应用平台

AI 原生应用引擎用户指南

文档版本 03

发布日期 2024-05-07

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。

目 录

1 AI 原生应用引擎简介

1.1 为什么使用 AI 原生应用引擎

AI原生应用引擎是一站式的企业专属AI原生应用开发平台，该平台面向企业的研发/技术人员，提供企业专属大模型开发和应用开发的整套工具链，包括数据准备、模型选择/调优、知识工程、模型编排、应用部署、应用集成等能力，降低智能应用开发门槛、提升开发效率。AI原生应用引擎助力企业客户将专属大模型能力融入自己的业务应用链路或对外应用服务中，实现降本增效、改进决策方式、提升客户体验、创新增长模式等经营目标，完成从传统应用到智能应用的竞争力转型。

企业构建 AI 原生应用过程中面临的痛点

- 管好大模型难：大模型百花齐放，能力各异，管好大模型难，为应用场景选择表现最佳模型难。
- 用好大模型难：在企业的复杂场景中，基础大模型效果不佳，且多个大模型结合缺乏有效手段。
- 获取高质量数据难：高质量数据决定AIGC的高度，企业缺少准备契合行业和企业的高质量数据集的能力。
- 数据及模型安全保障难：数据是企业的高价值资产，如何防止数据泄露、安全风险是企业的难题。

AI 原生应用引擎优势

- 提供企业专属大模型开发的整套工具链，包括数据准备、模型选择/调优、知识工程等能力，广泛纳入业界优秀大模型，快速接入模型，提供行业模型评测能力，对多系列、多规格、多版本、多领域、多场景的大模型完成分级分权等精细化管理。
- 提供基于大模型快速构建AI原生应用的整套工具链，支持可视化画布流程编排，开箱即用的RAG/Prompt模版应用，应用部署及应用集成能力，帮助企业用好大模型。
- 构建企业应用与大模型之间的安全隔离带，保障AI原生应用安全可信。

1.2 应用场景

面向不同的企业需求，AI原生应用引擎提供不同的功能服务。

例如，智能对话、以文搜图、NL2SQL等通用应用场景，可在AI原生应用引擎体验各大模型推理云服务，并通过可视化画布流程编排进行业务集成。

细分领域如金融、电网场景，需要对推理结果进行定制调整，则可在AI原生应用引擎使用模型在线微调训练功能，快速生成行业场景定制模型服务，满足用户特定需求。

- **对话沟通**

  针对客户服务和销售团队，通过对话沟通，快速理解并响应客户的需求，以提供高效的解决方案或产品信息。这包括了使用CRM系统进行客户管理、利用即时通讯工具与客户进行互动、进行销售拓展，以及提供定制化的服务方案，旨在提高客户满意度和忠诚度。

- **内容创作**

  可应用于市场营销和品牌传播部门。根据目标受众的偏好和需求，创作吸引人的营销文案、视频剧本和故事内容，包括市场研究、内容策划、以及利用各种数字媒体平台发布和推广内容。帮助企业增强品牌影响力，提高用户参与度和品牌忠诚度。

- **分析控制**

  针对数据分析和业务智能部门，利用先进的数据分析工具和算法，从海量数据中提取有价值的信息，帮助企业做出基于数据的决策。包括客户行为分析、市场趋势预测、以及优化业务流程等。帮助企业提高运营效率，降低成本，同时为客户提供更加个性化的服务。

1.3 功能介绍

AI原生应用引擎的主要功能如**表1-1**所示。

表 1-1 AI 原生应用引擎功能介绍

主要功能	功能简介
应用管理	提供自定义创建、开发、发布、取消发布AI应用，还可以对自己收藏的AI应用进行运行调试等。用户可以将自己在**AI资产中心**关注或后续计划使用的AI应用、技能（工具）进行收藏或取消收藏。
应用体验	将平台预置的应用和用户自己创建的应用进行API调测，帮助开发人员发现并解决应用接口上的问题和错误。
数据管理	平台纳管了用户自定义的和平台预置的数据集，用户使用这些数据集进行模型训练、知识库构建等，快速完成平台使用并验证模型训练效果。
模型管理	用户可以将平台预置模型通过创建模型微调流水线生成微调的模型，还可以创建模型服务及调测模型，检验模型的准确性、可靠性及反应效果。
提示语管理	用户可以将自己创建的、收藏的及平台预置的提示语模板进行优化和改进。
知识库管理	用户可以自定义创建并管理知识库，用于组织和管理大量的数据信息，且创建的知识库启用后可在**创建及管理应用**时引用。

1.4 使用流程

参考**编排应用的流程**、**调优大模型的流程**、**创建知识库的流程**可帮助您快速上手AI原生应用引擎的使用流程和核心功能。

编排应用的流程

图 1-1 编排应用的流程

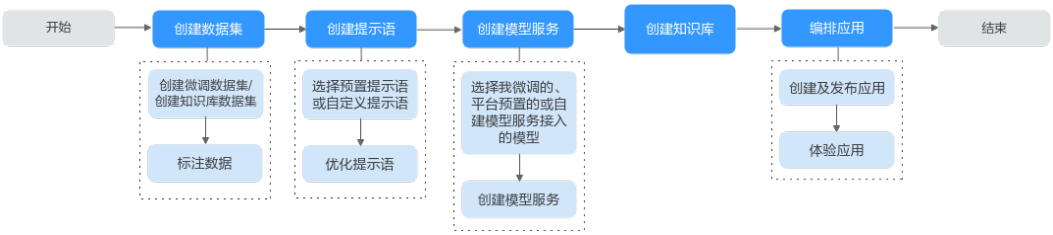

表 1-2 编排应用的流程详解

序号	流程环节		说明
1	创建数据集	**创建微调数据集/创建知识库数据集**	用户根据需要创建微调数据集、知识库数据集，分别用于模型微调、创建知识库。
		标注数据	用户可以将数据集中的某些元素进行标记或分类，以便模型可以更好地理解和使用这些数据。
2	创建提示语	**选择平台预置提示语或自定义提示语**	用户根据需要选择平台预置的提示语模板或自定义提示语模板，可在**创建应用**、**调测模型**中快速引用。
		优化提示语	针对提示语进行结构、排版、内容等维度的优化和改进，将大模型的输入限定在一个特定的范围中，进而更好地控制模型的输出。
3	**创建模型服务**		模型需要部署成功后才可正式提供模型推理服务，平台支持将微调后的模型、系统预置的模型以及通过自建模型服务接入的模型发布为模型服务。调测模型、应用调用均需先部署模型（即部署模型服务）。
4	**创建知识库**		自定义创建并管理知识库，创建的知识库启用后可在**创建应用**时引用。
5	编排应用	**创建及发布应用**	将准备好的模型服务、提示语、知识库等编排应用，以及将应用程序和相关的组件发布，使其能够正常运行。
		体验应用	对应用进行API调测，帮助开发人员发现并解决应用接口上的问题和错误。

调优大模型的流程

图 1-2 调优大模型的流程

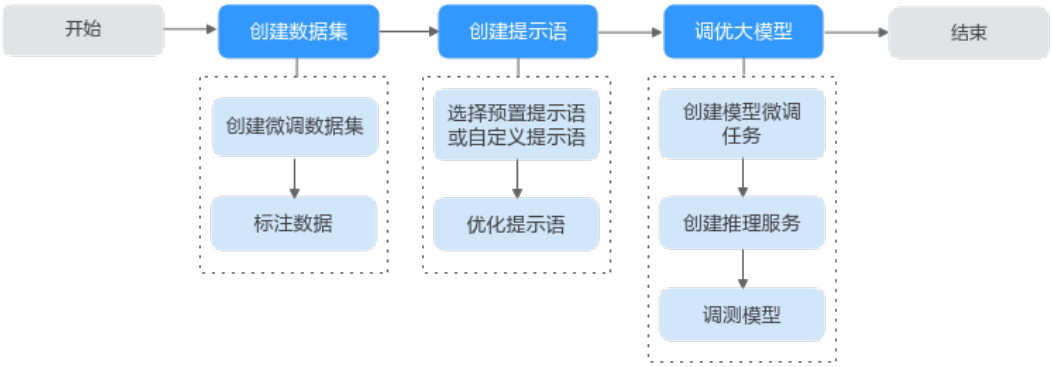

表 1-3 调优大模型的流程详解

序号	流程环节		说明
1	创建数据集	**创建微调数据集**	用户根据需要创建微调数据集，用于模型微调。
		标注数据	用户可以将数据集中的某些元素进行标记或分类，以便模型可以更好地理解和使用这些数据。
2	创建提示语	**选择平台预置提示语或自定义提示语**	用户根据需要选择平台预置的提示语模板或自定义提示语模板，可在**创建应用**、**调测模型**中快速引用。
		优化提示语	针对提示语进行结构、排版、内容等维度的优化和改进，将大模型的输入限定在一个特定的范围中，进而更好地控制模型的输出。
3	调优大模型	**创建模型微调流水线**	通过选择合适的数据集，调整参数，训练平台预置的模型以提高模型效果，可通过训练过程/结果指标初步判断训练效果。
		创建模型服务	训练好的模型需要部署后才可提供推理服务（在线测试模型、应用调用均需先部署模型）。
		调测模型	通过调测模型，检验模型的准确性、可靠性及反应效果，发现模型中存在的问题和局限性。

创建知识库的流程

图 1-3 创建知识库的流程

表 1-4 创建知识库的流程详解

序号	流程环节		说明
1	创建数据集	创建知识库数据集	用户根据需要创建知识库数据集，用于创建知识库。
2	创建知识库		自定义创建并管理知识库，创建的知识库启用后可在创建应用时引用。

1.5 基本概念

使用之前，请先了解表1-5中相关概念，从而更好的使用AI原生应用引擎。

表 1-5 基本概念说明

基本概念	说明
AI应用	AI应用通常指代一种软件程序，具有一定的智能和自主性，可以自主地执行任务、做出决策，并与其他AI应用进行通信和协作。
技能	技能是在自动化和人工智能领域的应用程序。能够自动地执行一些任务或提供一些服务，如客户服务、数据分析、信息传输、智能助手、自动回复等。
智能编排	智能编排是一种基于人工智能技术的自动化流程编排工具，通过分析业务流程，自动构建流程模型，并根据预设规则自动化执行流程，从而提高工作效率和准确性。
ClickHouse	ClickHouse是一个开源的分布式列式数据库管理系统，主要用于在线分析处理（OLAP）场景。它具有高性能、高可靠性、高可扩展性等特点，可以处理海量数据，支持复杂的查询和数据分析操作。ClickHouse支持SQL语言，同时还提供了许多扩展功能，如数据压缩、数据分区、分布式查询等。它被广泛应用于互联网企业、金融、电商、游戏等领域。
节点数	节点数是指在一个特定的环境中，例如测试或生产环境，需要部署的节点数量。
镜像名称	用于标识环境配置的镜像。
镜像版本	用于区分一个镜像库中不同的镜像文件所使用的标签。
资源规格	指根据不同的环境类型和用途，对服务器的 CPU 、内存、数据盘等硬件资源进行合理分配和管理的过程。例如，开发环境的资源规格可能会比生产环境的小，而性能测试环境的资源规格可能会更大，以满足其对硬件资源的需求。
容器端口	容器端口是指在容器内部运行的应用程序所监听的网络端口。容器是一种虚拟化技术，它可以将应用程序及其依赖项打包在一起，形成一个独立运行的环境。在容器内部，应用程序需要监听一个或多个网络端口，以便与外部系统进行通信。

基本概念	说明
服务端口	服务端口是计算机网络中用于标识应用程序的端口号，它是一个16位的整数，范围从0到65535。在一个计算机上，可以同时运行多个应用程序，每个应用程序都需要一个唯一的端口号来标识自己。当一个应用程序需要接受网络请求时，它会监听自己的端口号，等待来自网络的连接请求。当连接请求到达时，应用程序会接受连接并开始处理请求。
推理单元	推理单元是指计算机系统中的一个模块，用于进行逻辑推理和推断。其主要功能是根据已知的事实和规则，推导出新的结论或答案。 推理单元常常被用于解决问题、推理、诊断、规划等任务。它可以帮助计算机系统自动推理出一些结论，从而实现智能化的决策和行为。推理单元通常包括知识表示、推理机和推理策略三个部分。知识表示用于将事实和规则以一定的形式表示出来，推理机则用于实现推理过程，推理策略则用于指导推理机的搜索和推理方向。
大语言模型	大语言模型是一种能够理解和生成人类语言的人工智能模型。这些模型通常使用大量的数据进行训练，以便它们能够识别语言中的模式和规律。大语言模型的应用范围非常广泛，包括自然语言处理、机器翻译、语音识别、智能问答等领域。
向量化模型	向量化模型是将文本数据转换为数值向量的过程。常用于将文本转换为机器可以处理的形式，以便进行各种任务，如文本分类、情感分析、机器翻译等。
多模态模型	多模态模型是指能够处理多种类型数据（如文本、图像、音频等）的机器学习模型。这些模型可以将不同类型的数据进行融合和联合分析，从而实现更全面的理解和更准确的预测。多模态模型的应用非常广泛，例如在图像识别中，可以将图像和文本信息结合起来，提高图像识别的准确性；在自然语言处理中，可以将文本和语音信息结合起来，提高文本语义理解的准确性。
LoRA	Low-Rank Adaptation，低秩适应，是一种将预训练模型权重冻结，并将可训练的秩分解矩阵注入Transformer架构每一层的技术，该技术可减少下游任务的可训练参数数量。
Loss曲线	Loss曲线是一个用于评估模型训练效果的工具，它展示了模型在训练过程中产生的损失（Loss）随时间的变化情况。通过观察Loss曲线，可以了解模型的收敛效果、参数的敏感性和有效性。

2 进入 AI 原生应用引擎

前提条件

已开通AI原生应用引擎。

操作步骤

步骤1 登录**AppStage业务控制台**。

步骤2 在快捷入口选择“AI原生应用引擎”，进入AI原生应用引擎工作台。

----结束

3 管理账号信息

在账号信息页面，用户可以便捷的查看当前登录账号的账户信息（账号名、所属部门等），以及修改账号密码。为保障账号安全，建议定期更新密码。

查看账号信息

在AI原生应用引擎工作台，鼠标光标移至右上角登录的用户名，弹出“账号信息”页面可，可查看当前登录用户的账户信息：账号名、所属部门。

说明

- 如果该账号已同时绑定手机号码和邮箱，则可显示手机号码和邮箱信息。
- 如果该账号仅绑定手机号码或邮箱其中一个，则相应的仅显示已绑定的手机号码或邮箱信息。

修改成员账号密码（通过 OrgID 创建的成员账号）

适用于通过**添加成员**加入组织的成员账号修改密码。为保障账号安全，建议定期更新密码。

步骤1 在AI原生应用引擎工作台，鼠标光标移至右上角登录的用户名，弹出“账号信息”页面。

步骤2 在“账号信息”页面，单击“修改密码”。

步骤3 为确认本人操作需进行身份验证，可选择手机短信验证码方式或邮件验证码方式。

说明

- 如果该账号已同时绑定手机号码和邮箱，则可使用手机短信验证码方式或邮件验证码两种方式。
- 如果该账号仅绑定手机号码或邮箱其中一个，则相应的只需使用手机验证码方式或邮政验证码一种方式。

- 手机短信验证码验证方式的操作如下：
  a. 单击“获取验证码”。
  b. 输入手机上收到的短信验证码，单击“确定”。
- 邮件验证码验证方式的操作如下：
  a. 单击“选择其他验证方式”。

b. 勾选使用邮箱的方式，单击“下一步”。
c. 单击“获取验证码”。
d. 输入邮箱收到的邮件验证码，单击“确定”。

步骤4 在“重置帐号密码”页面，输入旧密码、新密码及再次输入新密码，单击“确定”。

说明

密码需满足以下要求：

- 至少8个字符。
- 至少包含字母和数字，不能包含空格。
- 密码强度：勿使用其他账号的密码。

如果忘记旧密码，可通过如下操作找回密码：

1. 单击“忘记旧密码”。
2. 在“找回密码”页面，输入华为账号（注册账号的手机号或邮件地址）。
3. 输入图形验证码，单击“下一步”。
4. 单击“获取验证码”，输入相应的邮件验证码或手机验证码，再单击“下一步”。
5. 设置新密码并确认新密码，单击“确定”。

说明

- 密码需满足以下要求：
  - 至少8个字符。
  - 至少包含字母和数字，不能包含空格。
  - 密码强度：勿使用其他账号的密码。
- 如果您有其他设备使用此账号，设置新密码后需重新登录，以确保正常使用华为服务。

----结束

修改个人华为账号的密码

适用于修改个人华为账号（包括购买AppStage的租户开通者的个人华为账号、通过**邀请成员**加入组织的个人华为账号）的密码。为保障账号安全，建议定期更新密码。

步骤1 鼠标光标移至右上角登录的用户名，弹出“账号信息”页面。

步骤2 在“账号信息”页面，单击“修改密码”，进入华为账号的“帐号与安全”页面。

步骤3 在“安全中心”区域单击“重置帐号密码”右侧“重置”。

步骤4 在“重置帐号密码”页面，输入旧密码、新密码及再次输入新密码，单击“确定”。

说明

密码需满足以下要求：

- 至少8个字符。
- 至少包含字母和数字，不能包含空格。
- 密码强度：勿使用其他账号的密码。

如果忘记旧密码，可通过如下操作找回密码：

1. 单击“忘记旧密码”。
2. 在“找回密码”页面，输入华为账号（注册账号的手机号或邮件地址）。
3. 输入图形验证码，单击“下一步”。
4. 单击“获取验证码”，输入相应的邮件验证码或手机验证码，再单击“下一步”。
5. 设置新密码并确认新密码，单击“确定”。

说明

- 密码需满足以下要求：
  - 至少8个字符。
  - 至少包含字母和数字，不能包含空格。
  - 密码强度：勿使用其他账号的密码。
- 如果您有其他设备使用此账号，设置新密码后需重新登录，以确保正常使用华为服务。

----结束

4 配置中心

4.1 平台租户鉴权

4.1.1 创建及管理 AK/SK 访问密钥

AK/SK访问密钥是每个用户单独的身份认证，是个人调用应用接口的依据，必须妥善保管。租户**开发的应用**在调用平台接口时需要进行平台鉴权认证，可以使用“AK/SK访问密钥”进行平台的鉴权认证。

操作须知

- 如果访问密钥泄露，会带来数据泄露风险。且每个访问密钥仅能下载一次，为了账号安全性，建议您定期更换并妥善保存访问密钥。
- 若您的访问密钥已丢失，您可创建新的访问密钥并停用原有的访问密钥。
- 每个租户最多只能拥有两个访问密钥。

创建 AK/SK 访问密钥

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“配置中心 > 平台租户鉴权”。

步骤2 在“平台租户鉴权”页面，选择“AK/SK访问密钥”页签。

步骤3 单击“新增访问密钥”，在“新增访问密钥”对话框，输入描述，单击“确定”。

说明

为了保证历史兼容性，会使用访问密钥创建作为初始值。

步骤4 创建成功后，在“创建成功”对话框，单击“立即下载”及时下载并保存访问密钥，否则弹窗关闭后将无法再次获取该密钥信息，但可重新创建新的密钥。

----结束

删除 AK/SK 访问秘钥

密钥删除后无法恢复，请谨慎删除。

步骤1 在“平台租户鉴权”页面，选择“AK/SK访问秘钥”页签。

步骤2 在AK/SK访问秘钥列表中，单击“操作”列“删除”。

步骤3 在“删除访问密钥”对话框，单击“确定”，即可删除不需要的访问密钥。

----结束

4.1.2 创建及管理平台 API Key

API Key是每个用户单独的身份认证，是个人调用应用接口的依据，必须妥善保管。租户**开发的应用**在调用平台接口时需要进行平台鉴权认证，可以使用“平台API Key”进行平台的鉴权认证。

背景信息

对于华为或者第三方运营的商业化模型服务，支持通过API接入到AI原生应用引擎。模型运营方负责模型能力及技术支持，租户自行在模型运营方订购服务，或者通过云市场订购模型服务后，在AI原生应用引擎创建平台API Key，即可通过AI原生应用引擎平台调用模型接口。

创建 API Key

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“配置中心 > 平台租户鉴权”。

步骤2 在“平台租户鉴权”页面，选择“平台API Key”页签，单击“新增平台API Key”。

步骤3 在“新增平台API Key”对话框中的输入框设置API Key名称，用以区分API Key。

说明

最多可添加10个平台API Key。

步骤4 单击“确定”，新建的API Key显示在API Key列表中。

步骤5 （可选）在列表中单击可快速复制各API Key。

----结束

删除 API Key

删除API Key后无法恢复，请谨慎删除。

步骤1 在“配置中心 > 平台租户鉴权 > 平台API Key”页面的API Key列表中，单击“操作”列“删除”。

步骤2 在“删除平台API Key”对话框，单击“确定”，即可删除不需要的API Key。

----结束

5 AI 工作空间

在AI原生应用引擎工作台左侧导航栏选择“AI工作空间”，进入AI工作空间页面，可获得系统中各资源数据概览及产品的相关快速指引。

AI工作空间页面分为数据统计、选择应用创建类型、操作指引三个区域，如图5-1所示，各区域的功能说明如表5-1所述。

图 5-1 AI 工作空间

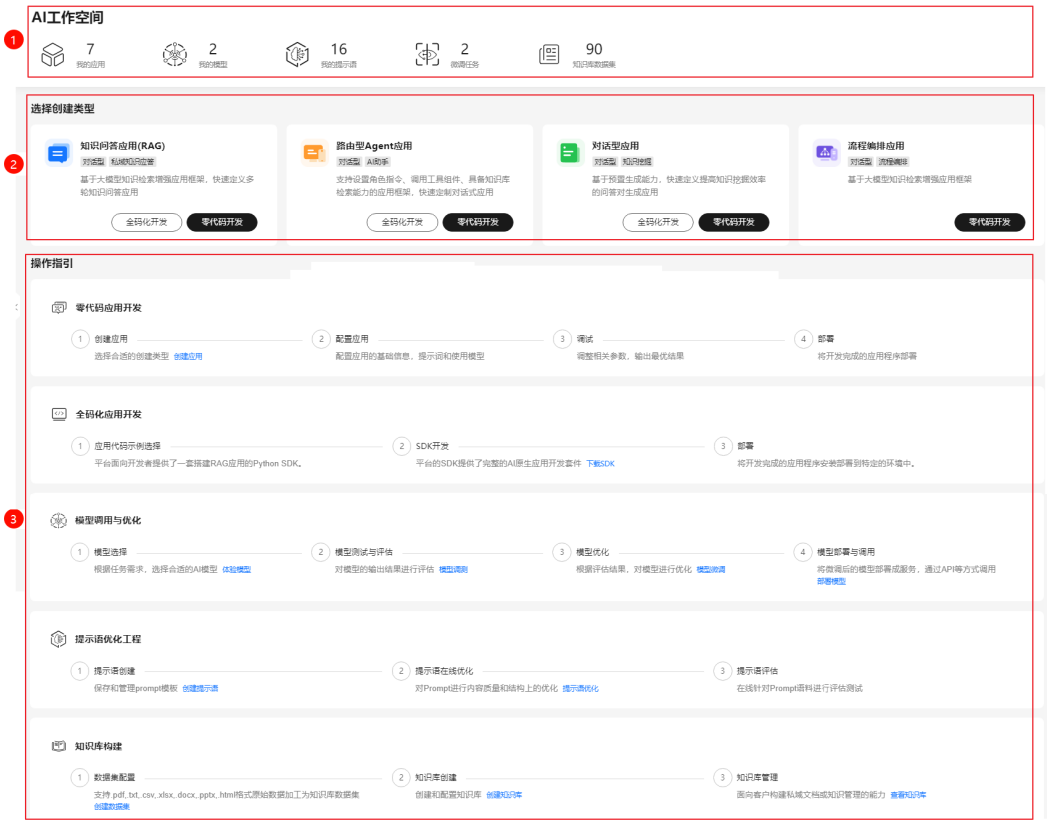

表 5-1 AI 工作空间页面说明

序号	区域	说明
1	数据总览	在数据总览区域可查看下述信息数据： ● 我的应用总数 ● 我的模型数 ● 我的提示语数 ● 微调任务数 ● 知识库数据集数
2	创建应用指引	在“选择创建类型”区域根据指引选择应用类型和开发方式（具体操作请参见**创建及管理应用**）： ● 知识问答应用（RAG）：基于大模型+提示语+知识库的RAG应用框架。支持全码化开发或零代码方式开发。 ● 路由型Agent应用：基于大模型+提示语+知识库+工具的应用框架。支持全码化开发或零代码方式开发。 ● 对话型应用：基于大模型+提示语的对话应用框架。支持全码化开发或零代码方式开发。 ● 流程编排应用：基于大模型知识检索增强应用框架。支持零代码方式开发。
3	操作指引	在“操作指引”区域，可概览各使用场景的流程指引： ● 零代码应用开发：详细流程说明请参见**零代码应用开发**。 ● 全码化应用开发：详细流程说明请参见**全码化应用开发**。 ● 模型调用与优化：详细流程说明请参见**创建模型微调流水线**、**调测模型**、**创建及管理模型服务**。 ● 提示语优化工程：详细流程说明请参见**创建提示语**、**优化提示语**。 ● 知识库构建：详细流程说明请参见**创建及管理知识库**。

6 AI 资产中心

资产中心是一个展示和推荐各种人工智能应用的平台，让用户可以方便地下载和使用不同的AI应用。AI应用广场有不同的类型，包括AI应用、技能等，用户可以直接使用或者二次开发后使用，享受AI带来的便利和乐趣。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“AI资产中心”。

步骤2 在“AI资产中心”页面可依次展开平台预置的AI应用、技能、大模型、数据集、提示语模板页签，可执行如**表6-1**所示的操作。

表 6-1 AI 应用广场支持的操作

分类	支持的操作	说明
AI应用	快速筛选	选择“AI应用”页签，在左侧“筛选”区域，进行不同维度的快速筛选、查看和搜索。
	查看应用详情	单击应用卡片上的应用名称，进入应用详情页面，可查看应用的基础信息、应用组成、接口信息、套件介绍、原理介绍等信息。
	收藏应用	通过如下两种方法，将自己关注或后续计划使用的应用收藏后，可便捷的在**创建及管理应用**中对应用进行运行调试等操作。 • 方法一：鼠标光标移至应用卡片上，单击卡片右上角  （单击可取消收藏）。 • 方法二：在查看应用详情的页面，单击右上角  （单击  可取消收藏）。

分类	支持的操作	说明
	体验应用	1. 鼠标光标移至应用卡片上单击“体验”，进入“应用体验”页面。 2. 在“应用体验”页面，进行以下相关参数和请求体配置。 – 选择应用部署：无需配置，默认为平台预置的应用部署。 – 选择应用：无需配置，默认为当前选择的应用。 – 选择接口API：在下拉列表选择调试应用的接口API。 – 请求体：输入应用接口中的请求体内容。示例如下： `{ "query": "请详细说明AppStage平台有哪些大模型", "file_id": [] }` 3. 在“应用体验”页面右侧“API调测”区域，单击  查看调测结果。 **说明** – 对话框中输入API调试语句也可进行调测。 – 单击右上角  可清空历史调试语句。
技能	快速筛选	选择“技能”页签，在左侧“筛选”区域，进行不同维度的快速筛选、查看和搜索。
	查看技能详情	单击技能卡片上的工具名称，进入工具详情页面，可查看工具的基础信息、组成、接口信息等详情。
	收藏技能	通过如下两种方法，将自己关注或后续计划使用的技能（工具）收藏后，可便捷的在**创建应用**中对技能进行运行调试、二次开发等操作。 方法一：鼠标光标移至技能卡片上，单击卡片右上角  （单击  可取消收藏）。 方法二：在查看技能（工具）详情的页面，单击右上角  （单击  可取消收藏）。

分类	支持的操作	说明
	体验技能	1. 鼠标光标移至技能卡片上单击“体验”，进入“应用体验”页面。 2. 在“应用体验”页面，进行以下相关参数和请求体配置。 – 选择应用部署：无需配置，默认为平台预置的应用部署。 – 选择应用：无需配置，默认为当前选择的应用。 – 选择接口API：在下拉列表选择调试应用的接口API。 – 请求体：输入应用接口中的请求体内容。示例如下： `{ "query": "请详细说明AppStage平台有哪些大模型", "file_id": [] }` 3. 在“应用体验”页面右侧“API调测”区域，单击  查看调测结果。 **说明** – 对话框中输入API调试语句也可进行调测。 – 单击右上角  可清空历史调试语句。
大模型	快速筛选	选择“大模型”页签，在左侧“筛选”区域，进行不同维度的快速筛选。
	查看大模型详情	单击模型卡片，进入模型详情页面，查看模型信息（模型类型、来源、发布者、上架状态、更新时间等）和模型介绍等。
	收藏大模型	通过如下两种方法，将自己关注或后续计划使用的模型收藏后，可便捷的在模型训练、模型部署时使用。 方法一：鼠标光标移至模型卡片上，单击卡片右上角  （单击  可取消收藏）。 方法二：在查看模型详情的页面，单击右上角  （单击  可取消收藏）。
	体验大模型	鼠标移至大模型卡片并单击“体验”，进入模型调测页面。
	部署大模型	鼠标移至大模型卡片并单击“部署”，进入“创建推理服务”页面，参见**创建及管理模型服务**将模型部署为在线服务，对在线服务进行预测和调用。
	调优大模型	鼠标移至大模型卡片并单击“调优”，进入“创建微调任务”页面，参见**创建模型微调流水线**调整大型语言模型的参数以适应特定任务。

分类	支持的操作	说明
	设置鉴权	第三方的大模型需要设置鉴权信息，鼠标移至第三方大模型卡片单击“设置鉴权”，弹出“设置鉴权信息”对话框，在对话框中可根据提示链接跳转至第三方模型官网获取相应API Key。 • 如果未设置API Key，在“设置鉴权信息”对话框输入API Key，单击“保存”。 • 如果已设置API Key，在“设置鉴权信息”对话框单击“移除”可清除已存在的API Key；鼠标移至模型卡片再次单击“设置鉴权”，在“设置鉴权信息”对话框重新输入API Key，单击“保存”即可。
数据集	快速筛选	选择“数据集”页签，在左侧“筛选”区域，进行不同维度的快速筛选、查看和搜索。
	查看数据集详情	鼠标移至数据集卡片并单击数据集名称，进入“数据概况”页面，可查看数据预览、基础信息（数据集用途、格式、来源、创建时间等）、数据介绍（如数据结构、数据使用注意事项等）信息。
	收藏数据集	通过如下两种方法，将自己关注或后续计划使用的数据集收藏后，可便捷的在模型训练、数据标注、创建知识库时使用。 方法一：鼠标光标移至数据集卡片上，单击卡片右上角  （单击  可取消收藏）。 方法二：在查看数据集详情的页面，单击右上角  （单击  可取消收藏）。
提示语模板	快速筛选	选择“提示语”页签，在左侧“筛选”区域，进行不同维度的快速筛选、查看和搜索。
	查看提示语模板详情	鼠标移至提示语卡片并单击模板名称，进入提示语详情页面，查看提示语的基础信息（适用行业、适用任务类型、更新时间等）及提示语信息（适用模型服务、提示语内容、变量）。
	收藏提示语模板	通过如下两种方法，将自己关注或后续计划使用的提示语收藏后，可便捷的在**创建应用**、**调测模型**中快速使用。 方法一：鼠标光标移至提示语卡片上，单击卡片右上角  （单击  可取消收藏）。 方法二：在查看提示语详情的页面，单击右上角  （单击  可取消收藏）。
	复制内容	鼠标移至提示语卡片并单击“复制内容”，可一键复制提示语模板的全部内容。

----结束

7 应用编排中心

7.1 应用编排中心概述

在应用编排中心，用户可以选择零代码应用开发，即通过提示语编辑的方式，结合大模型，提供行为说明，引入数据集，工具等能力，完成智能体开发；也可以选择全码化应用开发，即基于SDK或API的灵活开发模式，进行多范式开发。

操作指引

图 7-1 应用编排中心操作指引

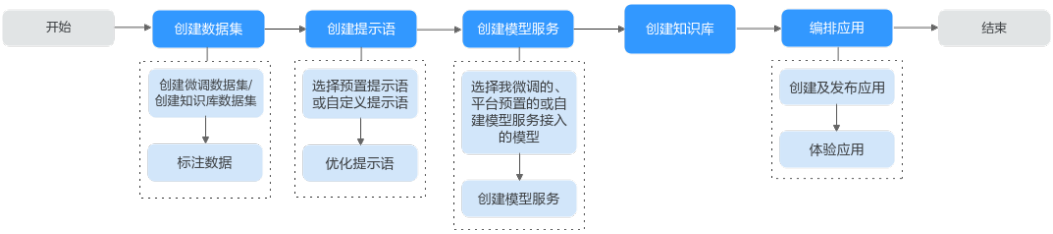

表 7-1 编排应用的流程详解

序号	流程环节		说明
1	创建数据集	创建微调数据集/创建知识库数据集	用户根据需要创建微调数据集、知识库数据集，分别用于模型微调、创建知识库。
		标注数据	用户可以将数据集中的某些元素进行标记或分类，以便模型可以更好地理解和使用这些数据。
2	创建提示语	选择平台预置提示语或自定义提示语	用户根据需要选择平台预置的提示语模板或自定义提示语模板，可在创建应用、调测模型中快速引用。
		优化提示语	针对提示语进行结构、排版、内容等维度的优化和改进，将大模型的输入限定在一个特定的范围中，进而更好地控制模型的输出。

序号	流程环节		说明
3	创建模型服务		模型需要部署成功后才可正式提供模型推理服务，平台支持将微调后的模型、系统预置的模型以及通过自建模型服务接入的模型发布为模型服务。调测模型、应用调用均需先部署模型（即部署模型服务）。
4	创建知识库		自定义创建并管理知识库，创建的知识库启用后可在创建应用时引用。
5	编排应用	创建及发布应用	将准备好的模型服务、提示语、知识库等编排应用，以及将应用程序和相关的组件发布，使其能够正常运行。
		体验应用	对应用进行API调测，帮助开发人员发现并解决应用接口上的问题和错误。

7.2 创建提示语

提示语是给大模型的指令。它可以是一个问题、一段文字描述，也可以是带有一堆参数的文字描述，用于在对话或文章中的一些简短的、不太明确的线索或暗示，推进引导对话的发展，或者增加故事的复杂性和深度。大模型会基于提示语所提供的信息，生成对应的文本或者图片。

提示语模板

AI资产中心的“提示语模板”页签中预置了多款提示语模板，用户可一键快速复制内容并收藏至自己的提示语管理中，这些模板是基于大量应用场景下的经验或者训练语料而总结出一些优质的提示语组成结构，将其抽离成为一种模板，支持一键快速复制内容、收藏、在线优化等功能。

用户创建的、收藏的以及平台预置的提示语模板都可在创建及管理应用、调测模型中快速引用。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的提示语”。

步骤2 在“我的提示语 > 我创建的”页面，单击右上角“创建提示语”。

步骤3 在“创建提示语”页面，参照表7-2进行基础配置后，单击“下一步”。

表 7-2 提示语基础配置参数说明

参数名称	参数说明
提示语名称	用户自定义提示语名称，命名要求：长度2~20，不能以下划线数字开头，只能由中文、字母、数字、下划线组成。

参数名称	参数说明
适用行业	提示语适用的行业领域，包括： ● 通用 ● 政务 ● 金融 ● 交通 ● 能源 ● 教育 ● 医疗 ● 制造 ● 零售 ● 其他
适用任务类型	提示语适用的任务类型，包括： ● 对话问答 ● 文案生成 ● 多模生成 ● 代码生成 ● NL2SQL ● 功能调用 ● 任务规划 ● 全功能
标签	为提示语选择标签分类。可从以下几个维度选择（支持多选）： ● 通用：适配模型（盘古、百川、千问、llama、chatglm、通用）、语言（中文、英文） ● 适用领域：对话问答、文案生成、多模生成、代码生成、NL2SQL、功能调用 ● 行业：通用、政务、金融、交通、能源、教育
变量标识符	用户可选择以下符号标识提示语内容中的变量。 ● 大括号{} ● 双大括号{{}} ● 双中括号[[]] ● 中括号[] ● 小括号() ● 双小括号(())

参数名称	参数说明
提示语内容	可通过以下两种方式定义提示语内容。 • 自定义提示语内容： 插值参数通过所选的变量标识符来填写定义，支持英文、数字、下划线（_）,不能以数字开头。 以变量标识符“双大括号{{}}”为例，提示语中的变量内容则填入双大括号{{}}中。 • 引用模板提示语内容： 单击输入框右侧的“引用模板”选择我创建的、我收藏的或平台预置的提示语模板。

步骤4 在“在线调优”页面，参照表7-3进行参数配置。

表 7-3 提示语在线调优参数说明

参数名称	参数说明
变量标识符	可选择以下符号标识提示语内容中的变量。 • 大括号{} • 双大括号{{}} • 双中括号[[]] • 中括号[] • 小括号() • 双小括号(())
提示语内容	可通过以下两种方式定义提示语内容。 • 自定义提示语内容： 插值参数通过所选的变量标识符来填写定义，支持英文、数字、下划线（_）,不能以数字开头。 以变量标识符“双大括号{{}}”为例，提示语中的变量内容则填入双大括号{{}}中。 • 引用模板提示语内容： 单击输入框右侧的“引用模板”选择我创建的、我收藏的或平台预置的提示语模板。
调测模型	将提示语应用于我创建的或平台预置的模型服务中，预览推理结果。

单击“更多参数配置”，可配置调测模型的相关参数，如表7-4所示。

表 7-4 调测模型更多参数配置说明

参数名称	参数说明
输入加输出最大token数	简称max_length，表示模型输入+输出的最大长度。
输出最大token数	简写为max_new_tokens，表示模型输出的最大长度（即max_length减去输入的那部分），设置了该参数就不用再设置“输入和输出最大token数”（max_length）。
重复惩罚	简写为repetition_penalty，使用通过对已生成的token增加惩罚，减少重复生成的现象，值越大表示惩罚越大。
多样性	简写为top_p，影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”（temperature）只设置1个。
温度	简写为temperature，较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”（top_p）只设置1个。

步骤5 单击“获取推理结果”，可查看提示语应用于调测模型的测试结果。

针对推理结果，用户可通过以下操作对提示语进行结构、排版、内容等维度进行优化和改进。

- 单击“执行优化”，系统将对提示语模板进行首次优化。
- 单击“重新优化”，系统将对提示语模板进行多轮优化。

步骤6 提示语内容优化达到需要结果后，单击“采纳”可将最终优化的提示语内容一键覆盖至提示语内容中；单击“复制”可复制最终优化的提示语内容，用户可自行根据需要使用。

步骤7 单击“保存”，创建提示语完成，在“我创建的”页面的提示语列表中可看到新建的提示语模板。

----结束

更多操作

创建提示语完成后，可执行如下**表7-5**所示的操作。

表 7-5 更多操作

操作	说明
修改提示语	1. 在“我的提示语 > 我创建的”页面的提示语列表中，单击“操作”列“修改”。 2. 参照**表7-2**，修改提示语的基础配置参数。
优化提示语	1. 在“我的提示语 > 我创建的”页面的提示语列表中，单击“操作”列“优化”。 2. 参照**表9-5**，配置提示语的调优参数。

操作	说明
删除提示语	• 单个删除：在“我的提示语 > 我创建的”页面的提示语列表中，单击提示语所在行的“操作”列的“ 删除”，单击“确认”。 • 批量删除：在“我的提示语 > 我创建的”页面的提示语列表中，勾选需要删除的提示语，单击“批量删除”，单击“确认”。

7.3 创建及管理模型服务

模型需要部署成功后才可正式提供模型服务，平台支持将微调后的模型、系统预置的模型以及通过自建模型服务接入的模型发布为模型服务，生成的模型服务可用于创建应用或调测模型。

前提条件

- 已购买推理单元资源，具体购买方法请参见**购买AI原生应用引擎包年包月资源**。
- 由于在线运行需消耗资源，请确保账户有可用资源，且用户费用状态正常。

创建模型服务

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的模型服务”。

步骤2 在“我的模型服务”页面右上角单击“创建模型服务”。

步骤3 在“创建模型服务”页面，配置基础信息参数，包括服务名称、服务描述（可选）。

- 服务名称：自定义模型名称，支持中文、英文、数字、下划线(_)，2-20个字符以内，不能以下划线为开头。
- 服务描述(可选)：用户自定义的模型服务相关描述。

步骤4 设置模型服务配置信息参数。

- 模型来源选择“微调的模型”或“收藏的模型”时，需配置如下**表7-6**所示参数：

表 7-6 微调的和收藏的模型来源配置信息参数说明

参数名称	参数说明
模型	在下拉列表选择相应来源的具体模型。
实例个数	设置服务部署的实例个数。 不同的模型部署1个实例需要的推理单元个数不同，比如，ChatGLM3-6B需要2个实例。 不同的模型因为模型参数量不同，模型参数量越多，需要消耗的资源越多，因此需要的推理单元个数越多。

参数名称	参数说明
推理单元资源	在下拉列表可以看到已购买的推理单元的可用个数，根据实际情况选择。 如果推理单元个数不足以满足实例个数，则需减少实例个数以使推理单元资源满足需求。 **说明** 在推理单元到期后，部署的模型将被下架，可通过购买推理单元资源恢复。

- 模型来源选择“自建模型服务接入”时，需配置如下表7-7所示参数：

表 7-7 自建模型服务接入的配置信息参数说明

参数名称	参数说明
模型	在下拉列表选择通过自建模型服务接入的具体模型。
模型功能分类	无需配置，默认为“文本对话”。
URL	接入的模型服务的URL。 当前仅支持https协议，例如appstage.huaweicloud.com/v1/xxx
鉴权方式	选择鉴权方式。 – 无鉴权 – API Key – AK/SK
API Key	当“鉴权方式”选择“API Key”时，需配置此参数，且输入的关键信息将进行加密保存，仅用于模型服务的调用。
AK	当“鉴权方式”选择“AK/SK”时，需配置此参数。
SK	当“鉴权方式”选择“AK/SK”时，需配置此参数，且输入的关键信息将进行加密保存，仅用于模型服务的调用。
API接口协议	可选如下协议： – 标准OpenAI协议 – 盘古大模型协议

步骤5 单击“创建”，创建推理服务完成，新创建的服务显示在模型服务列表中。

----结束

管理模型服务

创建推理服务完成后，可执行如下表7-8所示的管理模型服务相关操作。

表 7-8 更多操作

操作	说明
修改模型服务	1. 在“我的模型服务”页面的服务列表中，单击“操作”列“更多 > 修改”。 2. 参照**步骤3**和**步骤4**，修改基础信息和配置信息。
删除模型服务	1. 在“我的模型服务”页面的服务列表中，单击“操作”列“更多 > 删除”。 2. 单击“确认”。
模型调测	只有部署完成且“运行中”状态的模型服务才能进行模型调测。 1. 在“我的模型服务”页面服务列表中，单击“操作”列“模型调测”。 2. 参照**调测模型**的步骤，完成模型测试。
启用模型服务	在“我的模型服务”页面服务列表中，单击“操作”列“启用”。
停用模型服务	在“我的模型服务”页面服务列表中，单击“操作”列“停用”。

7.4 创建及管理知识库

知识库是一个组织、存储及管理知识的系统，包括文档、数据库、图表、表格等多种形式的信息的分类、整理和归纳，可以帮助用户组织和管理大量的信息，以便快速访问和使用，平台为用户提供了创建并管理知识库的能力，且创建的知识库启用后可在**创建及管理应用**时引用。

前提条件

已**创建知识库数据集**。

创建知识库

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的知识库”。

步骤2 在“我的知识库”页面，单击右上角“创建知识库”。

步骤3 在“创建知识库”页面，参照**表7-9**进行基础配置和知识库配置。

表 7-9 知识库参数说明

参数名称		参数说明
基础配置	知识库名称	自定义知识库的名称。名称要求：长度2~20，不能以下划线数字开头，只能由中文、字母、数字、下划线组成。
	知识库描述	知识库的相关信息描述。

参数名称		参数说明
	知识库类型	知识库的类型。当前仅支持clickHouse（关于clickHouse的解释详见**基本概念**）。
	知识库调用接口	知识库的调用接口。
	知识库数据集	单击“请选择知识库数据集”，在“我创建的知识库数据集”面板单目标数据集“操作”列“选择”。
	数据集版本号	选择知识库数据集后，该参数值默认为数据集最新版本号。
调度配置	调度类型	可选如下两种类型： ● 一次性调度 ● 定时调度
	版本模式	可选覆盖模式、多版本模式。
	执行周期	可选周期包括： ● CRON：通过特定的自动化运行命令或脚本指定时间间隔（例如每分钟、每小时、每天等）。 ● 天：每天执行。
	CRON表达式	“执行周期”选择“CRON”时，需配置此参数。 示例：0 0/5 * * * ?
	执行时间	“执行周期”选择“天”时，需配置此参数。 设置每日开始执行的时间。
	立即执行	选择是否立即执行。

步骤4 单击“提交”，保存知识库的参数配置；或单击“提交并启用”，创建知识库完成并启用该知识库。

----结束

管理知识库

创建知识库完成，可执行如下**表7-10**所示的管理知识库相关操作。

表 7-10 管理知识库

操作	说明
查看知识库详情	在知识库列表中单击知识库名称，进入知识库详情页，可查看该知识库数据概况和更新记录。
修改知识库	不能修改已启用的知识库；可先停用知识库后再修改。 1. 在知识库列表中“操作”列单击“修改”。 2. 在“修改知识库”页面，可修改知识库描述。

操作	说明
删除知识库	不能删除已启用的知识库；可先停用知识库后再删除。 1. 在知识库列表中“操作”列单击“删除”。 2. 在“删除知识库”对话框，单击“确认”。
启用知识库	在知识库列表中，对于“已停用”状态的知识库，可在“操作”列单击“启用”将其重新启用，启用后的知识库才可在**创建应用**时引用。
停用知识库	在知识库列表中，对于“已启用”状态的知识库，可在“操作”列单击“停用”将其暂停使用。

7.5 创建及管理应用

用户根据实际业务需要可以通过零代码应用开发或全码化应用开发的方式创建AI应用。如需快速开发AI应用推荐零代码方式，对于编码经验丰富的开发者可选择更加灵活的全码化方式开发AI应用。同时用户还可灵活管理自己创建的、收藏的、订阅的AI应用。

创建应用

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的应用”。

步骤2 在“操作指引”区域根据流程指引可以了解零代码应用开发和全码化应用开发的过程。

- **零代码应用开发**是通过提示语编辑的方式，结合大模型，提供行为说明，引入数据集、工具等能力，完成AI应用开发。
- **全码化应用开发**基于SDK或API的灵活开发模式，支持多范式开发，多能力集成，提供云端一体解决方案。

步骤3 在“应用列表”区域选择“我创建的”页签，根据实际需求选择通过创建应用、下载SDK或导入工具方式创建应用。

- **创建知识问答应用（RAG）**，具体操作如下：
  a. 在“我创建的”页签单击“创建应用”。
  b. 在“选择创建类型”对话框选择“知识问答应用（RAG）”，进入应用配置页面，配置如**表7-11**所示参数。

表 7-11 创建知识问答应用（RAG）参数说明

参数名称		参数说明
基本设置	应用图标	单击 + 选择本地图标文件自定义为应用的图标。
	应用名称	自定义应用的名称。命名要求：长度2~20，不能以下划线、数字开头，只能由中文、字母、数字、下划线组成。

参数名称		参数说明
	适用行业	在下拉列表中选择应用的所属行业，可选的行业包括： ■ 电力 ■ 互联网 ■ 政务 ■ 其它
	标签	标签是用来描述或标记应用的关键词或短语，帮助用户快速地找到相关的应用信息或资源。 在下拉列表选择相应的标签名称。
	应用描述	输入应用的功能或用途等应用相关信息描述。
大模型配置	模型服务	在下拉列表中应用包含的模型服务，可选择我已创建好的模型服务或平台预置的模型服务。
	输入加输出最大token数	简称max_length，表示模型输入+输出的最大长度。
	温度	简称temperature，较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”（top_p）只设置1个。
	多样性	简称top_p，影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”（temperature）只设置1个。
	存在惩罚	简称presence_penalty：介于-2.0和2.0之间的数字。正值会尽量避免重复已经使用过的词语，更倾向于生成新词语。
	频率惩罚	简称frequency_penalty，介于-2.0和2.0之间的数字。正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。
知识库	知识库来源	在下拉列表中选择应用包含的知识库，可选择我已创建好的知识库或平台预置的知识库。

c. 在右侧“应用预览”区域输入问题即可预览应用效果。
d. 单击右上角“API Key”，在“请输入API Key”对话框输入调用凭证**API Key**，后续才能正常调测应用。
e. 单击右上角“保存”。

在“我创建的”页签的应用列表中可查看到新建的应用“状态”为“草稿”，创建应用完成。

- **创建路由Agent应用**，具体操作如下：
  a. 在“我创建的”页签单击“创建应用”。
  b. 在“选择创建类型”对话框选择“路由Agent应用”，进入应用配置页面，配置如表7-12所示参数。

表 7-12 创建路由 Agent 应用参数说明

参数名称		参数说明
基本设置	应用图标	单击 + 选择本地图标文件自定义为应用的图标。
	应用名称	自定义应用的名称。命名要求：长度2~20，不能以下划线、数字开头，只能由中文、字母、数字、下划线组成。
	适用行业	在下拉列表中选择应用的所属行业，可选的行业包括： ■ 电力 ■ 互联网 ■ 政务 ■ 其它
	标签	标签是用来描述或标记应用的关键词或短语，帮助用户快速地找到相关的应用信息或资源。 在下拉列表选择相应的标签名称。
	应用描述	输入应用的功能或用途等应用相关信息描述。
Planner配置	模型服务	在下拉列表选择我创建的模型服务或平台预置的模型服务。
Action配置	选择技能	单击“添加Action”后，在“选择技能”下拉列表选择自己已创建的技能。
	技能描述	输入技能的功能等详细描述，以区分不同技能。

  c. 在右侧“应用预览”区域输入问题即可预览应用效果。
  d. 单击右上角“API Key”，在“请输入API Key”对话框输入调用凭证**AK/SK访问密钥或平台API Key**，才能正常调测应用。
  e. 单击右上角“保存”。
  在“我创建的”页签的应用列表中可查看到新建的应用“状态”为“草稿”，创建应用完成。

- **创建对话型应用**，具体操作如下：
  a. 在“我创建的”页签单击“创建应用”。
  b. 在“选择创建类型”对话框选择“对话型应用”，进入应用配置页面，配置如表7-13所示参数。

表 7-13 创建对话型应用参数说明

参数名称		参数说明
基本设置	应用图标	单击 + 选择本地图标文件自定义为应用的图标。
	应用名称	自定义应用的名称。命名要求：长度2~20，不能以下划线、数字开头，只能由中文、字母、数字、下划线组成。
	适用行业	在下拉列表中选择应用的所属行业，可选的行业包括： ■ 电力 ■ 互联网 ■ 政务 ■ 其它
	标签	标签是用来描述或标记应用的关键词或短语，帮助用户快速地找到相关的应用信息或资源。 在下拉列表选择相应的标签名称。
	应用描述	输入应用的功能或用途等应用相关信息描述等。
提示语配置	提示语内容	根据模板输入提示语内容。
大模型配置	模型服务	在下拉列表中选择应用包含的模型服务，可选择我已创建好的模型服务或平台预置的模型服务。
	输入加输出最大token数	简称max_length，表示模型输入+输出的最大长度。
	温度	简称temperature，较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”（top_p）只设置1个。
	多样性	简称top_p，影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”（temperature）只设置1个。

参数名称		参数说明
	存在惩罚	简称presence_penalty：介于-2.0和2.0之间的数字。正值会尽量避免重复已经使用过的词语，更倾向于生成新词语。
	频率惩罚	简称frequency_penalty，介于-2.0和2.0之间的数字。正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。

c. 在右侧“应用预览”区域输入问题即可预览应用效果。

d. 单击右上角“API Key”，在“请输入API Key”对话框输入调用凭证**API Key**，后续才能正常调测应用。

e. 单击右上角“保存”。

在“我创建的”页签的应用列表中可查看到新建的应用“状态”为“草稿”，创建应用完成。

- **创建流程编排应用**，具体操作如下：

a. 在“应用列表 > 我创建的”页签，单击“创建应用”，在“选择创建类型”对话框选择“流程编排应用”。

b. 参照**表7-14**配置基础设置参数。

表 7-14 创建流程编排应用基础设置参数说明

参数名称	参数说明
应用图标	支持bmp、jpeg、jpg、png、tiff、gif格式的图片，图片小于10MB。
应用名称	设置工具名称。
适用行业	可选如下行业： ■ 电力 ■ 互联网 ■ 政务 ■ 其它
标签	用来描述或标记应用的关键词或短语，帮助用户快速地找到相关的工具信息。
工具描述	输入应用的功能相关描述等。

c. 在“Action流程编排”区域单击“编辑流程编排”。

d. 在“Action流程编排”对话框，在“技能列表”筛选需要的技能，根据界面指引，将所选技能拖拽到右侧步骤下指定框中。

e. 单击“确定”。

- 通过下载SDK进行全码化应用开发：平台面向开发者提供了一套搭建RAG应用的Python SDK，在“应用列表 > 我创建的”页签，单击“下载SDK”即可获取完整的AI原生应用开发套件。
- 导入工具

  a. 在“应用列表 > 我创建的”页签中，单击“导入工具”，在“导入工具”面板，配置如表7-15所示参数。

  表 7-15 导入工具参数说明

参数名称	参数说明
工具图标	支持bmp、jpeg、jpg、png、tiff、gif格式的图片，图片小于10MB。
工具名称	设置工具名称。
适用行业	可选如下行业： ■ 电力 ■ 互联网 ■ 政务 ■ 其它
标签	用来描述或标记工具的关键词或短语，帮助用户快速地找到相关的工具信息。
工具描述	输入工具的功能相关描述等。
编辑API	单击“API”，在“编辑API Key”对话框按照如下步骤配置参数： 1. 在“基本信息”页签，设置接口地址，选择鉴权方式，可选无需鉴权、API key。选择“API key”时，需设置相应鉴权方式的密钥位置、密钥参数名、密钥值。基本信息设置完成后，单击“下一步”。 2. 在“请求参数”页签，单击“添加请求参数”，输入请求参数的名称，参数描述，参数类型、请求方式（BODY或HEADER）、设置是否必填。添加参数信息完成后，单击“下一步”。 3. 在“返回参数”页签，单击“添加返回参数”，输入返回参数的名称，参数描述，参数类型、设置是否必填。添加参数信息完成后，单击“下一步”。 4. 在“API调试”页签，填写请求参数后单击“运行调试”，查看返回结果。 5. 单击“保存”。

  b. 单击“发布”。

----结束

管理我创建的应用

创建应用完成后，可执行如表7-16所示操作。

表 7-16 管理我创建的应用

操作	说明
删除应用	不能删除已发布的应用；可先将应用取消发布再删除。 1. 在“我的应用 > 我创建的”页面的应用列表中的“操作”列，选择“更多 > 删除”。 2. 单击“确认”。
发布应用	发布后的应用才可进行应用体验。 1. 在“我的应用 > 我创建的”页面的应用列表中的“操作”列，单击“发布应用”。 2. 单击“确认”。
取消发布	1. 在“我的应用 > 我创建的”页面的应用列表中的“操作”列，单击“取消发布”。 2. 单击“确认”。

管理我收藏的应用

步骤1 在“我的应用”页面的“应用列表”区域，选择“我收藏的”页签。

步骤2 在收藏的应用列表中，单击应用所在行的“操作”列的“取消收藏”，可从收藏的应用列表中移除我已收藏的应用。

步骤3 单击“操作”列的“体验”，进入“应用体验”页面。

步骤4 在“应用体验”页面，参照表7-17进行相关参数和请求体配置。

表 7-17 应用体验参数配置

参数名称		参数说明
参数配置	API	无需配置，默认为调用应用的URL。
	选择应用部署	无需配置，由系统自动部署生成。
	选择应用	无需配置，默认为当前应用。
	选择接口API	仅体验平台预置的应用时，需要配置此参数。 无需配置，默认为当前应用的接口API。

请求体		输入应用接口中的请求体内容。
		示例如下：
		`{`
		`"query": "请详细说明AppStage平台有哪些大模型",`
		`"file_id": []`
		`}`

步骤5 在“应用体验”页面右侧“API调测”区域，单击  查看调测结果。

说明

对话框中输入API调试语句也可进行调测。

----结束

管理我订阅的应用

我订阅的应用即租户购买的问答AI服务，仅AI原生应用引擎管理员可对其进行初始化配置。

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的应用”。

步骤2 在“我订阅的”应用列表中，单击“操作”列的“初始化配置”。

步骤3 在“配置平台API KEY”对话框，输入API KEY值，单击“确定”。

步骤4 完成初始化配置后，单击“操作”列的“体验”，可体验相关相应的应用。

----结束

7.6 体验应用

应用体验是将用户发布的应用、收藏的平台预置应用以及订阅的应用进行API调测，以对话的形式进行AI应用的测试，发现并解决应用接口上的问题和错误。

前提条件

体验我创建的应用前，需先发布应用，具体操作请参见**管理我创建的应用**中的“发布应用”。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“应用编排中心 > 我的应用”。

步骤2 在“应用列表 > 我创建的”或“应用列表 > 我收藏的”中，单击应用所在行“操作”列“体验”。

步骤3 在“应用体验”页面，参照**表7-18**进行相关参数和请求体配置。

表 7-18 应用体验参数配置

参数名称		参数说明
参数配置	API	无需配置，默认为调用应用的URL。
	选择应用部署	无需配置，由系统自动部署生成。
	选择应用	无需配置，默认为当前应用。

参数名称		参数说明
	选择接口API	仅体验平台预置的应用时，需要配置此参数。 无需配置，默认为当前应用的接口API。

请求体		输入应用接口中的请求体内容。
		示例如下：
		{
		"query": "请详细说明AppStage平台有哪些大模型",
		"file_id": []
		}

步骤4 在“应用体验”页面右侧“API调测”区域，单击  查看调测结果。

说明

对话框中输入API调试语句也可进行调测。

----结束

8 模型中心

8.1 模型中心概述

模型中心是集中管理用户微调后的模型、模型微调流水线（即模型微调任务），以及调测模型。

操作指引

图 8-1 模型中心操作指引

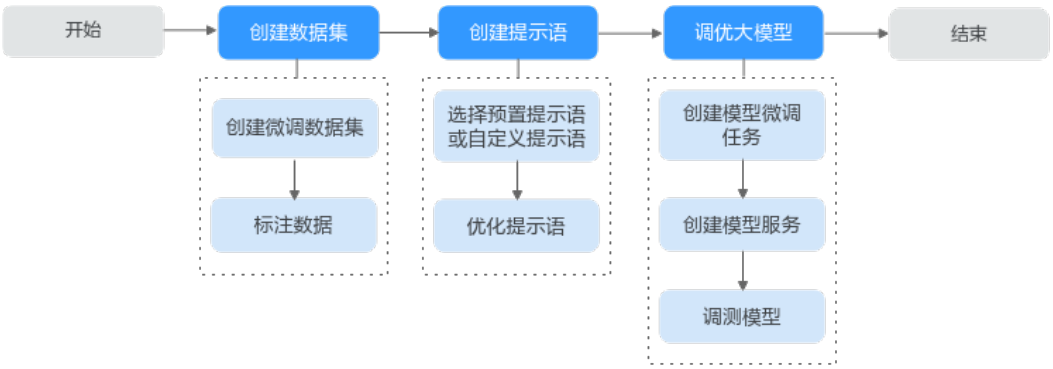

表 8-1 模型中心操作指引详解

序号	流程环节		说明
1	创建数据集	**创建微调数据集**	用户根据需要创建微调数据集，用于模型微调。
		标注数据	用户可以将数据集中的某些元素进行标记或分类，以便模型可以更好地理解和使用这些数据。
2	创建提示语	**选择平台预置提示语或自定义提示语**	用户根据需要选择平台预置的提示语模板或自定义提示语模板，可在**创建应用**、**调测模型**中快速引用。

序号	流程环节		说明
		优化提示语	针对提示语进行结构、排版、内容等维度进行优化和改进，将大模型的输入限定在了一个特定的范围之中，进而更好地控制模型的输出。
3	调优大模型	创建模型微调流水线	通过选择合适的数据集，调整参数，训练平台预置的模型以提高模型效果，可通过训练过程/结果指标初步判断训练效果。
		创建模型服务	模型需要部署成功后才可正式提供模型推理服务，平台支持将微调后的模型、系统预置的模型以及通过自建模型服务接入的模型发布为模型服务。调测模型、应用调用均需先部署模型（即部署模型服务）。
		调测模型	通过调测模型，检验模型的准确性、可靠性及反应效果，发现模型中存在的问题和局限性。

8.2 创建模型微调流水线

模型微调是指调整大型语言模型的参数以适应特定任务的过程，这是通过在与任务相关的数据集上训练模型完成，所需的微调量取决于任务的复杂性和数据集的大小。在深度学习中，微调用于改进预训练模型的性能。

前提条件

- 已订购大模型微调-SFT局部调优资源，订购方法请参见**购买AI原生应用引擎按需计费资源**。
- 已创建同时满足用途为“模型训练”、任务领域为“自然语言处理”、任务子领域为“文本生成”、数据集格式为“对话文本”四个条件的**微调数据集**。

创建微调任务

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型微调流水线”。

步骤2 在“模型微调流水线”页面右上角单击“创建任务”，进入“创建微调任务”页面。

步骤3 参照**表8-2**配置基础信息、模型及数据。

表 8-2 创建微调任务参数说明

参数名称		参数说明
基础信息	任务名称	自定义任务名称。
	任务描述(可选)	自定义任务相关的描述。
模型配置	微调前模型	在下拉列表中选择我创建的或我收藏的模型。

参数名称		参数说明
	训练模式	默认为“LoRA”。 LoRA（Low-Rank Adaptation，低秩适应），是一种将预训练模型权重冻结，并将可训练的秩分解矩阵注入Transformer架构每一层的技术，该技术可减少下游任务的可训练参数数量。
	微调后名称	自定义模型微调后的新名称。命名要求：以字母或下划线开头，由字母、数字、下划线(_)、短横线(-)组成，最多100个字符。
数据配置	数据集	在下拉列表中选择数据集。
	训练数据比例	训练数据比例是指用于训练模型的数据集与测试数据集的比例。通常情况下，会将数据集分成训练集和测试集两部分，其中训练集用于训练模型，测试集用于评估模型的性能。 在实际应用中，训练数据比例的选择取决于许多因素，例如可用数据量、模型复杂度和数据的特征等。通常情况下，会选择较大的训练数据比例，以便训练出更准确的模型。一般来说，训练数据比例在70%到90%之间是比较常见的选择。
	验证数据比例	验证数据比例是指在模型训练过程中，将数据集分为训练集、验证集和测试集三部分，其中验证集的比例是指在训练集和验证集的比例中，验证集所占的比例。 通常情况下，数据集会按照一定比例划分为训练集、验证集和测试集，比如常见的划分比例是60%训练集、20%验证集和20%测试集。在这种情况下，验证集的比例就是20%。 验证集的比例对于机器学习模型的性能评估非常重要。如果验证集的比例过小，可能导致模型在验证集上表现不够稳定，无法准确评估模型的性能。如果验证集的比例过大，可能会导致训练集的样本量不足，影响模型的训练效果。因此，在选择验证集的比例时，需要根据具体情况进行调整，以保证模型的性能评估和训练效果的准确性。
	测试数据比例	测试数据比例是指在模型训练中，将数据集分为训练集和测试集两部分，测试数据比例指测试集占总数据集的比例。 通常，测试数据比例在20%到30%之间较为常见，但具体比例取决于数据集的大小和质量，以及模型的复杂度和训练时间等因素。较小的测试数据比例可能导致过拟合，而过大的比例则可能导致欠拟合。因此，选择适当的测试数据比例对于训练出准确可靠的机器学习模型非常重要。
任务配置	资源池	选择执行任务的资源池，在下拉列表可以看到各资源池的可用卡数，根据实际情况选择。

步骤4 单击“下一步”，分别参照表8-4和表8-4配置基础参数、高阶参数。

表 8-3 基础参数配置说明

参数英文名	参数中文名	参数说明
gradient_accumulation_steps	梯度更新累积步数	使用累积梯度进行模型参数更新时，需要累计的步数。
learning_rate	学习率	学习率是每一次迭代中梯度向损失函数最优解移动的步长。
weight_decay	权重衰减因子	对模型参数进行正则化的一种因子，可以缓解模型过拟合现象。
num_train_epochs	训练epoch数	优化算法在完整训练数据集上的工作轮数。
lr_scheduler_type	学习率调度方法	调整学习率的方法，用于在模型训练时动态调整学习率。
target_modules	LoRA微调层	LoRA微调的layer名关键字。 baichuan系列： down_proj,gate_proj,up_proj,W_pack,o_proj chatglm系列： dense_4h_to_h,dense_h_to_4h,dense,query_key_value
lora_rank	秩	LoRA微调中的秩。
lora_alpha	缩放系数	LoRA微调中的缩放系数。
lora_dropout	遗忘率	LoRA微调中的dropout比例。

表 8-4 高阶参数配置说明

类别	参数英文名	参数中文名	参数说明
训练阶段超参	per_device_train_batch_size	单批次训练数据条数	训练的batch size。
	per_device_eval_batch_size	单批次验证数据条数	验证时的batch size。
	max_steps	训练最大步数	模型训练的最大步数。
	warmup_ratio	学习率热启动比例	学习率热启动参数，一开始以较小的学习率去更新参数，然后再使用预设学习率，有效避免模型震荡。
	warmup_steps	学习率热启动步数	学习率热启动的过程中预设的步数。
	bf16	计算精度	是否开启bf16。

类别	参数英文名	参数中文名	参数说明
	fp16	计算精度	是否开启fp16。
	gradient_checkpointing	梯度存档	是否开启梯度检查点。
	max_seq_length	最大token长度	训练样本最大token长度。
	seed	随机因子	随机种子。
LoRA参数	modules_to_save	全量微调的layer名	全量微调时，模型的layer名称。
验证日志及保存策略配置	evaluation_strategy	验证策略	模型验证策略。
	logging_strategy	日志策略	训练日志打印策略。
	save_strategy	存档策略	保存检查点的策略。
	eval_steps	验证步数	每多少步做一次验证。

步骤5 单击“创建”。新创建的微调任务显示在任务列表中。

----结束

更多操作

创建微调任务完成后，可执行如**表8-5**所示的操作。

表 8-5 更多操作

操作	说明
查看任务详情	在“模型微调流水线”页面的任务列表中，单击任务名称或单击“操作”列“更多 > 运行日志”，查看任务的基础信息、参数信息、运行日志和Loss曲线等详情。
重新创建任务	1. 在“模型微调流水线”页面的任务列表中，单击“操作”列“更多 > 重新创建”。 2. 在“修改微调任务”页面，参照**步骤3**~**步骤4**进行配置。
删除任务	1. 在“模型微调流水线”页面的任务列表中，单击“操作”列“更多 > 删除”。 2. 单击“确认”。
启用任务	1. 在“模型微调流水线”页面的任务列表中，单击“操作”列“启用”。 2. 单击“确认”。

操作	说明
停用任务	1. 在“模型微调流水线”页面的任务列表中，单击“操作”列“停用”。 2. 单击“确认”。
发布任务	运行完成后，点发布完成后，生成更优的新模型，展示在“模型管理 > 我微调的”列表中。

8.3 创建及管理模型

用户通过微调平台预置的模型生成微调的模型后，可进行部署、修改、删除的管理操作；对于我收藏的平台预置模型，也可进行体验、设置鉴权等。

创建模型

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型管理”。

步骤2 在“模型管理”页面，单击“创建模型”。

步骤3 在“创建模型”页面，参照表8-6配置模型相关参数。

表 8-6 模型配置参数说明

参数名称	参数说明
模型名称	支持中英文、数字、中划线(-)、下划线(_)、点(.)，2~64个字符，仅支持中英文开头。
模型类型	可选模型类型包括：文本对话、文本向量化、文本生图、图像理解、语音识别。
模型参数量	模型参数的数量。计量单位可选以下两种： B：表示Billion，即十亿。 M：表示Million，即一百万。
上下文长度	“模型类型”选择“文本对话”时，需配置此参数。 对话文本输入和输出的总长度。
标签（可选）	用来描述或标记模型的关键词或短语，帮助用户快速地找到相关的模型信息或资源。
模型描述（可选）	自定义模型相关描述信息。

步骤4 单击“创建模型”，新创建的模型显示在模型列表中。

----结束

管理我创建的模型

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型管理”。

步骤2 在“模型管理 > 我创建的”页面的模型列表中，可进行如表8-7所示模型管理相关操作。

表 8-7 管理我微调的模型

操作	说明
自建服务接入	1. 在模型列表“操作”列单击“修改”。 2. 在“创建推理服务”页面，配置基础信息参数，包括服务名称、服务描述（可选）。 – 服务名称：自定义模型名称，支持中文、英文、数字、下划线(_)，2-20个字符以内，不能以下划线为开头。 – 服务描述(可选)：用户自定义的服务相关描述。 3. 参照表7-7设置配置信息参数。
修改模型	1. 在模型列表“操作”列单击“修改”。 2. 在“编辑模型”页面，修改模型以下参数配置： – 模型名称：模型的命名，支持中英文、数字、下划线(_)，2-64个字符，不能以下划线为开头，数字开头。 – 模型类型：默认为“大语言模型”，无需配置。 – 模型版本：设置模型版本，格式为：模型名称-参数规模-类型。示例：model-14b-chat – 标签：标签是用来描述或标记模型的关键词或短语，帮助用户快速地找到相关的模型信息或资源。 – 模型描述（可选）：输入模型相关描述信息。 3. 单击“保存”。
删除模型	1. 在模型列表“操作”列单击“删除”。 2. 单击“确认”。
部署模型	1. 在模型列表“操作”列单击“部署”，进入“创建推理服务”页面。 2. 参照创建推理服务的步骤，完成模型服务的创建。

----结束

管理我收藏的模型

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型管理”。

步骤2 在“模型管理 > 我收藏的”页面的模型列表中，可进行如表8-8所示操作。

表 8-8 管理我收藏的模型

操作	说明
部署模型	1. 在模型列表中，鼠标移至模型卡片单击“部署”，进入“创建推理服务”页面。 2. 参照**创建模型服务**完成模型部署。
设置鉴权	在模型列表中，鼠标移至模型卡片单击“设置鉴权”，弹出“设置鉴权信息”对话框。 ● 如果未设置API Key，在“设置鉴权信息”对话框输入API Key，单击“保存”。 ● 如果已设置API Key，在“设置鉴权信息”对话框单击“移除”可清除已存在的API Key；鼠标移至模型卡片再次单击“设置鉴权”，在“设置鉴权信息”对话框重新输入API Key，单击“保存”即可。
调优模型	1. 在模型列表中，鼠标移至模型卡片单击“调优”，进入“创建微调任务”页面。 2. 参照**创建模型微调流水线**进行操作生成调优后的新模型。

----结束

8.4 调测模型

通过调测模型，可检验模型的准确性、可靠性及反应效果，发现模型中存在的问题和局限性，确保模型能够在实际应用中正常运行，并且能够准确地预测和处理数据。

前提矫健

已部署或接入模型服务，具体操作请参见**创建及管理模型服务**。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型调测”。

步骤2 在“模型调测”页面，可调测文本对话类型模型、文本生图类型模型、图像理解类型模型、语音转文本类型模型、文本向量化类型模型。

- **调测文本对话类型模型**，具体操作如下：
  a. 在“模型类型”下选择“文本对话”并配置**表8-9**所示参数。

表 8-9 调测文本对话类型模型参数说明

参数名称	参数说明
模型服务	选择要调测的模型服务，在下拉列表可选我部署的、我接入的、平台预置模型或三方模型服务。

参数名称	参数说明
输出方式	可选非流式、流式。二者区别如下： ■ 非流式：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 ■ 流式：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不用等待大语言模型生成完成。
输入加输出最大token数	简称max_length，表示模型输入+输出的最大长度。
温度	简称temperature，较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”（top_p）只设置1个。
多样性	简称top_p，影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”（temperature）只设置1个。
存在惩罚	简称presence_penalty：介于-2.0和2.0之间的数字。正值会尽量避免重复已经使用过的词语，更倾向于生成新词语。
频率惩罚	简称frequency_penalty，介于-2.0和2.0之间的数字。正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。

b. 在右侧“效果预览”区域，可通过以下两种方式进行模型测试。

- 在对话输入框输入测试语句后按Enter键或单击进行模型测试。
- 单击“引用已有模板”，弹出“选择模板”面板，可通过分类筛选我创建的、我收藏的或平台预置的提示语模板，然后按Enter键或单击进行模型测试。

- **调测文本生图类型模型**，具体操作如下：

a. 在“模型类型”下选择“文本生图”并配置表8-10所示参数。

表 8-10 调测文本生图类型模型参数说明

参数名称	参数说明
模型服务	选择要调测的模型服务，在下拉列表可选我部署的、我接入的、平台预置模型或三方模型服务。
输入内容	输入你希望生成的图片内容。
风格	选择生成的图片风格，可选“自然”或“超自然”。
图片比例	默认1:1，无需配置。

参数名称	参数说明
图片质量	选择标准或高清。
选择图片尺寸	可选512*512、1024*1024。
选择图片数量	设置生成图片的数量，可选数量为1~10。

b. 单击“生成图片”，在右侧“效果预览”区域即可收到生成的图片。

- **调测图像理解类型模型**，具体操作如下：

  a. 在“模型类型”下选择“图像理解”并配置以下参数。

  - 模型服务：选择要调测的模型服务，在下拉列表可选我部署的、我接入的、平台预置模型或三方模型服务。
  - 上传图片：单击，可上传本地图片。
  - 提示语内容：描述需要知道图片中什么信息，例如：图片里有什么?

  b. 单击“生成图像理解”，在右侧“效果预览”区域即可收到信息解答。

- **调测语音转文本类型模型**，具体操作如下：

  a. 在“模型类型”下选择“语音转文本”并配置表8-11所示参数。

表 8-11 调测语音转文本类型模型参数说明

参数名称	参数说明
模型服务	选择要调测的模型服务，在下拉列表可选我部署的、我接入的、平台预置模型或三方模型服务。
上传音频	单击“添加音频”，上传音频文件（只能上传MP3/AAC/WAV文件，且不能超过25MB）。
语言	在下拉列表选择转换成的语言种类“中文”或“英文”，默认为音频文件原语言，可以做语言翻译任务。
输出格式	可选格式包括： ■ json ■ verbose_json
分段粒度	当“输出格式”为“verbose_json”时，需配置此参数。 可选包括： ■ segment：较短的文本片段。 ■ word：单个的中文汉字或英文单词。
温度	temperature：较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。

b. 单击“生成语音转文本”，在右侧“效果预览”区域即可收到生成的文本。

- **调测文本向量化类型模型**，具体操作如下：

  a. 在“模型类型”下选择“文本向量化”并配置以下参数。

     - 模型服务：选择要调测的模型服务，在下拉列表可选我部署的、我接入的、平台预置模型或三方模型服务。
     - 请输入文本，可参照以下示例输入文本。
       - 示例1：那是个快乐的人
       - 示例2： ["那是个快乐的人", "那是个高兴的人", "那是个忧郁的人"]

  b. 单击“生成向量化”，在右侧“效果预览”区域即可收到生成结果。

----结束

8.5 查看模型调用记录

用户可以查看模型的调用记录，包括模型调用唯一ID、调用模型、调用方式、用时及调用时间等信息。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“模型中心 > 模型调用记录”。

步骤2 在“模型调用记录”页面，通过筛选时间范围、状态，或输入模型名称可快速查看模型调用记录信息，如模型调用唯一ID、调用模型、调用方式、用时及调用时间等信息。

----结束

9 知识中心

9.1 知识中心概述

数据是模型训练（含数据标注）以及知识库的基础，在整个模型、知识库中起着至关重要的作用。平台提供统一的数据管理功能，将分散的数据进行集中式纳管，从而节省了数据收集和管理的成本。

知识中心纳管了用户自定义的和平台预置的数据集，可供模型微调、数据标注、创建知识库时快捷使用。

9.2 创建微调数据集

数据集即数据的集合，微调数据集是模型训练的基础。用户可自主创建用于模型训练的数据集。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“知识中心 > 微调数据集”。

步骤2 在“微调数据集”页面，单击右上角“创建微调数据集”。

步骤3 在“创建数据集”页面，参照表9-1进行相关参数的配置。

表 9-1 数据集基础配置参数说明

参数名称		参数说明
基础配置	数据集名称	自定义数据集名称。命名要求：长度2~20，不能以下划线数字开头，只能由中文、字母、数字、下划线组成。
	数据集描述	输入数据集的相关描述。
	标签	在下拉列表选择数据集的分类标识。

参数名称		参数说明
	任务领域	可选择如下： ● 自然语言处理 ● 多模态任务
	任务子领域	“任务领域”选择“自然语言处理”时，需配置此参数。可选以下： ● 文本生成 ● 文本向量化

	数据集格式	“任务子领域”选择“文本生成”时，需配置此参数。支持以下选项：
		● 选择“对话文本”，文件格式建议为json，支持以下5种格式：
		– 1行1条数据，如下所示：
		{"input":"xxx","output":"xxx"}
		– 1行1条数据，结尾带逗号，如下所示：
		{"input":"xxx","output":"xxx"},
		– 1行1个json数组，包含多条数据，如下所示：
		[{"input":"xxx","output":"xxx"},
		{"input":"xxx","output":"xxx"}]
		– 1行1个json数组，包含多条数据，结尾带逗号，如下所示：[{"input":"xxx","output":"xxx"},
		{"input":"xxx","output":"xxx"}],
		– 标准json文件，1个json数组，多行，如下所示：
		[{
		"input": "xxx",
		"output": "xxx"
		},
		{
		"input": "xxx",
		"output": "xxx"
		}]
		● 选择“纯文本”，支持docx、txt 格式；文件大小<=50M，仅支持UTF-8编码。
		● 选择“文生图”，支持以下：
		支持 tar.gz、zip 格式；
		压缩包数量为1，大小 <= 100M。超过100M请先将压缩文件解压后整体上传OBS，通过数据接入创建数据集；
		压缩包内无目录，支持存放 jpg、png、bmp、jpeg 格式的图片；
		压缩包内需包含一个 csv 文件，名称固定为metadata.csv，标题必须为fileName，text；
		如超大文件（大于100M），请先将压缩文件解压后整体上传obs；

参数名称		参数说明
选择数据	数据来源	选择数据集的数据来源。支持以下两种来源： ● 文件上传 ● OBS接入
	数据文件上传	当“数据来源”选择“文件上传”时，需配置此参数。 单击“文件上传”选择本地JSON格式的文件进行上传（仅支持JSON格式）。
	OBS桶名	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的OBS桶名。
	OBS路径	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的具体OBS路径。
调度配置	调度类型	可选如下两种类型，其中本地文件上传仅支持一次性调度，OBS接入支持一次性调度或定时调度两种类型。 ● 一次性调度 ● 定时调度
	版本模式	可选覆盖模式、多版本模式。
	执行周期	可选周期包括： ● CRON：通过特定的自动化运行命令或脚本指定时间间隔（例如每分钟、每小时、每天等）。 ● 天：每天执行。
	CRON表达式	“执行周期”选择“CRON”时，需配置此参数。 示例：0 0/5 * * * ?
	执行时间	“执行周期”选择“天”时，需配置此参数。 设置每日开始执行的时间。
	立即执行	选择是否立即执行。

步骤4 单击“保存”。创建的数据集显示在“我创建的”页签的数据集列表中，创建数据集完成。

----结束

更多操作

创建数据集完成后，可执行如表9-2所示的操作。

表 9-2 更多操作

操作	步骤
修改数据集	1. 在“微调数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选数据集并单击“操作”列的“修改”。 3. 在“修改数据集”页面，仅支持修改数据集描述、修改标签名称。
删除数据集	• 单个删除数据集： 1. 在“我的数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选单个数据集，然后选择“操作”列的“删除”。 3. 单击“确认”。 • 批量删除数据集： 1. 在“我的数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选多个数据集，再单击列表上方“批量删除”。 3. 在“批量删除”对话框，单击“确认”。
标注数据集	**说明** 只有同时满足用途为“模型训练”、任务领域为“自然语言处理”、任务子领域为“文本生成”、数据集格式为“对话文本”四个条件的数据集才可进行标注。 1. 在“微调数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选单个数据集，然后选择“操作”列的“标注”。 3. 进入“数据标注”页面，参照**标注数据**进行数据标注。

9.3 创建知识库数据集

知识库是一个组织、存储及管理知识的系统，包括文档、数据库、图表、表格等多种形式的信息的分类、整理和归纳，可以帮助用户组织和管理大量的信息，以便快速访问和使用。数据集是知识库的组成元素。用户可自主创建用于知识库的数据集。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“知识中心 > 知识库数据集 ”。

步骤2 在“我的数据集”页面，单击右上角“创建知识库数据集”。

步骤3 在“创建数据集”页面，参照**表9-3**进行相关参数的配置。

表 9-3 数据集配置参数说明

参数名称		参数说明
基础配置	数据集名称	自定义数据集名称。命名要求：长度2~20，不能以下划线数字开头，只能由中文、字母、数字、下划线组成。
	数据集描述	输入数据集的相关描述。
	标签	在下拉列表选择数据集的分类标识。
	数据类型	根据实际需要可选以下两种： ● 文档 ● 图片 ● 图片-摘要 ● 视频-摘要 ● 图文PDF
	向量模型服务	可选我创建的模型服务或我收藏平台模型服务。
分段配置	数据分段模式	“知识库类型”选择“文档”时，需配置此参数。在下拉列表可选以下模式： ● 自动切分：按照系统默认预设的规则和分隔符切分。 ● 自定义切分：自定义分段规则，分隔符，以及分段长度等参数。 ● 标题切分：按标题级别分块，分块后的内容按照自定义规则切分(标题切分仅支持docx格式，非docx格式的文件会按照自动切分处理 ）。
	标题层级深度	知识库类型选择“文档”且数据分段模式为“标题切分”，或知识库类型选择“图片”时，需配置此参数。 例如文本包含最多5级标题，选择的标题层级深度为3，则会分别将所有3级标题下的内容合并成文本块，文本块作为一个整体执行后续切分操作。
	标题保存方式	知识库类型选择“文档”且数据分段模式为“标题切分”，或知识库类型选择“图片”时，需配置此参数。 ● 多标题组合：多级标题用特定符号组合：1级标题-2级标题-3级标题-···-文本 ● 最后一级标题：仅组合最后一级标题：最后一级标题-文本
	分段策略	知识库类型选择“文档”且数据分段模式为“自定义切分”、“标题切分”，或知识库类型选择“图片”时，需配置此参数。在下拉列表可选以下策略： ● 递归切分：所选分隔符先后作为优先级顺序，优先高的先切分，切分后大于最大长度的分段再用优先级低的分隔符切分，如此往复。 ● 等价切分：分隔符无优先级，使用所选的所有分隔符切割，合并至分段最大长度。

参数名称		参数说明
	分段分隔符	知识库类型选择“文档”且数据分段模式为“自定义切分”、“标题切分”，或知识库类型选择“图片”时，需配置此参数。 设置用于文本分段的分隔符号。在下拉列表可选以下分隔符号： ● 英文逗号 , ● 中文逗号 ， ● 换行 \n ● 空两行 \n\n ● 空格 ● 英文句号 . ● 中文句号 。 ● 英文问号 ？ ● 中文问号 ？ ● 英文感叹号 ！ ● 中文感叹号 ！
	分段最大长度	知识库类型选择“文档”且数据分段模式为“自定义切分”、“标题切分”，或知识库类型选择“图片”时，需配置此参数。 用于设置文本分段后每段的最大长度。
	分段重叠长度	知识库类型选择“文档”且数据分段模式为“自定义切分”、“标题切分”，或知识库类型选择“图片”时，需配置此参数。 用于设置当前分段开头与上一个分段结尾重叠部分的长度。
选择数据	数据来源	选择数据集的数据来源。支持以下两种来源： ● 文件上传 ● OBS接入
	数据文件上传	当“数据来源”选择“文件上传”时，需配置此参数。 单击“文件上传”选择本地JSON格式的文件进行上传（仅支持JSON格式）。
	OBS桶名	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的OBS桶名。
	OBS路径	当“数据集来源”选择“OBS接入”时，需配置此参数。 在下拉列表中选择数据所在的具体OBS路径。

参数名称		参数说明
调度配置	调度类型	可选如下两种类型，其中本地文件上传仅支持一次性调度，OBS接入支持一次性调度或定时调度两种类型。 ● 一次性调度 ● 定时调度
	版本模式	可选覆盖模式、多版本模式。
	执行周期	可选周期包括： ● CRON：通过特定的自动化运行命令或脚本指定时间间隔（例如每分钟、每小时、每天等）。 ● 天：每天执行。
	CRON表达式	“执行周期”选择“CRON”时，需配置此参数。 示例：0 0/5 * * * ?
	执行时间	“执行周期”选择“天”时，需配置此参数。 设置每日开始执行的时间。
	立即执行	选择是否立即执行。
高级配置（选配）	数据清洗配置	在下拉列表可选以下（支持多选）： ● 删除所有的URL和电子邮件地址 ● 清除连续的空格，换行符和制表符 ● 清除不可见字符 ● 规范化空格 ● 清除乱码 ● 清除网页标识符 ● 清除表情
	向量数据索引文件	根据实际需要可选以下： ● 文件上传 ● 从数据源指定
	文件上传	“向量数据索引文件”选择“文件上传”时，需配置此参数。 单击“文件上传”选择本地文件进行上传。
	索引文件OBS桶名	“向量数据索引文件”选择“从数据源指定”时，需配置此参数。 在下拉列表选择索引文件OBS桶名。
	索引文件OBS地址	“向量数据索引文件”选择“从数据源指定”时，需配置此参数。 在下拉列表选择索引文件OBS地址。

步骤4 单击“保存”。创建的数据集显示在“我创建的”页签的数据集列表中，创建数据集完成。

----结束

更多操作

创建数据集完成后，可执行如表9-4所示的操作。

表 9-4 更多操作

操作	步骤
修改数据集	1. 在“知识库数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选数据集并单击“操作”列的“修改”。 3. 在“修改数据集”页面，仅支持修改数据集描述、修改标签名称。
删除数据集	• 单个删除数据集： 1. 在“知识库数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选单个数据集，然后选择“操作”列的“删除”。 3. 单击“确认”。 • 批量删除数据集： 1. 在“知识库数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选多个数据集，再单击列表上方“批量删除”。 3. 在“批量删除”对话框，单击“确认”。
标注数据集	**说明** 只有同时满足用途为“模型训练”、任务领域为“自然语言处理”、任务子领域为“文本生成”、数据集格式为“对话文本”四个条件的数据集才可进行标注。 1. 在“知识库数据集”页面选择“我创建的”页签。 2. 在数据集列表勾选单个数据集，然后选择“操作”列的“标注”。 3. 进入“数据标注”页面，参照**标注数据**进行数据标注。

9.4 优化提示语

提示语优化是针对提示语进行结构、排版、内容等维度进行优化和改进，将大模型的输入限定在了一个特定的范围之中，进而更好地控制模型的输出。通过提供清晰和具体的指令，引导模型输出并生成高相关、高准确且高质量的文本对答内容，属于自然语言处理领域突破的重要技术，可以提升用户的使用体验和效率，减少用户的困惑和误解。

提示语简介

提示语是给大模型的指令。它可以是一个问题、一段文字描述，也可以是带有一堆参数的文字描述，用于在对话或文章中的一些简短的、不太明确的线索或暗示，推进引

导对话的发展，或者增加故事的复杂性和深度。大模型会基于提示语所提供的信息，生成对应的文本或者图片。

通过对提示语进行结构、内容等维度的优化，将大模型的输入限定在一个特定的范围之中，进而更好地控制模型的输出，它通过提供清晰和具体的指令，引导模型输出生成高相关、高准确且高质量的文本对答内容，属于自然语言处理领域突破的重要部分。

提示语模板

AI资产中心的“提示语模板”页签中预置了多款提示语模板，用户可一键快速复制内容并收藏至自己的提示语管理中，这些模板是基于大量应用场景下的经验或者训练语料而总结出一些优质的提示语组成结构，将其抽离成为一种模板，支持一键快速复制内容、收藏、在线优化等功能。

用户创建的、收藏的以及平台预置的提示语模板都可在**创建及管理应用**、**调测模型**中快速引用。

前提条件

已**创建提示语**。

操作步骤

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“知识中心 > 提示语优化”。

步骤2 在“在线调优”页面，参照**表9-5**进行参数配置。

表 9-5 提示语在线优化参数说明

参数名称	参数说明
变量标识符	可选择以下符号标识提示语内容中的变量。 • 大括号{} • 双大括号{{}} • 双中括号[[]] • 中括号[] • 小括号() • 双小括号(())
提示语内容	可通过以下两种方式定义提示语内容。 • 自定义提示语内容： 插值参数通过所选的变量标识符来填写定义，支持英文、数字、下划线（_）,不能以数字开头。 以变量标识符“双大括号{{}}”为例，提示语中的变量内容则填入双大括号{{}}中。 • 引用模板提示语内容： 单击输入框右侧的“引用模板”选择我创建的、我收藏的或平台预置的提示语模板。

参数名称	参数说明
调测模型	将提示语应用于我创建的或平台预置的模型服务中，预览推理结果。

单击“更多参数配置”，可配置调测模型的相关参数，如表9-6所示。

表 9-6 更多参数配置说明

参数名称	参数说明
输入加输出最大token数	简称max_length，表示模型输入+输出的最大长度。
输出最大token数	简称max_new_tokens，表示模型输出的最大长度（即max_length减去输入的部分），设置该参数后则不需再设置“输入和输出最大token数”（max_length）。
重复惩罚	简称repetition_penalty，使用通过对已生成的token增加惩罚，减少重复生成的现象，值越大表示惩罚越大。
多样性	简称top_p，影响输出文本的多样性，取值越大，生成文本的多样性越强。建议该参数和“温度”（temperature）只设置1个。
温度	简称temperature，较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。建议该参数和“多样性”（top_p）只设置1个。

步骤3 单击“获取推理结果”，可查看提示语应用于调测模型的测试结果。

针对推理结果，用户可通过以下操作对提示语进行结构、排版、内容等维度进行优化和改进。

- 单击“执行优化”，系统将对提示语模板进行首次优化。
- 单击“重新优化”，系统将对提示语模板进行多轮优化。

步骤4 提示语内容优化达到需要结果后，单击“采纳”可将最终优化的提示语内容一键覆盖至提示语内容中；单击“复制”可复制最终优化的提示语内容，用户可自行根据需要使用。

步骤5 提示语内容优化达到期望目标后，单击“采纳”可将最终优化的提示语内容一键覆盖并替代原本的提示语内容；单击“复制”可复制最终优化的提示语内容，用户可自行根据需要使用。

----结束

9.5 标注数据

数据标注是将数据集中的某些元素进行标记或分类，以便模型可以更好地理解和使用这些数据。例如，在自动驾驶的应用中，云数据可以被标注为包含建筑物、其他小物

体、交通工具等信息，以便模型可以识别和理解这些对象。在辅助数据标注的方法中，通过训练模型，可以实现标注结果，从而提高数据的质量和准确性。

前提条件

已创建同时满足用途为“模型训练”、任务领域为“自然语言处理”、任务子领域为“文本生成”、数据集格式为“对话文本”四个条件的数据集才可进行标注。

创建数据标注

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“知识中心 > 数据标注”。

步骤2 在“数据标注”页面，单击右上角“创建数据标注”。

步骤3 在“创建数据标注”对话框，选择数据集。

步骤4 单击“确认”。新创建的标注数据集显示在列表中，创建数据标注完成，默认进入“数据标注”页面，继续执行**标注数据集**。

----结束

标注数据集

步骤1 在“数据标注”页面的标注数据集列表中，单击“操作”列“标注”。

步骤2 在“数据标注”页面，在“数据集文件列表”下拉列表中选择文件。

步骤3 单击“创建对话”顺次生成一条不完整信息（对话样式），用户根据实际需要填写对话的instruction（指令）、input（输入/提问）、output（输出/回答），完成一条数据标注。

对于单条标注，还可执行以下操作：

- 单击标注右侧“一键自动生成”由平台内置的模型一键生成所有行的output信息。
- 单击标注右侧“添加回答”可继续添加多条output。
- 单击标注右侧“删除”，可删除标注。

对于标注中的output，，还可执行以下操作：

- 单击output所在行右侧的“自动生成”，由平台内置的模型自动生成当前行的output信息。
- 单击output所在行右侧的“重新生成”，由平台内置的模型重新生成当前行的output信息。

----结束

更多操作

一条数据标注完成后，可执行如下**表9-7**所示的操作。

表 9-7 更多操作

操作	说明
删除标注	在“数据标注”页面的标注数据集列表中，单击“操作”列“删除”。
发布标注	1. 在“数据标注”页面的标注数据集列表中，单击“操作”列“发布”。 2. 在“发布”对话框，有两种发布方式： ● 选择发布后“更新原始数据集”，单击“确认”，覆盖原数据集信息（数据集名称不变）。 ● 选择发布后“创建新数据集”，设置新数据集名称，然后单击“确认”。

10 修订记录

发布日期	修订记录
2024-05-07	第三次正式发布。 ● 新增**查看模型调用记录** ● 修改如下章节： – **使用流程** – **AI工作空间** – **创建及管理模型服务** – **创建及管理知识库** – **创建及管理应用** – **体验应用** – **创建及管理模型** – **调测模型** – **创建微调数据集** – **创建知识库数据集** ● 删除“接入数据”章节
2024-03-30	第二次正式发布。 同步产品框架变更刷新全文。
2023-02-08	第一次正式发布。