

盘古大模型

计费说明

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 计费概述.....	1
2 计费模式.....	2
3 计费项.....	4
4 变更计费模式.....	5
5 续费.....	6
6 欠费说明.....	7
7 停止计费.....	8
8 计费 FAQ.....	9
8.1 包年/包月和按需计费模式有什么区别.....	9
8.2 包年/包月和按需计费模式哪个更划算.....	9
8.3 同一资源是否同时支持包年/包月和按需计费两种模式.....	9
8.4 包年/包月和按需计费模式是否支持互相切换.....	9
8.5 资源到期了如何续费.....	10

1 计费概述

通过阅读本文，您可以快速了解盘古大模型的计费模式、计费项、续费、欠费等主要计费信息。

- 计费模式

盘古大模型提供包周期计费、按需计费两种计费模式，以满足不同场景下的用户需求。关于计费模式的详细介绍请参见[计费模式](#)。

- 包周期计费是一种预付费模式，即先付费再使用，按照订单的购买周期进行结算，因此在购买之前，您必须确保账户余额充足。

- 按需计费是一种后付费模式，即先使用再付费，按照实际使用时长计费。

在购买后，如果发现当前计费模式无法满足业务需求，您还可以变更计费模式。详细介绍请参见[变更计费模式](#)。

- 计费项

盘古大模型的计费项由模型订阅、数据资源、训练资源和推理资源费用组成。了解每种计费项的详细信息，请参考[计费项](#)。

- 续费

包周期资源到期后，如果您想继续使用服务，需要在保留期内进行手动续费，否则不能再对已过保留期的服务进行续费操作，需重新购买对应的服务。了解更多关于续费的信息，请参见[续费](#)。

- 欠费

在使用云服务时，账户的可用额度小于待结算的账单，即被判定为账户欠费。欠费后，可能会影响云服务资源的正常运行，需要及时充值。详细介绍请参见[欠费说明](#)。

- 停止计费

详细介绍请参见[停止计费](#)。

2 计费模式

包周期计费模式属于预付费模式，即先付费再使用。按需计费模式属于后付费模式，即费用根据服务实际消耗量计费，系统将每小时自动扣费。

盘古大模型的计费模式见[表2-1](#)。

表 2-1 计费模式表

计费类别	计费项	计费模式	计费量纲	付费方式	计费周期
数据服务	数据智算服务	按需计费	智算单元	后付费，根据服务实际消耗量计费	按实际任务时长，时长精确到秒。
		包周期计费	智算单元	预付费，按照订单的购买周期结算	1个月~1年
	数据通算服务	按需计费	通算单元	后付费，根据服务实际消耗量计费	按实际任务时长，时长精确到秒。
		包周期计费	通算单元	预付费，按照订单的购买周期结算	1个月~1年
	数据托管服务	包周期计费	托管单元	预付费，按照订单的购买周期结算	1个月~1年
训练服务	训练服务	按需计费	训练单元	后付费，根据服务实际消耗量计费	按实际任务时长，时长精确到秒。
		包周期计费	训练单元	预付费，按照订单的购买周期结算	1个月~1年

计费类别	计费项	计费模式	计费量纲	付费方式	计费周期
推理服务	推理服务	包周期计费	推理单元	预付费，按照订单的购买周期结算	1个月~1年

3 计费项

数据资源、训练资源和推理资源三个收费项。

- 数据智算单元、数据通算单元支持两种计费方式，包周期订购按使用数量和时长预付费，提供1个月到1年供客户选择，按需订购按单元使用数量和时长后付费，时长精确到秒，数据托管单元按订购数量和时长预付费，提供1个月到1年供客户选择。
- 模型训练资源支持两种计费方式，包周期按订购数量和时长预付费，提供1个月到1年供客户选择；按需订购按单元使用数量和时长后付费，时长精确到秒。
- 模型推理资源按推理单元订购数量和时长预付费，提供1个月到1年供客户选择。

4 变更计费模式

数据托管单元、推理单元默认采用包周期计费，训练单元、数据智算单元、数据通算单元采用包周期和按需计费两种方式。

说明

盘古大模型使用周期内不支持变更配置。

5 续费

包周期服务到期后，您可以通过手动续费来延长服务的有效期。

包周期服务到期后，如果在保留期结束前未完成续费，后续则不能再对已过保留期的服务进行续费操作，需重新购买对应的服务。

6 欠费说明

在使用云服务时，如果账户的可用额度低于待结算账单金额，即被判定为账户欠费。欠费可能会影响云服务资源的正常运行，因此需要及时充值。预付费模式购买后不涉及欠费。

服务按时长计费的，当余额不足以支付当前费用时，账户将被判定为欠费。由于盘古大模型不涉及物理实体资源，因此无宽限期。欠费后继续调用服务会导致账户冻结，并直接进入**保留期**，保留期按需资源不可调用。续费后可恢复正常使用，但续费的生效时间以原到期时间为准，需支付从进入保留期开始至续费时的费用。

账户欠费后，部分操作将受限，建议您尽快续费。具体受限操作如下：

- 按需方式的API接口不可调用。
- 无法开通服务。

7 停止计费

包年/包月资源

对于包年/包月计费模式的资源，用户在购买时会一次性付费，服务将在到期后自动停止使用。

- 如果在计费周期内不再使用包年/包月资源，您可以执行退订操作，系统将根据资源是否属于五天无理由退订、是否使用代金券和折扣券等条件返还一定金额到您的账户。详细的退订规则请参见[云服务退订规则概览](#)。

按需计费资源

对于按需计费模式的资源，若不再使用这些资源且需停止计费，请删除相应资源。

说明

计费话单统计说明：

数据工程按需场景存在话单计费逻辑，涉及到使用智算单元和通算单元的任务执行。

统计目标：用户占用资源的时长。

统计规则：

任务时长以训练任务实际开始训练时间开始计算，不包含任务的初始化、排队中、等待创建所消耗的时间，只包含训练任务和计算任务实际的运行时长，当训练任务和计算任务在任务失败、任务停止时，终止时长统计。

- 开始训练时间：start_time来自于ModelArts训练任务和MRS的计算任务的开始时间。
- 训练时长：duration来自于modelArts训练任务和MRS计算任务的实际运行时长。

8 计费 FAQ

8.1 包年/包月和按需计费模式有什么区别

包年/包月和按需计费模式的区别如下：

- **包年/包月计费模式：**包年/包月的计费模式是一种预付费方式，按订单的购买周期计费，适用于可预估资源使用周期的场景。
- **按需计费模式：**按需付费是后付费方式，可以随时开通/关闭对应资源，支持秒级计费，系统会根据云服务器的实际使用情况每小时出账单，并从账户余额里扣款。

8.2 包年/包月和按需计费模式哪个更划算

包年/包月和按需计费模式可针对不同业务需求进行选择：

- 对于长时间且稳定的业务需求，包年/包月模式通常更划算，因为它能提供更低的平均成本和一定的稳定性。
- 对于短期、突发或不可预测的业务需求，按需计费模式则更为合适，因为它提供了更高的灵活性和避免长期预付费可能带来的压力。

8.3 同一资源是否同时支持包年/包月和按需计费两种模式

盘古大模型的模型订阅、数据托管单元、推理单元默认采用包周期计费，训练单元、数据智算单元、数据通算单元采用包周期和按需计费两种方式。

📖 说明

两种计费方式不能共存，只支持按照一种计费方式进行订购。

8.4 包年/包月和按需计费模式是否支持互相切换

包年/包月和按需计费模式使用周期内不支持主动切换，需要联系客服进行变更配置。

盘古大模型的资源与资产计费模式不同，详见[计费模式](#)。

8.5 资源到期了如何续费

包年/包月方式购买的资源到期后，请在平台订购管理页面进行续订操作。具体步骤如下：

1. 登录ModelArts Studio大模型开发平台，单击页面右上角“订购管理”。
2. 在“订购管理”页面，单击“资源订购”页签。
3. 在“资源订购”页签可进行数据资源、训练资源、推理资源的续费操作。

Studio训练按需计费规则说明：

训练任务按需计费时训练时长统计规则：

训练时长以训练任务实际开始训练时间开始计算，不包含ModelArts Studio的初始化、排队中、等待创建所消耗的时间，只包含训练任务实际的运行时长，当训练任务失败、停止时，即终止训练时长的统计。

其中开始训练时间：start_time为训练任务开始运行的时间。

训练时长：duration训练任务运行时长。

特殊场景的处理规则：

任务重试：当新创建了工作流的子任务，则统计新任务的实际训练时长与原任务实际运行的时长，统计规则与上相同。

断点续训：选择断点进行续训或者重训，本质上都是新的训练任务，按照以上规则统计任务的实际运行时长。

故障恢复：当任务配置并且触发了故障恢复，如果发生pod级的任务重调度，原始任务只统计开始运行到因故障导致失败的时长，如果在pod重调度时，资源充足，新任务不进行排队，统计新任务实际的运行时长，当资源不足，任务会进入排队，直到运行中或者失败，如果排队调度成功，只统计任务实际运行时长。