单击“立即开通”。

- 概览页签：查看服务的介绍、关键能力、使用场景、MCP使用说明、常见问题等信息。
- 工具页签：查看MCP服务可用的工具、工具描述及参数说明。

5. 在“开通MCP服务”对话框，部分第三方MCP服务需要输入API Key（请以实际环境为准），然后单击“确认开通”。

   您可以参照第三方MCP服务的对应文档，获取API Key。开通后，MCP服务卡片中会显示“已开通”。

查看已开通的 MCP 服务

您可以在“MCP管理”页面，查看所有已开通的MCP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“应用中心 > MCP管理”。
3. 在“MCP管理”页面的“已开通的MCP”页签，查看所有已开通的MCP服务。

图 8-4 已开通的 MCP 服务

将 MCP 服务添加至应用进行发布并调用

开通预置MCP服务后，您可以在“应用管理”页面创建应用，添加已开通的MCP服务，将应用进行发布并调用。具体操作，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。

更新鉴权信息

当MCP服务的鉴权信息发生变更时，需要更新鉴权信息，并更新发布使用该服务的应用。

1. 在“MCP广场”页面，单击目标MCP服务卡片，在MCP详情右上角单击“更新鉴权信息”。
2. 在“更新鉴权信息”对话框，输入API Key，单击“确认更新”。
3. 更新MCP鉴权信息后，需要更新发布使用该服务的应用，使MCP鉴权信息生效。

   在左侧导航栏，单击“应用管理”，在使用该MCP服务的应用卡片单击“编辑”，在“编辑应用”页面右上角，单击“更新”。更多信息，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。

取消开通 MCP 服务

对于已开通但不再需要的MCP服务，您可以取消开通该服务。如果发布的应用已添加该MCP，则无法直接取消开通该服务。

1. 在“MCP广场”页面，单击目标服务卡片，查看服务的详情，然后在右上角单击“取消开通”。
2. 在“取消开通MCP服务”对话框，按需选择以下操作。
   - 如果发布的应用已添加该MCP服务：
     i. “取消开通MCP服务”对话框会提示无法取消开通该MCP服务，您可以单击应用名称，在“编辑应用”页面将需要取消的MCP进行删除，然后在右上角单击“更新”。
     ii. 重复以上操作，确保发布的应用均已删除该MCP服务。
     iii. 在“取消开通MCP服务”对话框，输入**YES**，单击“确定”。
   - 如果发布的应用未添加该MCP服务：输入**YES**，单击“确定”。

   取消开通MCP服务后，在“MCP广场”页面的目标卡片中，将不再显示“已开通”。

常见问题

MCP广场的所有MCP服务都可以正常部署使用吗?

MCP广场提供了获取MCP服务的渠道，但由于MCP服务提供商可能会修改或关闭服务，MaaS不保证这些服务总是可部署使用。

8.2.3 在 ModelArts Studio（MaaS）创建自定义 MCP 服务

除了开通预置的MCP服务，MaaS还支持部署开源社区和自行开发的MCP服务。自定义MCP服务会被部署到函数计算FunctionGraph中，无需配置和管理服务器等基础设施，函数以弹性、免运维、高可靠的方式运行。

操作场景

当预置的MCP服务无法满足个性化需求（如特殊业务逻辑处理、特定工具集成等）时，您可以基于开源社区资源或自主开发，通过本地（NPX/UVX）和远端（SSE）部署方式搭建专属MCP服务，实现业务流程的深度定制与高效运行。

约束限制

- 同一账户下不允许存在同名的自定义MCP Server。
- 使用开源社区方案或自主开发服务需遵循对应开源协议与法律法规。

计费说明

- 创建自定义MCP服务不涉及计费。
- 使用预留实例或调用自定义MCP服务可能会产生费用。实际计费请以FunctionGraph计费为准，详情请参见**FunctionGraph函数工作流计费规则**。您可以在**FunctionGraph控制台**查看应用的调用总量统计及资源用量统计。

  使用预留实例：按照部署时长额外计费，从创建开始计费，直到取消使用预留实例或者停用MCP服务。关于计费详情，请参见**预留实例计费规则**。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见**注册华为账号并开通华为云**和**实名认证**。
- 已完成ModelArts委托授权。具体操作，请参见**配置ModelArts Studio（MaaS）访问授权**。

创建自定义 MCP 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“应用中心 > MCP管理”。
3. 在“MCP管理”页面右上角，单击“创建MCP”。
4. 在“创建MCP”页面，配置相关信息。

   初次创建MCP时，会出现“授权提醒”弹窗，请单击“同意授权”完成授权避免影响应用发布。

图 8-5 授权提醒

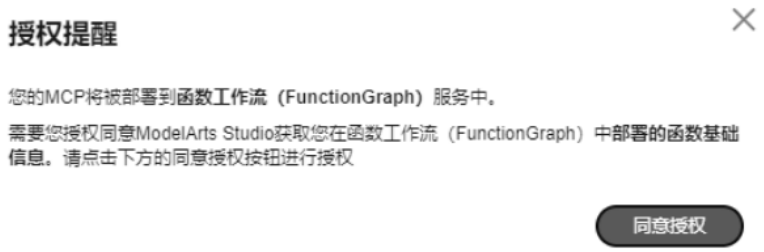


表 8-4 创建 MCP 参数说明

参数	说明
MCP名称	自定义MCP的名称。MCP名称具有**唯一性**，不能重复。支持1~256位只包含中英文，数字，下划线（_）、中划线（-）和半角句号（.）的名称。
描述	自定义MCP的描述，最多支持1024字符。
部署方式	支持NPX、UVX和Remote URL三种部署方式，请您按需选择。关于部署方式的更多信息，请参见**MCP部署方式**。 • 如果您需要托管本地MCP服务（stdio），可以选择NPX或UVX。 – NPX：部署Node.js环境下运行的MCP Server。 – UVX：部署Python环境下运行的MCP Server。 • 如果您需要连接远程MCP服务（SSE），可以选择Remote URL。 Remote URL：接入部署好的SSE协议通信的MCP Server。

参数	说明
使用预留实例	仅“部署方式”选择“NPX”或“UVX”时，支持设置该参数。 适用于对时延要求较高的场景，通过预留实例预热函数，从而消除冷启动对时延的影响。预留实例按照部署时长额外计费，从创建开始计费，直到停用MCP服务。

MCP服务配置	需符合所选部署方式的标准格式。在JSON中，NPX和UVX需要指定"command": "npx", 或 "command": "uvx"，Remote URL（SSE）方式需要有url字段。请确保JSON中只包含一个MCP Server，如果存在名称相同的MCP Server，仅最后一个会部署。
	您可以在“MCP广场”页面，“部署方式”选择“本地部署”，单击MCP应用卡片，在“概览”页签查看MCP服务对应的JSON配置文件。
	代码示例如下：
	● NPX
	{
	"mcpServers": {
	"amap-maps": {
	"command": "npx",
	"args": [
	"-y",
	"@amap/amap-maps-mcp-server"
	],
	"env": {
	"AMAP_MAPS_API_KEY": "******"
	}
	}
	}
	}
	● UVX
	{
	"mcpServers": {
	"MCP-timeserver": {
	"command": "uvx",
	"args": ["MCP-timeserver"]
	}
	}
	}
	● Remote URL
	{
	"mcpServers": {
	"amap-maps-sse": {
	"url": "https://mcp.amap.com/sse?key=高德开放平台上申请的Key"
	}
	}
	}

5. 确认配置信息和计费无误后，单击“立即创建”。

   创建完成后，在“MCP管理”页面的“自定义MCP”页签，可以看到新建的MCP服务，且状态为“已部署”。您可以单击MCP服务卡片查看详情，也可以进行编辑、停用、删除等操作。

MCP 部署方式

MaaS支持本地部署和云端部署MCP服务。

- 本地部署：不可以直接开通使用，仅提供元数据。您可以在“MCP广场”页面查看支持本地部署的MCP服务和JSON配置文件，然后在“MCP管理”页面通过NPX、UVX等方式进行部署。
- 云端部署：可以直接在“MCP广场”页面开通使用，包括MCP官方、三方平台以及MaaS云端部署的MCP服务，提供SSE访问方式。

表 8-5 部署方式说明

维度	本地部署	云端部署
定义	将MCP平台直接部署在企业本地数据中心或自有服务器上，系统完全由企业自主管理。	将MCP平台部署在公有云或第三方托管服务上，通过网络远程管理集群。
通信方式	Stdio（标准输出的本地通信方式）	SSE（远程通信）
部署方式	NPX、UVX	SSE（Remote URL）
优势	安全、自主可控，适合有敏感数据的场景。	• 弹性扩缩容，动态负载均衡。 • 无需企业运维。 • SSE的访问方式，更方便构建标准化的Agent。
适用场景	• 重视数据安全，且有成熟的运维团队。 • C端用户结合工具本地部署，部署后主要本机用户访问。	• 访问量较大，需要弹性扩缩容。 • B端用户，部署后，可提供多用户远程访问。
相关文档	**在ModelArts Studio（MaaS）创建自定义MCP服务**	**在ModelArts Studio（MaaS）MCP广场开通预置MCP服务**

将 MCP 服务添加至应用进行发布并调用

自定义MCP服务创建完成后，您可以在“应用管理”页面创建应用，添加已创建的自定义MCP服务，将应用进行发布并调用。具体操作，请参见**在ModelArts Studio（MaaS）应用管理创建应用**。

编辑自定义 MCP 服务

已部署的MCP服务支持修改MCP名称、MCP服务配置等信息。部署方式不支持修改。更新自定义MCP服务将会重新部署该服务。

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“应用中心 > MCP管理”。
3. 在“MCP管理 > 自定义MCP”页面，单击MCP卡片中的“编辑”。

4. 在“编辑MCP”页面，按需修改MCP名称、MCP服务配置等信息，然后单击“更新”。

   关于参数说明，请参见**表8-4**。

5. 在“更新MCP”对话框，按需选择以下操作。
   - 如果发布的应用已关联该MCP服务：
     i. “更新MCP”对话框会提示已关联的MCP应用不会自动同步更新，为确保可用性，请在MCP更新后，重新发布关联应用。
     ii. 在“更新MCP”对话框的“关联应用列表”，单击或记录已关联的应用名称。
     iii. 在“更新MCP”对话框，输入**YES**，单击“确认”。
     iv. 重新发布已关联该MCP的应用。
   - 如果发布的应用未添加该MCP服务：单击“确认”。

停用/启动自定义 MCP 服务

已部署的MCP服务支持停用操作，停用后支持重新启动。

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“应用中心 > MCP管理”。
3. 在“MCP管理 > 自定义MCP”页面，按需进行以下操作。
   - 停用自定义MCP服务：
     - 如果发布的应用已添加该MCP服务：
       1) “停用MCP”对话框会提示无法停止该MCP服务，您可以单击应用名称，在“编辑应用”页面将删除该MCP，然后在右上角单击“更新”。
       2) 重复以上操作，确保发布的应用均已删除该MCP服务。
       3) 在“停用MCP”对话框，输入**YES**，单击“确定”。
     - 如果发布的应用未添加该MCP服务：在“停用MCP”对话框，输入**YES**，单击“确定”。

     停用后，MCP卡片将显示为“停用”。
   - 启用自定义MCP服务：

     在已停用的MCP卡片单击“启用”。启用后，MCP卡片将显示为“已部署”。

删除自定义 MCP 服务

对于不需要的MCP服务，您可以进行删除操作。删除后将无法恢复，请谨慎操作。如果发布的应用已添加该MCP，则无法直接删除该自定义MCP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“应用中心 > MCP管理”。
3. 在“MCP管理 > 自定义MCP”页面，单击MCP卡片中的“删除”。
4. 在“删除MCP”对话框，按需进行以下操作。
   - 如果发布的应用已添加该MCP服务：

i. “停用MCP”对话框会提示无法停止该MCP服务，您可以单击应用名称，在“编辑应用”页面删除该MCP，然后在右上角单击“更新”。
ii. 重复以上操作，确保发布的应用均已删除该MCP服务。
iii. 在“删除MCP”对话框，输入**DELETE**，单击“确定”。

- 如果发布的应用未添加该MCP服务：在“删除MCP”对话框，输入**DELETE**，单击“确定”。

常见问题

是否只支持MaaS平台预置的本地部署的MCP服务?

不是，只要您准备好准确的NPX/UVX的JSON配置文件，即可创建自定义MCP服务。

8.3 在 ModelArts Studio（MaaS）应用体验中心查看应用解决方案

ModelArts Studio大模型即服务平台提供了MaaS应用体验中心，为具体的应用场景提供一整套解决方案。

应用中心介绍

“MaaS应用体验中心”提供基于行业客户应用场景的AI解决方案。MaaS提供的模型服务和华为云各AI应用层构建工具之间相互连通，通过灵活的组合方案，来帮助客户快速解决模型落地应用时所面临的业务及技术挑战。

MaaS应用体验中心结合KooSearch企业搜索服务、盘古数字人大脑和Dify，为具体的客户应用场景提供一整套解决方案。

- KooSearch企业搜索服务：基于在MaaS开源大模型部署的模型API，搭建企业专属方案、LLM驱动的语义搜索、多模态搜索增强。
- 盘古数字人大脑：基于在MaaS开源大模型部署的模型API，升级智能对话解决方案，含智能客服、数字人。
- Dify：支持自部署的应用构建开源解决方案，用于Agent编排、自定义工作流。

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. 在左侧导航栏，选择“应用体验中心”。
3. 在“大模型应用实践中心”页面，单击想要查看的应用方案，了解方案详情。

9 ModelArts Studio（MaaS）管理与统计

9.1 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。

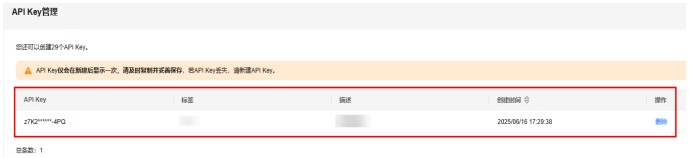
表 9-1 创建 API Key 参数说明

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，默认显示API Key的前四位和后四位。

图 9-1 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，贵阳一区域的API Key必须通过贵阳一控制台创建，且仅能在该区域内调用服务。其他区域的API Key同理。

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Dify、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**。

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

9.2.1 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、预置服务-免费服务、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

- 统计范围：
  - 仅统计预置服务-商用服务、预置服务-免费服务、自定义接入点、我的服务的调用数据。2025年8月21日前的商用服务历史调用数据，无法区分版本。
  - 调用统计数据仅统计通过API接口调用和在线体验产生的数据。
- 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

预置服务或我的服务满足以下任一条件：

- 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**。

- 预置服务-免费服务：已使用免费服务并产生调用记录。具体操作，请参见**在ModelArts Studio（MaaS）预置服务中体验免费服务**。
- 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 9-2 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**。
	预置服务-免费服务	在“预置服务 > 免费服务”页签提供的免费服务。更多信息，请参见**在ModelArts Studio（MaaS）预置服务中体验免费服务**。
	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		支持API Key调用和在线体验。 • API Key调用：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**。 • 在线体验：在线体验模型服务产生的调用数据。更多信息，请参见**ModelArts Studio（MaaS）在线体验**。

参数	说明
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 9-3 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、有效期内的预置服务-免费服务、已创建的自定义接入点或已部署成功的我的服务。

   表 9-4 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。

参数	说明
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、有效期内的预置服务-免费服务、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 9-5 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 9-6 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。

1. 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”。
2. 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9-5。
3. 在页面右上角，单击“导出”。
4. 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

1. 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。
2. 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。

9.2.2 在 CES 查看 ModelArts Studio（MaaS）调用数据和监控指标

云监控服务CES提供云服务监控功能，支持查看MaaS预置服务-商用服务、预置服务-免费服务、我的服务在指定时间段内的调用数据和监控指标详情，包括RPM、TPM、请求失败率、输入Tokens数、输出Tokens数等信息，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

- 统计范围：
  - 仅统计预置服务-商用服务、预置服务-免费服务、我的服务的调用数据。
  - 实例列表：如实例超过一定时长（大于3小时）未上报监控数据，将不会展示在实例列表中。
  - 实例列表-视图页面：如实例指标超过一定时长（大于1小时）未上报监控数据，则该指标将不会展示在视图页面。
- 时间范围限制：
  - 支持预设时间段：近15分钟、近30分钟、近1小时、近2小时、近3小时、近12小时、近24小时、近7天、近14天、近30天。
  - 自定义时间段：最长不超过30天。

计费说明

- 云服务监控功能本身不收费。
- 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

预置服务或我的服务满足以下条件：

- 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**。
- 预置服务-免费服务：已使用免费服务并产生调用记录。具体操作，请参见**在ModelArts Studio（MaaS）预置服务中体验免费服务**。
- 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**。

监控指标的命名空间

SYS.MaaS

查看服务调用的监控数据

1. 登录**云监控服务管理控制台**，在左侧导航栏单击“云服务监控”。
2. 在“云服务监控”页面，单击“MaaS MaaS”看板名称。
3. 在“资源详情”页签的实例列表，查看服务的整体情况。

4. 在实例列表的“操作”列，单击目标服务对应的“查看监控指标”。
5. 在“资源实例”页签或“API Key”页签，查看服务的监控指标详情。

- **首Token时延和增量Token时延仅统计流式响应**。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**。
- 不同监控周期对应聚合方式的聚合时间不同，详情请参见**查看监控视图**。
- 监控指标默认保留两位小数。

表 9-7 监控指标说明

指标ID	指标名称	指标含义	单位	进制	监控周期
rpm	RPM	每分钟处理的请求数。	count/min	-	1分钟
tpm	TPM	每分钟处理的Tokens数（输入+输出）。	thousand/min	-	1分钟
req_count_4xx	4XX数量	服务调用错误4XX次数。	count/min	-	1分钟
req_count_5xx	5XX数量	服务调用错误5XX次数。	count/min	-	1分钟
req_count	调用总量	调用的总量。	count/min	-	1分钟
req_count_2xx	调用成功次数	2XX成功的次数。	count/min	-	1分钟
req_count_error	调用失败次数	调用失败的次数。 • 调用失败次数可能会超过4XX和5XX错误的总和，因为还可能包含不属于4xx或5xx类别的错误。 • 调用失败次数仅涵盖模型服务产生的4XX和5XX错误，不包括租户在服务请求中因非模型服务因素导致的错误，例如鉴权失败等。	count/min	-	1分钟
req_error_rate	请求失败率	调用失败次数占调用总次数的比例。	%	-	1分钟

指标ID	指标名称	指标含义	单位	进制	监控周期
req_error_4xx_rate	请求4XX失败率	调用失败4XX次数/调用总次数。	%	-	1分钟
req_error_5xx_rate	请求5XX失败率	调用失败5XX次数/调用总次数。	%	-	1分钟
prompt_tokens	输入tokens数	服务调用输入Tokens数。	thousand	-	1分钟
completion_tokens	输出tokens数	服务调用输出Tokens数。	thousand	-	1分钟
total_tokens	调用总tokens数	服务调用总Tokens数。	thousand	-	1分钟
prompt_tokens_avg	平均输入token长度	输入Token平均长度。	thousand	-	1分钟
completion_tokens_avg	平均输出token长度	输出Token平均长度。	thousand	-	1分钟
prompt_tokens_p50	输入token TP 50	50%的输入Token大小低于该值。	thousand	-	1分钟
prompt_tokens_p80	输入token TP 80	80%的输入Token大小低于该值。	thousand	-	1分钟
prompt_tokens_p90	输入token TP 90	90%的输入Token大小低于该值。	thousand	-	1分钟
prompt_tokens_p99	输入token TP 99	99%的输入Token大小低于该值。	thousand	-	1分钟
completion_tokens_p50	输出token TP 50	50%的输出Token大小低于该值。	thousand	-	1分钟
completion_tokens_p80	输出token TP 80	80%的输出Token大小低于该值。	thousand	-	1分钟

指标ID	指标名称	指标含义	单位	进制	监控周期
completion_tokens_p90	输出token TP 90	90%的输出Token大小低于该值。	thousand	-	1分钟
completion_tokens_p99	输出token TP 99	99%的输出Token大小低于该值。	thousand	-	1分钟
prompt_tokens_max	最长输入token长度	输入Token最大值。	thousand	-	1分钟
completion_tokens_max	最长输出token长度	输出Token最大值。	thousand	-	1分钟
ttft	TTFT（AVG）	首Token时延，即从接收请求到生成第一个输出Token所需的时间。	ms	-	1分钟
tpot	TPOT（AVG）	增量Token时延，即生成后续每个输出Token所需的时间间隔。	ms	-	1分钟
latency_avg	平均响应时延	单位时间内成功请求的响应时间平均值。	ms	-	1分钟
ttft_p50	首token时延 TP50	50%的首Token时延低于该值。	ms	-	1分钟
ttft_p80	首token时延 TP80	80%的首Token时延低于该值。	ms	-	1分钟
ttft_p90	首token时延 TP90	90%的首Token时延低于该值。	ms	-	1分钟
ttft_p99	首token时延 TP99	99%的首Token时延低于该值。	ms	-	1分钟
ttft_max	最长首token时延	首Token时延最大值。	ms	-	1分钟
tpot_p50	增量 token时延 TP50	50%的增量Token时延低于该值。	ms	-	1分钟

指标ID	指标名称	指标含义	单位	进制	监控周期
tpot_p80	增量 token 时延 TP80	80%的增量Token时延低于该值。	ms	-	1分钟
tpot_p90	增量 token 时延 TP90	90%的增量Token时延低于该值。	ms	-	1分钟
tpot_p99	增量 token 时延 TP99	99%的增量Token时延低于该值。	ms	-	1分钟
tpot_max	最长增量 token时延	增量Token时延最大值。	ms	-	1分钟
average_generation_time	Average generation time	从输入到生成输出的平均时间。	s	-	1分钟
req_count_400	400 Quantity	服务调用错误400次数。	count/min	-	1分钟
req_count_401	401 Quantity	服务调用错误401次数。	count/min	-	1分钟
req_count_403	403 Quantity	服务调用错误403次数。	count/min	-	1分钟
req_count_404	404 Quantity	服务调用错误404次数。	count/min	-	1分钟
req_count_413	413 Quantity	服务调用错误413次数。	count/min	-	1分钟
req_count_429	429 Quantity	服务调用错误429次数。	count/min	-	1分钟
req_count_500	500 Quantity	服务调用错误500次数。	count/min	-	1分钟
req_count_503	503 Quantity	服务调用错误503次数。	count/min	-	1分钟

指标ID	指标名称	指标含义	单位	进制	监控周期
req_count_504	504 Quantity	服务调用错误504次数。	count/min	-	1分钟

10 ModelArts Studio（MaaS）模型能力

10.1 在 ModelArts Studio（MaaS）中通过 Function Calling 扩展大语言模型交互能力

10.1.1 Function Calling 介绍

使用场景

大语言模型在处理复杂任务时可能会遇到自身能力的局限性，例如需要调用实时数据、执行专业领域计算或进行特定服务操作时，模型本身的知识和能力可能无法满足需求。这种情况下，如何让模型突破自身限制，完成更复杂的任务成为了亟待解决的问题。大语言模型的Function Calling能力正是为了解决这一问题而设计的。通过Function Calling，模型能够调用外部函数或服务，从而扩展其自身的能力边界，执行它本身无法完成的任务。这种机制不仅提升了模型的实用性，还使其能够处理更复杂、更专业的场景，例如实时天气查询、数据分析、API调用等，从而显著提升了模型的准确性和任务处理效率。以下是一些Function Calling的使用场景：

表 10-1 Function Calling 使用场景说明

使用场景	说明
增强能力	大模型通过Function Calling可以调用外部工具或服务，例如实时数据检索、文件处理、数据库查询等，从而扩展其能力。
实时数据访问	由于大模型通常基于静态数据集训练，不具备实时信息。Function Calling允许模型访问最新的数据，提供更准确、更及时的回答。
提高准确性	在需要精确计算或特定领域知识时，大模型可以通过调用专门的函数来提高回答的准确性，例如调用数学计算函数、翻译服务或专业知识库。

支持模型

支持Qwen2.5系列预置服务：

- Qwen2.5-72B-32K-1128
- Qwen2.5-72B-Instruct-1128
- Qwen2.5-7B-Instruct-1128

计费影响

- 调用我的服务或者预置服务-商用服务时，会按照Token使用量计费，计费详情请参见**计费说明**。
- 部署Dify平台需要资源成本，具体费用请参考**资源和成本规划**。

使用方式

- 方式一：在请求体中添加相关函数。

```
"tools": [
    {
        'type': 'function',
        'function': {
            'name': '对应到实际执行的函数名称',
            'description': '此处是函数相关描述',
            'parameters': {
                '_comments': '此处是函数参数相关描述'
            },
        }
    },
    {
        '_comments': '其他函数相关说明'
    }
]
```

- 方式二：通过OpenAI库发起请求。

```
response = client.chat.completions.create(
    model="MaaS模型名称",
    messages = messages,
    tools=[
        {
            'type': 'function',
            'function': {
                'name': '对应到实际执行的函数名称',
                'description': '此处是函数相关描述',
                'parameters': {
                    // 此处是函数参数相关描述
                },
            }
        },
        {
            // 其他函数相关说明
        }
    ]
    // chat.completions 其他参数
)
```

应用示例

- 示例一：**在Dify中配置支持Function Calling的模型使用**
- 示例二：**通过Function Calling扩展大语言模型对外部环境的理解**

10.1.2 在 Dify 中配置支持 Function Calling 的模型使用

Dify是一个能力丰富的开源AI应用开发平台，为大型语言模型（LLM）应用的开发而设计。它巧妙地结合了后端即服务（Backend as Service）和LLMOps的理念，提供了一套易用的界面和API，加速了开发者构建可扩展的生成式AI应用的过程。

前提条件

用户已有可正常使用的Dify。

操作步骤

1. 在Dify界面右上角单击用户头像，选择“设置”。
2. 在“设置”页面左侧，选择“模型供应商”页签，找到“ OpenAI-API-compatible”供应商，单击添加模型。
3. 在弹窗中，配置MaaS对应的模型名称、API Key、API Endpoint URL、Function calling等信息。

   表 10-2 配置说明

配置项	说明
模型名称	MaaS“调用说明”页面显示的模型名称。
API Key	MaaS“API Key管理”页面中创建的API Key。具体操作，请参见**创建API Key**。
API Endpoint URL	服务调用界面中MaaS服务的基础API地址，需要去掉地址尾部的“/chat/completions” 。具体操作，请参见**步骤二：调用MaaS模型服务进行预测**。
Function calling	设置为“Tool Call”。
Stream function calling	暂不支持。

4. 在Dify中创建Agent进行编排，在右上角单击“Agent 设置”，选择上一步配置好的模型进行使用。

   在Agent设置中可以看到Dify已自动将Agent Mode切换到了Function Calling模式。

图 10-1 Agent 设置

5. 在“编排”页面的“提示词”文本框，输入以下信息。

```
你是一位乐于助人的AI助手。在回答用户问题时，你需要：1. 始终使用自然语言解释你将要采取的行动 2. 在调用工具之前，说明你要使用哪个工具以及原因 3. 在获取信息的过程中，清晰地描述你正在做什么 4. 永远不要返回空的回复 - 确保用自然语言解释你的每个步骤，比如当查询天气时，你应该先说'让我使用天气工具为您查询...'，然后再进行工具调用。记住：先表达你的理解和计划，再使用工具。每次回复都必须包含对用户的清晰解释。
```

图 10-2 输入提示词

6. 在“编排”页面的“工具”区域右侧，单击“添加”，按需添加工具并与模型进行对话调用。

   Dify内置有丰富的插件，同时支持自定义工具的创建。您可以按需使用。

图 10-3 添加工具

10.1.3 通过 Function Calling 扩展大语言模型对外部环境的理解

本示例将展示如何定义一个获取送货日期的函数，并通过LLM来调用外部API来获取外部信息。

操作步骤

1. 设置MaaS的API Key和模型服务地址URL。

```
import requests
from openai import OpenAI
client = OpenAI(
    api_key="您的 APIKEY", # 从ModelArts Studio（MaaS）控制台API Key管理页面获取。
    base_url="https://infer-modelarts.cn-east-4.myhuaweicloud.com/v1/infers/xxxxxx/v1"  # MaaS模型服务的基础url，不包含尾部的chat/completions部分。
)
```

2. 自定义一个获取送货日期的函数。

```
from datetime import datetime
def get_delivery_date(order_id: int) -> datetime:
    if order_id == 1:
        return datetime.strptime("2024-09-01 18:30", "%Y-%m-%d %H:%M")
    elif order_id == 2:
        return datetime.strptime("2024-10-20 12:00", "%Y-%m-%d %H:%M")
    else:
        return f"cannot find order_id {order_id}"

# Example usage
print(get_delivery_date(1))
```

3. 编写函数的描述。

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "get_delivery_date",
            "description": "Get the delivery date for a customer's order. Call this whenever you need to know the delivery date, for example when a customer asks 'Where is my package'",
            "parameters": {
                "type": "object",
                "properties": {
```

```
                "order_id": {
                    "type": "string",
                    "description": "The customer's order ID.",
                },
            },
            "required": ["order_id"],
            "additionalProperties": False,
        },
    }
}
]
```

4. 与LLM进行对话。

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "get_delivery_date",
            "description": "Get the delivery date for a customer's order. Call this whenever you need to
know the delivery date, for example when a customer asks 'Where is my package'",
            "parameters": {
                "type": "object",
                "properties": {
                    "order_id": {
                        "type": "string",
                        "description": "The customer's order ID.",
                    },
                },
                "required": ["order_id"],
                "additionalProperties": False,
            },
        }
    }
]

messages = [
    {"role": "system", "content": "You are a helpful customer support assistant. Use the supplied tools
to assist the user."},
    {"role": "user", "content": "Hi, can you tell me the delivery date for my order?"}
]
messages.append({"role": "assistant", "content": "Hi there! I can help with that. Can you please
provide your order ID?"})
messages.append({"role": "user", "content": "i think it is 1"})

response = client.chat.completions.create(
    model="Qwen2.5-72B-32K",
    messages=messages,
    tools=tools,
)

print(response)
```

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。

11 ModelArts Studio（MaaS）业务最佳实践

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

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

- Dify：使用MaaS（大模型即服务平台）的免费DeepSeek-R1 API接入Dify（开源Agent平台），快速构建AI对话机器人并将其嵌入在网页中。具体操作，请参见**基于ModelArts Studio（MaaS） DeepSeek API和Dify快速构建网站智能客服**。
- 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**。
- RAGFlow：使用RAGFlow调用部署在ModelArts Studio上的DeepSeek模型，快速构建AI助理。具体操作，请参见**基于ModelArts Studio（MaaS） DeepSeek API和RAGFlow快速构建AI助理**。
- Deep Research：使用Deep Research调用部署在ModelArts Studio上的DeepSeek模型，快速实现行业洞察。具体操作，请参见**基于ModelArts Studio（MaaS） DeepSeek API和Deep Research快速实现行业洞察**。