盘古大模型

产品介绍

文档版本 01

发布日期 2024-08-31

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 1
2 产品优势 2
3 应用场景 4
4 模型能力与规格 6
4.1 模型的基础信息 6
4.2 模型支持的区域 7
4.3 模型支持的操作 8
5 计费说明 10
6 安全 12
6.1 责任共担 12
6.2 身份认证与访问控制 13
6.3 数据保护技术 13
6.4 审计 13
6.5 监控安全风险 14
7 权限管理 15
8 约束与限制 17
9 与其他云服务的关系 18
10 基本概念 19

1 什么是盘古大模型

盘古大模型致力于深耕行业，打造多领域的行业大模型和能力集。其核心能力依托于盘古大模型套件平台，该平台是华为云推出的集数据管理、模型训练和模型部署为一体的一站式大模型开发与应用平台。平台提供了包括盘古大模型在内的多种大模型服务，支持大模型的定制开发，并提供覆盖全生命周期的大模型工具链。

盘古大模型为开发者提供了一种简单高效的方式来开发和部署大模型。通过数据工程、模型开发和应用开发等功能套件，帮助开发者充分发挥盘古大模型的强大功能。企业可根据自身需求选择合适的大模型相关服务和产品，轻松构建自己的模型。

- **数据工程套件**

  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。数据工程套件作为盘古大模型服务的重要组成部分，具备数据获取、清洗、配比和管理等功能。该套件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，数据工程套件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。

- **模型开发套件**

  模型开发套件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。该套件具备模型管理、训练、评估、压缩、部署、推理和迁移等功能，支持模型的自动化评估，确保模型的高性能和可靠性。通过高效的推理性能和跨平台迁移工具，模型开发套件能够保障模型在不同环境中的高效应用。

- **应用开发套件**

  应用开发套件是盘古大模型平台的关键模块，支持提示词工程和AI助手创建。该套件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。通过丰富的开发SDK，应用开发套件加速大模型应用的开发，满足复杂业务需求。

2 产品优势

海量训练数据

盘古大模型依托海量且多样化的训练数据，涵盖从日常对话到专业领域的广泛内容，帮助模型更好地理解和生成自然语言文本，适用于多个领域的业务应用。这些数据不仅丰富多样，还为模型提供了深度和广度的语言学习基础，使其能够生成更加自然、准确且符合语境的文本。

通过对海量数据的深入学习和分析，盘古大模型能够捕捉语言中的细微差别和复杂模式，无论是在词汇使用、语法结构，还是语义理解上，都能达到令人满意的精度。此外，模型具备自我学习和不断进化的能力，随着新数据的持续输入，其性能和适应性不断提升，确保在多变的语言环境中始终保持领先地位。

应用场景灵活

盘古大模型具备强大的学习能力，能够通过少量行业数据快速适应特定业务场景的需求。模型在微调后能够迅速掌握并理解特定行业的专业知识和术语，从而深刻把握行业特性。这种快速学习与适应能力，为各行业企业和机构带来了极大的便利。它们可以根据具体需求，利用盘古大模型构建或优化业务流程，提高工作效率，降低运营成本，并为客户提供更精准、个性化的服务。

模型效果优秀

经过海量数据训练，盘古大模型在各种自然语言处理任务中展现出卓越的性能。无论是文本分类、情感分析、机器翻译，还是问答系统，模型都能以高准确率完成任务，为用户提供高质量的输出结果。

这种卓越的表现源于其先进的算法和深度学习架构。盘古大模型能够深入理解语言的内在逻辑与语义关系，因此在处理复杂语言任务时展现出更高的精准度和效率。这不仅提高了任务的成功率，也大幅提升了用户体验，使盘古大模型成为企业和开发者构建智能应用的首选。

创作能力强

盘古大模型通过海量数据训练，能够捕捉更多语言规律和特征，在各类处理任务中表现出色。无论是生成文章、撰写报告，还是设计广告文案，盘古大模型都能根据输入需求灵活调整，生成符合预期的高质量内容。

推理速度快

盘古大模型采用了高效的深度学习架构和优化算法，显著提升了推理速度。在处理请求时，模型能够更快地生成结果，减少等待时间，从而提升用户体验。这种快速的推理能力使盘古大模型适用于广泛的应用场景。在需要实时反馈的业务中，如在线客服和智能推荐，盘古大模型能够迅速提供准确的结果。

迁移能力强

盘古大模型的迁移能力是其适应多变业务需求的关键。除了在已有领域中表现出色，它还能通过少量的新数据快速迁移到新的领域或场景。这种迁移能力使模型能够在面对新挑战时迅速调整和优化，提供适应新领域的服务。

通过微调技术，盘古大模型能够在保持原有优势的同时，融入新领域的特征和规律，实现对新任务的快速适应。这种能力极大地扩展了模型的应用范围，使其在更广泛的业务场景中发挥作用，为用户提供更加全面和深入的智能服务。

3 应用场景

智能客服

在政企场景中，传统的智能客服系统常受限于语义泛化能力和意图理解能力，导致用户需求难以准确捕捉，频繁转接至人工客服。这不仅增加了企业的运营成本，也影响了用户体验。盘古大模型的引入为这一问题提供了有效解决方案。

盘古大模型通过将客户知识数据转换为向量并存储在向量数据库中，利用先进的自然语言处理技术对用户输入的文本进行深度分析和理解。它能够精准识别用户的意图和需求，即使是复杂或模糊的查询，也能提供准确的响应。这种对话问答方式提高了知识获取效率，使智能客服系统更加人性化和有温度。

此外，盘古大模型还能够根据用户的行为和反馈不断学习和优化，进一步提升服务能力。它能识别用户的情绪和语气，调整回答的语调和内容，更贴近用户的实际需求。这种智能化、个性化的服务体验不仅减少了转人工的频率，还提升了用户满意度。

创意营销

在创意营销领域，企业常常需要投入大量的时间和资源来撰写吸引人的营销文案。然而，传统的人工撰写方式不仅效率低下，还受到写手个人素质的影响。盘古大模型的应用为这一问题提供了创新的解决方案。

盘古大模型通过学习用户所需的文案风格和内容，能够轻松完成广告文案、社交媒体帖子、新闻稿等多种写作任务。它不仅能提供创意丰富、语言生动的文案，还能根据不同产品特性和目标受众进行定制，帮助产品吸引更多的潜在客户。

此外，盘古大模型还能根据市场趋势和用户反馈不断优化文案的创作策略和内容。它能够分析用户的阅读习惯和偏好，调整文案结构和语言风格，以更好地吸引用户注意。这种智能化、个性化的营销文案创作，不仅提升了营销效果，也释放了企业的创作活力和创新潜力。

代码助手

在软件开发领域，编程语言的多样性和复杂性给程序员带来了巨大的挑战。盘古NLP大模型为程序员提供了强大的代码助手，显著提升了研发效率。

盘古大模型能够根据用户给定的题目，快速生成高质量的代码，支持Java、Python、Go等多种编程语言。它不仅能够提供完整的代码实现，还能够根据用户的需求，进行代码补全和不同编程语言之间的改写转化。

借助盘古大模型，程序员可以更加专注于创新和设计，而无需过多关注繁琐的编码工作。它不仅提升了代码的质量和稳定性，还缩短了开发周期，加速了产品的迭代和发布。

4 模型能力与规格

4.1 模型的基础信息

盘古大模型平台为用户提供了多种规格的模型，涵盖从基模型到功能模型的多种选择，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 4-1 NLP 大模型清单

模型类别	模型	token	简介
NLP大模型	盘古-NLP-N1-基础功能模型-32K	部署可选4096、32768	基于NLP-N1-基模型训练的基础功能模型，具备文案生成、多轮对话、实体抽取、翻译、知识问答等大模型通用能力，具有32K上下文能力。
	盘古-NLP-N1-基础功能模型-8K	8192 可外推：16384	基于NLP-N1-基模型训练的基础功能模型，具备文案生成、多轮对话、实体抽取、翻译、知识问答等大模型通用能力，具有8K上下文能力，可外推至16K。
	盘古-NLP-N2-基模型	-	预训练模型，擅长通用任务，擅长文本理解，可以高效进行文案生成与文本解析，高性能、时延低。
	盘古-NLP-N2-基础功能模型-4K	4098	基于NLP-N2-基模型训练的基础功能模型，具备文案生成、多轮对话、实体抽取、翻译、知识问答等大模型通用能力。
	盘古-NLP-N2-基础功能模型-32K	32768	基于NLP-N2-基模型训练的基础功能模型，具备文案生成、多轮对话、实体抽取、翻译、知识问答等大模型通用能力。

模型类别	模型	token	简介
	盘古-NLP-N2-应用增强模型-4K	4096	基于NLP-N2-基模型训练的应用增强模型，支持插件调用，支持多种开发套件，可部署集成至业务系统。
	盘古-NLP-N4-基模型	-	预训练模型，擅长逻辑推理，支持工具调用、自然语言生成SQL，可执行复杂任务，质量更高。
	盘古-NLP-N4-基础功能模型-4K	4096	基于NLP-N4-基模型训练的基础功能模型，具备文案生成、多轮对话、实体抽取、翻译、知识问答等大模型通用能力，具有4K上下文能力。
	盘古-NLP-BI专业大模型-4K	4096	基于NLP-N2-基础功能模型运用特定专业代码数据训练后的BI专业大模型，具有4K上下文能力。
	盘古-NLP-BI专业大模型-32K	32768	基于NLP-N2-基础功能模型运用特定专业代码数据训练后的BI专业大模型，具有32K上下文能力。
	盘古-NLP-N2单场景模型-4K	4096	基于NLP-N2-基模型训练的单场景模型，可支持选择一个场景进行推理，如：搜索RAG方案等，具有4K上下文能力。
	盘古-NLP-N2单场景模型-32K	32768	基于NLP-N2-基模型训练的单场景模型，可支持选择一个场景进行推理，如：搜索RAG方案等，具有32K上下文能力。

说明

基于盘古大模型打造的专业大模型包括BI专业大模型与单场景大模型，支持模型推理，但不支持模型训练。

4.2 模型支持的区域

区域是一个地理区域的概念。我国地域面积广大，由于带宽的原因，无法仅依靠一个数据中心为全国客户提供服务。因此，根据地理区域的不同将全国划分成不同的支持区域。

盘古大模型当前仅支持西南-贵阳一区域。

图 4-1 盘古大模型服务区域

4.3 模型支持的操作

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、模型评估、模型压缩和在线推理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。以下是各个模型支持的具体操作：

表 4-2 模型支持的操作

模型	预训练	微调	模型评估	模型压缩	在线推理
盘古-NLP-N1-基础功能模型-8K	-	√	√	-	√
盘古-NLP-N1-基础功能模型-32K	-	√	-	-	√
盘古-NLP-N2-基模型	-	-	-	-	-
盘古-NLP-N2-基础功能模型-4K	-	√	√	√	√
盘古-NLP-N2-基础功能模型-32K	-	√	√	-	√
盘古-NLP-N2-应用增强模型-4K	-	√	√	-	√
盘古-NLP-N4-基模型	√	-	-	-	-
盘古-NLP-N4-基础功能模型-4K	-	√	-	√	√
盘古-NLP-BI专业大模型-4K	-	-	-	-	√
盘古-NLP-BI专业大模型-32K	-	-	-	-	√
盘古-NLP-N2单场景模型-4K	-	-	-	-	√
盘古-NLP-N2单场景模型-32K	-	-	-	-	√

说明

当前支持评估操作的模型需要经过SFT（有监督微调）后方可进行模型评估。

5 计费说明

计费项

关于盘古大模型的详细费用信息，敬请咨询**华为云售前咨询**，我们将为您提供专业的解答和支持。

盘古NLP大模型分为模型订阅服务、训练服务和推理服务三个收费项。

- 模型订阅服务和推理服务按调用时长计费，时长精确到秒。
- 训练服务按实际消耗的Tokens数量计费，话单周期内的Tokens计算精确到1K Tokens，不足1K Tokens的部分舍去。
- 专业大模型按需推理计费仅支持OP账号使用，推理服务按实际调用的Tokens数量计费，不足1K Tokens则小数点保留至后四位计算。

计费模式

模型订阅服务和推理服务提供包周期计费模式，属于预付费模式，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，请确保您的账户余额充足。

训练与推理服务提供按需计费模式，属于后付费模式，即先使用再付费。费用根据训练服务实际消耗的Tokens数量乘以Tokens单价计费，系统将每小时自动扣费。

变更配置

盘古NLP大模型的模型订阅服务和推理服务默认采用包周期计费，训练服务则默认采用按需计费。使用周期内不支持变更配置。

欠费

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。

模型订阅服务和推理服务为预付费，购买后不涉及欠费。

训练服务按实际消耗的Tokens数量计费，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NLP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。

服务到期

包年包月服务到期后，保留期时长将根据“客户等级”定义。在保留期内的资源处理和费用请参见“**保留期**”。

按需计费模式下，若账户欠费，保留期时长同样依据“客户等级”定义。在保留期内的资源处理和费用请参见“**保留期**”。

如果保留期结束后仍未续订或充值，数据将被删除且无法恢复。

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 安全

6.1 责任共担

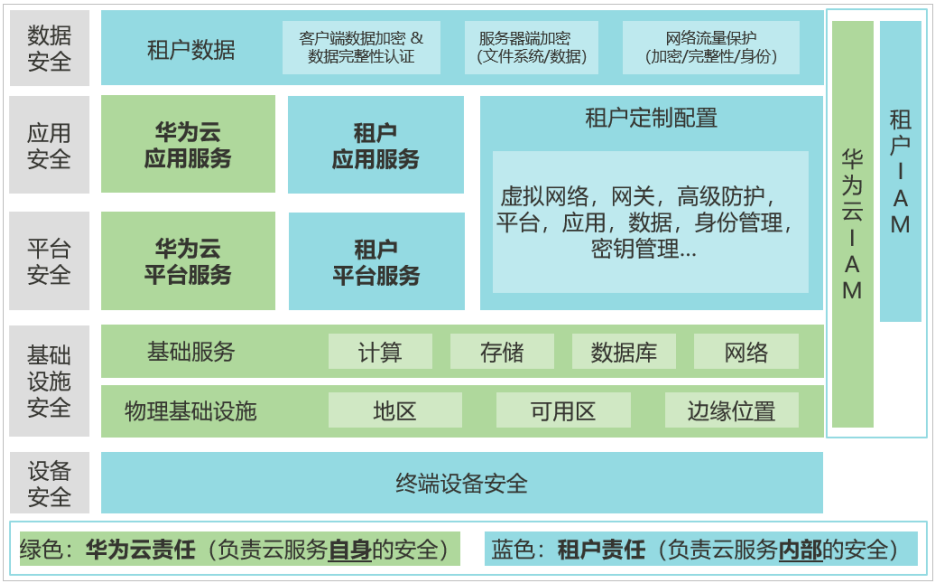
华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

安全性是华为云与您的共同责任，如**图6-1**所示。

- **华为云**：负责云服务**自身**的安全，提供安全的云。华为云的安全责任在于保障其所提供的IaaS、PaaS和SaaS类云服务自身的安全，涵盖华为云数据中心的物理环境设施和运行其上的基础服务、平台服务、应用服务等。这不仅包括华为云基础设施和各项云服务技术的安全功能和性能本身，也包括运维运营安全，以及更广义的安全合规遵从。
- **租户**：负责云服务**内部**的安全，安全地使用云。华为云租户的安全责任在于对使用的IaaS、PaaS和SaaS类云服务内部的安全以及对租户定制配置进行安全有效的管理，包括但不限于虚拟网络、虚拟主机和访客虚拟机的操作系统，虚拟防火墙、API网关和高级安全服务，各项云服务，租户数据，以及身份账号和密钥管理等方面的安全配置。

《华为云安全白皮书》详细介绍华为云安全性的构建思路与措施，包括云安全战略、责任共担模型、合规与隐私、安全组织与人员、基础设施安全、租户服务与租户安全、工程安全、运维运营安全、生态安全。

图 6-1 华为云安全责任共担模型



6.2 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。

6.3 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 6-1 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：https://support.huaweicloud.com/productdesc-obs/obs_03_0375.html

6.4 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**。

6.5 监控安全风险

盘古提供基于主机防护服务HSS的资源和操作监控能力，同时支持CTS审计日志，帮助用户监控自身企业账号下的管理操作。用户可以实时掌握服务使用过程中所产生的各类监控指标。

7 权限管理

如果您需要为企业员工设置不同的访问权限，以实现对华为云上购买的盘古大模型资源的权限隔离，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，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为员工创建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，对于负责软件开发的人员，您希望他们拥有接口的调用权限，但不希望他们拥有训练模型或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，控制他们对资源的使用范围。

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 7-1 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被赋予不同的角色，对平台资源进行精细化的控制。

表 7-2 盘古用户角色

角色	说明
系统管理员	购买平台的用户默认为系统管理员，具有所有操作的权限。
运营人员	具备资产订购（模型订购、资源订购、资源扩容、资源退订）的权限。
模型开发人员	具备推理服务接口调用、能力调测、模型开发套件（模型训练管理、模型部署管理）、数据工程（数据集管理）、应用开发（SDK）功能的使用权限。
推理服务API调用人员	具备推理服务接口调用、能力调测、应用开发（SDK）功能的使用权限。
Prompt工程人员	具备推理服务接口调用、能力调测、应用开发（Prompt）、应用开发（SDK）功能的使用权限。

8 约束与限制

受技术等多种因素制约，盘古大模型服务存在一些约束限制。

- 每个模型请求的最大Token数有所差异，详细请参见**模型的基础信息**。
- 模型所支持的训练数据量、数据格式要求请参见《 用户指南 》“准备训练数据集 > 模型训练所需数据量与数据格式要求 ”。

9 与其他云服务的关系

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。

10 基本概念

训练相关概念

表 10-1 训练相关概念说明

概念名	说明
Token	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者计算。 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”和“weight”。在中文中，有些汉字会根据语义被整合，如“等于”、“王者荣耀”。 例如，在盘古NLP大模型中，1token≈0.75个英文单词，1token≈1.5汉字。
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。

概念名	说明
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 10-2 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。

Prompt 工程相关概念

表 10-3 Prompt 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。