盘古大模型

产品介绍

文档版本 01

发布日期 2025-07-28

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 什么是盘古大模型

盘古大模型服务简介

ModelArts Studio是一站式大模型工具链平台，支持百模千态，打造数据、模型、应用三引擎的大模型开发平台。

- **数据工程工具链**

  数据是大模型训练的基础，为大模型提供了必要的知识和信息。数据工程工具链作为盘古大模型服务的重要组成部分，具备数据获取、清洗、数据合成、数据标注、数据评估、数据配比、数据发布和管理等功能。

  该工具链能够高效收集和处理各种格式的数据，满足不同训练和评测任务的需求。通过提供自动化的质量检测和数据清洗能力，对原始数据进行优化，确保其质量和一致性。同时，数据工程工具链还提供强大的数据存储和管理能力，为大模型训练提供高质量的数据支撑。

- **模型开发工具链**

  模型开发工具链是盘古大模型服务的核心组件，提供从模型创建到部署的一站式解决方案。

  该工具链具备模型训练、压缩、部署、评测、推理等功能，通过高效的推理性能和跨平台迁移工具，模型开发工具链能够保障模型在不同环境中的高效应用。

- **Agent开发工具链**

  Agent应用开发工具链是盘古大模型平台的关键模块，支持提示词工程和智能Agent应用创建。该工具链提供提示词设计和管理工具，优化大模型的输入提示，提升输出的准确性和相关性。通过可视化编排工具，应用开发工具链加速大模型应用的开发，满足复杂业务需求。

三方大模型与 ModelArts Studio 开发平台关系

第三方模型能力通过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。

ModelArts Studio大模型开发平台是盘古大模型服务推出的包括数据工程、模型开发、Agent开发为一体的一站式大模型开发平台及大模型应用开发平台，提供覆盖全生命周期的大模型工具链。数据工程套件预置了80+数据处理AI算子，可以满足用户多种任务场景需求，极大提升用户数据处理效率，为大模型训练提供高质量数据；模型开发套件支持对三方大模型进行微调、压缩、评测、部署，极大地降低了大模型开发门槛；Agent开发套件提供基于大语言模型的可视化流程编排能力，结合丰富的功能插件

配置，极大地提升了大模型应用开发效率。同时，ModelArts Studio平台，对数据、模型和Agent在统一的入口进行管理，可以快速掌握资产的使用情况、版本情况和溯源信息等，能够让用户更加方便地进行统一管控和资产管理。

2 产品优势

全面的大模型系列

盘古大模型服务支持三方大模型的推理与部署，目前已预置DeepSeek R1/V3模型，多种模态逐步融合。

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

客服助手

通过三方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。

客服助手：

- 提高服务效率：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。

4 产品功能

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。

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。

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：

- **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、气象类型的数据进行数据提取、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诉求。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。
- **数据评估：**平台支持对处理后的文本、图片、视频等多种格式数据进行质量评估，并预置了基础的评估标准，用户可以直接使用预置标准或创建自定义评估标准，以满足个性化的数据质量需求。最终生成详细的质量评估报告，这些报告能够帮助用户检验数据的准确性、完整性和一致性，确保数据在进行模型训练前的高质量标准，以保证模型在实际应用中的可靠性和稳定性。
- **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根溯源。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。

- **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，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。
- **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

Agent开发平台为开发者提供了一个全面的工具集，帮助您高效地开发、优化和部署应用智能体。无论您是新手还是有经验的开发者，都能通过平台提供的提示词工程、插件扩展、灵活的工作流设计和全链路调测功能，快速实现智能体应用的开发与落地，加速行业AI应用的创新与应用。

- **对于零码开发者（无代码开发经验的用户）**：
  - 平台提供了Prompt提示词工程和插件自定义等功能，帮助用户在无需编写代码的情况下，快速构建、调优并运行属于自己的大模型应用。通过简单的配置，用户可以轻松创建Agent应用，快速体验智能化应用的便捷性。
  - 平台提供导入知识功能，支持用户存储和管理数据，并与AI应用进行互动。支持多种格式的本地文档（如docx、pptx、pdf等），方便导入知识库，为Agent应用提供个性化数据支持。
  - 平台还提供全链路信息观测和调试工具，支持开发者深入分析Agent执行过程中的每个环节。通过对信息进行分层展示，帮助开发者优化AI应用的性能和稳定性，确保应用在不同环境下的顺畅运行。
- **对于低码开发者（具有一定代码开发经验的用户）**：
  - 基于上述功能，平台还提供了灵活的工作流设计功能，支持用户编写少量代码来构建逻辑复杂、稳定性要求高的Agent应用。通过拖拉拽方式，开发者可

以组合各种组件（如大模型、代码、意图识别等），快速搭建工作流，实现更高效的应用开发。

- 平台还提供全链路信息观测和调试工具，支持开发者深入分析工作流执行过程中的每个环节。通过对信息进行分层展示，帮助开发者优化AI应用的性能和稳定性，确保应用在不同环境下的顺畅运行。

5 模型能力与规格

5.1 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 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 结构类似，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。

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。

模型支持区域	模型名称	可处理最大上下文长度	可处理最大输出长度	说明
中国-香港	DeepSeek-R1-32K-0.0.2	32K	8K	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-V3-32K-0.0.2	32K	8K	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。

模型支持区域	模型名称	可处理最大上下文长度	可处理最大输出长度	说明
	DeepSeek-R1-Distil-Qwen-32B-0.0.1	32K	8K	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-0.0.1	32K	8K	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-0.0.1	32K	8K	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-0.0.1	32K	8K	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-0.0.1	32K	8K	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-0.0.1	32K	8K	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。
	通义千问3-14B-0.0.1	32K	8K	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-0.0.1	32K	8K	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。

模型支持区域	模型名称	可处理最大上下文长度	可处理最大输出长度	说明
	通义千问2.5-72B-0.0.1	32K	8K	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-QWQ-32B-0.0.1	32K	8K	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 5-1 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-0.0.2	√	√	√
DeepSeek-R1-32K-0.0.2	√	√	√
DeepSeek-R1-Distil-Qwen-32B-0.0.1	√	√	√
DeepSeek-R1-distill-LLama-70B-0.0.1	√	√	√
DeepSeek-R1-distill-LLama-8B-0.0.1	√	√	√
通义千问3-235B-A22B-0.0.1	√	√	√
通义千问3-32B-0.0.1	√	√	√
通义千问3-30B-A3B-0.0.1	√	√	√
通义千问3-14B-0.0.1	√	√	√
通义千问3-8B-0.0.1	√	√	√
通义千问2.5-72B-0.0.1	√	√	√

模型名称	模型评测	在线推理	体验中心能力调测
通义千问-QWQ-32B-0.0.1	√	√	√

三方大模型对资源的依赖

表 5-2 三方大模型对资源的依赖清单

模型名称	云上部署
	ARM+Snt9B3
DeepSeek-V3-32K-0.0.2	支持
DeepSeek-R1-32K-0.0.2	支持
DeepSeek-R1-Distil-Qwen-32B-0.0.1	支持
DeepSeek-R1-distill-LLama-70B-0.0.1	支持
DeepSeek-R1-distill-LLama-8B-0.0.1	支持
通义千问3-235B-A22B-0.0.1	支持
通义千问3-32B-0.0.1	支持
通义千问3-30B-A3B-0.0.1	支持
通义千问3-14B-0.0.1	支持
通义千问3-8B-0.0.1	支持
通义千问2.5-72B-0.0.1	支持
通义千问-QWQ-32B-0.0.1	支持

6 基础知识

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。开发一个大模型的流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。 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”。 在盘古大模型中，以N1系列模型为例，盘古1token≈0.75个英文单词，1token≈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

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**OBS数据保护技术说明**。

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 obs:object:DeleteObject obs:object:DeleteObjectVersion 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 obs:bucket:HeadBucket obs:bucket:ListAllMyBuckets obs:bucket:ListBucket obs:object:GetObject obs:object:GetObjectAcl obs:object:GetObjectVersion obs:object:GetObjectVersionAcl 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： ● “数据工程 > 数据加工 > 标注任务 > 任务管理”模块 ● “数据工程 > 数据加工 > 标注任务 > 标注作业”模块 ● “数据工程 > 数据加工 > 标注任务 > 标注审核”模块 ● “数据工程 > 数据管理 > 数据集”模块
标注作业员	拥有权限如下： ● “数据工程 > 数据加工 > 标注任务 > 标注作业”模块

角色名称	角色描述
标注审核员	拥有权限如下： ● “数据工程 > 数据加工 > 标注任务 > 标注审核”模块
评估管理员	拥有权限如下： ● “数据工程 > 数据管理 > 数据集”模块 ● “数据工程 > 数据管理 > 数据评估 > 人工评估”模块 ● “数据工程 > 数据管理 > 数据评估 > 人工评估标准”模块
评估作业员	拥有权限如下： ● “数据工程 > 数据管理 > 数据评估 > 人工评估”模块
数据导入员	拥有权限如下： ● “数据工程 > 数据获取 > 导入任务”模块 ● “数据工程 > 数据管理 > 数据集”模块
数据加工员	拥有权限如下： ● “数据工程 > 数据加工 > 加工任务”模块 ● “数据工程 > 数据加工 > 合成任务”模块 ● “数据工程 > 数据加工 > 配比任务”模块 ● “数据工程 > 数据管理 > 数据指令”模块 ● “数据工程 > 数据管理 > 数据集”模块
数据发布员	拥有权限如下： ● “数据工程 > 数据发布 > 发布任务”模块 ● “数据工程 > 数据管理 > 数据集”模块

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平台上，单个用户最多可创建和管理2000个模型实例。	是 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平台训练、评测不同模型时，存在不同数据量的限制。
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**模型能力与规格**。

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、高可靠和低成本的存储需求。

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。