应用平台

AI 原生应用引擎用户指南

文档版本 05

发布日期 2024-07-19

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
1.1 为什么使用 AI 原生应用引擎 1
1.2 AI 原生应用引擎应用场景 1
1.3 AI 原生应用引擎功能介绍 2
1.4 AI 原生应用引擎基本概念 3
2 AI 原生应用引擎使用流程 5
3 进入 AI 原生应用引擎 8
4 管理账号信息 9
5 配置中心 12
5.1 平台租户鉴权 12
5.1.1 创建及管理 AK/SK 访问密钥 12
5.1.2 创建及管理平台 API Key 13
5.2 设置模型鉴权 14
5.3 下载 SDK 开发 Agent 14
6 工作空间 15
7 资产中心 17
8 Agent 编排中心 21
8.1 Agent 编排中心概述 21
8.2 创建提示语 22
8.3 创建模型服务 25
8.3.1 创建部署服务 26
8.3.2 创建接入模型服务 28
8.4 创建及管理知识库 31
8.5 创建及管理工具 37
8.6 创建及管理工作流 39
8.6.1 工作流概述 39
8.6.2 工作流基础节点说明 39
8.6.2.1 LLM 39
8.6.2.2 知识库 40
8.6.2.3 变量 V2 41

8.6.2.4 控制 44
8.6.2.5 Code 代码 53
8.6.2.6 结束 54
8.6.3 创建工作流 54
8.6.4 管理工作流 56
8.7 创建及管理 Agent 58
9 模型中心 63
9.1 模型中心概述 63
9.2 创建模型微调流水线 64
9.3 调测模型 68
9.4 查看模型调用记录 71
10 知识中心 72
10.1 知识中心概述 72
10.2 创建微调数据集 72
10.3 创建知识数据集 74
10.4 优化提示语 78
10.5 标注数据 80
11 修订记录 82

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 AI 原生应用引擎应用场景

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 AI 原生应用引擎功能介绍

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
Agent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。
AI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Agent**时引用。

1.4 AI 原生应用引擎基本概念

使用之前，请先了解表1-2中相关概念，从而更好的使用AI原生应用引擎。

表 1-2 基本概念说明

基本概念	说明
Agent	Agent指具备自主智能的实体，具有一定的智能和自主性，可以自主地发现问题、设定目标、构思策略、执行任务等。
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。
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。

基本概念	说明
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 AI 原生应用引擎使用流程

参考**编排Agent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 Agent 的流程

图 2-1 编排 Agent 的流程

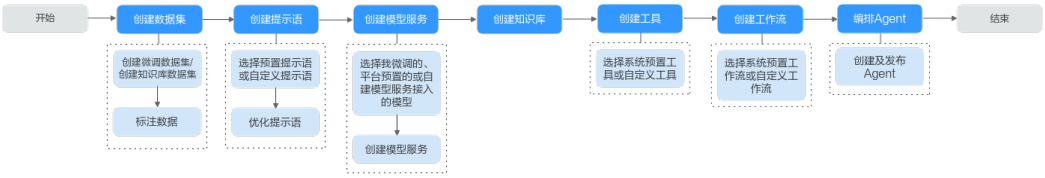


表 2-1 编排 Agent 的流程详解

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Agent**、**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Agent**时引用。

序号	流程环节		说明
5	**创建工具**		用户根据实际需求可自主创建用于实现特定功能的模块或组件或选择系统预置的通用工具。
6	**创建工作流**		用户根据实际需求可自主创建工作流（一系列有序执行的工具，完成复杂任务的过程）或选择系统预置的通用工作流。
5	编排Agent	**创建及发布Agent**	将准备好的模型服务、提示语、知识库等编排Agent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。

调优大模型的流程

图 2-2 调优大模型的流程

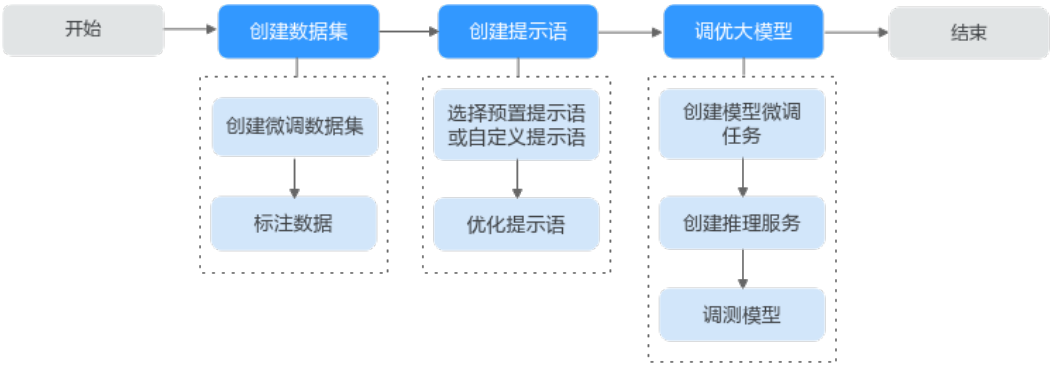


表 2-2 调优大模型的流程详解

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Agent**、**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）。

序号	流程环节		说明
		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 2-3 创建知识库的流程

表 2-3 创建知识库的流程详解

序号	流程环节	说明
1	创建知识数据集	用户根据需要创建知识数据集，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Agent时引用。

3 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

4 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），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。

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“重置账号密码”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“账号与安全”页面。

步骤3 在“安全中心”区域单击“重置账号密码”右侧“重置”。

步骤4 在“重置账号密码”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

5 配置中心

5.1 平台租户鉴权

5.1.1 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

5.1.2 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，单击“确定”。

说明

最多可添加10个平台API Key。

步骤4 在弹出的下载窗口中单击“立即下载”，将API Key下载到本地查看。

须知

下载后，请妥善保管密钥，弹窗关闭后将无法再次获取该密钥信息。

----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

5.2 设置模型鉴权

租户调用第三方模型服务前需设置鉴权，具体鉴权信息则需根据界面提示前往模型供应商官网进行申请。

操作步骤

步骤1 在AI原生应用引擎工作台左侧导航栏选择“配置中心 > 模型鉴权设置”。

步骤2 在“模型供应商列表”页面，单击模型供应商卡片上“设置鉴权”，针对不同的模型服务设置相应鉴权信息。

----结束

5.3 下载 SDK 开发 Agent

平台面向开发者提供了一套搭建Agent应用的Python SDK，对于编码经验丰富的开发者可下载该SDK后通过灵活的全码化方式开发Agent应用。

在AI原生应用引擎工作台左侧导航栏选择“配置中心 > 下载SDK”，可获取完整的AI原生应用开发套件。

6 工作空间

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

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6-1所示，各区域的功能说明如表6-1所述。

图 6-1 工作空间

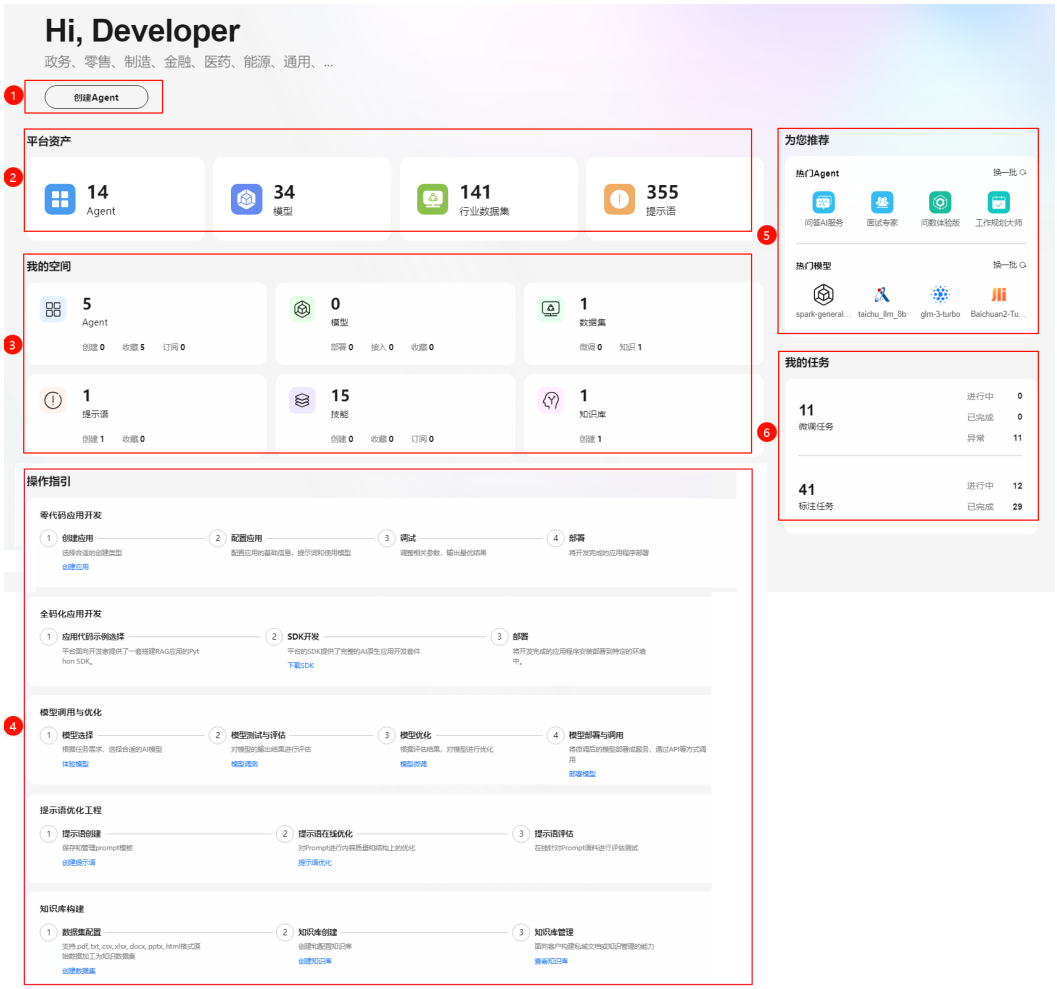


表 6-1 AI 工作空间页面说明

序号	区域	说明
1	创建Agent快捷入口	在该区域单击“创建Agent”可快速进入一站式创建Agent页面，详细介绍请参见**一站式创建Agent**。
2	平台资产	在“平台资产”区域，可查看下述信息数据： ● Agent数据 ● 模型数据 ● 数据集数据 ● 提示语数据
3	我的空间	在“我的空间”区域，可查看下述信息数据： ● Agent：当前账号创建的、收藏的、订阅的Agent个数 ● 模型：当前账号部署的、收藏的、接入的模型个数 ● 数据集：当前账号创建的微调数据集个数、知识库数据集个数 ● 提示语：当前账号创建的、收藏的提示语个数 ● 工具：当前账号创建的、收藏的、订阅的技能个数 ● 知识库：当前账号创建的知识库个数
4	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**、**创建部署服务**。 ● 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**。
5	为您推荐	为您推荐的热门Agent、热门模型。 单击“换一批”可查看更多推荐的热门Agent、热门模型。
6	我的任务	分别展示我创建的模型微调任务、数据标注任务状态统计，包括如下： ● 进行中、已完成、异常状态的**模型微调任务**数 ● 进行中、已完成状态的**标注数据**数

7 资产中心

AI原生应用引擎的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资产中心有不同的类型，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“资产中心”。

步骤2 在“资产中心”页面可依次展开平台预置的AI应用、工具、大模型、数据集、提示语模板页签，可执行如**表7-1**所示的操作。

表 7-1 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Agent**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": [] }` 1. 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。
工具	配置API key	系统预置了一些工具供用户配置“**创建Agent** > 技能”时调用，需对这些技能设置API Key。 鼠标光标移至工具卡片上，单击“配置API Key”。 ● 对于未设置API Key的技能，在“设置鉴权信息”对话框，输入API Key，单击“保存”。 ● 对于已设置API Key的技能，在“设置鉴权信息”对话框，单击“移除”，可重新设置API Key。 **说明** 移除API Key后将影响该技能的调用，需重新设置才能进行调用。
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。

分类	支持的操作	说明
	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**创建部署服务**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。
	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）。

分类	支持的操作	说明
	测试提示语	单击提示语卡片上的“去测试”，进入“模型调测”页面，在**调测文本对话类型模型**时，可引用提示语模板预览效果，具体请参见**引用已有提示语模板**。
	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

8 Agent 编排中心

8.1 Agent 编排中心概述

Agent指具备自主智能的AI实体应用，具有一定的智能和自主性，可以自主地发现问题、设定目标、构思策略、执行任务等。在Agent编排中心，用户可以选择一站式Agent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AI应用开发。

操作指引

图 8-1 Agent 编排中心操作指引

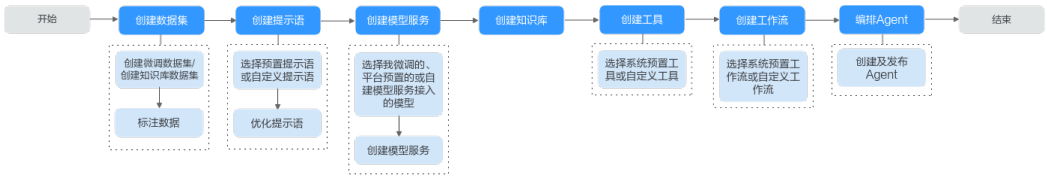


表 8-1 Agent 编排中心操作指引详解

序号	流程环节		说明
1	创建数据集	**创建微调数据集/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Agent**、**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Agent时引用。
5	创建工具		用户根据实际需求可自主创建用于实现特定功能的模块或组件或选择系统预置的通用工具。
6	创建工作流		用户根据实际需求可自主创建工作流（一系列有序执行的工具，完成复杂任务的过程）或选择系统预置的通用工作流。
7	编排Agent	创建及发布Agent	将准备好的模型服务、提示语、知识库等编排Agent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。

8.2 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Agent、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“Agent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8-2进行基础配置后，单击“下一步”。

表 8-2 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： ● 交通 ● 能源 ● 制造 ● 公共事业 ● 金融 ● 互联网 ● 政务 ● 通用行业
适用任务类型	提示语适用的任务类型，包括： ● 对话问答 ● NL2SQL ● 多模生成 ● 任务规划 ● 文案生成 ● 功能调用 ● 代码生成 ● 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）： ● 行业 ● 适用领域 ● 通用
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{{}} ● 中括号[] ● 双中括号[[]] ● 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“在线优化”页面，参照表8-3进行参数配置。

表 8-3 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{{}} • 中括号[] • 双中括号[[]] • 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。
推理模型	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。 选择推理模型后，可配置推理模型的相关参数，如表8-4所示。

表 8-4 推理模型参数配置说明

参数名称	参数说明
最大token数	影响推理返回内容的最大长度，取值范围：1-10000。

参数名称	参数说明
温度	影响结果的随机性，取值越大，随机性越高，取值范围：0-2.0。
多样性	影响结果的多样性，取值越大，结果的多样性越强，取值范围：0-1.0。
存在惩罚	影响结果中词语重复率，取值越大，重复率越高，取值范围：-2.0-2.0

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“创建”，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8-5所示的操作。

表 8-5 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8-2，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8-3，配置提示语的调优参数。
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”。

8.3 创建模型服务

8.3.1 创建部署服务

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。
- 已准备好模型，具体请参见**创建接入模型服务**。

部署模型服务

步骤1 在AI原生应用引擎工作台的左侧导航栏选择“Agent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“部署模型服务”。

步骤3 在“创建部署服务”页面，配置模型信息，参数说明如**表8-6**所示。

表 8-6 模型信息参数说明

参数名称	参数说明
模型来源	• 微调的模型 • 平台模型
选择模型	在下拉列表选择相应来源的具体模型。
服务名称	自定义模型名称，支持中英文、数字、中划线(-)、下划线(_)、点(.)，长度2-64个字符，仅支持以中英文开头。
模型服务描述	用户自定义的模型服务相关描述。
标签	为模型服务选择标签分类。可从以下几个维度选择（支持多选）： • 行业 • 适用领域 • 通用

步骤4 配置部署模型参数，参数说明如**表8-7**所示。

表 8-7 微调的模型部署参数说明

参数名称	参数说明
实例个数	设置模型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	在下拉列表可以查看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。
流控配置	超出流控值，则触发限流，用户的请求会因为流控而失败。 ● 无限制 ● 10次/秒 ● 50次/秒 ● 100次/秒 ● 200次/秒

步骤5 单击“保存”，部署模型服务，新部署的服务显示在“我部署的”页签中。

----结束

管理模型服务

部署模型服务完成后，可执行如下表8-8所示的管理模型服务相关操作。

表 8-8 更多操作

操作	说明
修改模型服务	1. 在“我部署的”页签的服务列表中，单击“操作”列“更多 > 修改”。 2. 参照3和步骤4，修改基础信息和配置信息。
删除模型服务	1. 在“我部署的”页签的服务列表中，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“我部署的”页签服务列表中，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“我部署的”页签服务列表中，单击“操作”列“启用”。
停用模型服务	在“我部署的”页签服务列表中，单击“操作”列“停用”。

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“Agent编排中心 > 我的模型服务”。

步骤2 选择“我收藏的”页签，可进行如**表8-9**所示操作。

表 8-9 管理我收藏的模型

操作	说明
体验模型	将鼠标移至模型卡片单击“体验”，参照**调测模型**进行模型调测。
部署模型	将鼠标移至模型卡片单击“部署”，参照**部署模型服务**完成模型部署。
微调模型	将鼠标移至模型卡片单击“微调”，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

8.3.2 创建接入模型服务

支持自建模型接入并发布为模型服务，生成的模型服务可用于创建应用或调测模型。对于我收藏的平台预置模型，也可进行体验、部署等。

创建接入模型服务

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

步骤2 选择“我接入的”页签，单击“接入模型服务”。

步骤3 在“创建接入模型服务”页面，参照**表8-10**配置模型信息。

表 8-10 模型信息参数说明

参数名称	参数说明
模型名称	自定义模型名称。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。
模型参数量	模型参数的数量。计量单位B，表示Billion，即十亿。
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。
模型描述（可选）	自定义模型相关描述信息。
服务名称	自定义服务名称。支持中英文、数字、中划线(-)、下划线(_)、点(.)，长度2-64个字符，仅支持中英文开头
模型服务描述（可选）	自定义模型服务相关描述信息。

参数名称	参数说明
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。

步骤4 配置模型服务API配置相关参数，参数说明如表8-11所示。

表 8-11 模型服务 API 配置参数说明

参数名称	参数说明
URL(POST)	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	• 无鉴权 • Api-key：Api-key认证方式，通过请求header的Authentication字段携带Bearer <Api-key> 进行认证，需要提供Api-key。 • AK/SK：适用于盘古大模型的AK/SK认证方式，通过AK（ Access Key ID ）/SK（ Secret Access Key)加密调用请求，需要提供AK和SK。
API key	鉴权方式为“Api-key”时，配置此参数。 API密钥所需的字段，以及该验证所必须的字段值。
AK	鉴权方式为“AK/SK”时，配置此参数。 访问密钥Id。
SK	鉴权方式为“AK/SK”时，配置此参数。 密钥。
API接口协议	• 标准OpenAI协议 • 盘古大模型协议
流控配置	超出流控值，则触发限流，用户的请求会因为流控而失败。 • 无限制 • 10次/秒 • 50次/秒 • 100次/秒 • 200次/秒

步骤5 单击“保存”，在模型调测区域，参考表8-12调测模型。

表 8-12 模型调测参数说明

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，不需等待大语言模型生成完成。
输出最大token数	简称max_tokens，表示模型输出的最大长度。
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。

步骤6 在右侧“模型效果预览”区域查看效果。

步骤7 单击“发布”，模型服务发布成功。

----结束

管理我接入的模型服务

模型服务发布完成后，可执行如下**表8-13**所示的管理模型服务相关操作。

表 8-13 管理我接入的模型服务

操作	说明
取消发布模型服务	在模型列表“操作”列单击“取消发布”。
模型调测	1. 在“我接入的”页签服务列表中，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。
修改模型服务	在我接入的”页签服务列表“操作”列选择“更多 > 修改”。

操作	说明
删除模型服务	1. 在模型列表“操作”列选择“更多 > 删除”。 2. 单击“确认”。

8.4 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Agent**时引用。

前提条件

- 已**创建知识数据集**。
- 通过OBS接入数据时，需**同意服务授权**以获得OBS（对象存储服务）只读权限。

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“Agent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8-14**进行基础配置和知识库配置。

表 8-14 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2~50，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。
	知识数据集	单击“请选择知识数据集”，在“我创建的知识库数据集”面板单击目标数据集“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。
刷新配置	刷新类型	可选如下两种类型： ● 一次性 ● 周期性
	刷新时间	刷新类型为“周期性”时，配置此参数。 设置每天刷新的时间。
	立即执行	刷新类型为“周期性”时，配置此参数。 选择是否立即执行。

参数名称		参数说明
索引配置	索引文件	• 无索引：无需配置索引文件。 • 文件级别索引：单击“下载模板”配置索引文件，具体操作请参见**配置索引文件**。 • 切片级别索引：：单击“下载模板”配置索引文件，具体操作请参见。
	数据来源	选择索引文件数据的来源。 索引文件仅支持csv文件，且编码为UTF-8格式，名称为固定格式：数据集名称+下划线+版本id（例如：name_versionId），其中文件名可通过下载模板处获取。本地文件上传文件最大为100M，OBS接入文件大小最大为500MB。 支持以下两种来源： • 文件上传 • OBS接入
	本地上传	当“数据来源”选择“本地上传”时，需配置此参数。 单击“上传文件”选择本地csv格式的文件进行上传。 仅支持csv文件，且编码为UTF-8格式，名称为固定格式：数据集名称+下划线+版本id（例如：name_versionId），其中文件名可通过下载模板处获取。本地文件上传文件最大为100M，obs接入文件大小最大为500MB。
	OBS桶名	当“数据来源”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。

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

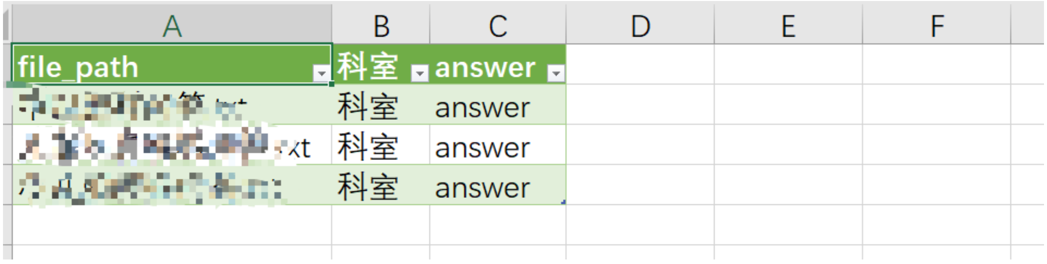
配置索引文件

下载模板时，可以选择添加索引列，模板会自动生成对应列，并填充空内容。其中文件级别索引列系统自带file_path；切片级别索引列系统自带file_path、segment_order、document。

步骤1 正常csv文件打开如**图8-2**示例：

- 文件级别：其中科室、answer为用户自定义索引列，file_path为模板自带索引列，不允许更改。

图 8-2 csv 文件（文件级别）



- 切片级别：其中科室、answer为用户自定义索引列，file_path、segment_order、document为模板自带索引列，不允许更改。

图 8-3 csv 文件（切片级别）

步骤2 如果使用notepad++或记事本打开，或使用代码生成，请查看以下事项：

- 请注意，csv文件使用竖线分隔，因此**文件索引内容请不要带有竖线**，以免程序解析有误。
- 如果内容需换行，请将索引列对应的内容**用英文双引号包围**，且内容中不要存在英文双引号，以免程序校验时报错。
- 请注意，平台支持csv文件有**固定命名规则**，且编码为**UTF-8**格式，请下载模板，以免程序校验报错。
- 索引文件列不应以ki_、 ko_开头或包含平台固定列：file_name, file_id, path, order, document, base64, segment_order。
- csv文件索引列及其内容请一一对应，若平台报错为：文件确少部分列，则请查看文件每一行数据是否有换行，若有换行，确定是否使用英文双引号包围，且英文双引号内部内容不应有英文双引号。

步骤3 如果使用excel打开模板时，可能会显示如**图8-4**所示乱码，则需在菜单栏选择“数据 > 从文本/CSV ”打开文件，如**图8-5**所示。

图 8-4 模板乱码

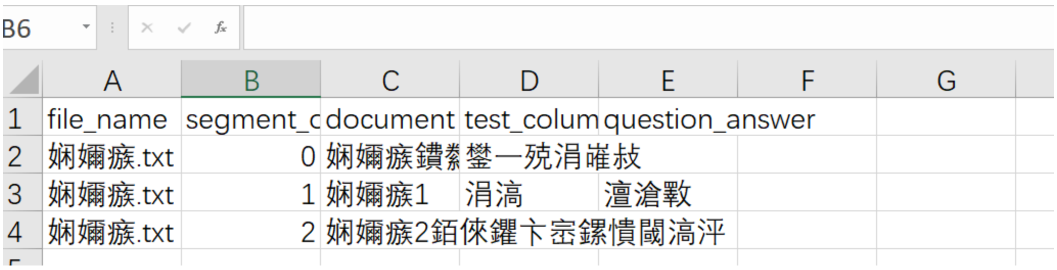
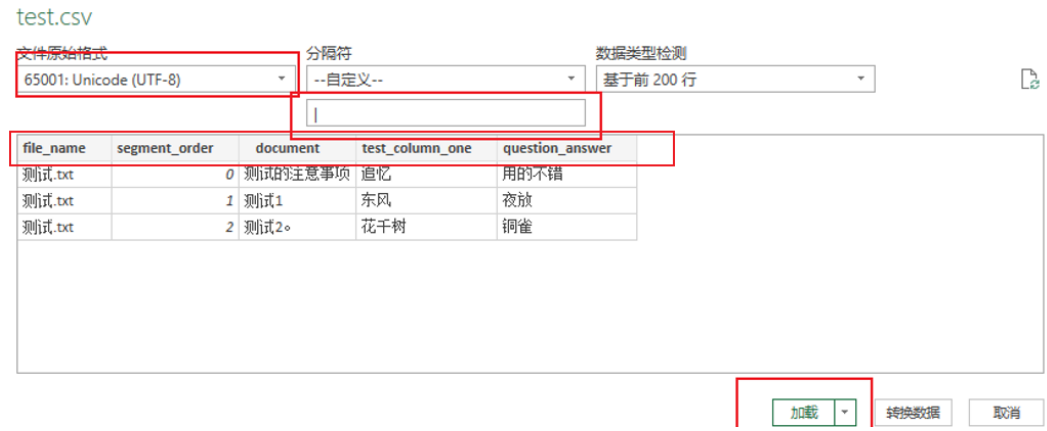


图 8-5 从文本/CSV 打开文件

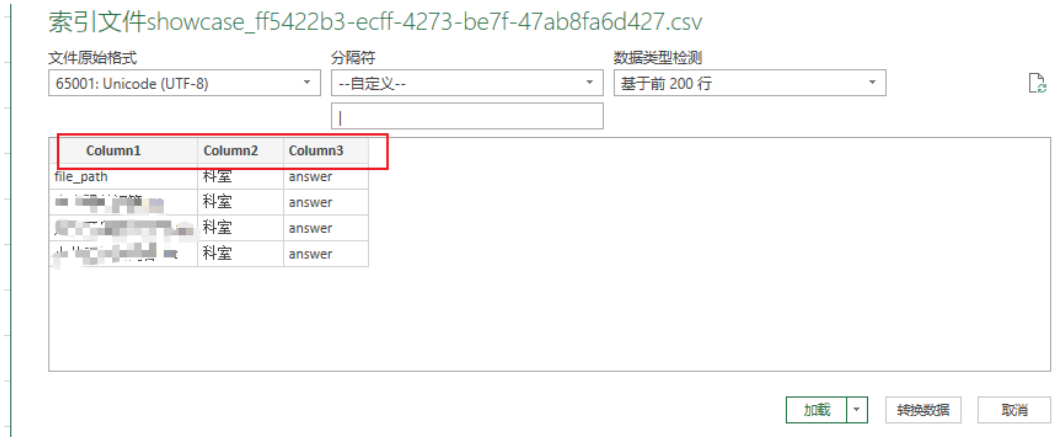
打开文件时，文件头必须如图8-6所示。

图 8-6 正常文件头



如果出现如图8-7所示情况，请参照步骤1-3修复文件。

图 8-7 文件头异常情况

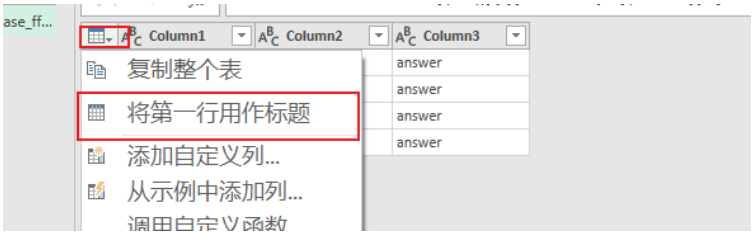


1. 选择编码格式后，单击“转换数据”，如图8-8所示。

图 8-8 转换数据

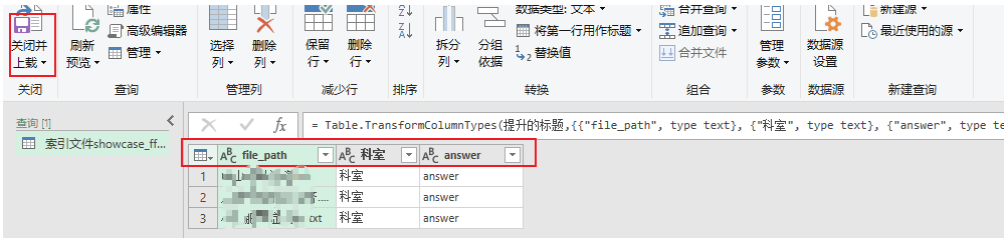
2. 选择“将第一行用作标题”，如图8-9所示。

图 8-9 将第一行用作标题



3. 单击左上角“关闭并上载”，即可正常打开文件，如图8-10所示。

图 8-10 关闭并上载文件



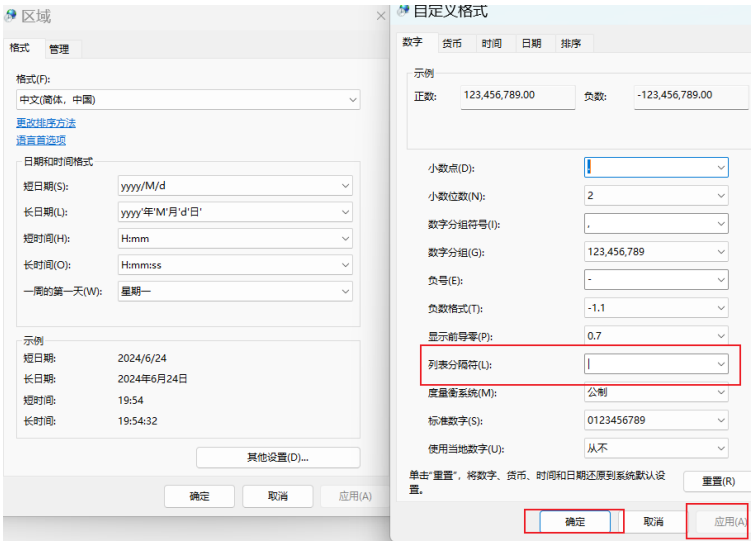
步骤4 完成索引文件修改后，**请将csv文件另存为以竖线“|”分隔的csv文件**，操作步骤如下：

1. 进入“调整计算机的设置”页面，单击“更改日期、时间或数字格式”，如图8-11所示。

图 8-11 调整计算机的设置页面

2. 选择调整系统分隔符为竖线“|”，如图8-12所示。

图 8-12 调整系统分隔符



3. 将修改后的csv文件另存为同名的csv文件，如图8-13所示。

图 8-13 另存 csv 文件

说明

- csv文件使用竖线分隔，因此**文件索引内容请不要带有竖线**，以免程序解析有误。
- 如果内容需换行，请将索引列对应的内容**用英文双引号包围**，且内容中不要存在英文双引号，以免程序校验时报错。
- 请注意，平台支持csv文件有**固定命名规则**，且编码为**UTF-8**格式，请下载模板，以免程序校验报错。
- 索引文件列不应以ki_、 ko_开头或包含平台固定列：file_name、file_id、path、order、document、base64、segment_order。
- csv文件索引列及其内容请一一对应，若平台报错“文件确少部分列”，则需查看文件每一行数据是否有换行，若有换行，确定是否使用英文双引号包围，且英文双引号内部内容不应有英文双引号。

----结束

管理知识库

创建知识库完成后，可执行如下表8-15所示的管理知识库相关操作。

表 8-15 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。

操作	说明
命中测试	命中测试即测试检索的命中率。 1. 在知识库列表中“操作”列单击“命中测试”。 2. 在“命中测试”页面根据界面提示输入测试文本，设置“相似度阈值”（相似度阈值的取值范围[0, 1]，例如配置为0.5，则返回相似度大于等于0.5的结果）、“查询数量”。 3. 单击“测试”。 4. 在页面右侧“测试结果”区域可查看测试效果。相似度越大则表示检索命中率越高。 5. 在页面左侧“测试历史”区域可查看该知识库的测试历史记录，每个知识库测试记录最多保留50条。
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。
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。

8.5 创建及管理工具

系统在资产中心预置了部分工具，同时也支持用户创建需要的工具；这些工具可在用户“**创建Agent** > 技能 > 工具”配置时调用，可对这些工具配置API Key。

创建工具

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

步骤2 在“我的工具”页面，单击“创建工具”。

步骤3 在“创建工具”页面，配置如**表8-16**所示参数。

表 8-16 工具参数配置说明

参数名称	参数说明
名称	自定义工具名称。命名要求：名称长度不能超过32字符，可包含中文、大小写字母、数字及下划线、中划线、英文小括号、开头不能是下划线、中划线、英文小括号。
描述	填写工具功能或作用等描述。

参数名称	参数说明
图标	支持上传PNG和JPG两种格式，文件不能超过1MB。
协议	• https • http
主机地址	输入主机地址。示例：huaweicloud.com 通过开关  可设置是否支持修改。
基准URL	即Base URL，域名的根路径。
验证方式	• 基本认证：用户在创建连接时提供有效的用户名（Username）和密码（Password）即可，此处无需定义。 • API key：用户在使用连接器前需提供API密钥所需的字段，以及该验证所必须的字段值。 • OAuth 2.0：使用 Oauth 2.0 身份验证框架对服务进行身份验证。 在使用此身份验证类型之前，需要向服务注册应用程序，以便它可以接收用户的访问Token。 • IAM：该认证用于通过用户名/密码的方式来获取IAM用户的Token。华为IAM认证的使用方式参考**获取IAM用户Token**。 • AK/SK：使用访问密钥Id（Ak，Access Key Id）和密钥（Sk，Secret Access Key）对请求进行签名，在请求时将签名信息添加到消息头，从而通过身份验证。用户在创建连接时输入值即可，此处无需定义。 Apig的App认证则需提供AppKey以及AppSecret。 • 自定义：自定义用户在创建连接时的身份验证方式。 • 无验证：用户不需要任何身份验证即可创建与连接器的连接。无验证时，任何用户都可以使用您的连接器。

步骤4 单击“创建”。新创建的工具显示在“我的工具”页面的工具列表中。

----结束

编辑工具

步骤1 在“我的工具”页面的工具列表中，鼠标光标移至工具卡片上，单击“编辑”。

步骤2 在“编辑工具”页面，参照**表8-16**编辑参数信息。

----结束

配置工具的 API Key

在“我的工具”页面的工具列表中，鼠标光标移至技能卡片上，单击“配置API Key”。

- 对于未设置API Key的工具，在“设置鉴权信息”对话框，输入API Key，单击“保存”。

- 对于已设置API Key的工具，在“设置鉴权信息”对话框，单击“移除”，可重新设置API Key。

  说明

  移除API Key后将影响该工具的调用，需重新设置才能进行调用。

8.6 创建及管理工作流

8.6.1 工作流概述

工作流体现的是一个具体的业务场景，通过一系列不同功能节点中的触发事件和执行动作编排而成，且开启流之后，当起始节点发生，可自动执行后续动作。AI原生应用引擎通过将传统工具API和大模型编排在一起实现复杂的工作流。

8.6.2 工作流基础节点说明

8.6.2.1 LLM

LLM（Large Language Model，大语言模型）即大模型，仅包含“chat”一个执行动作。

运行动作

- 输入参数

  用户配置运行动作执行动作，相关参数说明如表8-17所示。

  表 8-17 运行动作属性配置输入参数说明

参数	是否必填项	说明
模型service_key	是	从“Agent编排中心 > 我的模型服务”列表中“service_key”列获取。 对于我收藏的模型及资产中心的模型可以直接使用模型名称，其他的需要使用列表中的service_key。
消息	是	输入对话文本。选择数组类型的节点输出。 单击 可切换为数组样式后，配置对话内容（输入）。

- 输出参数

  该执行动作是根据用户定义的内容输出指定参数。

- 节点实例

  a. 单击“新增实例”在“创建实例”面板，配置表8-18参数信息。

表 8-18 创建实例参数说明

参数名称		参数说明
基本信息	实例名称	必填项，自定义实例名称。
	描述	选填项，输入实例相关描述信息。
验证信息	API Key	必填项，单击“获取API key”跳转至AI原生应用引擎的“配置中心 > 平台租户鉴权 > 平台API Key”获取。

b. 单击“保存”，创建实例成功。

8.6.2.2 知识库

知识库仅包含“查询知识库”一个执行动作。

运行动作

- 输入参数
  用户配置运行动作执行动作，相关参数说明如表8-19所示。

  表 8-19 运行动作属性配置输入参数说明

参数	是否必填项	说明
知识库ID	是	从“知识中心 > 我的知识库”列表中找到需要使用的知识库，复制知识库ID。
similarity_min	否	搜索的关键字和返回的内容的相似度阈值，取值范围是0~1。 示例：如果输入0.5，则返回大于等于0.5。
limit	是	数据条目限制，默认为10条。
keyword	否	搜索的关键字。

- 输出参数
  该执行动作是根据用户定义的内容输出指定参数。
- 节点实例
  a. 单击“新增实例”在“创建实例”面板，配置表8-20参数信息。

  表 8-20 创建实例参数说明

参数名称		参数说明
基本信息	实例名称	必填项，自定义实例名称。
	描述	选填项，输入实例相关描述信息。

参数名称		参数说明
验证信息	API Key	必填项，单击“获取API key”跳转至AI原生应用引擎的“配置中心 > 平台租户鉴权 > 平台API Key”获取。

b. 单击“保存”，创建实例成功。

8.6.2.3 变量 V2

变量定义，变量V2连接器包含“追加到数组变量”、“追加到字符串变量”、“数值递减”、“数值递增”、“变量定义 ”、“变量赋值”六个执行动作。

连接参数

变量连接器无需认证，无连接参数。

追加到数组变量

需要先定义一个数组变量，可将值内填写的数据，以字符串的形式追加到数组变量中。例如，先定义一个变量名为data的变量，类型为数组，值为【“123”】，使用追加到数组变量后，可在下拉框内选择data，传入值456，运行即可获得变量data，类型为数组，值为【“123”，“456”】。

- 输入参数

  用户配置追加到数组变量执行动作，相关参数说明请参考**表8-21**。

  表 8-21 追加到数组变量输入参数说明

参数	说明
变量名	选择参数类型（暂无数据）。
值	设定参数的预设值。

- 输出参数

  该执行动作无输出参数。

追加到字符串变量

需要先定义一个字符串变量，可将值内填写的数据，以字符串的形式追加到字符串变量中。例如，先定义一个变量名为data的变量，类型为字符串，值为Str，使用追加到字符串变量后，可在下拉框内选择data，传入值ing，运行即可获得变量data，类型为字符串，值为String。

- 输入参数

  用户配置追加到字符串变量执行动作，相关参数说明请参考**表2 追加到字符串变量输入参数说明**。

表 8-22 追加到字符串变量输入参数说明

参数	说明
变量名	选择参数类型（暂无数据）。
值	设定参数的预设值。

- 输出参数

  该执行动作无输出参数。

数值递增

需要先定义一个整数变量，可按填写的值进行递增。例如，先定义参数data为整数1，数值递增值为1，递增后可以得到data的值为2，如果放在循环内执行，可以得到递增次数为循环次数的数值data。

- 输入参数

  用户配置数值递增执行动作，相关参数说明请参考**数值递增输入参数说明**。

  表 8-23 数值递增输入参数说明

参数	说明
变量名	选择参数类型（暂无数据）。
值	设定参数的预设值。

- 输出参数

  该执行动作无输出参数。

数值递减

需要先定义一个整数变量，可按填写的值进行递减。例如，先定义参数data为整数10，数值递减值为2，递减后可以得到data的值为8，如果放在循环内执行，可以得到递减次数为循环次数的数值data。

- 输入参数

  用户配置数值递减执行动作，相关参数说明请参考**数值递减输入参数说明**。

  表 8-24 数值递减输入参数说明

参数	说明
变量名	选择参数类型（暂无数据）。
值	设定参数的预设值。

- 输出参数

  该执行动作无输出参数。

变量定义

- 输入参数
  用户配置初始化变量执行动作，相关参数说明请参考**初始化变量输入参数说明**。

  表 8-25 变量定义参数说明

参数	是否必选	说明	示例
变量名	是	用于指定将要命名的变量的名称。	re
类型	是	变量的类型。目前包含字符串、整数、布尔、浮点数、数组、对象。	• 字符串 • 整数 • 布尔 • 浮点数 • 数组 • 对象
值	否	用于指定该变量的值。	• 这是一句话 • 12345 • true • 3.1415 • [1,2,3,4,5] • {"key":"value"}

- 输出参数
  该执行动作无输出参数。

变量赋值

使用变量赋值前需进行变量的定义，即在“初始化变量”动作定义完成后，变量赋值的侧边栏参数“变量名”的下拉列表中才能选取到参数。在变量名的最右侧会展示变量的类型。

- 输入参数
  用户配置变量赋值执行动作，相关参数说明请参考**变量赋值输入参数说明**。

  表 8-26 变量赋值输入参数说明

参数	说明
变量名	选择参数类型（暂无数据）。
值	设定参数的预设值。通过填入值，实现对该参数值的改动。

- 输出参数

  该执行动作无输出参数。

8.6.2.4 控制

控制连接器包含“中断”“条件判断”“继续”“遍历集合元素”“分支”“数据分片”“多分支条件”“终止”“流程块”“循环”“异常监控和处理”执行动作。

连接参数

控制连接器无需认证，无连接参数。

中断

中断（break），设置了中断节点，流运行到中断节点后，不会再往后面执行，并跳出循环。如下图所示，当满足条件进入中断节点后，跳出本次循环，并结束整个循环。

- 输入参数

  该执行动作无输出参数。

- 输出参数

  该执行动作无输出参数。

条件判断

用户选择条件判断后，侧边栏会展示该动作包含的参数，同时画布上会展示两条分支，以下图为例：

用户首先需要填写判断条件的相关参数，包括：

1. 选择满足条件（全部满足/任意一项满足）
2. 输入待判断的参数
3. 选择判断条件（包含、不包含、等于、不等于、大于、大于等于、小于、小于或等于、为空）
4. 输入将要判断的值

如果包含多个判断语句，可以通过单击“添加条件”按钮进行添加。

当参数填写完成后，如果逻辑判断正确，则会走向“如果是”分支，反之则会走“如果不是”分支进行后续操作。

- 输入参数

  用户配置条件判断执行动作，相关参数说明如**表8-27**所示。

表 8-27 条件判断输入参数说明

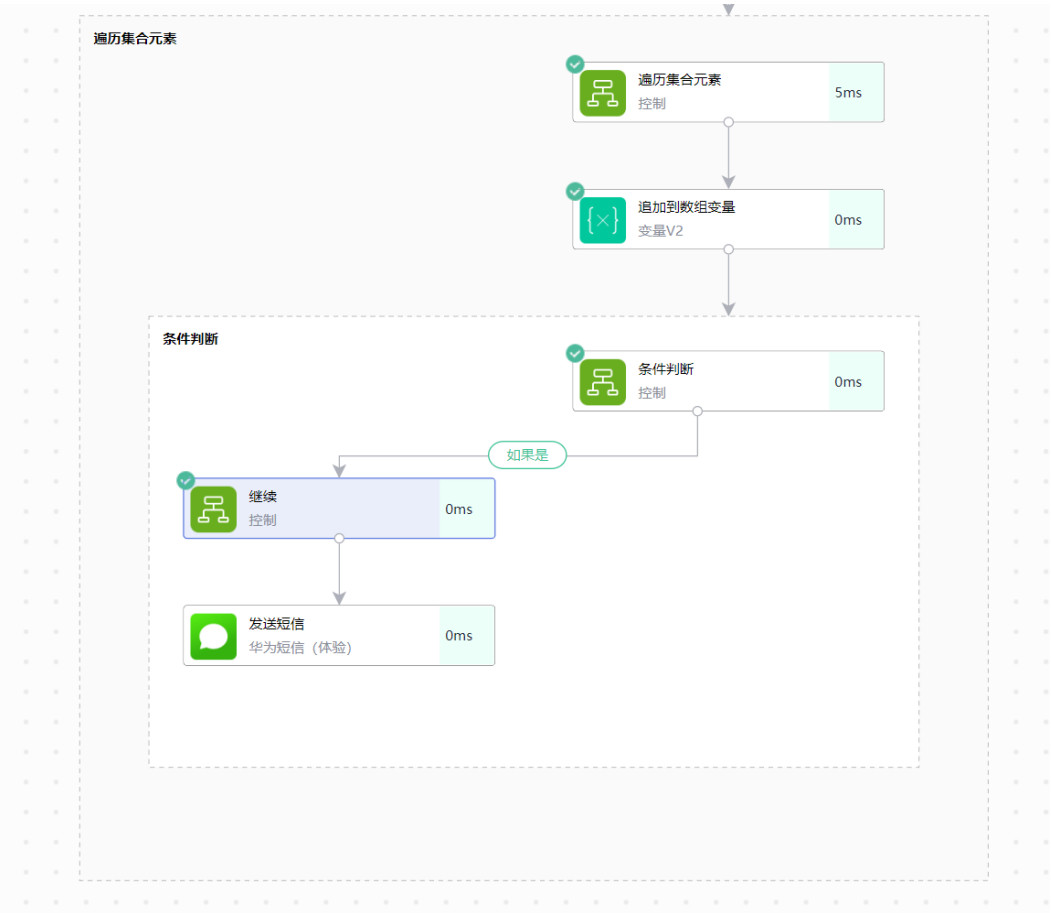
参数	必填	说明
满足条件	是	选择满足条件（全部满足/任意一项满足）。
待判断的参数	是	输入待判断的参数。
判断条件	是	选择判断条件（包含、不包含、等于、不等于、大于、大于等于、小于、小于或等于、为空）。
将要判断的值	是	输入将要判断的值。

- 输出参数

  该执行动作无输出参数。

继续

继续（continue），设置了继续节点，流运行到继续节点后，不会再往后面执行，而是跳出循环进入下一次循环。



- 输入参数

  该执行动作无输出参数。

- 输出参数
  该执行动作无输出参数。

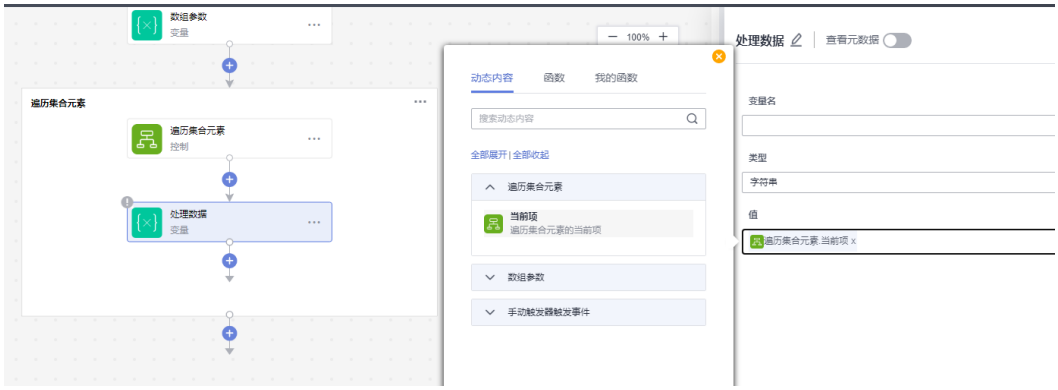
遍历集合元素

说明

添加执行动作时，如果选择了“计划”执行动作，则流编排能映射上遍历集合元素里面需要选到遍历项的子节点。

用户选择该执行动作后，侧边栏弹出输入框：

用户需在输入框内填入数组参数，如[1,2,3]或引用数组参数。



如果想对遍历的当前项进行数据处理，可在上下文引用中获取到当前项。

- 输入参数
  用户配置遍历集合元素执行动作，相关参数说明如表8-28所示。

  表 8-28 遍历集合元素输入参数说明

参数	必须	说明
条件	是	数据集合需要满足的条件。

- 输出参数
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考表8-29。

表 8-29 遍历集合元素输出参数说明

参数	说明
当前项	遍历集合元素的当前项。

分支

用户选择分支后，画布上会展示一个分支的白色框，单击“添加分支”，添加业务场景所需要的各个分支，同时画布上会展示各个分支的信息，最多可以配置五个分支，分支里面可以再添加其他分支。以下图为例：

如果想对分支的当前项进行数据处理，可在分支中进行业务编排。流开始运行时，会先运行分支里面的各个执行动作，之后运行分支以外的执行动作。

- 输入参数

  该执行动作无输入参数。
- 输出参数

  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考**表8-30**。

表 8-30 分支输出参数说明

参数	说明
当前项	分支中的当前项。

数据分片

数据分片执行动作可以将数组类型的变量按指定策略进行分组。例如：输入参数为[1,2,3,4]，按固定数量策略进行分组，期望分片数量为2，那么最终结果为[[1,2],[3,4]]。如果不能整分，则每小组数量为入参数组长度与期望分片数量相除，结果向上取整，例如：输入参数为[1,2,3,4,5]，按固定数量策略进行分组，期望分片数量为2，那么最终结果为[[1,2,3],[4,5]]。

- 输入参数

  用户配置数据分片执行动作，相关参数如**表8-31**所示。

表 8-31 数据分片输入参数说明

参数	必填	说明
分片对象	是	数组类型的自定义变量或之前节点的出参。
分片策略	是	设置分片策略，目前策略仅有“固定数量”一种。
分片数量	是	获得数组数量的期望值。

- 输出参数

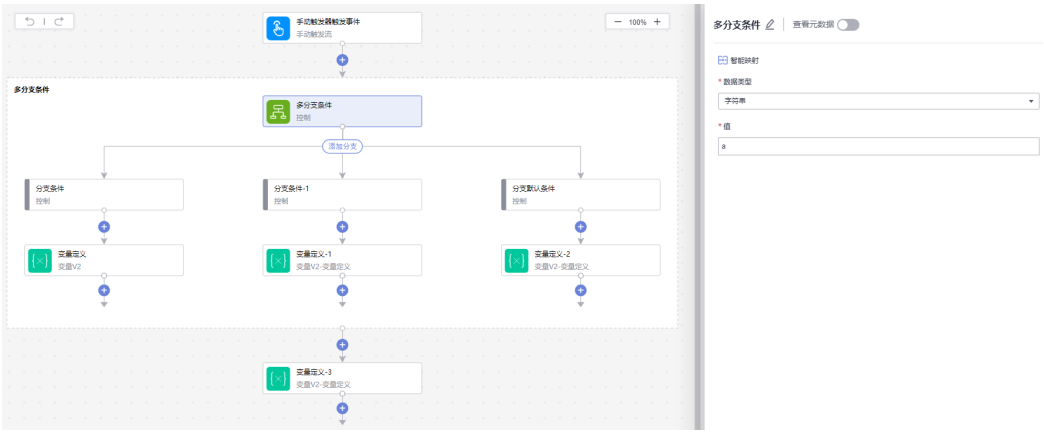
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考表8-32。

表 8-32 数据分片输出参数说明

参数	说明
执行结果	数据分片是否运行成功。 • true：表示成功。 • false：表示失败。
数据分片结果	数据分片后的结果，返回值为二维数组。

多分支条件

用户选择多分支条件后，侧边栏会展示该动作包含的参数，画布上会展示一个多分支条件的白色框，单击“添加分支”，添加业务场景所需要的各个分支条件，同时画布上会展示各个分支条件的信息，分支条件里面可以再添加其他分支。以下图为例：



如果想对多分支条件的当前项进行数据处理，可在多分支条件中进行业务编排。流开始运行时，会先运行多分支条件里面的各个执行动作，之后运行分支以外的执行动作。如果多分支条件运行没有结果，系统默认会按照分支默认条件去执行。

- 输入参数

  用户配置多分支条件执行动作，相关参数说明如表8-33所示。

表 8-33 多分支条件输入参数说明

参数	必填	说明
多分支条件	是	多分支条件需要满足的条件。

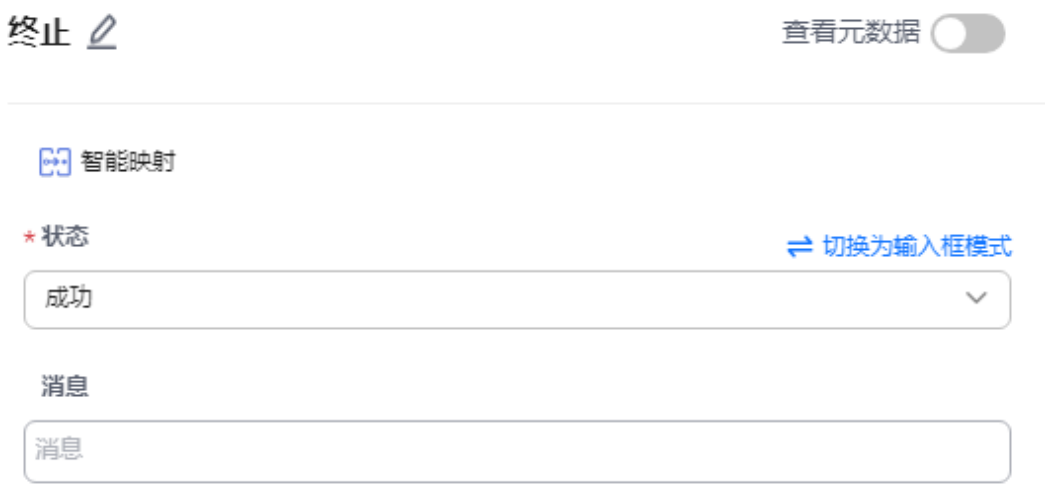
- 输出参数
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考表8-34。

表 8-34 多分支条件输出参数说明

参数	说明
当前项	多分支条件中的当前项。

终止

用户选择该动作后，会弹出侧边栏参数供用户进行填写，包括状态(成功/失败)消息。当流运行到该步骤时，流程运行终止（如果有后续步骤不会再继续执行）。



- 输入参数
  用户配置终止执行动作，相关参数说明如表8-35所示。

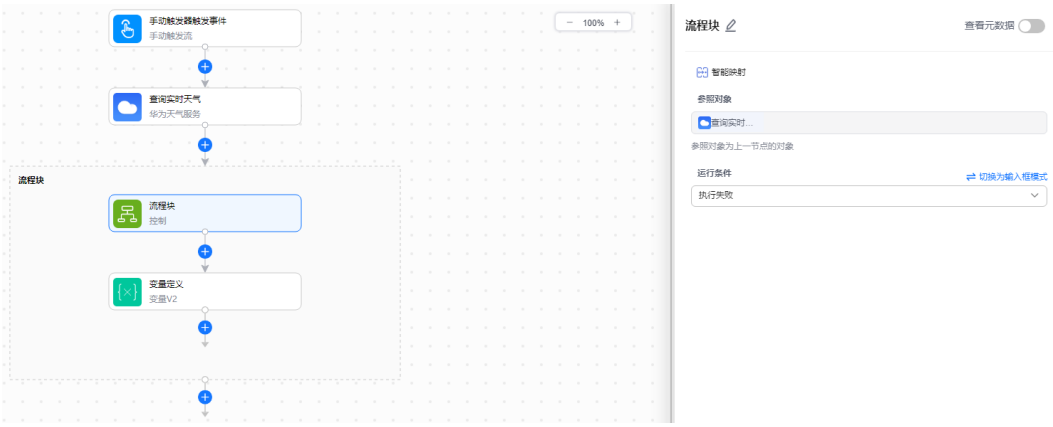
表 8-35 终止输入参数说明

参数	必填	说明
状态	是	选择运行状态，成功或者失败。
消息	否	运行到该步骤时，发送信息。

- 输出参数
  该执行动作无输出参数。

流程块

流程块（flow block），流程块用来监控上个节点的状态，并进入到流程块中执行流程块里面的逻辑。



- 输入参数

  用户配置终止执行动作，相关参数说明如**流程块输入参数说明**所示。

  表 8-36 流程块输入参数说明

参数	必填	说明
参照对象	是	默认上个节点。
运行条件	是	执行成功/执行失败。

- 输出参数

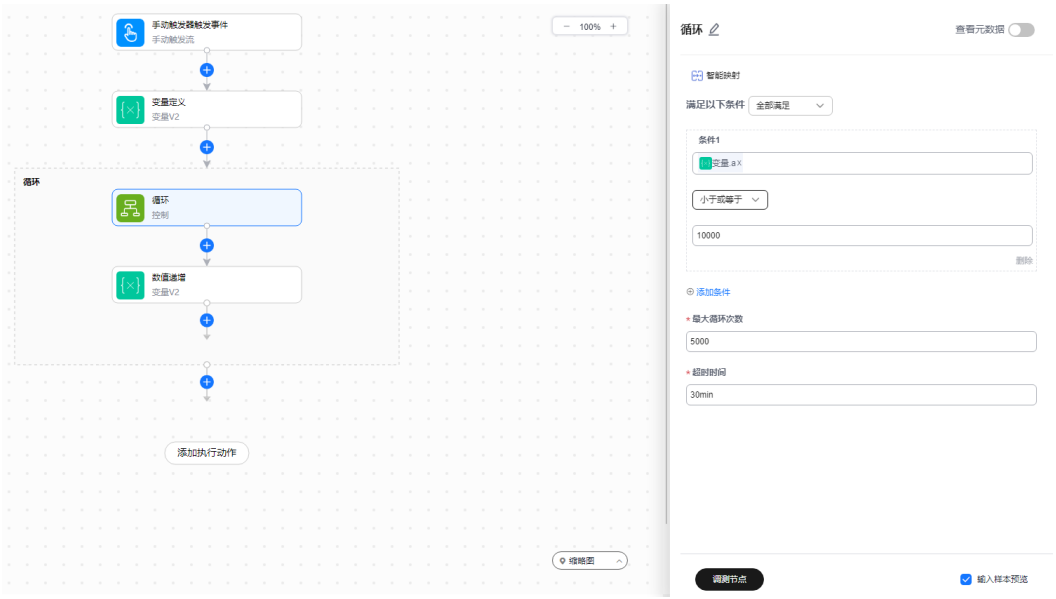
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考**流程块输出参数说明**。

  表 8-37 流程块输出参数说明

参数	说明
是否执行	流程块内步骤是否执行。
异常响应内容	异常响应内容。
异常连接器	异常连接器。
异常执行动作	异常执行动作。

循环

循环（while），当满足条件时，重复执行循环块内的逻辑，直到不满足条件或者超出最大循环次数，或者超出超时时间。



- 输入参数
  用户配置终止执行动作，相关参数说明如**循环输入参数说明**所示。

  表 8-38 循环输入参数说明

参数	说明
条件1	是否循环的条件。
最大循环次数	默认值循环5000。
超时时间	默认值30min。

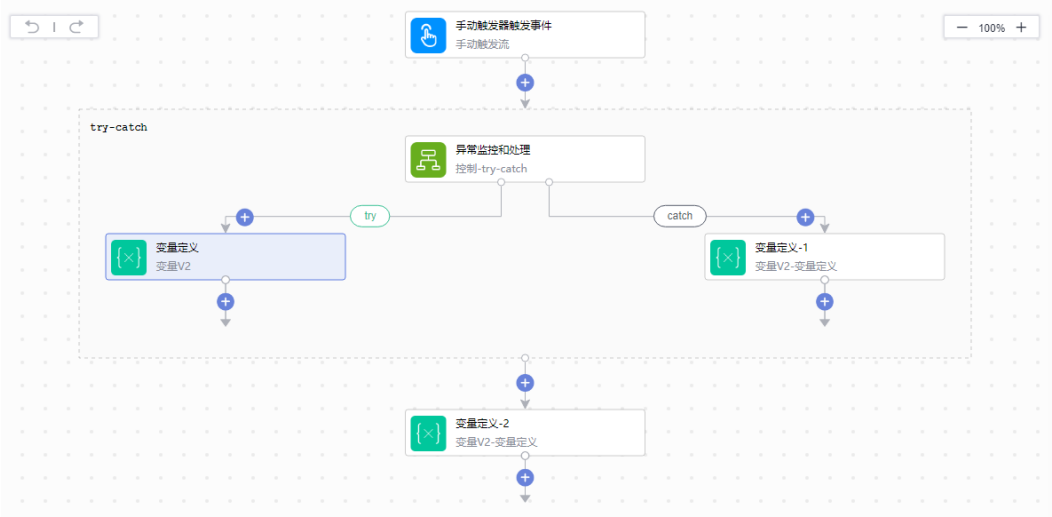
- 输出参数
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考**循环输出参数说明**。

  表 8-39 循环输出参数说明

参数	说明
循环次数	循环次数。

异常监控和处理

异常监控和处理（try-catch），左侧为try，右侧为catch。当左侧try分支出现错误时，会进入右侧catch分支，执行右侧的逻辑，如果左侧try逻辑无错误，则继续向下执行。



- 输入参数
  该执行动作无输出参数。
- 输出参数
  用户可以在之后的执行动作中调用该输出参数，输出参数说明请参考**循环输出参数说明**。

表 8-40 循环输出参数说明

参数	说明
异常响应内容	左侧try分支中异常连接器节点的报错信息。
异常连接器	左侧try分支中异常连接器节点的连接器名称。
异常执行动作	左侧try分支中异常连接器节点的执行动作名称。
响应头	存放以上三个异常信息的集合。

8.6.2.5 Code 代码

Code代码也被称为函数连接器，仅包含“运行动作”一个执行动作。

连接参数

Code代码连接器无需认证，无连接参数。

运行动作

- 输入参数
  用户配置运行动作执行动作，相关参数说明如**表8-41**所示。

表 8-41 运行动作属性配置输入参数说明

参数	必须	说明
函数名称	是	选择下拉列表中的函数，一般是之前已定义保存的函数，也可以进行以下操作。 • 单击⊕：可以直接在弹出的“创建函数”页面快速创建函数，参数配置完成后可单击“创建”保存函数。 • 单击✎：选择函数后，单击该图标可以在弹出的“编辑函数”页面中快速编辑函数，参数编辑完成后可单击“更新”保存函数。 • 单击↻：创建或编辑函数后，单击该图标，可刷新下拉列表中的函数列表。 • 单击⧉：选择函数后，单击该图标可快速复制函数。
连接	否	可选项，选择了连接，在Code代码时可通过连接器认证参数mssiAuthData获取该连接认证信息。

说明

创建流时，当选择该执行动作后，根据选择的函数不同，需要输入的参数不同，具体请以界面为准。

- 输出参数

  该执行动作是根据用户定义的函数内容输出指定参数。

8.6.2.6 结束

此节点将作为整条工作流的输出返回，包含“结束节点”一个执行动作，需配置响应体、状态码、响应头参数。

请在响应体中填入或选择已配节点的输出参数，只接受对象或数组类型，基本类型请组装成对象，例如：{"result":"exampleString"}

8.6.3 创建工作流

根据业务场景需求配置起始节点（相当于触发事件）和选择执行动作创建工作流。开启工作流流之后，当起始节点发生，可自动执行后续动作。

AI原生应用引擎提供的**一站式创建Agent**中配置“技能”参数时可选择“工具”和“工作流”两种类型。

创建工作流

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

步骤2 在“我的工作流”页面，单击“创建工作流”，进入编排工作流页面。

步骤3 在“基本信息”对话框，设置工作流名称、工作流相关描述。

步骤4 单击“起始节点”，配置表8-42所示参数。

表 8-42 起始节点配置参数说明

参数名称		参数说明
API请求方式		在下拉列表中可选择以下API请求方式： • get：get请求，用于从服务器获取数据，通常使用URL参数传递数据。 • post：post请求，用于向服务器提交数据，通常将数据放在请求体中。 • delete：delete请求，用于删除服务器上的资源，通常使用URL参数指定要删除的资源。 • put：put请求，用于更新服务器上的资源，通常将更新的数据放在请求体中。
API请求体架构	请求头	HTTP请求消息的组成部分之一，请求头负责通知服务器有关于客户端请求的信息。 单击“添加header参数”可添加多行请求头；单击即可删除不需要的请求头。
	请求参数	查询参数会追加到URL。例如，在 /items?id=#### 中，查询参数为ID。 单击“添加query参数”可添加多行请求参数；单击即可删除不需要的请求参数。
	请求体	HTTP请求消息的组成部分之一，请求体呈现发送给服务器的数据。

步骤5 添加执行动作。根据业务需求在画布中单击“ > 添加执行动作”或“添加执行动作”，在“选择节点”对话框中选择需要的节点作为执行动作，各类型节点的详细配置说明请参见工作流基础节点说明。

步骤6 添加执行动作完成后，单击页面右上角“保存”。

步骤7 （可选）工作流保存后，在“流保存成功”提示框单击“确定”，可以开启流。

也可以在“我的工作流”页面的工作流卡片上单击“开启”。

开启后，也可以单击“关闭”，关闭工作流。

----结束

更多画布操作说明

在左侧画布中还可以执行的操作如表8-43所示。

表 8-43 执行动作的画布操作

操作	说明
删除	删除该执行动作。 **说明** 工作流编排支持快速删除，鼠标光标移至动作节点上，单击 进行快速删除。
单击画布中的	撤回当前操作，最多可撤回10步。
单击画布中的	恢复撤回操作，最多可恢复10步。

8.6.4 管理工作流

创建AI流后，可以对AI流进行编辑、测试或删除的管理操作。

前提条件

已**创建AI流**。

查看工作流详情

在AI原生应用引擎“Agent编排中心 > 我的工作流”页面的流列表中，单击流名称，进入流详情页面。在流详情页面，可以查看流的节点信息、基本信息、运行历史、历史版本详情。

- 节点信息：在流详情页面左侧，展示最近一次流运行预览图，单击页面右上角“编辑”，可进入流编辑页面。
- 基本信息：在流详情页面右上方，显示流的基本信息，包括名称、状态、最近运行状态等。
- 运行历史：在流详情页面右下方，可以查看近24小时、近7天、近28天的运行历史记录，也可以自定义时间段进行查询。
- 历史版本：在流详情页面右下方，展示流的历史版本。

编辑工作流

步骤1 在“我的工作流”页面的流列表中，鼠标光标移至流卡片上单击“编辑”，进入流编辑页面。

步骤2 参照**步骤4**~**步骤7**编辑工作流的配置。

----结束

测试工作流

说明

当前仅支持post请求调用测试，也可以使用其他调测工具进行调测。

步骤1 在“我的工作流”页面的流列表中，鼠标光标移至流卡片上单击“测试”。

步骤2 在“测试”面板，输入post请求后，单击“提交测试”，示例如图8-14所示。

图 8-14 测试 AI 流

----结束

删除工作流

> **注意**
>
> 删除流将导致流停止运行并删除流相关记录，且操作不可撤消。

步骤1 在“我的工作流”页面的流列表中，鼠标光标移至流卡片上单击“删除”。

步骤2 在删除确认框确认是否删除流并单击“确定”。

----结束

8.7 创建及管理 Agent

用户根据实际业务需要可以选择一站式Agent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AI应用开发。

一站式创建 Agent

一站式创建Agent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。

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

步骤2 单击“创建Agent”，默认弹出“Agent生成”对话框。

可选择以下任一方式生成Agent应用。

- 方式一：由系统智能生成Agent，具体操作如下：
  a. 在**步骤2**中“Agent生成”对话框，根据页面参数配置引导，在“Agent名称”输入框输入想要的Agent名称，在“想要的Agent”输入框中描述想要的Agent的功能或用途等信息。
  b. 单击“生成”，系统将智能生成Agent配置及Agent。
  c. 在“Agent预览”区域单击“开始体验”，然后在对话输入框输入语句后按Enter键或单击  预览Agent效果。
- 方式二：配置Agent相关参数信息，生成Agent，具体操作如下：
  a. 关闭**步骤2**中“Agent生成”对话框，在“创建Agent”页面，参照**表8-44**配置基础信息、选择模型及设定角色。

表 8-44 创建 Agent 参数说明

参数名称	参数说明
基础信息	设置应用名称、应用相关信息说明。 也可以先输入应用功能描述等信息，单击  后智能生成基础信息。

参数名称	参数说明
模型选择	■ 方式一：单击+在弹框中设置如下参数： ○ 模型服务：在下拉列表选择模型服务商PI或预置模型API。 ○ 输出最大token数：简称max_length，表示模型输出的最大长度。 ○ 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。 ■ 方式二：单击 由系统智能生成模型。
角色设定	输入希望角色完成的任务目标、具备的组件能力以及对输出答案的要求与限制等（也可单击“智能修改”由系统智能修改内容）。 示例： #角色设定 作为一个电影剧本创作助手，你的任务是协助编剧创作电影剧本，提供创作灵感和故事构思。 #组件能力 你具备智能生成电影剧本、提供创作灵感和故事构思的能力。 #要求与限制 1. 输出内容的风格要求符合电影剧本的风格，具有吸引力和想象力。 2. 输出结果的格式需按照电影剧本的标准格式进行，包括场景描述、对话、动作等。 3. 输出内容的字数限制不超过5000字。

b. 在“能力拓展”区域，参照**表8-45**可拓展性的添加技能、知识库、开场白以及输入推荐问题。

表 8-45 能力拓展参数说明

参数名称	参数说明
技能	通过添加工具可以使Agent具备更多能力、添加工作流提高任务处理的效率和灵活性： ■ 工具：用于实现特定功能的模块或组件。单击 ••• 在“添加工具”对话框选择我创建的工具或系统预置的通用工具。 ■ 工作流：即一系列有序执行的工具，完成复杂任务的过程。单击 ••• 在“添加工作流”对话框选择我创建的工作流或系统预置的通用工作流。
知识库	单击+，在“添加知识库”对话框中的下拉列表选择现有知识库，单击“确认”。
开场白	可通过两种方式进行设置： ■ 在输入框自定义设置开场白语句。 示例：你好，我是差旅助手！我能为你规划行程、提供实时交通信息，助你出行无忧。请问有什么关于出行的问题我可以帮助你解答? ■ 单击 由系统智能生成开场白语句。
推荐问题	可通过两种方式进行设置： ■ 单击+，在输入框输入推荐的问题语句。 ■ 单击 由系统智能生成推荐的问题语句。

c. 在“Agent预览”区域单击“开始体验”，然后在对话输入框输入语句后按Enter键或单击 预览Agent效果。

----结束

管理我创建的应用

创建应用完成后，可执行如表8-46所示操作。

表 8-46 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Agent > 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”。

操作	说明
发布应用	发布后的应用才可进行应用体验。 1. 在“我的Agent > 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Agent > 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Agent”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8-47进行相关参数和请求体配置。

表 8-47 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“Agent编排中心 > 我的Agent”。

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

9 模型中心

9.1 模型中心概述

用户在模型中心可以集中管理自己创建的模型、微调后的模型、创建的模型微调流水线（即模型微调任务），以及调测模型。

操作指引

图 9-1 模型中心操作指引

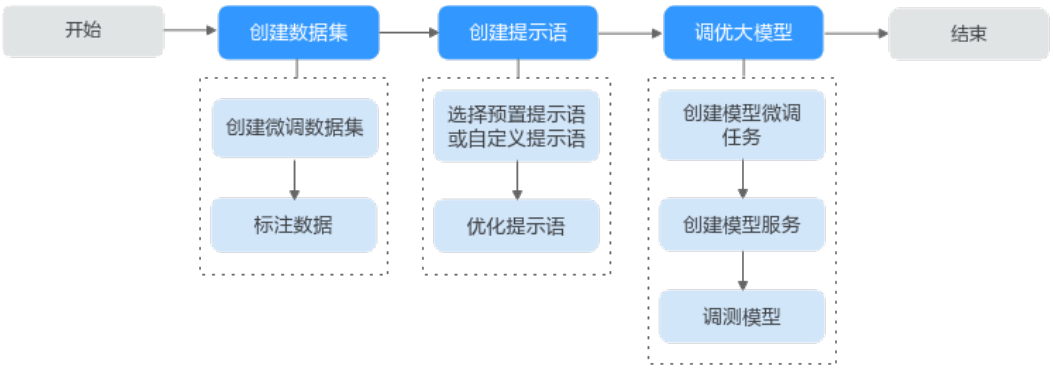


表 9-1 模型中心操作指引详解

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调优模型效果，可通过训练过程/结果指标初步判断调优效果。
		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。

9.2 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，适用于需要个性化定制模型或者在特定任务上追求更高性能表现的场景。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9-2**配置基础信息、模型及数据。

表 9-2 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。支持英文、数字、中划线(-)、下划线(_)，长度1-64个字符，仅支持字母或下划线开头。
	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。支持中英文、数字、中划线(-)、下划线(_)、点(.)，长度2-64个字符，仅支持中英文开头。
数据配置	数据集	在下拉列表中选择数据集。
	数据集版本	在下拉列表中选择数据集版本。
	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9-4和表9-4配置基础参数、高阶参数。

表 9-3 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 9-4 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。

类别	参数英文名	参数中文名	参数说明
	bf16	计算精度	是否开启bf16。
	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9-5**所示的操作。

表 9-5 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。

9.3 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。

前提条件

已部署或接入模型服务，具体操作请参见**创建部署服务**。

调测模型

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以及文本转语言类型模型。

- **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9-6**所示参数。

表 9-6 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。
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，不需等待大语言模型生成完成。
输出最大token数	表示模型输出的最大token数。
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。

参数名称	参数说明
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“引用已有提示语模板”，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9-7所示参数。

表 9-7 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。
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9-8所示参数。

表 9-8 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。

- **调测文本转语音类型模型**，具体操作如下：

a. 在“模型类型”下选择“文本转语音”并配置表9-9所示参数。

表 9-9 调测文本转语音类型模型参数说明

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。
音频格式	MP3
音色类型	在下拉列表选择音色类型，单击“试听”，可以试听音色。
速度	语速，参数范围：0.25-4，该值越大，语速越快。
请输入文本	输入待转为语音的文本。

b. 单击“文本生成语音”，在右侧“效果预览”区域即可收到生成结果。

----结束

9.4 查看模型调用记录

用户可以通过查看模型（包括平台预置模型及自建模型）的调用记录，获取模型调用方式、用时及调用时间等信息。

查看模型调用记录

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，如图9-2所示。

图 9-2 模型调用记录

----结束

10 知识中心

10.1 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。知识中心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。

10.2 创建微调数据集

数据集是数据的集合，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。

前提条件

通过OBS接入数据时，需**同意服务授权**以获得OBS（对象存储服务）只读权限。

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“创建微调数据集”页面，参照**表10-1**进行相关参数的配置。

表 10-1 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。支持中英文、数字、下划线（_），长度2-50个字符，以中英文、数字开头。
	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	无需配置，默认为“自然语言处理”。
	数据集格式	可选以下两种格式： ● 对话文本：文件内容要求为标准json数组，例如：[{"instruction": "aaa", "input": "aaa", "output": "aaa"}, {"instruction": "bbb", "input": "bbb", "output": "bbb"}] ● 纯文本：支持docx、txt 格式；文件大小 <=50M，txt文件仅支持UTF-8编码。
数据接入	数据来源	选择数据集的数据来源。支持以下两种来源： ● 本地上传 ● OBS接入
	本地上传	当“数据来源”选择“本地上传”时，需配置此参数。 单击“上传文件”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	当“调度类型”选择“定时调度”时，需配置此参数。 ● 覆盖模式：每次调度成功，会覆盖唯一的版本。 ● 多版本模式：每次调度成功，会生成一个新版本。
	执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 设置每日执行时间。
	立即执行	当“调度类型”选择“定时调度”时，需配置此参数。 选择是否立即执行。

步骤4 单击“提交”。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，可根据需要执行如表10-2所示的操作。

表 10-2 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。

10.3 创建知识数据集

知识数据集是用于**创建知识库**时使用的数据集，是组成知识库的重要元素。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。

前提条件

通过OBS接入数据时，需**同意服务授权**以获得OBS（对象存储服务）只读权限。

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 > 知识数据集 ”。

步骤2 在“知识数据集”页面，单击右上角“创建知识数据集”。

步骤3 在“创建知识数据集”页面，参照**表10-3**进行相关参数的配置。

表 10-3 创建知识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	自定义知识数据集名称。命名要求：支持中英文、数字、下划线(_)，2~50个字符，以中英文、数字开头。
	数据集描述	输入数据集的功能等相关描述。只能包含英文，中文，数字，下划线，中划线，空格及,.?:;"'：；“”’ ‘，。？ 、()（ ）/
	标签	在下拉列表选择数据集的分类标识。
	知识类型	无需配置，默认为“通用”。
	数据类型	根据实际需要可选以下格式： ● 文档 ● 图片 ● 图片-摘要 ● 视频-摘要
数据接入	数据来源	选择数据集的数据来源。支持以下两种来源： ● 本地上传 ● OBS接入
	数据文件	当“数据来源”选择“本地上传”时，需配置此参数。 单击“文件上传”选择本地文件进行上传。支持.pdf,.txt,.csv（只支持UTF-8）,.xlsx,.docx,.pptx,.html；单个文件最大为100MB；总上传大小最大为200MB。
	OBS桶名	当“数据来源”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。
	调度类型	当“数据来源”选择“OBS接入”时，需配置此参数。 可选如下两种类型： ● 一次性调度 ● 定时调度
	版本模式	当“调度类型”选择“定时调度”时，需配置此参数。 ● 覆盖模式：每次调度成功，会覆盖唯一的版本。 ● 多版本模式：每次调度成功，会生成一个新版本。
	执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 设置每日开始执行的时间。
	立即执行	当“调度类型”选择“定时调度”时，需配置此参数。 选择是否立即执行。

参数名称		参数说明
数据加工	数据清洗（可选）	在下拉列表可选以下（支持多选）： ● 清除URL和邮件地址 ● 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
	富文本处理	数据类型选择为“文档”时，显示此参数。 无需配置，默认为“无处理”。
	数据切分	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。 ● 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）。 ● 自定义切分：自定义分段规则，分隔符，以及分段长度等参数。
	标题层级深度	数据切分模式为“标题切分”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	数据切分模式为“标题切分”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
	文本切分策略	数据切分模式为“自定义切分”、“标题切分”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	数据切分模式为“自定义切分”、“标题切分”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 ！
	分段长度	数据切分模式为“自定义切分”、“标题切分”时，需配置此参数。 用于设置文本分段后每段的长度。
	分段重叠长度	数据切分模式为“自定义切分”、“标题切分”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。
	向量模型	在下拉列表中选择向量模型。

步骤4 单击“提交”。创建的数据集显示在“知识数据集”页面的数据集列表中，创建数据集完成。

----结束

更多操作

创建数据集完成后，可根据需要执行如表10-4所示的操作。

表 10-4 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。

操作	步骤
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”。

10.4 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Agent**、**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“在线优化”页面，参照**表10-5**进行参数配置。

表 10-5 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{{}} • 中括号[] • 双中括号[[]] • 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。
推理模型	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。 选择推理模型后，可配置推理模型的相关参数，如**表10-6**所示。

表 10-6 推理模型参数配置说明

参数名称	参数说明
最大token数	影响推理返回内容的最大长度，取值范围：1-10000。
温度	影响结果的随机性，取值越大，随机性越高，取值范围：0-2.0。
多样性	影响结果的多样性，取值越大，结果的多样性越强，取值范围：0-1.0。
存在惩罚	影响结果中词语重复率，取值越大，重复率越高，取值范围：-2.0-2.0

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。

----结束

10.5 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，选择微调数据集。

步骤4 单击“确定”。新创建的标注数据集显示在列表中，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“标注信息”页面，在“数据集文件列表”下拉列表中选择文件。

步骤3 单击“全部信息”页签下的“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“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。
- 单击output所在行右侧的“删除”，可删除当前行的output信息。

----结束

更多操作

一条数据标注完成后，可执行如下**表10-7**所示的操作。

表 10-7 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“确定”，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“确定”。

11 修订记录

发布日期	修订记录
2024-07-19	第五次正式发布。 修改如下章节： • **资产中心** • **创建及管理知识库** • **创建及管理工具** • **创建工作流** • **创建及管理Agent**
2024-06-20	第四次正式发布。 • 新增如下章节： – **设置模型鉴权** – **下载SDK开发Agent** – **创建及管理工具** – **创建及管理工作流** • 修改如下章节： – **进入AI原生应用引擎** – **工作空间** – **资产中心** – **创建及管理Agent** – **创建微调数据集** – **创建知识数据集**

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： – **AI原生应用引擎使用流程** – **进入AI原生应用引擎** – **创建部署服务** – **创建及管理知识库** – **创建及管理Agent** – **创建接入模型服务** – **调测模型** – **创建微调数据集** – **创建知识数据集** ● 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。