

盘古大模型

# 产品介绍

文档版本

01

发布日期

2025-09-16



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## 商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## 注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

## 华为云计算技术有限公司

地址：贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址：<https://www.huaweicloud.com/>

目 录

1 什么是盘古大模型服务.....1

2 产品优势.....4

3 应用场景.....5

3.1 大模型应用场景.....5

3.2 Agent 应用场景.....6

4 产品功能.....8

4.1 空间管理.....8

4.2 数据工程.....9

4.3 模型开发.....12

4.4 Agent 开发.....12

5 模型能力与规格.....19

5.1 盘古 NLP 大模型.....19

5.2 盘古 CV 大模型.....23

5.3 盘古科学计算大模型.....25

5.4 盘古预测大模型能力与规格.....30

5.5 盘古多模态大模型.....36

5.6 盘古专业大模型.....37

5.7 盘古行业大模型.....39

5.8 盘古图像搜索模型.....41

5.9 三方大模型.....42

6 基础知识.....47

6.1 大模型开发基本流程介绍.....47

6.2 大模型开发基本概念.....48

7 安全.....51

7.1 责任共担.....51

7.2 身份认证与访问控制.....52

7.3 数据保护技术.....53

7.4 审计.....53

8 权限管理.....54

9 约束与限制.....57

10 与其他服务的关系..... 59

# 1 什么是盘古大模型服务

## 盘古大模型服务简介

盘古大模型服务以“模型能力+开发平台”为核心架构，整体涵盖盘古系列大模型与 ModelArts Studio 大模型开发平台两大板块，形成从模型供给到全流程开发的完整服务闭环，具体内容如下：

### 一、盘古系列大模型：多维度覆盖场景需求

盘古系列大模型围绕“基础能力支撑 - 行业场景适配 - 专业需求落地 - 外部补充拓展”的逻辑，构建了多层次模型矩阵，满足不同领域的应用需求：

#### 1. 盘古基础大模型：核心能力底座

作为整个模型体系的基础支撑，盘古基础大模型提供五大核心基础能力模型，为上层模型开发与应用提供底层技术保障，具体包括：

- NLP（自然语言处理）模型：支撑文本理解、生成、翻译等语言类任务；
- 多模态模型：实现文本、图像、音频等多类型信息的融合处理；
- CV（计算机视觉）模型：覆盖图像识别、检测、分割等视觉类应用；
- 预测模型：提供趋势预测、数据研判等分析能力；
- 科学计算模型：助力物理、化学、工程等领域的科学问题求解。

#### 2. 盘古行业大模型：垂直领域定制化方案

基于盘古基础大模型的核心能力，结合各行业的业务特性与数据特征，打造行业定制化大模型，已覆盖多个关键领域，包括金融、政务、油气、矿山、电力、制造、交通、医学等，可直接适配行业场景中的核心业务需求，降低企业应用门槛。

#### 3. 盘古专业大模型：聚焦特定场景落地

针对细分场景的专业化需求，提供专项模型支持，目前主要面向 RAG（检索增强生成）场景，提供两类核心模型：

- 搜索规划模型：优化检索流程中的需求解析与搜索路径规划；
- 向量 & 重排模型：提升检索结果的准确性与相关性排序效果，强化 RAG 场景的输出质量。

#### 4. 盘古 AIKits 图像搜索大模型：细分领域专项工具

聚焦服装行业的特定需求，提供专业化图像搜索能力，可支持服装款式识别、相似款检索等场景，助力服装企业实现商品高效匹配与管理。

### 5. 三方大模型：拓展能力边界

为进一步丰富模型选择、满足多样化场景需求，引入外部优质模型资源，目前已涵盖 DeepSeek、通义千问系列模型，与盘古自有模型形成互补，覆盖更多细分应用场景。

## 二、ModelArts Studio 大模型开发平台：全流程开发支撑

ModelArts Studio 定位为“一站式大模型工具链平台”，以“支持百模千态”为目标，构建了**数据、模型、Agent 三引擎驱动**的开发体系，为大模型的全生命周期开发提供工具支撑，具体工具链如下：

### 1. 数据工程工具链：保障训练数据质量

数据是大模型训练的核心基础，该工具链围绕“数据全流程管理”设计，具备七大核心功能：数据获取、数据清洗、数据合成、数据标注、数据评估、数据配比、数据发布与管理。

通过自动化质量检测与数据清洗能力，可对原始数据进行优化，确保数据的质量与一致性；同时支持多格式数据的高效收集与处理，满足不同训练、评测任务的需求，并提供强大的数据存储与管理能力，为大模型训练提供稳定、高质量的数据支撑。

### 2. 模型开发工具链：核心开发全流程覆盖

作为盘古大模型服务的核心组件，该工具链提供“从模型创建到部署”的一站式解决方案，涵盖五大关键功能：模型训练、模型压缩、模型部署、模型评测、模型推理。

依托高效的推理性能优化技术与跨平台迁移工具，可保障模型在不同硬件环境、应用场景中的高效运行，降低模型开发与落地的技术难度。

### 3. Agent 开发工具链：加速智能应用落地

该工具链是盘古大模型平台的关键模块，聚焦智能 Agent 应用的开发需求，提供两大核心能力：

- 提示词工程支持：提供提示词设计与管理工具，可优化大模型的输入提示，提升输出结果的准确性与相关性；
- 可视化编排能力：通过可视化工具简化 Agent 应用的开发流程，加速大模型在复杂业务场景中的应用落地，满足企业多样化的业务需求。

## 服务架构

盘古大模型服务产品架构请参考图1-1。

图 1-1 盘古大模型服务产品架构



## 盘古大模型与 ModelArts Studio 大模型开发平台关系

盘古大模型通过ModelArts Studio大模型开发平台承载，用户在平台上可以使用自己的数据增训和微调模型，可对训练的模型进行压缩、评测、部署，也可以在平台上创建自己的Agent应用。

## 服务访问方式

- 管理控制台方式  
盘古大模型服务支持通过管理控制台访问，包含数据工程、模型开发、Agent开发等功能，您可以在管理控制台端到端完成您的AI开发，具体操作请参见《[用户指南](#)》。
- API方式  
如果您需要将盘古大模型服务集成到第三方系统，用于二次开发，请使用API方式访问盘古大模型服务，具体操作和API详细描述，请参见《[API参考](#)》。

# 2 产品优势

## 全面的大模型系列

盘古大模型服务支持盘古NLP、CV、多模态、预测、科学计算五大基础模型能力，支持三方大模型的推理与部署，目前已预置DeepSeek R1/V3及Qwen等模型，多种模态逐步融合。

## 成熟的行业服务

首创大模型分层架构，基于盘古基础大模型训练行业大模型，支持在矿山、电力、油气、政务、金融、铁路、医药等专业领域商业化落地，打造工业级盘古解决方案。

## 最全工具链平台

ModelArts Studio大模型开发平台打造业界最强、最全的大模型工具链平台，数据工程、评测中心、模型开发、Agent开发等沉淀华为大模型开发先进经验，做到国内行业首选工具链平台。

## 零代码开发平台

ModelArts Studio大模型开发平台支持零代码与低代码开发方式，应用可外载知识库与插件，开箱即用，工作流可进行拖拉拽的可视化编辑，小白也可轻易上手创建应用。

## 高性能、低成本

ModelArts Studio大模型开发平台基于昇腾高性能推理框架，支持数据加速、训练加速和推理加速，支持分布式高效训练和推理，提供高性价比算力。

## 稳固的安全工程

ModelArts Studio大模型开发平台推理服务调用时支持英文等多语种内容安全拦截，支持恶意算子脚本/恶意Prompt拦截，全面保护大模型平台上AIGC内容的安全访问。



# 3 应用场景

大模型应用场景

Agent应用场景

## 3.1 大模型应用场景

### 客服助手

通过NLP大模型对传统的客服系统进行智能化升级，提升智能客服的效果。

**企业原智能客服系统：**

- 仅支持回复基础的FAQ，无语义泛化能力，意图理解能力弱，转人工频率极高。
- 面对活动等时效性场景，智能客服无回答能力。

**使用NLP大模型对客服系统升级后：**

- 大模型智能客服可以7x24小时不间断服务，相较于人工客服，可以处理更多的客户咨询，且响应速度快；
- 降低运营成本：企业可以通过智能客服处理大部分的常规问题，将人工客服释放出来处理更复杂、更个性化的客户需求；
- 个性化服务：基于大模型的智能客服能够学习和适应用户的行为模式和偏好，提供更加个性化的服务。

### 证件审核

通过多模态大模型对电子材料进行审核，尤其是图片类电子材料，利用人工智能相关算法自动进行合规性审查和智能分类，对事项办理申请人提交材料的准确性进行核验。通过材料合规性审核，实现事件处理全过程智能化管理，方便对各环节数据的查询、统计、分类和审查，使业务人员从繁重的数据整理、汇总和审查的工作中解脱出来，提高事务处理效率。

### 城市治理

传统算法在治理交通拥堵下是基于人为预先设定的规则逻辑进行判断，灵活性缺失，例如交通拥堵是基于视频画面的车辆检测+跟踪，综合判断是否有大量车停留等待，从

而判断交通拥堵。但如果要进一步分析交通拥堵的原因，传统视觉算法就犯难了，盘古CV大模型支持对同一目标多个标签描述，通过检测模型对目标物体进行检测，分割模型对目标进行分割，能有效识别城市道路积水检测、违停车辆检测、占道经营检测等场景，对城市治理起到了关键作用。

## 智慧配煤

传统人工配煤方式具有以下难点：

- **成本保障难：**传统方式依赖专家经验，人工配煤在平衡质量和成本价格上难以做到最优，配比取值保守，增加炼烧原料成本；
- **监管回溯难：**采用Excel表格计算，需手工输入原料煤参数，难以支撑数据分析和溯源，不同煤种煤质差异大。

盘古预测大模型通过评估和预测焦炭生产过程中焦炭成分及其质量来准确预测焦炭的成分和质量，通过合理选择和搭配不同种类、性质和质量的煤炭，降低生产成本，对于钢铁企业的生产效率、成本控制和环境保护具有重要意义。

## 降水预测

科学计算大模型包括全球中期天气要素模型和降水模型，可以对未来一段时间的天气和降水进行预测，全球中期天气要素模型和降水模型能够在全局范围内进行预测，不仅仅局限于某个地区。它的分辨率相当于赤道附近每个点约25公里x25公里的空间。通过降水模型预测未来的降雨情况，农民和农业管理者可以更有效地规划灌溉时间和频率，也能为可能发生的干旱提供预警，使农业部门能够及时采取措施，如推广节水技术或调整种植计划。

## 3.2 Agent 应用场景

当前，基于Agent开发平台可以构建两种类型的Agent(智能体)，一种是针对文本生成、文本检索的知识型Agent，如搜索问答助手、代码生成助手等，执行主体在大模型；另一种是针对复杂 workflows 场景的流程型Agent，如金融分析助手、网络检测助手等。

### Agent 应用

以大模型为任务执行核心，用户通过配置Prompt、知识库、插件等信息，实现工具自主规划与调用，优点是可零码开发，对话过程更为智能，缺点是当大模型受到输入限制，难以执行链路较长且复杂的流程。

**应用场景：**

#### 1. 智能客服

- **场景：**电商售前咨询、售后问题解答。
- **应用方式：**通过配置Prompt和知识库，大模型直接理解用户问题（如退货政策、商品参数查询），调用插件获取订单信息，快速生成自然语言回复。
- **优势：**无需复杂流程设计，能处理开放式提问，对话体验更接近真人。

#### 2. 教育辅导

- **场景：**个性化学习答疑、知识点讲解。
- **应用方式：**结合学科知识库和大模型推理能力，学生通过对话提问（如数学题解析），模型实时生成解题步骤并反馈。

- **优势：**零代码快速覆盖多学科需求，灵活适应不同提问方式。
3. **医疗咨询**
- **场景：**症状初步分析、健康建议。
  - **应用方式：**用户描述症状后，模型基于医学知识库生成可能病因，调用插件获取药品信息或推荐就医科室。
  - **优势：**交互自然，可快速迭代更新知识库以适配最新医疗指南。

Agent workflow

以 workflow 为任务执行核心，用户通过在画布上对节点进行“拖拉拽”即可搭建出任务流程，场景的节点包括大模型节点、意图识别节点、提问器节点、插件节点、判断节点、代码节点、消息节点，优点是可扩展能力强，用户适当使用低码开发，缺点是对话交互智能度不高，复杂场景下分支多，难以维护。

应用场景：

1. **金融风控**
- **场景：**贷款审批、反欺诈审核。
  - **workflow 设计：**通过节点串联实现多步骤审核（如调用征信接口→大模型分析收入证明→判断节点触发人工复核）。
  - **优势：**流程透明可控，分支条件清晰（如不同额度触发不同审批路径），避免大模型幻觉干扰关键决策。
2. **供应链管理**
- **场景：**库存预警与自动补货、物流异常处理。
  - **workflow 设计：**通过代码节点对接 ERP 系统，判断节点识别库存阈值，触发补货插件或邮件通知节点。
  - **优势：**复杂规则（如供应商优先级、物流延迟策略）可通过分支节点精准实现，稳定性高于纯模型驱动。
3. **政务审批**
- **场景：**企业资质申报、惠民补贴发放。
  - **workflow 设计：**串联意图识别节点（分类用户需求）、提问器节点（补全材料）、代码节点（验证数据合规性），最终由消息节点反馈结果。
  - **优势：**标准化流程减少人为错误，分支节点可覆盖政策差异化要求（如不同地区补贴标准）。

# 4 产品功能

空间管理

数据工程

模型开发

Agent开发

## 4.1 空间管理

ModelArts Studio大模型开发平台为用户提供了灵活且高效的空间资产管理方式。

- 平台支持用户根据不同的使用场景、项目类别或团队需求，自定义创建多个工作空间。
- 每个工作空间都是完全独立的，确保了工作空间内的资产不受其他空间的影响，从而保障数据和资源的隔离性与安全性。用户可以根据需求灵活划分工作空间，实现资源的有序管理与优化配置，确保各类资源在不同场景中的最大化利用。
- 为进一步优化资源的管理，平台还提供了多种角色权限体系。用户可以根据自身角色从管理者到各模块人员进行不同层级的权限配置，确保每个用户在其指定的工作空间内，拥有合适的访问与操作权限。这种精细化的权限管理方式，既保证了数据的安全性，又提高了资源的高效利用。

### 空间资产管理

在平台中，空间资产指的是存储在工作空间中的所有资源，包括数据资产和模型资产。这些资产是用户在平台上进行开发和管理的基础，集中存储和统一管理的方式有助于提升操作效率，并确保资源的规范性与安全性。

- **数据资产：**数据资产是指用户在平台上发布的所有数据集。这些数据集会被存储在数据资产中，用户可以随时查看数据集的详细信息，如数据格式、大小、配比比例等，同时平台会自动记录每个数据集的操作历史，例如创建、发布及上线等过程。为了进一步简化管理，平台还支持数据集的删除功能，使用户能够对数据集进行灵活管理和调整。在模型训练和数据分析过程中，用户可以根据需求调用这些数据集，确保数据的准确性与安全性，从而提升数据资产的利用率。同时支持数据集发布到Gallery，支持从Gallery订阅数据集。
- **模型资产：**模型资产包括用户试用、订购或在平台上训练后发布的模型，这些模型统一存储在模型资产中，便于集中管理。用户可以查看模型的所有历史版本及

操作记录，从而了解模型的演变过程。同时，平台支持一系列便捷的模型操作，如模型训练、压缩和部署，帮助用户简化模型开发和应用流程。此外，平台还提供了导入和导出功能，支持用户将其他局点的盘古大模型迁移到本地局点，这使得模型资产在不同局点间的共享和管理变得更加灵活高效。同时支持模型发布到Gallery，支持从Gallery订阅模型。

- **组件资产：**组件资产包括脚本、算法包、后处理代码，支持从OBS导入后处理代码；支持预置组件，包括预置全域感知服务组件、决策优化服务组件。同时支持组件发布到Gallery，支持从Gallery订阅组件。

通过统一管理空间资产，平台不仅帮助用户高效组织和利用资源，还保障了资产的安全性、一致性与灵活性。这些功能的结合，确保了平台上资源的高效利用与智能配置，为用户提供了更为便捷的开发和管理体验。

## 4.2 数据工程

ModelArts Studio大模型开发平台提供了全面的数据工程功能。该模块涵盖数据获取、加工、标注、评估和发布等关键环节，帮助用户高效构建高质量的训练数据集，推动AI应用的成功落地。具体功能如下：

图 4-1 数据流程

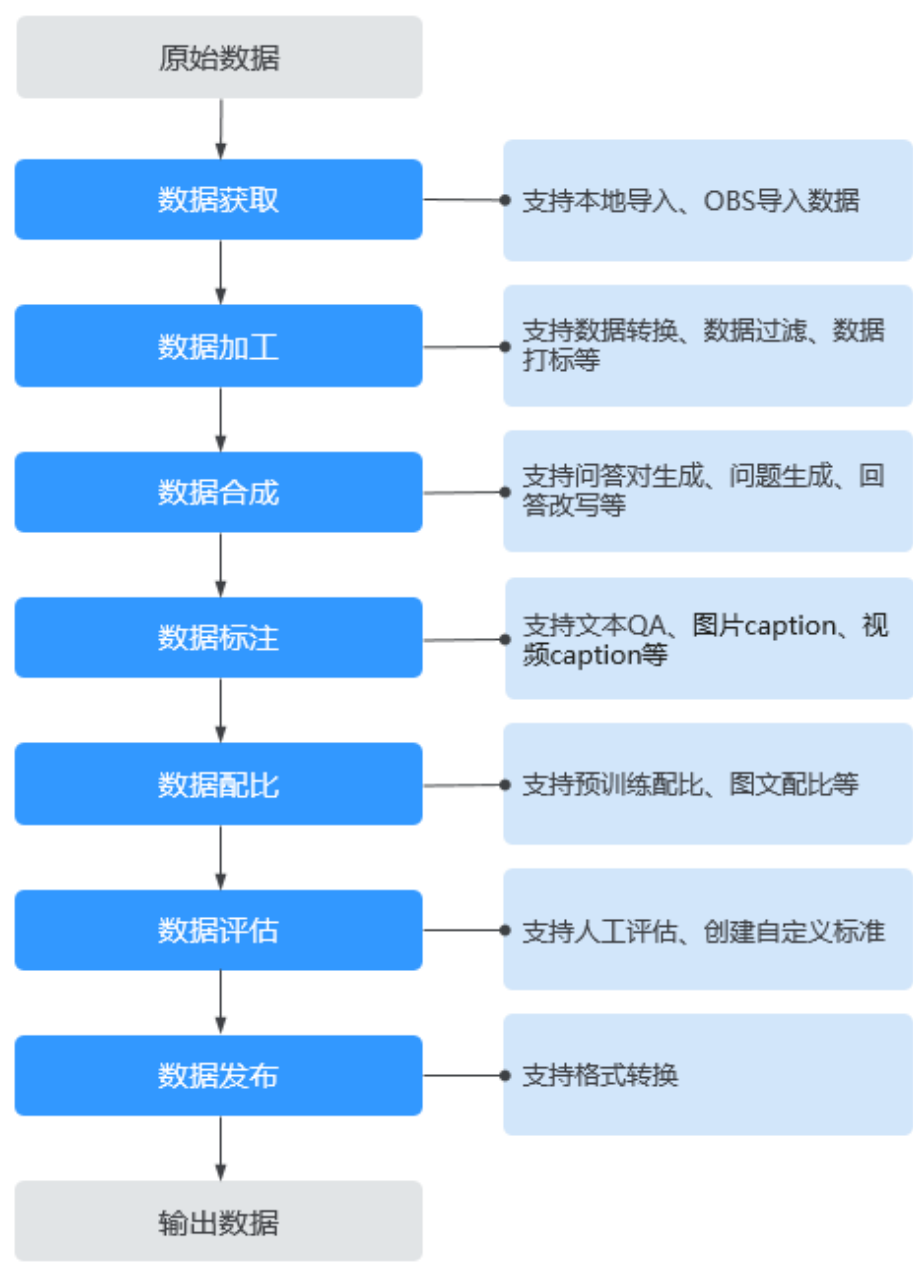


表 4-1 各类数据支持的操作

| 数据类型 | 数据获取 | 数据加工 | 数据合成 | 数据标注 | 数据配比 | 数据评估 | 数据发布 |
|------|------|------|------|------|------|------|------|
| 文本类  | √    | √    | √    | √    | √    | √    | √    |
| 图片类  | √    | √    | -    | √    | √    | √    | √    |
| 视频类  | √    | √    | -    | √    | √    | √    | √    |
| 音频类  | √    | √    | -    | √    | -    | -    | √    |

| 数据类型 | 数据获取 | 数据加工 | 数据合成 | 数据标注 | 数据配比 | 数据评估 | 数据发布 |
|------|------|------|------|------|------|------|------|
| 气象类  | √    | √    | -    | -    | -    | -    | √    |
| 预测类  | √    | √    | -    | -    | √    | -    | √    |
| 其他类  | √    | √    | -    | -    | -    | -    | √    |

- **数据获取：**用户可以轻松将多种类型的数据导入ModelArts Studio大模型开发平台，支持的数据类型包括文本、图片、视频、气象、预测数据以及用户自定义的其他类型数据。平台提供灵活的数据接入方式以及支持多种文件格式导入，确保不同业务场景下的数据获取需求得到满足。
- **数据加工：**平台提供强大的数据加工功能，可以对文本、视频、图片、气象类数据进行数据提取、过滤、转换、打标签和评分等加工处理。针对不同类型的数据集，平台提供了专用的清洗算子以及支持用户创建自定义算子实现个性化的数据清洗需求。确保生成高质量的训练数据以满足业务需求和模型训练的要求。用户还可以灵活地调整算子编排顺序以及自定义清洗模板，有效提升数据清洗效率并支持大规模数据处理，确保生成的数据集符合训练的标准。
- **数据合成：**平台支持利用预置或自定义的数据指令对预训练文本、单轮问答、单轮问答（人设）数据集类型进行处理，并根据设定的轮数生成新数据。通过数据合成技术，可以生成大量高质量的训练数据，这些数据可以用于大模型的预训练，增强模型的泛化能力和性能。
- **数据标注：**平台支持对无标签的数据添加标签或对现有的标签进行重新标注，以提升数据集的标注质量。用户可以针对不同的数据集灵活地选择对应的标注项，还可以自定义选择多人标注、审核以及标注任务移交。针对文本和图片类数据集，平台还提供AI预标注功能。利用盘古大模型的智能能力，显著降低人工标注的工作量和成本，从而显著地提高标注效率。
- **数据配比：**平台支持对文本、图片类数据进行数据配比。用户在勾选数据集时可以勾选多条，通过调整不同来源或类型数据的比例，以优化模型训练过程。通过数据配比可以确保模型能够更全面地学习和理解数据的多样性，提高模型的泛化能力和性能。
- **数据发布：**平台支持数据集发布。用户可以将处理后的数据集发布为多种格式，包括标准格式和盘古格式。尤其对于文本类和图片类数据集，平台支持将其转换为专门用于训练盘古大模型的盘古格式，为后续模型训练提供高效的数据支持。
- **数据管理：**平台支持数据全链路血缘追溯，用户单击数据集名称可以在“数据血缘”页签，查看该数据集所经历的操作。全链路血缘追溯可以帮助用户正向实现数据集影响分析，逆向实现快速问题追踪，提升数据运维和数据治理的效率，帮助用户更好地对数据进行追溯溯源。另外平台还提供了完善的标签体系、支持数据按行业标准进行分类、按行业标准进行安全分级、内置场景分类标签。帮助用户进行数据分类、数据质量控制和数据资产管理，提升数据治理的效率和效果。

通过整合上述功能，数据工程在AI研发中不仅帮助用户高效构建高质量的训练数据集，还通过全流程的数据处理和管理，探索数据与模型性能的内在联系，为模型训练和应用提供坚实的数据基础，推动了模型的精确训练与持续优化，提升了AI应用开发的效率和成果的可靠性。

## 4.3 模型开发

ModelArts Studio大模型开发平台提供了模型开发功能，涵盖了从模型训练到模型调用的各个环节。平台支持全流程的模型生命周期管理，确保从数据准备到模型部署的每一个环节都能高效、精确地执行，为实际应用提供强大的智能支持。

图 4-2 模型开发



- 模型训练：**在模型开发的第一步，ModelArts Studio大模型开发平台为用户提供了丰富的训练工具与灵活的配置选项。用户可以根据实际需求选择合适的模型架构，并结合不同的训练数据进行精细化训练。平台支持分布式训练，能够处理大规模数据集，从而帮助用户快速提升模型性能。该模块提供预训练、全量微调、LoRA微调、DPO等。
- 模型评测：**为了确保模型的实际应用效果，平台提供了多维度的模型评测功能。通过自动化的评测机制，用户可以在训练过程中持续监控模型的精度、召回率等关键指标，及时发现潜在问题并优化调整。评测功能能够帮助用户在多种应用场景下验证模型的准确性与可靠性。支持基于规则的自动评测方式，NLP模型展示准确率，F1分数，BLEU、ROUGE等自动评测指标，支持人工评测自定义配置评测指标；并且支持基于人工评价操作界面，对模型表现从不同评价指标进行打分。
- 模型压缩：**在模型部署前，进行模型压缩是提升推理性能的关键步骤。通过压缩模型，能够有效减少推理过程中的显存占用，节省推理资源，同时提高计算速度。当前，平台支持对NLP大模型进行压缩，目前支持INT8、INT4量化压缩。
- 模型部署：**平台提供了一键式模型部署功能，用户可以轻松将训练好的模型部署到云端或本地环境中。平台支持多种部署模式，能够满足不同场景的需求。通过灵活的API接口，模型可以无缝集成到各类应用中。
- 模型调用：**在模型部署后，用户可以通过模型调用功能快速访问模型的服务。平台提供了高效的API接口，确保用户能够方便地将模型嵌入到自己的应用中，实现智能对话、文本生成等功能。

## 4.4 Agent 开发

### 功能概述

Agent开发平台提供一站式智能体构建解决方案，通过零代码编排工具让开发者快速创建AI应用，支持画布式节点设计应对复杂业务场景。平台可以集成盘古大模型及



DeepSeek等第三方模型，提供角色设定、插件扩展、 workflow编排等功能，支持知识库管理、RAG检索和智能提示词优化，确保交互精准可靠。同时具备短期对话记忆和长期个性化存储能力，并支持通过API、网页多渠道发布，助力开发者高效打造专业级智能体应用。

• 编排能力

表 4-2 Agent 平台编排能力

| 功能   | 说明                                                                                                                                             |
|------|------------------------------------------------------------------------------------------------------------------------------------------------|
| 编排模式 | 支持用户对话式的快捷调用和创建agent，让业务人员以零代码操作方式，5分钟完成1个原生AI应用创建。<br>提供灵活的画布式节点编排，让agent能够准确解决复杂的业务场景问题。<br>提供多智能体编排画布，实现多个专家Agent的组合，系统自动调度分配，让最合适的智能体解决问题。 |
| 模型选择 | 平台提供盘古大模型，支持多款第三方多款模型，DeepSeek等第三方深度思考模型在Agent开发平台已完成适配。                                                                                       |

• 角色指令

Agent平台可通过角色指令设定拟人化特征，提升交互真实感。平台预置了角色指令模板，也可以通过智能添加的方式让大模型输出一个更佳的角色prompt，角色的设定便于开发者打造高度拟人化的语音交互场景。

• 技能

Agent的核心能力源于其技能体系，开发者可通过集成插件、设计 workflow等方式不断扩展模型的功能范围。

表 4-3 Agent 平台技能

| 功能  | 说明                                                                              |
|-----|---------------------------------------------------------------------------------|
| 插件  | 您可以通过API无缝连接各类平台和服务，快速扩展智能体的功能，平台提供了丰富的内置插件，开箱即用；同时也支持自定义插件开发，将任意API封装成工具，灵活调用。 |
| 工作流 | 工作流是构建复杂功能逻辑的可视化工具，通过灵活组合多个任务节点，能够设计多步骤的自动化流程，从而显著增强智能体应对复杂任务的能力。               |

• 知识库

提供开箱即用的企业级RAG服务，覆盖管理、测试、检索策略配置全功能。

表 4-4 Agent 平台知识中心

| 功能      | 说明                                                                                     |
|---------|----------------------------------------------------------------------------------------|
| 知识库列表管理 | 支持展示知识库名称、知识库ID、知识库描述、创建时间，支持知识库列表操作（查看、删除、编辑、发布）。                                     |
| 创建知识库   | 支持填写知识库名称、描述、知识库上传类型及文档上传；支持选择内置知识（txt、docx、pdf、doc、xlsx，读取url链接并导入知识库）和引用知识（链接已有知识库）。 |
| 上传知识库文档 | 支持文档列表展示，支持导入新文档，支持知识库列表操作（查看、删除）。                                                     |
| 切片策略    | 支持文本切分自定义策略配置（分句字符标识、分段最大长度、分段重叠最大字数占比）。                                               |
| 切片详情    | 支持查看文档基础信息、切片、及切片知识点详情。                                                                |

- 提示词工程**  
平台提供prompt模板与开发工具，使任何人都可轻松开发出高质量的prompt指令。

表 4-5 提示词功能

| 功能      | 说明                                                                   |
|---------|----------------------------------------------------------------------|
| 提示词撰写   | 撰写提示词，并将评估表现较好的提示词设为候选。支持导入提示词示例；支持模型设置；支持提示词中变量定义；支持效果预览；支持历史记录。    |
| 提示词比较   | 支持选择候选提示词进行比较（差异性比较、效果比较）。                                           |
| 提示词评估   | 支持构建评估用例集，采用多种评估方法对prompt质量进行评估，选择最优的prompt。                         |
| 提示词优化   | 提示工程平台提供提示词自动优化功能，基于启发式算法的提示词自优化技术、提示优选梯度优化技术，可以根据评估用例自动对现有的提示词进行优化。 |
| 提示词自动生成 | 借助模板智能匹配与布局优化技术，根据用户输入内容，结合提示词原模板和模型能力，自动生成高质量提示词模板。                 |
| 提示词使用   | 指令配置和工作流大模型节点，支持保存和引用prompt模板。                                       |

- 记忆能力**  
Agent平台对话型工作流记忆能力短期变量保障对话连贯性，长期记忆实现深度个性化，参数配置平衡性能与效果。

表 4-6 Agent 平台记忆能力

| 功能   | 说明                                      |
|------|-----------------------------------------|
| 变量   | 用于存储临时会话数据（如用户当前输入、计算中间值），仅在单次对话中有效。    |
| 记忆参数 | 决定保留多少轮历史对话作为上下文。                       |
| 长期记忆 | 持久化存储用户个性化数据（如偏好、历史行为），将内容编码为向量，支持语义检索。 |

- 对话体验**  
Agent开发对话体验支持全程可视化，通过调试能力快速定位与优化配置。

表 4-7 Agent 平台对话体验

| 功能         | 说明                                          |
|------------|---------------------------------------------|
| Agent Logo | Agent的品牌标识，用于视觉识别，通常体现其功能或个性。               |
| 开场白        | Agent与用户交互时的初始欢迎语，设定对话基调并引导用户，支持用户自定义配置。    |
| 推荐问题       | 每次对话开始预设的典型提问示例，帮助用户快速了解智能体的能力范围，支持用户自定义配置。 |
| 追问         | 与Agent对话时主动提出的跟进问题，用于澄清需求或深化交互。             |
| 预览调试       | 实时测试和优化Agent功能的工具，展示运行的结果与调用详情。             |

- 全局配置**  
对话型工作流中增加全局配置的入口，支持工作流对话体验、默认模型、全局特性开关和定义的配置能力。
  - a. 支持在对话型工作流中配置开场白/推荐问，支持自动生成开场白/推荐问；
  - b. 支持配置默认模型，为新拖入节点提供默认模型选项；支持全局模型一键修改，提升模型配置效率；
  - c. 支持长期记忆开关和记忆变量定义的配置。
- MCP服务**  
平台工具调用支持MCP协议，开发者可以通过集成MCP服务快速拓展智能体的功能。

表 4-8 MCP 服务说明

| 功能      | 说明                                                 |
|---------|----------------------------------------------------|
| MCP服务市场 | 平台预置了"高德地图"、"车票查询工具"、"必应搜索"等多种实用MCP服务，开通后可以一键集成调用。 |

| 功能       | 说明                                       |
|----------|------------------------------------------|
| 自定义MCP服务 | 平台开放自定义MCP服务创建能力，开发者可依据MCP服务地址快速创建MCP服务。 |

● **发布渠道**

Agent支持不同的发布方式和渠道，支持发布成API或发布至网页，通过这两种发布方式，Agent可快速落地到实际业务中，兼顾技术集成与终端用户体验。

表 4-9 Agent 平台发布渠道

| 功能    | 说明                                                           |
|-------|--------------------------------------------------------------|
| API   | 将Agent封装成标准化API接口，供开发者直接调用，适用于需要与其他系统（如企业内部应用、第三方服务）深度集成的场景。 |
| 应用百宝箱 | 管理账号将Agent应用发布至平台内置的应用百宝箱中，方便平台用户快速调用。                       |

为什么选择 Agent 开发平台

● **丰富的插件集成能力**

Agent开发平台通过丰富的插件生态持续扩展智能体的能力边界。平台预置了"文件处理"、"代码解释器"等十几种实用插件，做到开箱即用，更开放自定义插件创建能力，开发者可将企业内部API、第三方工具或自研函数封装为插件，通过简单的参数配置即可让智能体调用。

● **灵活的知识库管理**

平台提供智能化的知识库系统，支持开发者构建专属数据资产库。无论是本地文档（支持PDF、Doc、PPT等格式）、网页数据还是结构化数据库，均可快速导入知识库。智能体会自动学习这些数据并建立索引，实现精准的知识检索与内容生成（例如上传行业报告后，Agent可自动生成竞品分析报告）。

● **模块化 workflow 设计**

平台采用“零代码+低代码”的可视化开发模式，支持开发者通过拖拽节点快速搭建复杂任务流。提供包括自然语言处理、逻辑判断、代码执行等近百种标准化组件，满足从简单问答到多步骤任务的开发需求（例如构建一个包含需求分析、方案设计、成本核算的智能客服工作流）。

● **全链路Agent开发支持**

平台内置完整的开发调试体系，从代码编写、插件测试到工作流验证，提供全生命周期支持。开发者可实时观测Agent执行过程中的每一步操作，查看上下文记忆、插件调用结果及推理过程，并通过分层日志分析快速定位问题（例如追踪智能客服在处理用户投诉时的对话逻辑与数据调用路径）。

● **丰富的MCP拓展能力**

平台工具调用支持MCP协议，并预置了"高德地图"、"车票查询工具"、"必应搜索"等多种实用MCP服务，开通后可以一键集成调用，更开放自定义MCP服务创建能力，开发者可依据MCP服务地址快速创建MCP服务，提升开发效率。

## 平台架构

Agent开发平台提供了一站式AI应用搭建能力，平台提供多种能力组合，支持构建不同复杂度的Agent开发场景。

### 基础通用能力

Agent开发平台支持多个租户共同使用并统一进行管理，支持会话的管理，查看历史对话记录，Agent支持发布成API。

### 核心功能模块

- 工作台：工作台支持直接创建零代码应用、支持通过工作流编排应用，支持通过插件管理创建自定义插件并接入应用。
- 应用百宝箱：应用百宝箱包含了应用市场、工作流市场、外部应用，支持各种应开箱即用。
- 插件市场：插件市场预置了10+各种类型的插件，如"文件处理"、"代码解释器"等，做到开箱即用。
- MCP服务市场：预置多种外部MCP接口，便于用户便捷调用。
- 知识库：支持上传和管理知识库，并支持设置知识库的切片策略等。
- 提示词工程：prompt工程支持创建prompt模板，支持对prompt进行比较、评估、优化。

### 应用场景

Agent支持多种应用场景：

- 知识工程类场景，通过挂载知识库应用于如金融、医疗等专业领域。
- 动态规划类场景，应用于聊天机器人、工具调用等场景。
- 流程编排类场景，通过组件库、工作流编排应用于复杂业务场景。
- 多智能体编排场景，通过Agent进行精准意图分发。
- 创新探索场景，可以将Agent应用于具身智能等场景。

## 亮点特性 1：零代码轻松创建 Agent 应用

- 提示词智能生成：平台内置智能提示词生成功能，即使没有大模型基础，也能快速生成适合的提示词，轻松指引智能体完成任务。
- 丰富技能配置：提供多样化的技能配置选项，涵盖知识库、插件、工作流等。添加技能如同搭积木，只需简单单击和选择，就能让智能体掌握新本领，比如接入特定领域的知识库，使其成为该领域的专家。
- 快速创建与运行：整个创建流程极其简便，几步操作就能让智能体上线工作。从构思想法到实际应用，大幅缩短时间，让零代码用户也能快速拥有专属的智能助手。

## 亮点特性 2：低代码灵活搭建复杂工作流

- 可视化工作流设计：采用直观的图形化界面，将各种功能封装成节点，如大模型节点、意图识别节点等。开发者只需像拼图一样拖拽和连接节点，就能构建复杂的工作流，所见即所得，逻辑清晰可见。

- 代码增强灵活性：在节点间适当支持代码编写，为有编程经验的开发者提供更多定制空间。比如在代码节点添加几行数据处理代码，就能实现特殊的数据转换，满足个性化需求。

# 5 模型能力与规格

[盘古NLP大模型](#)

[盘古CV大模型](#)

[盘古科学计算大模型](#)

[盘古预测大模型能力与规格](#)

[盘古多模态大模型](#)

[盘古专业大模型](#)

[盘古行业大模型](#)

[盘古图像搜索模型](#)

[三方大模型](#)

## 5.1 盘古 NLP 大模型

### 盘古 NLP 大模型规格

盘古NLP大模型是业界首个超千亿参数的中文预训练大模型，结合了大量数据预训练和多源知识，借助持续学习不断吸收海量文本数据，持续提升模型性能。除了实现行业知识检索、文案生成、阅读理解等基础功能外，盘古NLP大模型还具备模型调用等高级特性，可在智能客服、创意营销等多个典型场景中，提供强大的AI技术支持。

ModelArts Studio大模型开发平台为用户提供了多种规格的NLP大模型，以满足不同场景和需求。不同模型在处理上下文token长度和功能上有所差异，以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-1 盘古 NLP 大模型规格清单

| 模型支持区域 | 模型名称                               | 可处理最大序列长度 | 说明                                                                                                    |
|--------|------------------------------------|-----------|-------------------------------------------------------------------------------------------------------|
| 西南-贵阳一 | Pangu-NLP-N1-32K-3.2.36            | 32K       | 2025年1月发布的版本，支持32K序列长度训练，4K/32K序列长度推理。全量微调、LoRA微调8个训练单元起训，1个推理单元即可部署，4K支持256并发，32K支持256并发。            |
|        | Pangu-NLP-N1-128K-3.2.36           | 128K      | 2025年1月发布的版本，仅支持128K序列长度推理，4个推理单元8并发。                                                                 |
|        | Pangu-NLP-N1-Reasoner-128K-3.0.1.1 | 128K      | 2025年6月发布的版本，仅支持128K序列长度推理，4个推理单元384并发。                                                               |
|        | Pangu-NLP-N2-32K-3.1.35            | 32K       | 2025年1月发布的版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。此模型版本差异化支持DPO训练特性。 |
|        | Pangu-NLP-N2-32K-5.0.1.1           | 32K       | 2025年8月发布的版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。                   |
|        | Pangu-NLP-N2-128K-5.0.1.1          | 128K      | 2025年8月发布的版本，仅支持128K序列长度推理。4个推理单元即可部署，支持128并发。                                                        |
|        | Pangu-NLP-N2-Reasoner-32K-3.0.1.1  | 32K       | 2025年6月发布的版本，支持32K序列长度训练和推理。全量微调16个训练单元起训，LoRA微调8个训练单元起训，GRPO训练24个训练单元起训。                             |
|        | Pangu-NLP-N2-Reasoner-128K-5.0.0.1 | 128K      | 2025年6月发布的版本，仅支持32K序列长度推理。全量微调64个训练单元起训，LoRA微调32个训练单元起训。                                              |
|        | Pangu-NLP-N4-4K-3.2.36             | 4K        | 2025年3月发布的版本，支持4K序列长度训练，4K序列长度推理。预训练、全量微调、DPO训练64个训练单元起训，LoRA微调8个训练单元起训，8个推理单元即可部署，支持128并发。           |
|        | Pangu-NLP-N4-32K-2.5.35            | 32K       | 2025年1月发布的版本，仅支持32K序列长度推理部署，8个推理单元128并发。                                                              |



| 模型支持区域 | 模型名称                               | 可处理最大序列长度 | 说明                                                                                                                               |
|--------|------------------------------------|-----------|----------------------------------------------------------------------------------------------------------------------------------|
|        | Pangu-NLP-N4-Reasoner-32K-3.0.1.2  | 32K       | 2025年6月发布的版本，支持32K序列长度训练和推理。全量微调96个训练单元起训，LoRA微调64个训练单元起训，8个推理单元104并发。                                                           |
|        | Pangu-NLP-N4-Reasoner-128K-3.0.1.2 | 128K      | 2025年6月发布的版本，仅支持128K序列长度推理部署，16个推理单元32并发。                                                                                        |
|        | Pangu-RAG-N1-32K-4.3.2             | 32K       | 2025年4月份发布的版本，支持32K序列长度推理，1个推理单元可部署，最高支持1QPS。该模型是基于盘古NLP大模型微调得到的RAG场景模型，提供对话问答能力。                                                |
|        | Pangu-AgentExpert-N1-0.0.2         | 32K       | 25年6月发布的版本，32K序列长度，1个推理单元可部署，256并发。该模型是基于盘古NLP大模型微调得到的Agent场景模型，提供意图识别、提问器。agent专业模型经过算法优化，相比传统模型有较大提升。                          |
|        | Pangu-AgentExpert-N2-0.0.2         | 32K       | 2025年6月份发布的版本，支持32K序列长度推理，1个推理单元可部署，最高支持1QPS。该模型是基于盘古NLP大模型微调得到的Agent场景模型，提供意图识别、提问器、AutoPlanning能力。agent专业模型经过算法优化，相比传统模型有较大提升。 |

盘古 NLP 大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古NLP大模型支持的具体操作：

表 5-2 盘古 NLP 大模型支持的能力清单

| 模型名称                    | 模型训练 | 微调   |      | 强化学习 |     |      | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-------------------------|------|------|------|------|-----|------|------|------|------|----------|
|                         |      | 全量微调 | LORA | DPO  | RFT | GRPO |      |      |      |          |
| Pangu-NLP-N1-32K-3.2.36 | √    | √    | √    | √    | -   | -    | √    | -    | √    | √        |

| 模型名称                               | 模型训练 | 微调   |      | 强化学习 |      |      | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|------------------------------------|------|------|------|------|------|------|------|------|------|----------|
|                                    |      | 全量微调 | LORA | DPO  | RF T | GRPO |      |      |      |          |
| Pangu-NLP-N1-128K-3.2.36           | -    | -    | -    | -    | -    | -    | √    | -    | √    | √        |
| Pangu-NLP-N1-Reasoner-128K-3.0.1.1 | √    | √    | √    | -    | -    | -    | -    | -    | √    | √        |
| Pangu-NLP-N2-32K-3.1.35            | √    | √    | √    | √    | -    | -    | √    | √    | √    | √        |
| Pangu-NLP-N2-32K-5.0.1.1           | √    | √    | √    | -    | -    | -    | √    | -    | √    | √        |
| Pangu-NLP-N2-128K-5.0.1.1          | -    | -    | -    | -    | -    | -    | √    | -    | √    | √        |
| Pangu-NLP-N2-Reasoner-32K-3.0.1.1  | √    | √    | √    | -    | -    | √    | -    | √    | √    | √        |
| Pangu-NLP-N2-Reasoner-128K-5.0.0.1 | √    | √    | √    | -    | -    | -    | -    | √    | √    | √        |
| Pangu-NLP-N4-4K-3.2.36             | √    | √    | √    | √    | -    | -    | √    | √    | √    | √        |
| Pangu-NLP-N4-32K-2.5.35            | √    | √    | √    | √    | √    | -    | √    | -    | √    | √        |
| Pangu-NLP-N4-Reasoner-32K-3.0.1.2  | √    | √    | √    | -    | -    | -    | -    | -    | √    | √        |
| Pangu-NLP-N4-Reasoner-128K-3.0.1.2 | -    | -    | -    | -    | -    | -    | -    | √    | √    | √        |
| Pangu-RAG-N1-32K-4.3.2             | -    | -    | -    | -    | -    | -    | -    | -    | √    | -        |
| Pangu-AgentExpert-N1-0.0.2         | -    | -    | -    | -    | -    | -    | -    | -    | √    | √        |
| Pangu-AgentExpert-N2-0.0.2         | -    | -    | -    | -    | -    | -    | -    | -    | √    | √        |

盘古 NLP 大模型对资源池的依赖

表 5-3 盘古 NLP 大模型对资源池的依赖清单

| 模型名称                               | 云上部署         | 边缘部署        |
|------------------------------------|--------------|-------------|
| Pangu-NLP-N1-32K-3.2.36            | 支持，1个推理单元部署  | -           |
| Pangu-NLP-N1-128K-3.2.36           | 支持，4个推理单元部署  | -           |
| Pangu-NLP-N1-Reasoner-128K-3.0.1.1 | 支持，4个推理单元部署  | -           |
| Pangu-NLP-N2-32K-3.1.35            | 支持，4个推理单元部署  | 支持，4个推理单元部署 |
| Pangu-NLP-N2-Reasoner-32K-3.0.1.1  | 支持，4个推理单元部署  | -           |
| Pangu-NLP-N2-Reasoner-128K-5.0.0.1 | 支持，4个推理单元部署  | -           |
| Pangu-NLP-N4-4K-3.2.36             | 支持，8个推理单元部署  | -           |
| Pangu-NLP-N4-32K-2.5.35            | 支持，8个推理单元部署  | -           |
| Pangu-NLP-N4-Reasoner-32K-3.0.1.2  | 支持，8个推理单元部署  | -           |
| Pangu-NLP-N4-Reasoner-128K-3.0.1.2 | 支持，16个推理单元部署 | -           |
| Pangu-RAG-N1-32K-4.3.2             | 支持，1个推理单元部署  | -           |
| Pangu-AgentExpert-N1-0.0.2         | 支持，1个推理单元部署  | -           |
| Pangu-AgentExpert-N2-0.0.2         | 支持，1个推理单元部署  | -           |

5.2 盘古 CV 大模型

盘古 CV 大模型规格

盘古CV大模型基于海量图像、视频数据和盘古独特技术构筑的视觉基础模型，赋能行业客户利用少量场景数据对模型微调即可实现特定场景任务。

ModelArts Studio大模型开发平台为用户提供了多种规格的CV大模型，如图像分类、目标检测、图像分割、姿态估计等以满足不同场景和需求。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-4 盘古 CV 大模型规格清单

| 模型支持区域 | 模型名称                  | 应用场景 | 行业 | 说明                                    |
|--------|-----------------------|------|----|---------------------------------------|
| 西南-贵阳一 | Pangu-CV-物体检测-S-3.1.0 | 物体检测 | 通用 | 从高精度、低时延、小目标检测三方面进行定向优化，为业界性价比最优检测模型。 |
|        | Pangu-CV-物体检测-S-2.1.0 | 物体检测 | 通用 | 小参数量物体检测，使用盘古自有海量数据进行预训练。             |
|        | Pangu-CV-物体检测-N-2.1.0 | 物体检测 | 通用 | 千万到亿级参数量物体检测，支持模型抽取，使用盘古自有海量数据进行预训练。  |
|        | Pangu-CV-图像分类-2.1.0   | 图像分类 | 通用 | 使用盘古统一预训练基模型底座，添加分类任务头构建图像分类模型。       |
|        | Pangu-CV-语义分割-2.1.0   | 语义分割 | 通用 | 使用盘古统一预训练基模型底座，添加分割任务头构建语义分割模型。       |
|        | Pangu-CV-实例分割-1.1.0   | 实例分割 | 通用 | 使用盘古统一预训练基模型底座，添加分类任务头构建语义分割模型。       |

盘古 CV 大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、模型压缩、在线推理和能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古CV大模型支持的具体操作：

表 5-5 盘古 CV 大模型支持的操作清单

| 模型名称                  | 微调 | L2云上部署 | L2边缘部署 | L0部署 |
|-----------------------|----|--------|--------|------|
| Pangu-CV-物体检测-S-3.1.0 | √  | √      | √      | -    |

| 模型名称                  | 微调 | L2云上部署 | L2边缘部署 | L0部署 |
|-----------------------|----|--------|--------|------|
| Pangu-CV-物体检测-S-2.1.0 | √  | √      | √      | -    |
| Pangu-CV-物体检测-N-2.1.0 | √  | √      | √      | -    |
| Pangu-CV-图像分类-2.1.0   | √  | √      | √      | √    |
| Pangu-CV-语义分割-2.1.0   | √  | √      | √      | -    |
| Pangu-CV-实例分割-1.1.0   | √  | √      | √      | -    |

盘古 CV 大模型对资源池的依赖

表 5-6 盘古 CV 大模型对资源池的依赖清单

| 模型名称                  | 云上部署        | 边缘部署        |
|-----------------------|-------------|-------------|
| Pangu-CV-物体检测-S-3.1.0 | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-CV-物体检测-S-2.1.0 | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-CV-物体检测-N-2.1.0 | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-CV-图像分类-2.1.0   | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-CV-语义分割-2.1.0   | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-CV-实例分割-1.1.0   | 支持，4个推理单元部署 | 支持，4个推理单元部署 |

5.3 盘古科学计算大模型

盘古科学计算大模型规格

盘古科学计算大模型面向气象、医药、水务、机械、航天航空等领域，融合了AI数据建模和AI方程求解方法。该模型从海量数据中提取数理规律，利用神经网络编码微分方程，通过AI模型更快速、更精准地解决科学计算问题。

ModelArts Studio大模型开发平台为用户提供了多种规格的科学计算大模型，以满足不同场景和需求。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-7 盘古科学计算大模型规格清单

| 模型支持区域 | 模型名称                        | 模型版本及时间分辨率        | 说明                                                               |
|--------|-----------------------------|-------------------|------------------------------------------------------------------|
| 西南-贵阳一 | Pangu-AI4S-Global-Ocean     | v2.1.4（时间分辨率：24h） | 2025年9月发布的版本，用于海洋基础要素预测，在预测海洋基础要素的同时，可同步输出基础天气要素，仅支持部署，1个推理单元部署。 |
|        | Pangu-AI4S-Regional-Ocean   | v2.1.4（时间分辨率：24h） | 2025年3月发布的版本，用于区域海洋基础要素预测，1个训练单元起训及1个实例部署。                       |
|        | Pangu-AI4S-Global-Ecology   | v2.1.4（时间分辨率：24h） | 2025年9月发布的版本，用于全球海洋生态要素预测，仅支持部署，1个训练单元起训及1个推理单元部署。               |
|        | Pangu-AI4S-Regional-Ecology | v1.1.4（时间分辨率：24h） | 2025年9月发布的版本，用于区域海洋生态预测，1个训练单元起训及1个推理单元部署。                       |
|        | Pangu-AI4S-Global-Swell     | v2.1.4（时间分辨率：3h）  | 2025年9月发布的版本，用于全球海浪预测，仅支持部署，1个推理单元部署。                            |
|        | Pangu-AI4S-Regional-Swell   | v1.1.4（时间分辨率：3h）  | 2025年9月发布的版本，用于区域海浪预测，1个训练单元起训及1个推理单元部署。                         |
|        | Pangu-AI4S-Global-Weather   | v1.1.4（时间分辨率：1h）  | 2025年3月发布的版本，用于天气基础要素预测，时间分辨率为1小时，1个训练单元起训及1个实例部署。               |
|        | Pangu-AI4S-Global-Weather   | v1.2.4（时间分辨率：3h）  | 2025年3月发布的版本，用于天气基础要素预测，时间分辨率为3小时，1个训练单元起训及1个实例部署。               |
|        | Pangu-AI4S-Global-Weather   | v1.3.4（时间分辨率：6h）  | 2025年3月发布的版本，用于天气基础要素预测，时间分辨率为6小时，1个训练单元起训及1个实例部署。               |
|        | Pangu-AI4S-Global-Weather   | v1.4.4（时间分辨率：24h） | 2025年3月发布的版本，用于天气基础要素预测，时间分辨率为24小时，1个训练单元起训及1个实例部署。              |
|        | Pangu-AI4S-Global-Weather   | v2.1.4（时间分辨率：6h）  | 2025年3月发布的版本，用于天气基础要素预测，时间分辨率为6小时，预报准确度更高，1个实例部署。                |

| 模型支持区域 | 模型名称                              | 模型版本及时间分辨率       | 说明                                                                       |
|--------|-----------------------------------|------------------|--------------------------------------------------------------------------|
|        | Pangu-AI4S-Global-Weather         | v4.1.4（时间分辨率：6h） | 2025年9月发布的版本，用于天气基础要素预测，相比于旧版本的精度有30%提升，时间分辨率为6小时，仅支持部署，至少8个推理单元部署。      |
|        | Pangu-AI4S-Regional-Weather       | v1.1.4（时间分辨率：1h） | 2025年9月发布的版本，用于区域中期天气要素预测场景，1个训练单元起训及1个推理单元部署。                           |
|        | Pangu-AI4S-Global-Precip          | v1.1.4（时间分辨率：6h） | 2025年3月发布的版本，用于降水预测，支持1个实例部署推理。                                          |
|        | Pangu-AI4S-Regional-Precip        | v1.1.4（时间分辨率：1h） | 2025年9月发布的版本，用于区域降水预测场景，1个训练单元起训及1个推理单元部署。                               |
|        | Pangu-AI4S-Regional-Pollution     | v1.1.4（时间分辨率：1h） | 2025年3月发布的版本，用于区域污染物基础要素预测，1个训练单元起训及1个实例部署。                              |
|        | Pangu-AI4S-Regional-Wind          | v1.1.4（时间分辨率：1h） | 2025年6月发布的版本，用于新能源风要素预测，如百米风，时间分辨率为1小时，支持预训练、微调、部署、在线推理特性，基于Snt9B3推理。    |
|        | Pangu-AI4S-Regional-Photovoltaics | v1.1.4（时间分辨率：1h） | 2025年6月发布的版本，用于区域新能源光要素预测，如短波辐射、时间分辨率为1小时，支持预训练、微调、部署、在线推理特性，基于Snt9B3推理。 |

📖 说明

**时间分辨率**是盘古科学计算大模型中的核心概念，其核心定义为：当相关科学数据输入模型后，模型可输出预测结果的最小时间间隔单位。

以气象预测场景为例，若盘古科学计算大模型的时间分辨率设定为 1 小时（1H），则表示输入实时气象数据（如温度、气压、风速等）后，模型可输出未来“1 小时”的气象结果。

盘古科学计算大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、模型压缩、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古科学计算大模型支持的具体操作：

表 5-8 盘古科学计算大模型支持的操作清单

| 模型名称                          | 模型版本及时间分辨率        | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-------------------------------|-------------------|-----|----|------|------|------|----------|
| Pangu-AI4S-Global-Ocean       | v2.1.4（时间分辨率：24h） | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Ocean     | v1.1.4（时间分辨率：24h） | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Ecology     | v2.1.4（时间分辨率：24h） | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Ecology   | v1.1.4（时间分辨率：24h） | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Swell       | v2.1.4（时间分辨率：3h）  | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Swell     | v1.1.4（时间分辨率：3h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v1.1.4（时间分辨率：1h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v1.2.4（时间分辨率：3h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v1.3.4（时间分辨率：6h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v2.1.4（时间分辨率：6h）  | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v4.1.4（时间分辨率：6h）  | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Weather     | v1.4.4（时间分辨率：24h） | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Weather   | V1.1.4（时间分辨率：1h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Global-Precip      | v1.1.4（时间分辨率：6h）  | -   | -  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Precip    | v1.1.4（时间分辨率：1h）  | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Pollution | v1.1.4（时间分辨率：1h）  | √   | √  | -    | -    | √    | √        |



| 模型名称                              | 模型版本及时间分辨率       | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-----------------------------------|------------------|-----|----|------|------|------|----------|
| Pangu-AI4S-Regional-Wind          | v1.1.4（时间分辨率：1h） | √   | √  | -    | -    | √    | √        |
| Pangu-AI4S-Regional-Photovoltaics | v1.1.4（时间分辨率：1h） | √   | √  | -    | -    | √    | √        |

盘古科学计算大模型对资源池的依赖

盘古科学计算的全部模型的训练、在线部署与边缘部署依赖ARM+Snt9B3资源。

表 5-9 科学计算大模型对资源池的依赖清单

| 模型名称                        | 模型版本及时间分辨率        | 云上部署        | 边缘部署        |
|-----------------------------|-------------------|-------------|-------------|
| Pangu-AI4S-Global-Ocean     | v2.1.4（时间分辨率：24h） | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Ocean   | v2.1.4（时间分辨率：24h） | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Ecology   | v2.1.4（时间分辨率：24h） | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Ecology | v1.1.4（时间分辨率：24h） | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Swell     | v2.1.4（时间分辨率：3h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Swell   | v1.1.4（时间分辨率：3h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Weather   | v1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Weather   | v1.2.4（时间分辨率：3h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Weather   | v1.3.4（时间分辨率：6h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Weather   | v2.1.4（时间分辨率：6h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Weather   | v4.1.4（时间分辨率：6h）  | 支持，4个推理单元部署 | 支持，4个推理单元部署 |

| 模型名称                              | 模型版本及时间分辨率        | 云上部署        | 边缘部署        |
|-----------------------------------|-------------------|-------------|-------------|
| Pangu-AI4S-Global-Weather         | v1.4.4（时间分辨率：24h） | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Weather       | V1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Global-Precip          | v1.1.4（时间分辨率：6h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Precip        | v1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Pollution     | v1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Wind          | v1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |
| Pangu-AI4S-Regional-Photovoltaics | v1.1.4（时间分辨率：1h）  | 支持，1个推理单元部署 | 支持，1个推理单元部署 |

## 5.4 盘古预测大模型能力与规格

### 盘古预测大模型规格

盘古预测大模型是面向结构化数据，通过任务理解、模型推荐、模型融合技术，构建通用的预测能力。

ModelArts Studio大模型开发平台为用户提供了多种规格的预测大模型，以满足不同场景和需求。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-10 盘古预测大模型规格清单

| 模型支持区域 | 模型名称                          | 说明                                                                                                                                                                                                                                                             |
|--------|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 西南-贵阳一 | Pangu-Predict-Cla-TS-2.1.2    | <p><b>版本说明</b></p> <p>盘古分类预测大模型分类任务微调工作流，支持时序数据的分类预测，预测目标数据需要是有限的状态类型，数据类型一般是离散枚举类。该版本基于时序预测基模型实现分类预测能力。时序分类预测有很多应用场景，例如：基于工业设备传感器一段时间采集的连续数据，实现设备正常或异常状态的预测。</p> <p><b>训练特性</b></p> <p>支持基于盘古预训练预测模型的微调、支持指定协变量。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p> |
|        | Pangu-Predict-Cla-Table-2.0.1 | <p><b>版本说明</b></p> <p>盘古分类预测大模型分类任务微调工作流，支持表格（非时序）数据的分类预测，预测目标数据需要是有限的状态类型，数据类型一般是离散枚举类。该版本以融合推荐预测模型为底座，实现了分类预测模型的一站式开发工作流。该任务支持特征重要性排序及训练Loss曲线展示。</p> <p><b>训练特性</b></p> <p>支持微调、提供训练特征重要性。</p> <p><b>推理特性</b></p> <p>提供分类置信度、支持云上部署、支持边缘部署。</p>             |
|        | Pangu-Predict-Cla-Table-2.1.0 | <p><b>版本说明</b></p> <p>盘古分类预测大模型分类任务微调工作流，支持表格（非时序）数据的分类预测，预测目标数据需要是有限的状态类型，数据类型一般是离散枚举类。该版本以融合推荐预测模型（极简模式）为基础，实现分类预测模型的一站式高效开发。</p> <p><b>训练特性</b></p> <p>支持微调、提供训练特征重要性。</p> <p><b>推理特性</b></p> <p>提供分类置信度、支持云上部署、支持边缘部署。</p>                                 |

| 模型支持区域 | 模型名称                          | 说明                                                                                                                                                                                                                                                        |
|--------|-------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | Pangu-Predict-Cla-Table-2.2.0 | <p><b>版本说明</b></p> <p>盘古分类预测大模型分类任务微调工作流，支持表格数据的分类预测，预测目标数据需要是有限的状态类型，数据类型一般是离散枚举类。该版本基于统一编码大模型实现分类预测能力。</p> <p><b>训练特性</b></p> <p>支持基于盘古预训练预测模型的微调、支持指定协变量。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p>                                                 |
|        | Pangu-Predict-Reg-TS-2.1.2    | <p><b>版本说明</b></p> <p>盘古预测大模型回归任务微调工作流，支持时序任务的连续值预测，预测目标数据需要是连续数值，数据类型一般是数值类。该版本基于时序预测基模型实现对未来一组观测值的预测，最多支持128个未来时间点的预测。</p> <p><b>训练特性</b></p> <p>支持基于盘古预训练预测模型的微调、支持多目标预测支持基于盘古预训练预测模型的微调、支持增量预训练、支持指定协变量。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p> |
|        | Pangu-Predict-Reg-Table-2.0.1 | <p><b>版本说明</b></p> <p>盘古预测大模型回归任务微调工作流，支持表格（非时序）任务的连续值预测，预测目标数据需要是连续数值，数据类型一般是数值类。该版本以推荐融合预测模型为基础实现了回归预测能力，支持部分场景下的特征重要性及Loss曲线展示。</p> <p><b>训练特性</b></p> <p>支持微调、提供训练特征重要性。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p>                                 |

| 模型支持区域 | 模型名称                           | 说明                                                                                                                                                                                                               |
|--------|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | Pangu-Predict-Reg-Table-2.1.0  | <p><b>版本说明</b></p> <p>盘古预测大模型回归任务微调工作流，支持表格（非时序）任务的连续值预测，预测目标数据需要是连续数值，数据类型一般是数值类。该版本以推荐融合预测模型（极简模式）为基础，实现回归预测模型的一站式高效开发。</p> <p><b>训练特性</b></p> <p>支持微调、提供训练特征重要性。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p> |
|        | Pangu-Predict-Reg-Table-2.2.0  | <p><b>版本说明</b></p> <p>盘古回归预测大模型回归任务微调工作流，支持表格数据的回归预测，预测目标数据需要是有限的状态类型，数据类型一般是离散枚举类。该版本基于统一编码大模型实现回归预测能力。</p> <p><b>训练特性</b></p> <p>支持基于盘古预训练预测模型的微调、支持增量预训练。</p> <p><b>推理特性</b></p> <p>支持云上部署、支持边缘部署。</p>        |
|        | Pangu-Predict-Anom-Table-2.0.1 | <p><b>版本说明</b></p> <p>盘古预测大模型异常检测任务微调工作流，支持表格（非时序）异常检测任务，预测目标是正常类和异常类，该版本基于结构化预测模型的分类能力实现对异常状态和正常状态的分析。</p> <p><b>训练特性</b></p> <p>支持微调。</p> <p><b>推理特性</b></p> <p>提供分类置信度、支持云上部署、支持边缘部署。</p>                     |

| 模型支持区域 | 模型名称                              | 说明                                                                                                                                                                                                                                              |
|--------|-----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | Pangu-Predict-Anom-TS-2.1.0       | <b>版本说明</b><br>盘古预测大模型异常检测任务预训练、微调 workflow，支持时序异常检测任务。预训练 workflow 支持使用设备历史正常数据对时序异常检测大模型进行全量训练，使其更好地学习并重构正常数据的潜在特征分布；微调 workflow 支持使用历史正异常数据对模型进行微调，实现对异常状态和正常状态的识别。<br><b>训练特性</b><br>支持基于盘古预训练预测模型的预训练、微调。<br><b>推理特性</b><br>支持云上部署、支持边缘部署。 |
|        | Pangu-Predict-FewShot-Table-3.0.0 | <b>版本说明</b><br>免微调直推预测模型，适用于小型表格预测任务（小于1W条样本），当前版本支持表格数据的回归、分类预测任务。暂未开放微调和预训练能力, 基于TabPFN V2架构实现。<br><b>训练特性</b><br>不支持微调。<br><b>推理特性</b><br>支持云上部署、支持边缘部署。                                                                                     |

盘古预测大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、模型压缩、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古预测大模型支持的具体操作：

表 5-11 盘古预测大模型支持的操作清单

| 模型                            | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-------------------------------|-----|----|------|------|------|----------|
| Pangu-Predict-Cla-TS-2.1.2    | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Cla-Table-2.0.1 | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Cla-Table-2.2.0 | -   | √  | √    | -    | √    | -        |

| 模型                                | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-----------------------------------|-----|----|------|------|------|----------|
| Pangu-Predict-Cla-Table-2.1.0     | -   | √  | -    | -    | √    | -        |
| Pangu-Predict-Reg-TS-2.1.2        | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Reg-Table-2.0.1     | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Reg-Table-2.1.0     | -   | √  | -    | -    | √    | -        |
| Pangu-Predict-Reg-Table-2.2.0     | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Anom-Table-2.0.1    | -   | √  | √    | -    | √    | -        |
| Pangu-Predict-Anom-TS-2.1.0       | √   | √  | √    | -    | √    | -        |
| Pangu-Predict-FewShot-Table-3.0.0 | -   | -  | -    | -    | √    | -        |

盘古预测大模型对资源池的依赖

表 5-12 盘古预测大模型对资源池的依赖清单

| 模型名称                          | 云上部署 | 边缘部署 |
|-------------------------------|------|------|
| Pangu-Predict-Cla-TS-2.1.2    | 支持   | 支持   |
| Pangu-Predict-Cla-Table-2.0.1 | 支持   | 支持   |
| Pangu-Predict-Cla-Table-2.1.0 | 支持   | 支持   |
| Pangu-Predict-Cla-Table-2.2.0 | 支持   | 支持   |
| Pangu-Predict-Reg-TS-2.1.2    | 支持   | 支持   |
| Pangu-Predict-Reg-Table-2.0.1 | 支持   | 支持   |
| Pangu-Predict-Reg-Table-2.1.0 | 支持   | 支持   |

| 模型名称                              | 云上部署 | 边缘部署 |
|-----------------------------------|------|------|
| Pangu-Predict-Reg-Table-2.2.0     | 支持   | 支持   |
| Pangu-Predict-Anom-Table-2.0.1    | 支持   | 支持   |
| Pangu-Predict-Anom-TS-2.1.0       | 支持   | 支持   |
| Pangu-Predict-FewShot-Table-3.0.0 | 支持   | 支持   |

## 5.5 盘古多模态大模型

### 盘古多模态大模型规格

盘古多模态大模型融合文本、图像、视频等数据，利用深度学习实现跨模态语义对齐，实现多种模态的理解。

ModelArts Studio大模型开发平台为用户提供了多种规格的多模态大模型，以满足不同场景和需求。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-13 盘古多模态大模型规格清单

| 模型支持区域 | 模型名称                             | 说明                                                                                                                                   |
|--------|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| 西南-贵阳一 | Pangu-MM-M2-Text2Video-1.0.0     | 2025年4月发布的版本，支持16:9、9:16、1:1、4:3、3:4等5种长宽比、时长5s的视频生成，需要8个推理单元部署。                                                                     |
|        | Pangu-MM-M2-AIGVideo-1.0.0       | 2025年4月发布的版本，图生视频支持16:9、9:16、1:1、4:3、3:4、原始比例等6种长宽比，时长5s的视频生成，视频续写支持16:9、9:16、1:1、4:3、3:4、原始比例等6种长宽比，续写96帧（24帧率的视频续写约4s），需要8个推理单元部署。 |
|        | Pangu-MM-M1-Txt2Img-1.0.0        | 2025年4月发布的版本，支持1024*1024分辨率的图像生成，支持写实、油画、动漫等数十种生成风格，需要8个推理单元部署。                                                                      |
|        | Pangu-MM-M2-Img2Txt-16K-V5.0.4.2 | 2025年6月发布的多模态理解大模型首版本。该模型具有百亿级参数量，支持图像理解，支持预训练、微调。                                                                                   |



盘古多模态大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、模型压缩、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古多模态大模型支持的具体操作：

表 5-14 盘古多模态大模型支持的平台操作清单

| 模型                               | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|----------------------------------|-----|----|------|------|------|----------|
| Pangu-MM-M2-Text2Video-1.0.0     | -   | -  | -    | -    | √    | √        |
| Pangu-MM-M2-AIGVideo-1.0.RC2     | √   | √  | -    | -    | √    | √        |
| Pangu-MM-M1-Txt2Img-1.0.0        | √   | √  | -    | -    | √    | √        |
| Pangu-MM-M2-Img2Txt-16K-V5.0.4.2 | √   | √  | √    | -    | √    | √        |

盘古多模态大模型对资源池的依赖

表 5-15 盘古多模态大模型对资源池的依赖清单

| 模型名称                             | 云上部署        | 边缘部署      |
|----------------------------------|-------------|-----------|
|                                  | ARM+Snt9B3  | ARM+D310P |
| Pangu-MM-M2-Text2Video-1.0.0     | 支持，8个推理单元部署 | -         |
| Pangu-MM-M2-AIGVideo-1.0.0       | 支持，8个推理单元部署 | -         |
| Pangu-MM-M1-Txt2Img-1.0.0        | 支持，8个推理单元部署 | -         |
| Pangu-MM-M2-Img2Txt-16K-V5.0.4.2 | 支持，8个推理单元部署 | -         |

5.6 盘古专业大模型

盘古向量&重排大模型规格

盘古向量&重排大模型（Pangu-EmbeddingRank模型），用于大模型RAG场景。在文本处理阶段，将文本转换为向量化表示，以便进行检索。在文本检索阶段，对query进

行相关度匹配，根据相关度在海量知识库中召回对应的切片，并对召回的切片进行精细化排序，以确保召回相关度top切片。

以下是Pangu-EmbeddingRank-zh大模型支持的模型清单，您可以根据实际需求选择模型进行开发和应用。

表 5-16 盘古向量&重排大模型规格清单

| 模型支持区域 | 模型名称                         | 说明                                           |
|--------|------------------------------|----------------------------------------------|
| 西南-贵阳一 | Pangu-EmbeddingRank-zh-2.0.3 | 2024年12月份发布的版本，支持中文文档向量生成，1个推理单元可部署，最高支持6QPS |

盘古向量&重排大模型支持的平台操作

在选择和使用Pangu-EmbeddingRank-zh时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型压缩、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古专业大模型支持的具体操作：

表 5-17 盘古向量&重排大模型支持的平台操作清单

| 模型                           | 预训练 | 微调 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|------------------------------|-----|----|------|------|----------|
| Pangu-EmbeddingRank-zh-2.0.3 | -   | -  | -    | √    | -        |

盘古向量&重排模型对资源池的依赖

表 5-18 向量&重排模型对资源池的依赖

| 模型名称                         | 云上部署        |
|------------------------------|-------------|
| Pangu-EmbeddingRank-zh-2.0.3 | 支持，1个推理单元部署 |

盘古搜索规划大模型规格

盘古搜索规划大模型（Pangu-SearchPlan模型），用于RAG场景，提供通用意图分类/多轮查询改写/复杂查询分解/时间抽取等功能，在RAG任务中生产用于检索的query，以及根据query分类路由到后续不同的流程。

ModelArts Studio大模型开发平台为用户提供了一种规格的Pangu-SearchPlan-ZH\_EN大模型。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-19 盘古搜索规划大模型规格清单

| 模型支持区域 | 模型名称                           | 说明                                             |
|--------|--------------------------------|------------------------------------------------|
| 西南-贵阳一 | Pangu-SearchPlan-ZH_EN-7.1.5.1 | 2024年12月份发布的版本，支持4K序列长度推理，1个推理单元可部署，最高支持10QPS。 |

盘古搜索规划大模型支持的平台操作

在选择和使用Pangu-SearchPlan-ZH\_EN时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型压缩、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古专业大模型支持的具体操作：

表 5-20 盘古搜索规划大模型支持的平台操作清单

| 模型                             | 预训练 | 微调 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|--------------------------------|-----|----|------|------|----------|
| Pangu-SearchPlan-ZH_EN-7.1.5.1 | -   | -  | -    | √    | -        |

盘古搜索规划模型对资源池的依赖

表 5-21 搜索规划模型对资源池的依赖

| 模型名称                           | 云上部署        |
|--------------------------------|-------------|
| Pangu-SearchPlan-ZH_EN-7.1.5.1 | 支持，1个推理单元部署 |

5.7 盘古行业大模型

盘古 NLP 行业大模型

盘古行业大模型在盘古NLP大模型的基础之上结合了垂直行业的海量数据，配套MA Studio平台和工具链，可化解特定行业数据构建难、模型效果调优难、场景能力评测难三大难题，满足行业通用场景开箱即用。除了实现行业知识检索、文案生成、阅读理解等基础功能外，盘古行业NLP大模型可在行业领域场景提供更高精度的AI技术支持。

ModelArts Studio大模型开发平台为用户提供了多种规格及行业的NLP大模型，以满足不同场景和需求。不同模型在处理上下文token长度和功能上有所差异，以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-22 盘古 NLP 行业大模型规格清单

| 模型支持区域 | 行业 | 模型名称                                       | 可处理最大上下文长度 | 说明                                                                                                            |
|--------|----|--------------------------------------------|------------|---------------------------------------------------------------------------------------------------------------|
| 西南-贵阳一 | 医学 | Pangu-Medical-NLP-N2-32K-3.1.35            | 32K        | 2025年3月发布版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。此模型版本差异化支持INT8量化特性。         |
|        |    | Pangu-Medical-NLP-N2-Reasoner-32K-3.0.0.1  | 32K        | 2025年6月发布Reasoner版本，支持32K序列长度训练，32K序列长度推理。全量微调32个训练单元起训，LoRA微调8个训练单元起训，4个推理单元即可部署，支持128并发。此模型版本差异化支持INT8量化特性。 |
|        | 金融 | Pangu-Finance-NLP-Reasoner-N2-128K-3.0.1.1 | 128K       | 2025年6月发布Reasoner版本，支持128K序列长度推理，4个推理单元即可部署。                                                                  |

盘古 NLP 行业大模型支持的平台操作

在选择和使用盘古大模型时，了解不同模型所支持的操作行为至关重要。不同模型在预训练、微调、模型评测、在线推理和体验中心能力调测等方面的支持程度各不相同，开发者应根据自身需求选择合适的模型。以下是盘古NLP大模型支持的具体操作：

表 5-23 盘古 NLP 大模型支持的能力清单

| 模型                                        | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|-------------------------------------------|-----|----|------|------|------|----------|
| Pangu-Medical-NLP-N2-32K-3.1.35           | √   | √  | √    | √    | √    | √        |
| Pangu-Medical-NLP-N2-Reasoner-32K-3.0.0.1 | √   | √  | √    | √    | √    | √        |

| 模型                                         | 预训练 | 微调 | 模型评测 | 模型压缩 | 在线推理 | 体验中心能力调测 |
|--------------------------------------------|-----|----|------|------|------|----------|
| Pangu-Finance-NLP-Reasoner-N2-128K-3.0.1.1 | -   | -  | -    | -    | √    | √        |

盘古 NLP 行业大模型对资源池的依赖

表 5-24 盘古 NLP 行业大模型对资源池的依赖清单

| 模型名称                                       | 云上部署        | 边缘部署        |
|--------------------------------------------|-------------|-------------|
| Pangu-Medical-NLP-N2-32K-3.1.35            | 支持，4个推理单元部署 | 支持，4个推理单元部署 |
| Pangu-Medical-NLP-N2-Reasoner-32K-3.0.0.1  | 支持，4个推理单元部署 | -           |
| Pangu-Finance-NLP-Reasoner-N2-128K-3.0.1.1 | 支持，4个推理单元部署 | -           |

5.8 盘古图像搜索模型

盘古图像搜索模型规格

盘古图像搜索模型，用于服装标签场景，基于transformer结构，其中Large模型的参数量31M，模型推理使用固定的尺寸，推理速度快，兼具模型效果和推理性能。

表 5-25 盘古图像搜索模型规格清单

| 模型支持区域 | 模型名称                                         | 说明                                                                           |
|--------|----------------------------------------------|------------------------------------------------------------------------------|
| 西南-贵阳一 | Pangu-AIKits-ImageSearch-Large-Tagging-1.0.0 | 2025年5月发布的版本，基于transformer结构，其中Large模型的参数量31M，模型推理使用固定的尺寸，推理速度快，兼具模型效果和推理性能。 |

盘古图像搜索模型支持的平台操作

表 5-26 盘古图像搜索模型支持的平台操作清单

| 模型名称                                         | 预训练 | 云上部署 |
|----------------------------------------------|-----|------|
| Pangu-AIKits-ImageSearch-Large-Tagging-1.0.0 | √   | √    |

盘古图像搜索模型对资源的依赖

| 模型名称                                         | 云上部署         |
|----------------------------------------------|--------------|
| Pangu-AIKits-ImageSearch-Large-Tagging-1.0.0 | 支持，1个推理单元起部署 |

5.9 三方大模型

三方大模型规格

除了盘古自研模型外，当前ModelArts Studio大模型开发平台还面向NLP领域，集成热门的开源三方NLP模型以供客户选择使用。

例如：DeepSeek V3发布于2024年12月26日，是一个MoE架构的LLM模型，总共671B参数量，在数学、代码类相关评测集上取得了超过GPT-4.5的得分成绩。DeepSeek R1与DeepSeek V3模型结构相同，其于2025年1月20号正式开源，其作为强推理能力模型的杰出代表，引起了极大的关注。DeepSeek R1在数学推理、代码生成等核心任务上追平甚至超过GPT-4o和o1等顶尖闭源模型的效果，成为业界公认的LLM领先模型。近期，DeepSeek开源了更新后的版本模型DeepSeek V3-0324和DeepSeek-R1-0528，模型能力得到了进一步的提升，ModelArts Studio大模型开发平台也进行了集成。

除了DeepSeek外，当前还集成了Qwen3系列模型（Qwen3-8B/14B/30B-A3B/32B/235B-A22B）、DeepSeek-R1-Distill-Qwen-32B、DeepSeek-R1-Distill-LLama-70B/8B以及Qwen2.5-72B、QwQ-32B、Qwen2.5-VL-32B模型。

ModelArts Studio大模型开发平台为用户提供了多种规格的三方NLP大模型，以满足不同场景和需求。以下是当前支持的模型清单，您可以根据实际需求选择最合适的模型进行开发和应用。

表 5-27 三方大模型规格清单

| 模型支持区域 | 模型名称                          | 可处理最大上下文长度 | 说明                                                                                                 |
|--------|-------------------------------|------------|----------------------------------------------------------------------------------------------------|
| 西南-贵阳一 | DeepSeek-V3-32K-0.0.1         | 32K        | 2025年3月发布的版本，支持32K序列长度推理。16个推理单元即可部署，32K支持256并发。                                                   |
|        | DeepSeek-V3-32K-0.0.2         | 32K        | 2025年6月发布的版本，支持32K序列长度推理。16个推理单元即可部署，32K支持256并发。该版本基模型为DeepSeek V3-0324开源版本模型。                     |
|        | DeepSeek-R1-32K-0.0.1         | 32K        | 2025年3月发布的版本，支持32K序列长度推理。16个推理单元即可部署，32K支持256并发。                                                   |
|        | DeepSeek-R1-32K-0.0.2         | 32K        | 2025年6月发布的版本，支持32K序列长度推理。16个推理单元即可部署，32K支持256并发。该版本基模型为DeepSeek R1-0528开源版本模型。                     |
|        | DeepSeek-R1-distill-Qwen-32B  | 32K        | DeepSeek-R1-Distill-Qwen-32B是基于开源模型Qwen2.5-32B，使用DeepSeek-R1生成的数据微调得到的模型。                          |
|        | DeepSeek-R1-distill-Llama-70B | 32K        | DeepSeek-R1-Distill-Llama-70B是基于开源模型Llama-3.1-70B，使用DeepSeek-R1生成的数据微调得到的模型。                       |
|        | DeepSeek-R1-distill-Llama-8B  | 32K        | DeepSeek-R1-Distill-Llama-8B是基于开源模型Llama-3.1-8B，使用DeepSeek-R2生成的数据微调得到的模型。                         |
|        | 通义千问3-235B-A22B               | 32K        | Qwen3-235B-A22B实现思考模式和非思考模式的有效融合，可在对话中切换模式。推理能力显著超过QwQ、通用能力显著超过Qwen2.5-72B-Instruct，达到同规模业界SOTA水平。 |
|        | 通义千问3-32B                     | 32K        | Qwen3-32B实现思考模式和非思考模式的有效融合，可在对话中切换模式。推理能力显著超过QwQ、通用能力显著超过Qwen2.5-32B-Instruct，达到同规模业界SOTA水平。       |
|        | 通义千问3-30B-A3B                 | 32K        | Qwen3-30B-A3B实现思考模式和非思考模式的有效融合，可在对话中切换模式。推理能力显著超过QwQ、通用能力显著超过Qwen2.5-32B-Instruct，达到同规模业界SOTA水平。   |

| 模型支持区域 | 模型名称           | 可处理最大上下文长度 | 说明                                                                                                                                                                                    |
|--------|----------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | 通义千问3-14B      | 32K        | Qwen3-14B实现思考模式和非思考模式的有效融合，可在对话中切换模式。推理能力达到同规模业界SOTA水平、通用能力显著超过Qwen2.5-14B。                                                                                                           |
|        | 通义千问3-8B       | 32K        | Qwen3-8B实现思考模式和非思考模式的有效融合，可在对话中切换模式。推理能力达到同规模业界SOTA水平、通用能力显著超过Qwen2.5-7B。                                                                                                             |
|        | 通义千问2.5-72B    | 32K        | Qwen2.5系列72B模型，相较于Qwen2，Qwen2.5获得了显著更多的知识，并在编程能力和数学能力方面有了大幅提升。此外，新模型在指令执行、生成长文本、理解结构化数据（例如表格）以及生成结构化输出特别是JSON方面取得了显著改进。                                                               |
|        | 通义千问2.5-VL-32B | 32K        | Qwen2.5-VL系列32B模型，具备图像识别、精准视觉定位、文字识别和理解、文档解析、视频理解等能力。                                                                                                                                 |
|        | 通义千问-QWQ-32B   | 32K        | 基于Qwen2.5-32B模型训练的QwQ推理模型，通过强化学习大幅度提升了模型推理能力。模型数学代码等核心指标（AIME 24/25、livecodebench）以及部分通用指标（IFEval、LiveBench等）达到DeepSeek-R1满血版水平，各指标均显著超过同样基于Qwen2.5-32B的DeepSeek-R1-Distill-Qwen-32B。 |

三方大模型支持的平台操作

表 5-28 三方大模型支持的平台操作清单

| 模型名称                          | 模型评测 | 在线推理 | 体验中心能力调测 |
|-------------------------------|------|------|----------|
| DeepSeek-V3-32K-0.0.1         | √    | √    | √        |
| DeepSeek-V3-32K-0.0.2         | √    | √    | √        |
| DeepSeek-R1-32K-0.0.1         | √    | √    | √        |
| DeepSeek-R1-32K-0.0.2         | √    | √    | √        |
| DeepSeek-R1-distil-Qwen-32B   | √    | √    | √        |
| DeepSeek-R1-distill-LLama-70B | √    | √    | √        |



| 模型名称                         | 模型评测 | 在线推理 | 体验中心能力调测 |
|------------------------------|------|------|----------|
| DeepSeek-R1-distill-LLama-8B | √    | √    | √        |
| 通义千问3-235B-A22B              | √    | √    | √        |
| 通义千问3-32B                    | √    | √    | √        |
| 通义千问3-30B-A3B                | √    | √    | √        |
| 通义千问3-14B                    | √    | √    | √        |
| 通义千问3-8B                     | √    | √    | √        |
| 通义千问2.5-72B                  | √    | √    | √        |
| 通义千问2.5-VL-32B               | √    | √    | √        |
| 通义千问-QWQ-32B                 | √    | √    | √        |

三方大模型对资源的依赖

表 5-29 三方大模型对资源的依赖清单

| 模型名称                          | 云上部署         |
|-------------------------------|--------------|
| DeepSeek-V3-32K-0.0.1         | 支持，16个推理单元部署 |
| DeepSeek-V3-32K-0.0.2         | 支持，16个推理单元部署 |
| DeepSeek-R1-32K-0.0.1         | 支持，16个推理单元部署 |
| DeepSeek-R1-32K-0.0.2         | 支持，16个推理单元部署 |
| DeepSeek-R1-distil-Qwen-32B   | 支持，2个推理单元部署  |
| DeepSeek-R1-distill-LLama-70B | 支持，4个推理单元部署  |
| DeepSeek-R1-distill-LLama-8B  | 支持，1个推理单元部署  |
| 通义千问3-235B-A22B               | 支持，16个推理单元部署 |
| 通义千问3-32B                     | 支持，4个推理单元部署  |
| 通义千问3-30B-A3B                 | 支持，2个推理单元部署  |
| 通义千问3-14B                     | 支持，1个推理单元部署  |

| 模型名称           | 云上部署        |
|----------------|-------------|
| 通义千问3-8B       | 支持，1个推理单元部署 |
| 通义千问2.5-72B    | 支持，4个推理单元部署 |
| 通义千问2.5-VL-32B | 支持，4个推理单元部署 |
| 通义千问-QWQ-32B   | 支持，4个推理单元部署 |

# 6 基础知识

大模型开发基本流程介绍

大模型开发基本概念

## 6.1 大模型开发基本流程介绍

大模型（Large Models）通常指的是具有海量参数和复杂结构的深度学习模型，广泛应用于自然语言处理（NLP）等领域。开发一个大模型的过程可以分为以下几个主要步骤：

- **数据集准备**：大模型的性能往往依赖于大量的训练数据。因此，数据集准备是模型开发的第一步。首先，需要根据业务需求收集相关的原始数据，确保数据的覆盖面和多样性。例如，若是自然语言处理任务，可能需要大量的文本数据；如果是计算机视觉任务，则需要图像或视频数据。
- **数据预处理**：数据预处理是数据准备过程中的重要环节，旨在提高数据质量和适应模型的需求。常见的数据预处理操作包括：
  - 去除重复数据：确保数据集中每条数据的唯一性。
  - 填补缺失值：填充数据中的缺失部分，常用方法包括均值填充、中位数填充或删除缺失数据。
  - 数据标准化：将数据转换为统一的格式或范围，特别是在处理数值型数据时（如归一化或标准化）。
  - 去噪处理：去除无关或异常值，减少对模型训练的干扰。数据预处理的目的是保证数据集的质量，使其能够有效地训练模型，并减少对模型性能的不利影响。
- **模型开发**：模型开发是大模型项目中的核心阶段，通常包括以下步骤：
  - 选择合适的模型：根据任务目标选择适当的模型。
  - 模型训练：使用处理后的数据集训练模型。
  - 超参数调优：选择合适的学习率、批次大小等超参数，确保模型在训练过程中能够快速收敛并取得良好的性能。开发阶段的关键是平衡模型的复杂度和计算资源，避免过拟合，同时保证模型能够在实际应用中提供准确的预测结果。
- **应用与部署**：当大模型训练完成并通过验证后，进入应用阶段。主要包括以下几个方面：

- 模型优化与部署：将训练好的大模型部署到生产环境中，可能通过云服务或本地服务器进行推理服务。此时要考虑到模型的响应时间和并发能力。
- 模型监控与迭代：部署后的模型需要持续监控其性能，并根据反馈进行定期更新或再训练。随着新数据的加入，模型可能需要进行调整，以保证其在实际应用中的表现稳定。

在应用阶段，除了将模型嵌入到具体业务流程中外，还需要根据业务需求不断对模型进行优化，使其更加精准和高效。

## 6.2 大模型开发基本概念

### 大模型相关概念

| 概念名                  | 说明                                                                                                                                                                                                                                                                                      |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 大模型是什么               | 大模型是大规模预训练模型的简称，也称预训练模型或基础模型。所谓预训练模型，是指在一个原始任务上预先训练出一个初始模型，然后在下游任务中对该模型进行精调，以提高下游任务的准确性。大规模预训练模型则是指模型参数达到千亿、万亿级别的预训练模型。此类大模型因具备更强的泛化能力，能够沉淀行业经验，并更高效、准确地获取信息。                                                                                                                           |
| 大模型的计量单位 token 指的是什么 | 令牌（Token）是指模型处理和生成文本的基本单位。token 可以是词或者字符的片段。模型的输入和输出的文本都会被转换成 token，然后根据模型的概率分布进行采样或计算。<br><br>例如，在英文中，有些组合单词会根据语义拆分，如 overweight 会被设计为 2 个 token：“over”、“weight”。在中文中，有些汉字会根据语义被整合，如“等于”、“王者荣耀”。<br><br>在盘古大模型中，以 N1 系列模型为例，盘古 1 个 token≈0.75 个英文单词，1 个 token≈1.5 汉字。不同模型的具体情况详见表 6-1。 |

表 6-1 token 比

| 模型规格    | token 比（token/英文单词） | token 比（token/汉字） |
|---------|---------------------|-------------------|
| N1 系列模型 | 0.75                | 1.5               |
| N2 系列模型 | 0.88                | 1.24              |
| N4 系列模型 | 0.75                | 1.5               |

训练相关概念

表 6-2 训练相关概念说明

| 概念名   | 说明                                                                                                                                                  |
|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| 自监督学习 | 自监督学习（ Self-Supervised Learning，简称SSL）是一种机器学习方法，它从未标记的数据中提取监督信号，属于无监督学习的一个子集。该方法通过创建“预设任务”让模型从数据中学习，从而生成有用的表示，可用于后续任务。它无需额外的人工标签数据，因为监督信号直接从数据本身派生。 |
| 有监督学习 | 有监督学习是机器学习任务的一种。它从有标记的训练数据中推导出预测函数。有标记的训练数据是指每个训练实例都包括输入和期望的输出。                                                                                     |
| LoRA  | 局部微调（ LoRA ）是一种优化技术，用于在深度学习模型的微调过程中，只对模型的一部分参数进行更新，而不是对所有参数进行更新。这种方法可以显著减少微调所需的计算资源和时间，同时保持或接近模型的最佳性能。                                              |
| 过拟合   | 过拟合是指为了得到一致假设而使假设变得过度严格，会导致模型产生“以偏概全”的现象，导致模型泛化效果变差。                                                                                                |
| 欠拟合   | 欠拟合是指模型拟合程度不高，数据距离拟合曲线较远，或指模型没有很好地捕捉到数据特征，不能够很好地拟合数据。                                                                                               |
| 损失函数  | 损失函数（ Loss Function ）是用来度量模型的预测值 $f(x)$ 与真实值 $Y$ 的差异程度的运算函数。它是一个非负实值函数，通常使用 $L(Y, f(x))$ 来表示，损失函数越小，模型的鲁棒性就越好。                                      |

推理相关概念

表 6-3 训练相关概念说明

| 概念名     | 说明                                                                                                                                  |
|---------|-------------------------------------------------------------------------------------------------------------------------------------|
| 温度系数    | 温度系数（ temperature ）控制生成语言模型中生成文本的随机性和创造性，调整模型的softmax输出层中预测词的概率。其值越大，则预测词的概率的方差减小，即很多词被选择的可能性增大，利于文本多样化。                            |
| 多样性与一致性 | 多样性和一致性是评估LLM生成语言的两个重要方面。多样性指模型生成的不同输出之间的差异。一致性指相同输入对应的不同输出之间的一致性。                                                                  |
| 重复惩罚    | 重复惩罚（ repetition_penalty ）是在模型训练或生成过程中加入的惩罚项，旨在减少重复生成的可能性。通过在计算损失函数（用于优化模型的指标）时增加对重复输出的惩罚来实现的。如果模型生成了重复的文本，它的损失会增加，从而鼓励模型寻找更多样化的输出。 |

提示词工程相关概念

表 6-4 提示词工程相关概念说明

| 概念名           | 说明                                                                          |
|---------------|-----------------------------------------------------------------------------|
| 提示词           | 提示词（ Prompt ）是一种用于与AI人工智能模型交互的语言，用于指示模型生成所需的内容。                             |
| 思维链           | 思维链（ Chain-of-Thought ）是一种模拟人类解决问题的方法，通过一系列自然语言形式的推理过程，从输入问题开始，逐步推导至最终输出结论。 |
| Self-instruct | Self-instruct是一种将预训练语言模型与指令对齐的方法，允许模型自主生成数据，而不需要大量的人工标注。                    |

# 7 安全

责任共担

身份认证与访问控制

数据保护技术

审计

## 7.1 责任共担

华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

与传统的本地数据中心相比，云计算的运营方和使用方分离，提供了更好的灵活性和控制力，有效降低了客户的运营负担。正因如此，云的安全性无法由一方完全承担，云安全工作需要华为云与您共同努力，如[图7-1](#)所示。

- **华为云**：无论在任何云服务类别下，华为云都会承担基础设施的安全责任，包括安全性、合规性。该基础设施由华为云提供的物理数据中心（计算、存储、网络等）、虚拟化平台及云服务组成。在PaaS、SaaS场景下，华为云也会基于控制原则承担所提供服务或组件的安全配置、漏洞修复、安全防护和入侵检测等职责。
- **客户**：无论在任何云服务类别下，客户数据资产的所有权和控制权都不会转移。在未经授权的情况，华为云承诺不触碰客户数据，客户的内容数据、身份和权限都需要客户自身看护，这包括确保云上内容的合法合规，使用安全的凭证（如强口令、多因子认证）并妥善管理，同时监控内容安全事件和账号异常行为并及时响应。

图 7-1 华为云安全责任共担模型



云安全责任基于控制权，以可见、可用作为前提。在客户上云的过程中，资产（例如设备、硬件、软件、介质、虚拟机、操作系统、数据等）由客户完全控制向客户与华为云共同控制转变，这也就意味着客户需要承担的责任取决于客户所选取的云服务。如图7-1所示，客户可以基于自身的业务需求选择不同的云服务类别（例如IaaS、PaaS、SaaS服务）。不同的云服务类别中，每个组件的控制权不同，这也导致了华为云与客户的责任关系不同。

- 在On-prem场景下，由于客户享有对硬件、软件和数据等资产的全部控制权，因此客户应当对所有组件的安全性负责。
- 在IaaS场景下，客户控制着除基础设施外的所有组件，因此客户需要做好除基础设施外的所有组件的安全工作，例如应用自身的合法合规性、开发设计安全，以及相关组件（如中间件、数据库和操作系统）的漏洞修复、配置安全、安全防护方案等。
- 在PaaS场景下，客户除了对自身部署的应用负责，也要做好自身控制的中间件、数据库、网络控制的安全配置和策略工作。
- 在SaaS场景下，客户对客户内容、账号和权限具有控制权，客户需要做好自身内容的保护以及合法合规、账号和权限的配置和保护等。

## 7.2 身份认证与访问控制

用户可以通过调用REST网络的API来访问盘古大模型服务，有以下两种调用方式：

- Token认证：通过Token认证调用请求。
- AK/SK认证：通过AK（Access Key ID）/SK（Secret Access Key）加密调用请求。经过认证的请求总是需要包含一个签名值，该签名值以请求者的访问密钥（AK/SK）作为加密因子，结合请求体携带的特定信息计算而成。通过访问密钥（AK/SK）认证方式进行认证鉴权，即使用Access Key ID（AK）/Secret Access Key（SK）加密的方法来验证某个请求发送者身份。



## 7.3 数据保护技术

盘古大模型服务通过多种数据保护手段和特性，保障存储在服务中的数据安全可靠。

表 7-1 盘古大模型的数据保护手段和特性

| 数据保护手段       | 简要说明                                                   |
|--------------|--------------------------------------------------------|
| 传输加密（HTTPS）  | 盘古服务使用HTTPS传输协议保证数据传输的安全性。                             |
| 基于OBS提供的数据保护 | 基于OBS服务对用户的数据进行存储和保护。请参考 <a href="#">OBS数据保护技术说明</a> 。 |

## 7.4 审计

云审计服务（Cloud Trace Service，CTS）是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

用户开通云审计服务并创建、配置追踪器后，CTS可记录用户使用盘古的管理事件和数据事件用于审计。

CTS的详细介绍和开通配置方法，请参见[CTS快速入门](#)。

# 8 权限管理

如果您需要对华为云上购买的盘古大模型资源，为企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（IAM）和盘古角色管理功能进行精细的权限管理。

如果华为云账号已经能满足您的要求，不需要创建独立的IAM用户（子用户）进行权限管理，您可以跳过本章节，不影响您使用服务的其他功能。

通过IAM，您可以在华为云账号中给员工创建IAM用户（子用户），并授权控制他们对华为云资源的访问范围。例如，您的员工中有负责软件开发的人员，您希望他们拥有接口的调用权限，但是不希望他们拥有训练模型或者访问训练数据的权限，那么您可以先创建一个IAM用户，并设置该用户在盘古平台中的角色，控制对资源的使用范围。

## IAM 权限

默认情况下，管理员创建的IAM用户（子用户）没有任何权限，需要将其加入用户组，并对用户组授权，才能使得用户组中的用户获得对应的权限。授权后，用户就可以基于被授予的权限对云服务进行操作。

服务使用OBS存储训练数据和评估数据，如果需要对OBS的访问权限进行细粒度的控制。可以在盘古服务的委托中增加Pangu OBSWriteOnly、Pangu OBSReadOnly策略，控制OBS的读写权限。

表 8-1 策略信息

| 策略名称               | 拥有细粒度权限/Action                                                                                                       | 权限描述        |
|--------------------|----------------------------------------------------------------------------------------------------------------------|-------------|
| Pangu OBSWriteOnly | obs:object:AbortMultipartUpload<br>obs:object:DeleteObject<br>obs:object:DeleteObjectVersion<br>obs:object:PutObject | 拥有用户OBS桶写权限 |

| 策略名称              | 拥有细粒度权限/Action                                                                                                                                                                                                                                                           | 权限描述         |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|
| Pangu OBSReadOnly | obs:bucket:GetBucketLocation<br>obs:bucket:HeadBucket<br>obs:bucket:ListAllMyBuckets<br>obs:bucket:ListBucket<br>obs:object:GetObject<br>obs:object:GetObjectAcl<br>obs:object:GetObjectVersion<br>obs:object:GetObjectVersionAcl<br>obs:object:ListMultipartUploadParts | 拥有用户OBS桶只读权限 |

盘古用户角色

盘古大模型的用户可以被赋予不同的角色，对平台资源进行精细化的控制。

表 8-2 角色定义

| 角色名称    | 角色描述                                                                                                                                                                                                                                 |
|---------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 超级管理员   | 订购服务的用户，具备当前平台下对所有工作空间的所有权限。                                                                                                                                                                                                         |
| 管理员     | 对工作空间有完全访问权，包括查看、创建、编辑或删除（适用时）工作空间中的资产，同时拥有添加、移除所在空间成员以及编辑所在空间成员角色的权限。                                                                                                                                                               |
| 模型开发工程师 | 可以执行模型开发工具链模块的所有操作，但是不能创建或者删除计算资源，也不能修改所在空间本身。                                                                                                                                                                                       |
| 应用开发工程师 | 应用开发工程师具备执行应用开发工具链模块所有操作的权限，其余角色不具备。                                                                                                                                                                                                 |
| 标注管理员   | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据加工 &gt; 标注任务 &gt; 任务管理” 模块</li><li>“数据工程 &gt; 数据加工 &gt; 标注任务 &gt; 标注作业” 模块</li><li>“数据工程 &gt; 数据加工 &gt; 标注任务 &gt; 标注审核” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据集” 模块</li></ul> |
| 标注作业员   | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据加工 &gt; 标注任务 &gt; 标注作业” 模块</li></ul>                                                                                                                                      |

| 角色名称  | 角色描述                                                                                                                                                                                                                                         |
|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 标注审核员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据加工 &gt; 标注任务 &gt; 标注审核” 模块</li></ul>                                                                                                                                              |
| 评估管理员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据管理 &gt; 数据集” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据评估 &gt; 人工评估” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据评估 &gt; 人工评估标准” 模块</li></ul>                                                       |
| 评估作业员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据管理 &gt; 数据评估 &gt; 人工评估” 模块</li></ul>                                                                                                                                              |
| 数据导入员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据获取 &gt; 导入任务” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据集” 模块</li></ul>                                                                                                                   |
| 数据加工员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据加工 &gt; 加工任务” 模块</li><li>“数据工程 &gt; 数据加工 &gt; 合成任务” 模块</li><li>“数据工程 &gt; 数据加工 &gt; 配比任务” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据指令” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据集” 模块</li></ul> |
| 数据发布员 | 拥有权限如下： <ul style="list-style-type: none"><li>“数据工程 &gt; 数据发布 &gt; 发布任务” 模块</li><li>“数据工程 &gt; 数据管理 &gt; 数据集” 模块</li></ul>                                                                                                                   |

# 9 约束与限制

本节介绍盘古大模型服务在使用过程中的约束和限制。

## 规格限制

盘古大模型服务的规格限制详见[表9-1](#)。

表 9-1 规格限制

| 资产、资源类型             | 规格                                 | 说明                           |
|---------------------|------------------------------------|------------------------------|
| 模型资产、数据资源、训练资源、推理资源 | 所有按需计费、包年/包月中的模型资产、数据资源、训练资源、推理资源。 | 购买的所有类型的资产与资源仅支持在西南-贵阳一区域使用。 |

## 配额限制

盘古大模型服务的配额限制详见[表9-2](#)。

表 9-2 配额限制

| 资源类型 | 默认配额限制                                          | 是否支持调整                 |
|------|-------------------------------------------------|------------------------|
| 模型实例 | ModelArts Studio大模型开发平台上，单个用户最多可创建和管理2000个模型实例。 | 是<br>如果希望申请提升配额，请联系客服。 |

## 功能限制

盘古大模型服务的功能限制详见[表9-3](#)。

表 9-3 功能限制

| 功能类型                   | 使用限制                                                                                                 |
|------------------------|------------------------------------------------------------------------------------------------------|
| 数据工程-数据格式要求            | ModelArts Studio大模型开发平台支持接入的数据需要满足格式要求，包括文件格式、单个文件大小、所有文本大小以及文件数量等，请参考《用户指南》“使用数据工程构建数据集 > 数据集格式要求”。 |
| 模型开发-训练、评测最小数据量要求      | 使用ModelArts Studio大模型开发平台训练、评测不同模型时，存在不同数据量的限制。以NLP大模型为例，请参考《用户指南》“开发盘古NLP大模型 > 使用数据工程构建NLP大模型数据集”。  |
| 模型开发-模型最小训练单元          | 不同模型的最小训练单元有所不同，具体信息请参见 <a href="#">模型能力与规格</a> 。                                                    |
| 模型开发-NLP大模型请求的最大Token数 | 不同系列的NLP大模型支持请求的最大Token数有所不同，具体信息请参见 <a href="#">模型能力与规格</a> 。                                       |

# 10 与其他服务的关系

---

## 与对象存储服务的关系

盘古大模型使用对象存储服务（Object Storage Service，简称OBS）存储数据和模型，实现安全、高可靠和低成本存储需求。

## 与 ModelArts 服务的关系

盘古大模型使用ModelArts服务进行算法训练部署，帮助用户快速创建和部署模型。

## 与云搜索服务的关系

盘古大模型使用云搜索服务CSS，加入检索模块，提高模型回复的准确性、解决内容过期问题。