应用平台

AI 原生应用引擎用户指南

文档版本	10
发布日期	2025-02-14

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 什么是 AI 原生应用引擎

AI原生应用引擎是企业专属的一站式大模型开发及应用构建平台。无论是研发技术人员还是业务人员，都可通过简易的界面化操作，快速开发大模型应用或训练专属模型。

AI原生应用引擎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，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。

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。

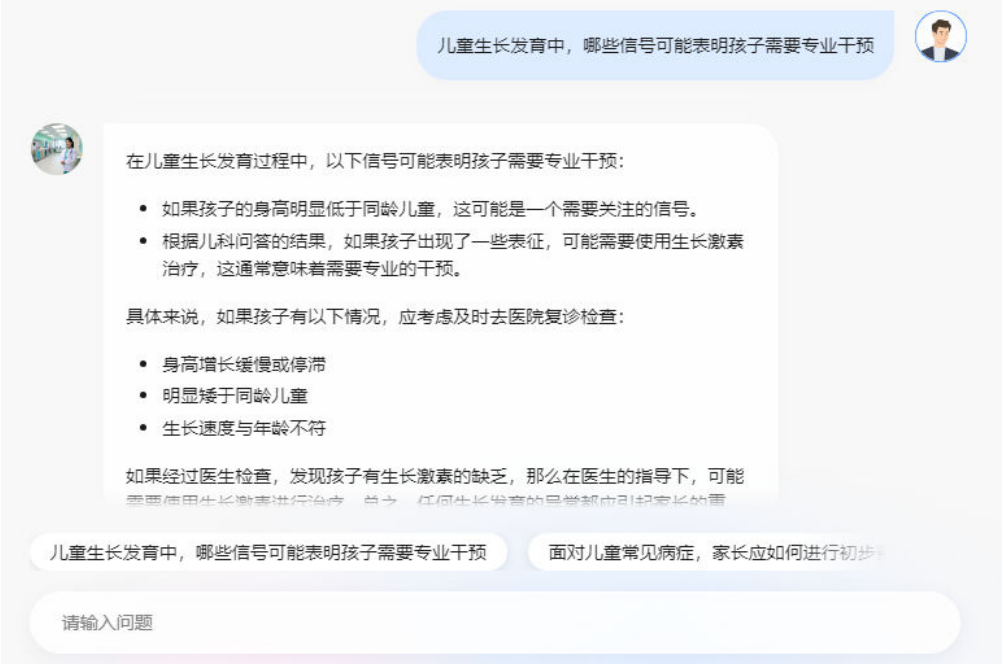
细分领域如金融、电网场景，需要对推理结果进行定制调整，则可在AI原生应用引擎使用模型在线微调训练功能，快速生成行业场景定制模型服务，满足用户特定需求。

- **对话沟通**

  通过对话沟通，快速理解并响应客户的需求，提供高效的解决方案或信息。对于涉及行业和领域的专业知识或技术，平台的知识库能够有效地补充相关知识，确保提供专业性的指导或建议。

  举例：儿科知识问答Agent，不仅可以迅速响应用户问题，还可为患者提供专业且权威的儿科医学知识。

图 1-1 对话沟通

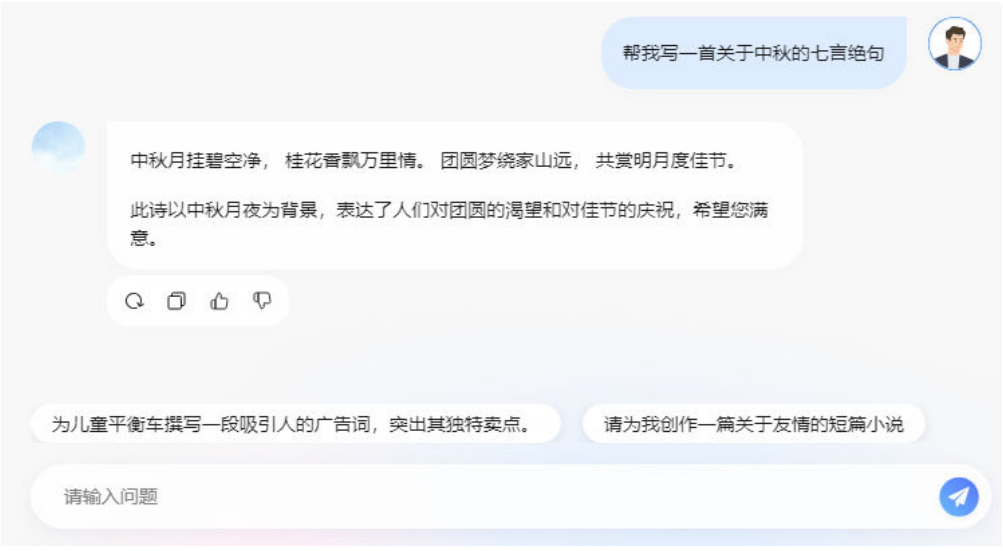


- **内容创作**

  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。

  举例：根据关键词或主题自动生成诗歌、推广文案、报告、故事等。

图 1-2 内容创作

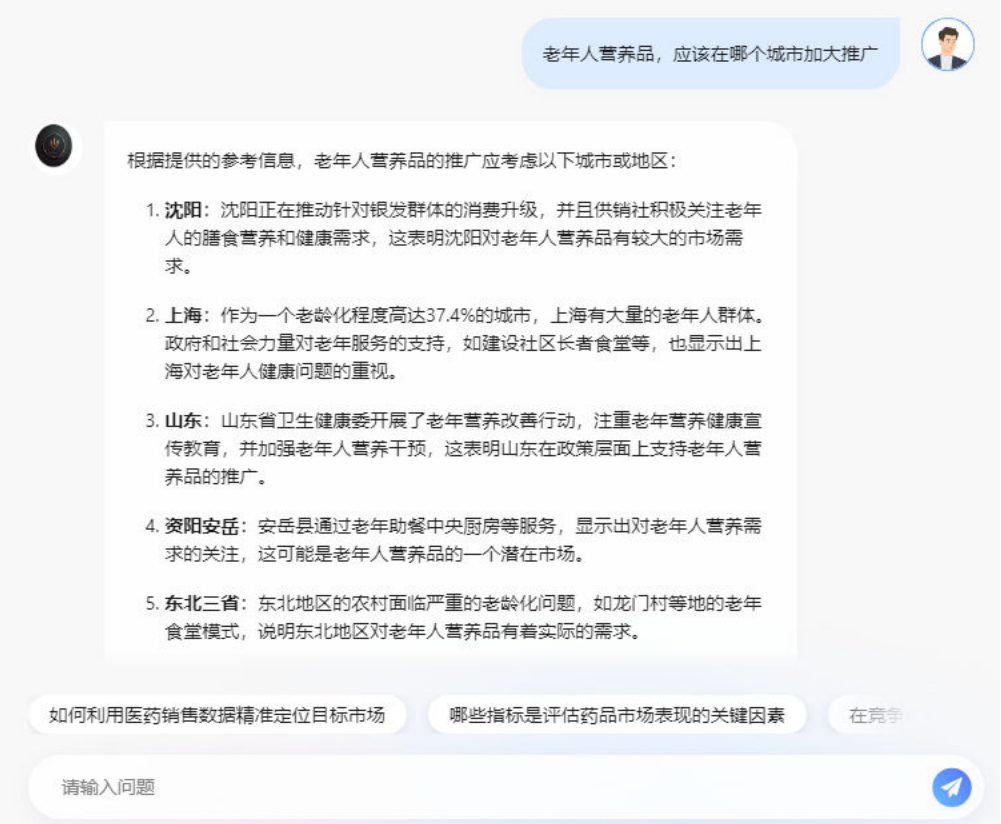


- **分析控制**

  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。

举例：在策划营销活动时，可以帮助营销人员分析市场数据并制定合理的推广方案。

图 1-3 分析控制



为什么使用 AI 原生应用引擎

- **一站式AI原生应用平台**

  平台提供数据准备、模型选择/调优、知识工程、可视化画布流程编排、开箱即用的Prompt模板应用、应用部署及应用集成能力，为企业打造专属的AI原生应用。
- **丰富多样的模型选择**

  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。
- **安全可信**

  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。

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	支持一站式创建专属AI原生应用，对于创建的Agent进行体验调测，并通过API或Web方式发布后即可对外提供服务。 同时，您可以体验平台预置的Agent，享受AI带来的便利和乐趣。
数据管理	除平台预置的数据集外，同时还支持创建知识数据集和微调数据集。丰富的知识数据集及强大的索引配置是构建专业化、结构化知识库的基础；微调数据集是模型微调的基础，通过在微调数据集上进行训练，您可以获得改进后的新模型以适应特定任务。
模型管理	支持通过API接入模型服务，同时支持将平台预置模型进行微调后，部署为模型服务，为检验模型的准确性及反应效果，您可以通过调测模型能力进行体验调测，确保模型能够在实际应用中正常运行。
提示语管理	平台预置了丰富的提示语模板，并支持用户自创建提示语模板。同时，平台提供的提示语优化及推理结果获取等功能，有效地提升了提示语模板的准确性，使得提示语模板更符合情境，引导Agent提供更加精准的回答。
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**创建并发布Agent**时引用。

1.2 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。
LLM	大语言模型（Large Language Model，简称LLM）是通过深度学习技术训练的人工智能模型，具备理解、生成和处理人类语言的能力。
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。

基本概念	说明
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。
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	LoRA（Low-Rank Adaptation）是一种轻量级大模型微调技术，通过低秩矩阵分解技术显著减少了微调所需的参数，降低了微调过程中所需的存储和计算资源，可灵活地运用于不同的预训练模型和任务。

基本概念	说明
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 原生应用引擎使用前准备

使用AI原生应用引擎前，需要先准备如**表2-1**所示内容。

表 2-1 准备事项

准备事项	说明
购买AI原生应用引擎	首次使用需要先购买AI原生应用引擎，具体操作请参见**购买AppStage**。
为AppStage关联组织	首次购买AppStage后，其账号需创建并关联使用AppStage的组织（仅可关联一个组织），才能使用AppStage服务及后续购买AppStage相关产品套餐或增量包等，具体操作请参见**为AppStage关联组织**。
添加部门/成员信息	为已关联的组织添加部门及成员，完善组织架构，具体操作请参见**管理已关联组织的部门及成员**。
录入产品	将企业产品信息录入AppStage系统中，信息录入成功后，AppStage将同步产品信息至AI原生应用引擎，具体操作请参见**在AppStage中管理产品**。
申请权限	已添加成员在使用AI原生应用引擎前需要先申请AI原生应用引擎权限，具体操作请参见**申请权限**。

3 AI 原生应用引擎概览页介绍

进入 AI 原生应用引擎

步骤1 登录AppStage。

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。

----结束

概览页介绍

在AI原生应用引擎的左侧导航栏选择“概览”，进入概览页，可获得系统中各资源数据概览及产品的相关快速指引。

概览页分为数据统计、选择应用创建类型、操作指引三个区域，如图3-1所示，各区域的功能说明如表3-1所述。

图 3-1 概览

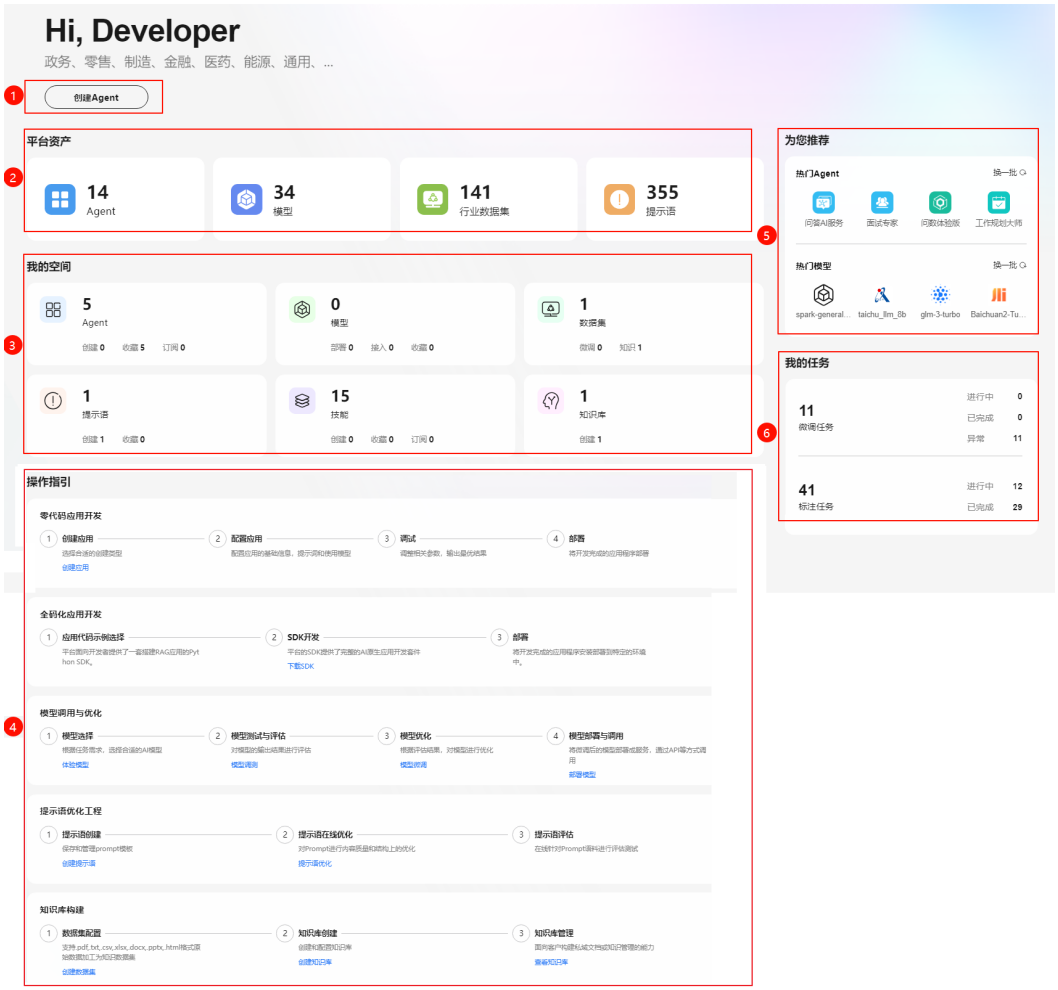


表 3-1 概览页说明

序号	区域	说明
1	创建Agent快捷入口	在该区域单击“创建Agent”可快速进入创建Agent页面，详细介绍请参见**创建Agent（LLM模式）**。
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

序号	区域	说明
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、订阅的工具个数。 ● 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： ● 零码Agent开发：详细流程说明请参见**创建并发布Agent**。 ● 全码化应用开发：详细流程说明请参见**下载AI原生应用引擎SDK**。 ● 模型调用与优化：详细流程说明请参见**调测/体验模型**、**基于微调数据集进行模型微调**、**生成模型服务**。 ● 提示语创建和优化：详细流程说明请参见**创建提示语**、**对创建的提示语进行优化**。 ● 知识库构建：详细流程说明请参见**创建知识数据集**、**创建知识库**。
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。

4 AI 原生应用引擎资产中心介绍

在AI原生应用引擎的左侧导航栏选择“资产中心”，进入资产中心页面，资产中心页面分为搜索、快速筛选、卡片展示三个区域，如图4-1所示。

资产中心提供了AI应用、工具、大模型、数据集及提示语模板各类资产，请参考表4-1了解各类资产及使用说明。

- ①：搜索区域，输入资产名称关键字进行搜索。
- ②：快速筛选区域，支持按照行业、类型、适用领域等各个维度快速筛选资产。
- ③：卡片展示区域，以卡片形式展示资产，单击卡片，进入详情页面查看各资产的详情，包括基本信息、基本配置、Agent对话日志、模型介绍、数据集的数据概况等；另外，卡片上还提供了收藏、部署、体验、去创建等入口，以便对各类资产进行操作。

图 4-1 资产中心

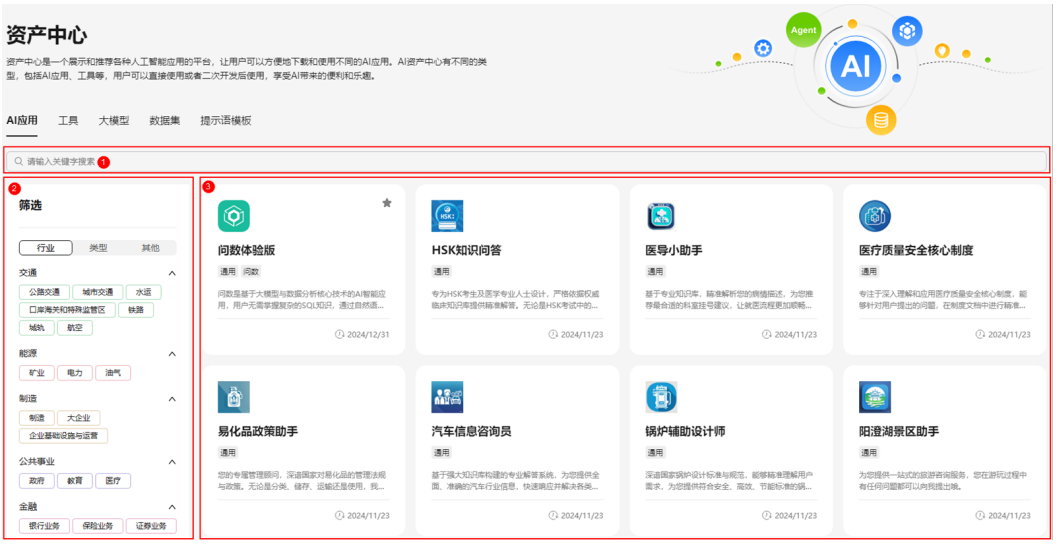


表 4-1 资产介绍

资产	说明
AI应用	AI应用页签下展示平台预置的AI应用。其中，部分AI应用支持用户进行**体验**，另一部分仅供查看，不支持体验。具体请以界面为准。
工具	工具页签下展示平台预置的第三方工具及租户上架的工具。 • 平台预置的第三方工具 请参考**调用资产中心工具前设置认证鉴权**设置鉴权，再进行调用。 • 租户上架的工具 请参考**收藏上架的工具**和**调用资产中心工具前设置认证鉴权**进行收藏及鉴权，再进行调用。
大模型	大模型页签下展示平台预置的大模型和平台接入的第三方模型服务。 • 平台预置的开源模型 – 开源模型Qwen系列、deepseek-coder系列等 请参考**将已有模型部署为模型服务**进行部署，部署后即可进行**调测/体验**、调用。 – 开源模型chatglm3-6b 平台提供了对应的模型服务API，但是该模型能力有限，只能作为问答模型，不能作为思考模型，首次使用该模型服务API需要订购免费的“ChatGLM3-6B大模型服务API在线调用”资源，订购后即可进行**调测/体验**、调用，订购操作请参见**购买AppStage**。 – 开源模型bge-reranker-large、bge-large-zh-v1.5、whisper-large-v3 平台提供了对应的模型服务API，可直接进行**调测/体验**、调用。其中，bge-reranker-large可以在知识检索流中作为重排序模型调用，bge-large-zh-v1.5可以在知识库中作为向量化模型调用。 • 平台接入的第三方模型服务 第三方厂商闭源模型，例如glm系列、moonshot、deepseek系列等。 请先参考**如何对模型供应商提供的模型服务设置鉴权**设置鉴权，再进行**调测/体验**、调用。
数据集	数据集页签下展示平台预置的微调数据集和知识数据集。 • 微调数据集 可用于模型微调任务，使用前请先参考**收藏预置微调数据集**收藏数据集。 • 知识数据集 仅展示，不支持使用。
提示语模板	提示语模板页签下展示平台预置的提示语模板。支持收藏、测试，基于模板创建新的提示语，具体操作请参见**管理资产中心预置提示语**。

5 通过运营看板查看 AI 原生应用引擎资产总览

5.1 查看资产总览

在资产总览页面可以查看当前租户所在的根部门以及租户下二级子部门的资产（Agent、模型、数据集、提示语、工具及知识库）统计数据，并支持通过各资产面板，筛选查看各子部门资产的创建、收藏及订阅数据。

前提条件

需要具备AI原生应用引擎租户运营管理员权限，权限申请操作请参见**AppStage组织成员申请权限**。

查看资产总览

在AI原生应用引擎的左侧导航栏选择“运营看板 > 资产总览”，资产总览页面如**图5-1**所示。

图 5-1 资产总览

表 5-1 资产总览页面说明

区域	说明
①	展示当前租户所在的根部门以及租户下二级子部门的资产（Agent、模型、数据集、提示语、工具及知识库）统计数据。
②	各类资产面板，支持筛选查看各子部门资产的创建、收藏及订阅数据。 您可以在各资产面板右上角下拉框选择需要展示的子部门。 TOP10：默认勾选，系统默认展示右上角下拉框中排序前10名的子部门资产数据。

5.2 查看模型调用统计

模型调用统计页面展示当前租户所在的根部门以及租户下二级子部门的模型调用情况。

前提条件

需要具备AI原生应用引擎租户运营管理员权限，权限申请操作请参见**AppStage组织成员申请权限**。

查看模型调用统计

1. 在AI原生应用引擎的左侧导航栏选择“运营看板 > 模型调用统计”。
2. 在模型调用统计页面，默认展示当前租户所在的根部门以及租户下二级子部门的模型调用情况，可以查看调用服务名称、调用部门、调用成功率、消耗Token等信息。
3. 模型调用统计列表支持通过高级搜索来查询模型调用情况，您可以在筛选器组合一个或多个筛选条件：
   - 部门名称：选择按租户所在根部门或某个二级子部门查询。
   - 时间范围：选择查询最近7天、最近30天、最近一年的模型调用记录。

5.3 查看 Agent 调用统计

Agent调用统计页面展示当前租户所在的根部门以及租户下二级子部门的Agent调用情况。

前提条件

需要具备AI原生应用引擎租户运营管理员权限，权限申请操作请参见**AppStage组织成员申请权限**。

查看模型调用统计

1. 在AI原生应用引擎的左侧导航栏选择“运营看板 > Agent调用统计”。

2. 在Agent调用统计页面，默认展示当前租户所在的根部门以及租户下二级子部门的Agent调用情况，可以查看调用服务名称、调用部门、用户总反馈、平均响应时长等信息。
3. Agent调用统计列表支持通过高级搜索来查询Agent调用情况，您可以在筛选器组合一个或多个筛选条件：
   - 部门名称：选择按租户所在根部门或某个二级子部门查询。
   - 时间范围：选择查询最近7天、最近30天、最近一年的Agent调用记录。

6 AI 原生应用引擎使用流程

AI原生应用引擎是企业专属的一站式大模型开发及应用构建平台，其核心是将自创建或平台预置的模型服务、工具、工作流及知识库等编排成具有一定智能性和自主性的Agent。本章节梳理了AI原生应用引擎使用流程，可帮助您快速了解AI原生应用引擎的核心功能。

图 6-1 AI 原生应用引擎使用流程

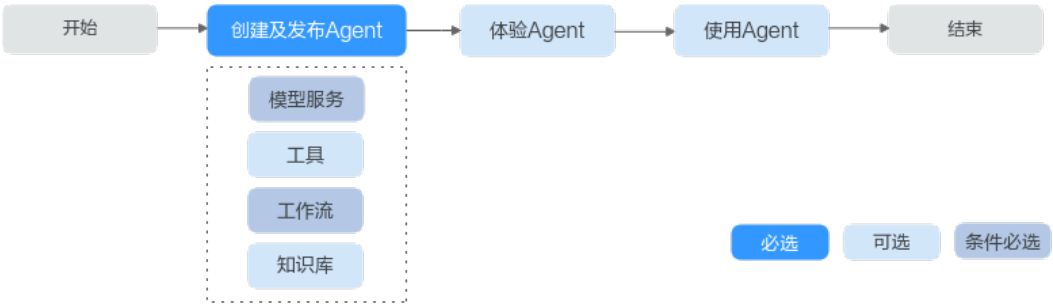


表 6-1 AI 原生应用引擎使用流程详解

序号	流程环节	说明
1	创建及发布Agent	一站式创建专属Agent，并将应用程序及相关组件进行发布，使其能够正常运行。当前支持创建LLM模式和工作流模式两种类型的Agent。 • LLM模式下，将准备好的模型服务（必选）、工具、工作流及知识库等编排成Agent。 • 工作流模式下，用户与工作流进行对话，因此必须添加工作流，不支持添加模型、工具、知识库等配置。
2	体验Agent	以对话的形式，对创建的Agent或平台资产中心预置的AI应用进行体验调测，以发现并解决Agent接口上的问题和错误。
3	使用Agent	支持通过API接口调用或Web界面访问两种方式使用Agent。

7 管理模型

7.1 模型使用指引

操作指引

图 7-1 模型使用操作指引

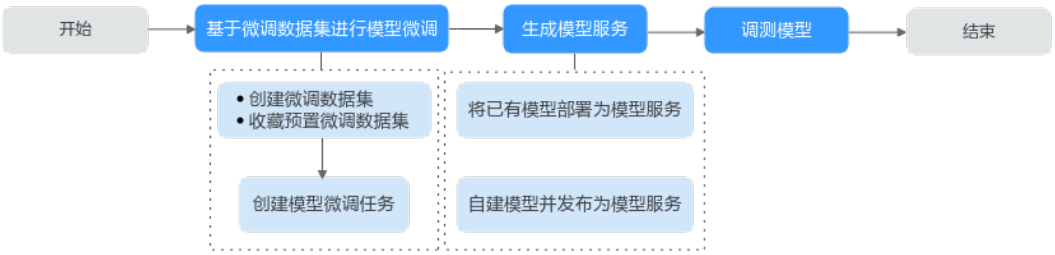


表 7-1 模型使用流程详解

序号	流程环节		说明
1	基于微调数据集进行模型微调	• 创建微调数据集 • 收藏预置微调数据集	对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，往往需要对大语言模型进行模型微调以适应特定任务。微调数据集是模型微调的基础，通过在微调数据集上进行训练从而获得改进后的新模型。
		创建模型微调任务	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微调数据集上训练模型来实现的，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。
2	生成模型服务	• 将已有模型部署为模型服务 • 接入模型服务	支持通过API接入模型服务，同时支持将平台预置模型进行微调后，部署为模型服务，模型服务可以在创建Agent时使用或通过模型调用接口调用。

序号	流程环节	说明
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。

7.2 基于微调数据集进行模型微调

7.2.1 创建微调数据集

对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，往往需要对大语言模型进行模型微调以适应特定任务。微调数据集是模型微调的基础，通过在微调数据集上进行训练从而获得改进后的新模型。

平台在资产中心预置了部分微调数据集，同时也支持用户根据需求自定义创建微调数据集。本文介绍如何创建微调数据集。

前提条件

- 通过OBS（对象存储服务）接入数据时，操作账号需获得OBS只读权限，具体操作请参见**对其他账号授予桶的读写权限**。
- 需具备充足的知识库容量包资源（包含OBS存储配额和向量库存储配额，两者比例为5:1），每个租户默认具备5G的OBS存储配额，默认配额用完后，请参考**购买AppStage**购买知识库容量包。
- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建微调数据集

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 微调数据集”。

步骤2 在“微调数据集”页面，单击“创建微调数据集”。

步骤3 参照**表7-2**进行相关参数的配置。

表 7-2 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。 当创建的微调数据集用于Functioncall能力增强类型的微调任务时，标签需选择为“功能调用”。

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：只支持json格式，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：数据文件在本地，从本地选择文件进行上传。 ● OBS接入：数据文件存放在华为云OBS桶，从OBS桶接入数据。 仅支持使用区域位置为北京四的OBS桶接入数据。
	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”选择本地文件进行上传。
	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
	版本更新模式	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：当OBS桶内内容发生变化时，调度成功后会生成一个新版本。
	执行周期	当“调度类型”选择“定时调度”时，需配置此参数。 设置执行周期，支持选择为天、周。
	执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 ● 当执行周期为“天”时，设置每日开始执行的时间。 ● 当执行周期为“周”时，指定每周周几，并设置当日开始执行的时间。
	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。

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7-3所示的操作。

表 7-3 更多操作

操作	步骤
查看数据集详情	1. 在微调数据集页面选择“我创建的”页签。 2. 在数据集列表中单击数据集名称，在微调数据集详情页面查看数据概况、调度历史，并支持对数据集进行溯源。
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. 在修改页面编辑数据集描述、修改标签，单击“保存”。
删除数据集	• 单个删除数据集： 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： 被标注的数据集无法删除。 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	• 只有格式为“对话文本”的数据集才可进行标注。 • 调度类型为“一次性调度”的数据集才可进行标注。 • 需要先在对微调数据集进行数据标注中创建标注任务，才能在当前页面执行数据标注。 1. 在微调数据集页面选择“我创建的”页签。 2. 在数据集列表中，单击数据集记录前的 > 。 3. 单击版本列表操作列的“标注”，参照对微调数据集进行数据标注进行数据标注。

7.2.2 收藏预置微调数据集

支持将平台预置微调数据集进行收藏，收藏后可便捷地在模型微调任务中使用。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

收藏预置微调数据集

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“数据集”页签。

步骤3 选择“微调数据集”子页签，将鼠标光标移至数据集卡片上，单击卡片右上角 。

单击工具卡片右上角的，可以取消收藏。

步骤4 收藏成功后，您可以在“知识中心 > 微调数据集”页面“数据集列表”页签的“我收藏的”子页签下，查看收藏结果。

单击数据集列表操作列的“取消收藏”，可以取消收藏。

步骤5 单击数据集列表中的数据集名称，可以便捷地查看数据集详情，包括基本信息、调度信息及版本记录。

----结束

7.2.3 对微调数据集进行数据标注

数据标注是将微调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。

约束与限制

- 只有格式为“对话文本”的微调数据集才可进行标注。
- 调度类型为“一次性调度”的微调数据集才可进行标注。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建数据标注

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 微调数据集”。

步骤2 在“微调数据集”页面，单击“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、数据集版本，填写标注名称。

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”列表中，单击操作列的“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7-4所示的操作。

表 7-4 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，单击操作列的“删除”。
发布标注	1. 在“数据标注”页签下的列表中，单击操作列的“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”。

7.2.4 创建模型微调任务

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微调数据集上训练模型来实现的，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。

支持将平台资产中心的部分模型作为微调前基础模型，也支持选择微调后的新模型作为基础模型再次进行微调。

前提条件

- 已订购大模型微调服务API在线调用-SFT局部调优，订购方法请参见**购买AI原生应用引擎按需计费资源**。
- 已具备格式为“对话文本”的微调数据集，具体请参考**创建微调数据集**或**收藏预置微调数据集**。
- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建微调任务

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型微调流水线”，单击“创建微调任务”。

如果选择资产中心的模型作为基础模型，您也可以在左侧导航栏单击“资产中心”，选择“大模型”页签，将鼠标移至目标模型卡片并单击“微调”，进入任务创建页面。

步骤2 选择“Functioncall能力增强微调”或“通用能力增强微调”。

- Functioncall能力增强微调：具备Functioncall能力的模型能够识别并执行函数调用或API调用，通过功能调用微调数据集增强模型的扩展性。
- 通用能力增强微调：广泛应用于各个领域，针对特定任务或行业需求，通过领域微调数据集增强模型领域能力。

步骤3 在创建微调任务页面，参照**表7-5**配置基础信息、模型及数据。

表 7-5 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	在下拉列表中选择微调的模型或平台的模型。
	训练模式	默认为“LoRA”。 LoRA（Low-Rank Adaptation）是一种轻量级大模型微调技术，通过低秩矩阵分解技术显著减少了微调所需的参数，降低了微调过程中所需的存储和计算资源，可灵活地运用于不同的预训练模型和任务。
	微调后名称	自定义模型微调后的新名称。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。
数据配置	选择微调数据集	单击“选择微调数据集”，选择“我创建的”或“我收藏的”数据集。
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7-6和表7-7配置基础参数、LoRA参数。

表 7-6 基础参数配置说明

参数英文名	参数中文名	参数说明
global_bs	各设备batch size总合	表示多个设备上使用的总样本数量。
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。
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。
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。

表 7-7 LoRA 参数配置说明

参数英文名	参数中文名	参数说明
lora_rank	秩	LoRA微调中的秩。
lora_alpha	缩放系数	LoRA微调中的缩放系数。
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

步骤5 单击“创建”。

新创建的微调任务显示在任务列表中，任务状态为“待启动”，请参考表7-8启用任务。

----结束

更多操作

创建微调任务完成后，可执行如表7-8所示的操作。

表 7-8 更多操作

操作	说明
启用任务	1. 在模型微调流水线任务列表中，单击操作列的“启用”，启动微调任务。当任务拥塞时，状态显示为“等待中”，待任务状态变为“运行中”时，表示正在执行微调任务。 2. 当任务状态变为“已完成”时，表示微调任务已执行完成。如果任务失败，任务状态显示为“运行失败”，您可以检查配置后重新启用。
停用任务	1. 在模型微调流水线任务列表中，单击操作列的“停用”，停用微调任务。当任务拥塞时，状态显示为“等待中”，待任务状态变为“停止中”时，表示正在停止微调任务。 2. 当任务状态变为“已停止”时，表示微调任务已停止执行。如果停止失败，任务状态显示为“运行失败”，您可以检查配置后重新停用。
发布微调后的模型	微调任务执行完成后，可以将微调后的模型部署为模型服务，模型部署后才能进行模型调测以及在创建Agent时调用。 在模型微调流水线任务列表中，单击操作列的“发布”，当任务状态显示为“已发布”，表示模型部署完成。如果部署失败，任务状态显示为“发布失败”，您可以检查配置后重新发布。
查看任务详情	在模型微调流水线任务列表中，单击任务名称，查看任务的基础信息、参数信息、运行日志以及Loss曲线等详情，并支持对模型之间的关系进行溯源。
重新创建任务	1. 在模型微调流水线任务列表中，选择操作列“更多 > 重新创建”。 2. 在修改微调任务页面，参照**3**~**4**进行配置。
删除任务	如果任务状态为“已发布”，需要先取消发布，才能删除。 1. 在模型微调流水线任务列表中，选择操作列的“更多 > 删除”。 2. 单击“确认”。

7.3 生成模型服务

7.3.1 将已有模型部署为模型服务

模型需要部署成功后才可正式提供模型服务。部署成功后，可以对模型服务进行模型调测，并支持在创建Agent时使用或通过**模型调用**接口调用。

本文介绍如何将微调后的模型或部分平台资产中心的模型部署为模型服务。

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。

- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

部署模型服务

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 我的模型服务”，单击“部署模型服务”。

对资产中心的模型进行部署时，您也可以在左侧导航栏单击“资产中心”，选择“大模型”页签，将鼠标移至目标模型卡片并单击“部署”，进入创建部署服务页面。

步骤2 配置模型信息，参数说明如**表7-9**所示。

表 7-9 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 仅支持模型类型为“文本对话”。 • 平台模型 仅支持模型类型为“文本对话”和“文本向量化”。
选择模型	在下拉列表选择待部署的模型。
服务名称	自定义模型服务名称，支持中英文、数字、中划线（-）、下划线（_）、点（.），长度2-36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3 配置部署模型参数，参数说明如**表7-10**所示。

表 7-10 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。 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4 单击“保存”，开始部署模型服务，在右侧模型效果预览区域，您可以看到模型服务状态为“部署中”。

部署完成后，模型服务状态变为“运行中”，此时才可进行模型调测及模型效果预览。

步骤5 （可选）在模型调测区域，参考**调测模型**进行模型调测。

步骤6 （可选）在右侧“模型效果预览”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。

----结束

更多操作

部署模型服务完成后，可执行如下**表7-11**所示的管理模型服务相关操作。

表 7-11 更多操作

操作	说明
启用模型服务	启用后的模型服务才能进行调测以及在创建Agent时调用。 1. 在“我部署的”模型服务列表中，单击操作列的“启用”，开始部署模型，此时模型服务状态显示为“部署中”。 2. 当模型状态变为“运行中”时，表示已部署完成，模型成功启用。如果部署失败，模型状态显示为“失败”，您可以检查配置后重新启用。

操作	说明
停用模型服务	1. 在“我部署的”模型服务列表中，单击操作列的“停用”，此时模型服务状态显示为“停止中”。 2. 当模型状态变为“停止”时，表示模型服务已停用；如果停用失败，模型状态显示为“失败”。
修改模型服务	运行中的模型服务需要先停用，才能修改。 1. 在“我部署的”模型服务列表中，选择操作列“更多 > 修改”。 2. 参照**步骤2**和**步骤3**，修改基础信息和配置信息。
删除模型服务	状态为“部署中”或“运行中”的模型服务需要先停用，才能删除。 1. 在“我部署的”模型服务列表中，选择操作列“更多 > 删除”。 2. 单击“确认”。
模型调测	只有部署完成的，状态为“运行中”的模型服务才能进行模型调测。 1. 在“我部署的”模型服务列表中，单击操作列的“模型调测”。 2. 参照**调测模型**的步骤，完成模型测试。

7.3.2 接入模型服务

支持通过API接入模型服务，模型服务接入后，可以进行模型调测，并支持在创建Agent时使用或通过**模型调用**接口调用。

前提条件

- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。
- 模型API接入之前，请确保符合相对应的接口规范，具体规范要求请参见**模型API接入接口规范**。

创建接入模型服务

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 我的模型服务”，单击“接入模型服务”。

步骤2 在“创建接入模型服务”页面，参照**表7-12**配置模型信息。

表 7-12 模型信息参数说明

参数名称	参数说明
模型名称	填写的模型名称必须为该模型的模型ID/模型编码（登录第三方模型厂商官网查看），例如：Baichuan4、deepseek-chat、glm-4-air，否则会导致模型不可用。 支持中英文、数字、中划线（-）、下划线（_）、点（.），长度2-36个字符，仅支持以中英文开头。 通过API调用模型服务时，该模型名称将用于OpenAPI调用请求体的model字段，详细介绍请参见**模型调用**。
模型类型	可选模型类型包括：文本对话、文本向量化、文本排序。
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2-36个字符，仅支持以中英文开头。
模型服务描述（可选）	自定义模型服务相关描述信息。 如果要在创建Agent时选择该模型作为思考模型，需要在模型服务描述中填写“SupportFunctionCall，AdaptFunctionCall”进行适配。
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3 配置模型服务API相关参数，参数说明如**表7-13**所示，配置完单击“保存”。

表 7-13 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，支持HTTPS、HTTP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。 • App-code：APP认证方式，通过请求header的X-Apig-Appcode字段携带App-code进行认证，需要提供App-code。

参数名称	参数说明
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。 ● 请通过API提供者或模型供应商获取API Key。 ● 输入的关键信息将进行加密保存，仅用于模型服务的调用。如果API Key发生变化，更新此处信息后，设置将于2分钟后生效。
AK/SK	鉴权方式为“AK/SK”时，配置此参数。 AK：访问密钥ID。 SK：密钥。 ● 请通过API提供者或模型供应商获取AK/SK。 ● 输入的关键信息将进行加密保存，仅用于模型服务的调用。
App code	鉴权方式为“App-code”时，配置此参数。 ● 请通过API提供者或模型供应商获取App code。 ● 输入的关键信息将进行加密保存，仅用于模型服务的调用。如果APP code发生变化，更新此处信息后，设置将于2分钟后生效。
API接口协议	模型类型为“文本对话”或“文本向量化”时，选择“标准OpenAI协议”。 模型类型为“文本排序”时，选择“AI引擎标准协议”。
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4 在模型调测区域调测模型。

- 调测文本对话类型模型，请参考**表7-14**配置参数。

表 7-14 文本对话类型模型调测参数说明

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	表示模型输出的最大长度。
温度	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”只设置1个。
多样性	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”只设置1个。
存在惩罚	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	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。
内容安全监测配置	当“输出方式”为“非流式”时，显示此参数。 选择是否打开开关，开启后，可对返回内容中的文本和图片进行安全监测。

- 调测文本向量化类型模型
  a. 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”。
- 调测文本排序类型模型
  a. 配置表7-15所示参数。

     表 7-15 调测文本排序类型模型参数说明

参数名称	参数说明
待排序文本	输入待排序文本。单击+添加文本，最多可以添加10条。
被展示文本条数	文本排序完成后，展示的条数。取值范围为1~10。
我的问题	描述想要解决的问题。

  b. 单击“开始排序”。

步骤5 在右侧预览模型效果。

步骤6 单击“发布”，模型服务发布成功。

----结束

更多操作

模型服务发布完成后，可执行如**表7-16**所示的相关操作。

表 7-16 更多操作

操作	说明
取消发布模型服务	在“我接入的”页签下的服务列表中，单击“操作”列的“取消发布”。
模型调测	1. 在“我接入的”页签的服务列表中，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”页签下的服务列表中，单击“操作”列的“更多 > 修改”。
删除模型服务	1. 在“我接入的”页签下的服务列表中，单击“操作”列的“更多 > 删除”。 2. 单击“确认”。

7.3.3 创建路由策略用于提供模型服务

通过配置路由策略，可以实现模型故障自动切换能力，当模型A因故障等原因无法正常工作时，可以自动切换为另一个可用的模型提供服务，从而提高模型服务的稳定性和可用性。

路由策略创建完成后，可以进行模型调测，并支持在创建Agent时使用或通过接口调用。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建路由策略

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 我的模型服务”，单击“创建路由策略”。

步骤2 在“创建路由策略”页面，参照**表7-17**配置策略信息，配置完单击“保存”。

表 7-17 路由策略参数说明

参数	说明
策略名称	自定义路由策略的名称，支持中英文、数字、中划线（-）、下划线（_）、点（.），长度2~36个字符，仅支持中英文开头。
AI模型	在“模型A”下拉框中选择模型。 单击“+ AI模型”，还可以增加2个AI模型。 路由策略提供模型服务时，模型调用顺序为：模型A > 模型B > 模型C，当模型A无法正常工作时，可以自动依次切换为模型B、模型C。
策略总超时时间	模型路由策略的总体超时时间，取值范围为1000-1000000ms。
模型重试次数	路由策略中单个模型服务的重试次数，取值范围为0-100次。
策略描述	路由策略的描述信息。

步骤3 在模型调测区域，参考**表7-18**调测模型。

表 7-18 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	表示模型输出的最大长度。
温度	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”只设置1个。
多样性	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”只设置1个。
存在惩罚	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	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。
内容安全监测配置	当“输出方式”为“非流式”时，显示此参数。 选择是否打开开关，开启后，可对返回内容中的文本和图片进行安全监测。

步骤4 在右侧“模型效果预览”区域查看效果。

----结束

更多操作

模型路由策略创建完成后，可执行如表7-19所示的操作。

表 7-19 更多操作

操作	说明
修改路由策略	在“我的路由策略”页签的列表中，单击“操作”列的“修改”，可以调整模型数量，编辑总超时时间、模型重试次数、描述信息。
删除路由策略	在“我的路由策略”页签的列表中，单击“操作”列的“删除”。

7.4 调测/体验模型

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。

支持对我的模型（我部署的、我接入的）、我的路由策略、平台预置的模型以及平台接入的第三方模型进行调测。

前提条件

对平台接入的第三方模型进行调测前，需要先设置鉴权，具体操作请参见**如何对平台接入的第三方模型服务设置鉴权**。

调测模型

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型调测”。

如果对资产中心的模型进行调测，您也可以在左侧导航栏单击“资产中心”，选择“大模型”页签，将鼠标移至目标模型卡片并单击“体验”，进入模型调测页面。

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、文本转语言类型模型以及文本排序类型模型。

- **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7-20所示参数。

表 7-20 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	选择待调测的模型服务，在下拉列表可选： ■ 模型服务商API（平台接入的第三方模型） ■ 预置模型API（平台预置的开源模型） ■ 我的模型API（我部署的、我接入的） ■ 我的路由策略
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	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”只设置1个。
多样性	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”只设置1个。
存在惩罚	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	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。
内容安全监测配置	当“输出方式”为“非流式”时，显示此参数。 选择是否打开开关，开启后，可对返回内容中的文本和图片进行安全监测。

b. 在右侧“效果预览”区域，可通过以下两种方式进行模型调测。

- 在对话输入框输入测试语句后按Enter键或单击进行模型调测。
- 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键或单击进行模型调测。

- **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7-21所示参数。

表 7-21 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，在下拉列表可选模型服务商API（平台接入的第三方模型服务）。
输入内容	输入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。
内容安全监测配置	选择是否打开开关。 开启后，可对返回内容中的文本和图片进行安全监测。
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，在下拉列表可选模型服务商API。
  - 上传图片：单击，可上传本地图片。
  - 内容安全监测配置：选择是否打开开关，开启后，可对返回内容中的文本和图片进行安全监测。
  - 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7-22所示参数。

表 7-22 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： ■ 模型服务商API（平台接入的第三方模型） ■ 预置模型API（平台预置的开源模型）
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）。

参数名称	参数说明
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	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。
内容安全监测配置	选择是否打开开关。 开启后，可对返回内容中的文本和图片进行安全监测。

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：
  - 模型服务商API（平台接入的第三方模型服务）
  - 预置模型API（平台预置的开源模型）
  - 我的模型API（我部署的、我接入的）
- 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7-23所示参数。

表 7-23 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，在下拉列表可选模型服务商API（平台接入的第三方模型服务）。

参数名称	参数说明
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。

- **调测文本排序类型模型**，具体操作如下：

a. 在“模型类型”下选择“文本排序”并配置**表7-24**所示参数。

表 7-24 调测文本排序类型模型参数说明

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： ■ 预置模型API（平台预置的开源模型） ■ 我的模型API（我部署的、我接入的）
待排序文本	输入待排序文本。单击+添加文本，最多可以添加10条。
被展示文本条数	文本排序完成后，展示的条数。取值范围为1~10。
我的问题	描述想要解决的问题。

b. 单击“开始排序”，在右侧“效果预览”区域即可收到生成结果。

----结束

7.5 评测模型

平台支持从多个维度对模型的能力、性能进行评估，以保证模型效果，为模型选型提供可靠依据。

约束与限制

仅支持对文本对话类型的模型服务进行评测。

前提条件

评测模型前，请先通过**调测/体验模型**功能确认模型可用。

创建评测任务

步骤1 在AI原生应用引擎的左侧导航栏选择“模型中心 > 模型评测”，单击“创建评测任务”。

步骤2 在“创建评测任务”弹框中选择“通用维度评测”。

步骤3 在创建评测任务页面，参照表7-25配置模型信息。

表 7-25 评测任务参数说明

参数	说明
任务名称	自定义评测任务的名称。 支持中英文、数字、中划线（-）、下划线（_）、点（.），长度2-36个字符，仅支持以中英文开头。
任务描述	评测任务的描述信息。
选择模型	选择待评测的模型，最多可选择3个模型，支持以下模型： • 我的模型API（我部署的、我接入的） • 预置模型API • 模型服务商API 单击“模型参数配置”，配置如下参数： • 输出最大token数：表示模型输出的最大token数。 • 温度：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”只设置1个。 • 多样性：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”只设置1个。 • 存在惩罚：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：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。
评测维度	支持通过以下三个维度评测模型，每个维度下又细分了多个子维度，子维度类别请以页面展示为准。 • 通用智能 • 专业技能 • 可信与AI治理

步骤4 单击“创建”。

新创建的任务显示在模型评测任务列表中，任务状态为“草稿”，请参考表7-26运行任务。

----结束

更多操作

创建评测任务完成后，可执行如表7-26所示的操作。

表 7-26 相关操作

操作	说明
运行评测任务	评测过程需要消耗大量大模型调用token，在评测前请确保余量充足，否则可能导致评测任务执行失败，您可以参考购买AI原生应用引擎订购。 在模型评测任务列表中，单击操作列的“运行”，任务状态显示为“运行中”，当任务状态变为“运行完成”时，表示评测任务已执行完成。
取消运行评测任务	当任务状态为“运行中”时，在模型评测任务列表中，单击操作列的“取消”，取消任务运行。
查看评测任务详情	在模型评测任务列表中，单击任务名称，进入模型评测详情页面，详情页面根据任务所包含的模型，分页签展示各评测维度的评测执行情况，包括评测数据集、评测执行量/评测总数量、执行成功数、执行失败数及分数等，您可以执行如下操作： • 失败用例重试：对于当前模型下执行失败的评测数据集，单击详情页面右上角的“失败用例重试”，重新基于失败的评测数据集进行模型评测。 • 查看评测数据：在详情列表中，单击操作列的“查看评测数据”，查看该模型某个维度的部分评测数据。
编辑评测任务	当任务状态为“草稿”时，在模型评测任务列表中，选择操作列的“更多 > 编辑”，修改任务参数。
删除评测任务	当任务状态为“运行中”时，请先取消任务，再进行删除。 在模型评测任务列表中，选择操作列的“更多 > 删除”，删除任务。
下载报告	评测任务执行完成后，在模型评测任务列表中，选择操作列的“更多 > 下载报告”，下载模型评测报告。

7.6 查看模型调用记录

通过查看模型的调用记录，可以获取模型调用方式、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，通过筛选调用方式、日期、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、调用方式、调用状态、用时及调用时间等信息。

----结束

7.7 收藏平台资产中心的模型

支持收藏平台资产中心提供的模型，包括预置的模型和平台接入的第三方模型。将自己关注的或后续计划使用的模型收藏后，可便捷地在收藏列表中进行查看。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

收藏平台资产中心的模型

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“大模型”页签。

步骤3 将鼠标光标移至模型卡片上，单击卡片右上角  。

单击模型卡片右上角的  ，可以取消收藏。

步骤4 收藏成功后，您可以在“模型中心 > 我的模型服务”页面“我收藏的”页签下，查看收藏结果，可以便捷地对收藏的模型进行部署、微调、体验。

并非所有模型都支持部署、微调和体验，实际可执行的操作请以界面为准。

----结束

7.8 模型 API 接入接口规范

当前模型网关支持文本对话（Chat）、文本向量化（Embeddings）、文本排序（Rerank）三种类型的API接入。模型API接入之前，请确保符合相对应的接口规范，其中Chat接口和Embeddings接口需要符合OpenAI接口规范，Rerank接口需要符合AI引擎标准协议。

文本对话（Chat）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-27 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **ChatCompletionRequestMessage** objects	文本对话消息体类。

参数	是否必选	参数类型	描述
model	是	String	文本对话使用的模型名称。
frequency_penalty	否	Number	介于-2.0和2.0之间的数字。 正值会根据文本中新Token的现有频率对其进行惩罚，从而降低模型重复相同行的可能性。 最小值：-2 最大值：2 缺省值：0
logit_bias	否	Map<String,Integer>	该参数接受一个JSON对象，将标记映射到从-100（禁止）到100（独占选择标记）的关联偏差值。 像-1和1这样的适度值将以较小的程度改变选择标记的概率。 使用logit_bias参数时，偏差被添加到模型生成的logits之前进行抽样。
max_tokens	否	Integer	返回体允许的最大token数。
n	否	Integer	返回体中包含的choices数量，建议默认设置为1，最大限度地降低成本。 最小值：1 最大值：128 缺省值：1

参数	是否必选	参数类型	描述
presence_penalty	否	Number	介于-2.0和2.0之间的数字。 正值会根据它们是否出现在文本中来惩罚得到新的Token，从而增加模型谈论新主题的可能性。 最小值：-2 最大值：2 缺省值：0
stream	否	Boolean	布尔类型。 设为true时，返回结果为流式。 设为false时，返回结果为JSON格式结构化数据。 缺省值：false
temperature	否	Number	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。 最小值：0 最大值：2 缺省值：1
top_p	否	Number	影响输出文本的多样性，取值越大，生成文本的多样性越强。 最小值：0.0 最大值：1.0 缺省值：1
tools	否	Array of **FunctionCallTool** objects	可供模型调用的工具。
tool_choice	否	String	用于控制模型是如何选择要调用的函数，仅当工具类型为function时补充。 默认为auto，且当前仅支持auto。

表 7-28 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
role	是	String	消息体对应的角色。 如果是系统则为system。 如果是用户则为user。 枚举值： • system • user
content	是	String	消息具体内容。
name	否	String	对话参与者的可选名称，提供给模型信息以区分相同角色的不同对话参与者。

表 7-29 FunctionCallTool

参数	是否必选	参数类型	描述
type	否	String	调用工具类型，目前仅支持function。
function	否	**function** object	仅当工具类型为function时补充。

表 7-30 function

参数	是否必选	参数类型	描述
name	否	String	函数名称，只能包含a-z、A-Z、0-9、下划线和中横线。最大长度限制为64个字符。
description	否	String	用于描述函数功能。 模型会根据这段描述决定函数调用方式。

参数	是否必选	参数类型	描述
parameters	否	Object	Json Schema对象，用于定义函数所接受的参数。

- 非工具调用请求示例

```
{
   "model": "my-chat-model",
   "messages": [
      {
         "role": "system",
         "content": " You are a helpful assistant. "
      },
      {
         "role": "user",
         "content": "你好！ "
      }
   ],
   "max_tokens": 20,
   "presence_penalty": 1.2,
   "frequency_penalty": 1.0,
   "temperature": 0.5,
   "top_p": 0.95,
   "stream": false
}
```

- 工具调用请求示例

```
{
   "model": "my-chat-model",
   "messages": [
      {
         "role": "user",
         "content": "请帮我查询南京的天气"
      }
   ],
   "tools": [
      {
         "type": "function",
         "function": {
            "name": "get_weather",
            "description": "获取给定地点的天气",
            "parameters": {
               "type": "object",
               "properties": {
                  "location": {
                     "type": "string",
                     "description": "地点，例如北京、上海。"
                  }
               },
               "required": ["location"]
            }
         }
      }
   ],
   "max_tokens": 200,
   "presence_penalty": 1.2,
   "frequency_penalty": 1.0,
   "temperature": 0.5,
```



```
    "top_p": 0.95,
    "stream": false
}
```

响应体参数

表 7-31 响应体参数

参数	参数类型	描述
id	String	文本对话唯一标识符。
choices	Array of **choices** objects	返回体列表。 如果'n'大于1，则结果为多个。
created	Integer	问答发生的时间（格式为时间戳）。
model	String	文本对话使用的模型名称。
object	String	固定值'chat.completion'。
usage	**CompletionUsage** object	文本对话用量统计。

表 7-32 choices

参数	参数类型	描述
index	Integer	返回多个choices时，每个choice对应的顺序。
message	**ChatCompletionResponseMessage** object	模型服务返回的具体消息体内容。

参数	参数类型	描述
finish_reason	String	返回结束的原因。 • stop：模型达到自然停止点或提供的停止序列。 • length：达到请求中指定的最大令牌数。 • content_filter：由于内容过滤器的标志而省略了内容。 • tool_calls：模型选择了某个工具。 枚举值： • stop • length • content_filter • tool_calls

表 7-33 ChatCompletionResponseMessage

参数	参数类型	描述
content	String	返回消息体的内容，与tool_calls二选一。
role	String	返回消息体的角色。 枚举值： • assistant
tool_calls	Array of **ToolCall** objects	工具调用消息，与content二选一。

表 7-34 ToolCall

参数	参数类型	描述
id	String	工具调用唯一标识符。
type	String	工具类型，当前仅支持function。
function	**CallFunction** Object	调用函数的详细信息。

表 7-35 CallFunction

参数	参数类型	描述
name	String	函数名。
arguments	String	调用函数的参数，Json格式。

表 7-36 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	回答包含的token数。
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

- 非流式响应示例
  - 非工具调用

```
{
   "id": "chatcmpl-xxx",
   "object": "chat.completion",
   "created": 1718772336,
   "model": "my-chat-model",
   "choices": [
      {
         "index": 0,
         "message": {
            "role": "assistant",
            "content": "你好，有什么我可以帮助你的吗？"
         },
         "finish_reason": "stop",
         "logprobs": null
      }
   ],
   "usage": {
      "prompt_tokens": 5,
      "completion_tokens": 10,
      "total_tokens": 15
   }
}
```

  - 工具调用

```
{
   "id": "chatcmpl-xxx",
   "object": "chat.completion",
   "created": 1718772336,
   "model": "my-chat-model",
   "choices": [
      {
         "index": 0,
         "message": {
            "role": "assistant",
            "content": null,
```



```
            "tool_calls": [
               {
                  "id": "call_123",
                  "type": "function",
                  "function": {
                     "name": "get_weather",
                     "arguments": "{\"location\": \"南京\"}"
                  }
               }
            ]
         },
         "finish_reason": "tool_calls",
         "logprobs": null
      }
   ],
   "usage": {
      "prompt_tokens": 5,
      "completion_tokens": 10,
      "total_tokens": 15
   }
}
```

- 流式响应示例
  - 非工具调用

```
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"role":"assistant","content":""},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你好"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"，"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"有"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"什么"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"我"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"可以"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"帮助"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
```


model","choices":[{"index":0,"delta":{"content":"的
"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-
xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
model","choices":[{"index":0,"delta":{"content":"吗
"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-
xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
model","choices":[{"index":0,"delta":{"content":"?
"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-
xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"stop"}]}

- 工具调用

  流式返回的工具调用信息必须在一条消息内，不能分拆返回。

{"id":"chatcmpl-
xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
model","choices":[{"index":0,"delta":{"role":"assistant","content":null,"tool_calls":
[{"id":"call_123","type":"function","function":
{"name":"get_weather","arguments":"{\"location\":\"南京
\"}"}}]}"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-
xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-
model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"tool_calls"}]}

文本向量化（Embeddings）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-37 请求体参数

参数	是否必选	参数类型	描述
input	是	Array of strings	输入支持2种格式： ● 纯文本（string），例如："你好" 。 ● 文本列表（array），例如：["你","好"] 。 数组长度：1-2048
model	是	String	向量化模型名称。

请求示例：

```
{
   "model": "my-embedding-model",
```



```
  "input": "你好"
}
```

响应体参数

表 7-38 响应体参数

参数	参数类型	描述
data	Array of **Embedding** objects	向量化结果。
model	String	向量化模型名称。
object	String	固定值‘list’。
usage	**usage** object	每次请求的用量统计。

表 7-39 Embedding

参数	参数类型	描述
index	Integer	向量在向量列表中的排序。
embedding	Array of numbers	向量数组（Float类型）。
object	String	固定值‘embedding’。

表 7-40 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问包含的token数。

响应示例：

```
{
  "data": [
    {
      "index": 0,
      "embedding": [
        0.02513289265334606,
        -0.017512470483779907,
        -0.029955564066767693,
        ...
      ],
      "object": "embedding"
    }
```



```
    ],
    "usage": {
        "prompt_tokens": 5,
        "total_tokens": 5
    },
    "model": "my-embedding-model",
    "object": "list"
}
```

文本排序（Rerank）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 7-41 请求体参数

参数	是否必选	参数类型	描述
query	是	String	原始请求问题，基于该问题对候选文本进行排序。
top_n	是	Integer	返回排序靠前的n个结果。
docs	是	Array of strings	候选文本。
model	是	String	排序模型名称。

请求示例：

```
{
    "model": "my-rerank-model",
    "query": "请问AI原生应用引擎提供了什么能力？",
    "docs": ["AI原生应用引擎提供了应用开发、模型网关等能力。", "AI原生应用引擎正在逐步完善、提高竞争力。"],
    "top_n": 3
}
```

响应体参数

表 7-42 响应体参数

参数	参数类型	描述
model	String	排序模型名称。
usage	**usage** object	每次请求的用量统计。
results	Array of **RankDocument** objects	排序结果

表 7-43 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问包含的token数。

表 7-44 RankDocument

参数	参数类型	描述
index	Integer	文本排序后对应的序号。
document	**Document** object	文本
relevance_score	Number	文本的排序分数。

表 7-45 Document

参数	参数类型	描述
text	String	文本内容。

响应示例：

```
{
   "model": "my-rerank-model",
   "usage": {
      "prompt_tokens": 5,
      "total_tokens": 5
   },
"results": [
    {
        "index": 0,
        "document": {"text": "AI原生应用引擎提供了应用开发、模型网关等能力。"},
        "relevance_score": 0.9
    },
    {
        "index": 1,
        "document": {"text": "AI原生应用引擎正在逐步完善、提高竞争力。"},
        "relevance_score": 0.5
    }
  ]
}
```

7.9 如何对平台接入的第三方模型服务设置鉴权

平台资产中心接入了第三方供应商的闭源模型，例如GLM系列、Moonshot系列等，这些模型服务在调测（体验）、调用前，需要先设置模型鉴权。

前提条件

需要具备AI原生应用引擎管理员权限，权限申请操作请参见**AppStage组织成员申请权限**。

操作步骤

步骤1 在AI原生应用引擎左侧导航栏选择“凭证管理 > 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。

具体鉴权信息需根据界面提示前往模型供应商官网进行申请。

----结束

8 构建知识库

8.1 创建知识数据集

知识数据集是构建和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。

前提条件

- 通过OBS（对象存储服务）接入数据时，操作账号需获得OBS只读权限，具体操作请参见**对其他账号授予桶的读写权限**。
- 需具备充足的知识库容量包资源（包含OBS存储配额和向量库存储配额，两者比例为5:1），每个租户默认具备5G的OBS存储配额，默认配额用完后，请参考**购买AppStage**购买知识库容量包。
- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建知识数据集

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 知识库”。

步骤2 选择页面右上角的“ ··· > 知识数据集”。

步骤3 单击“创建知识数据集”，参照**表8-1**进行相关参数的配置。

表 8-1 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，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。

参数名称		参数说明
	数据类型	根据实际需要可选以下格式： ● 文档：支持.pdf、.txt（只支持UTF-8）、.csv（只支持UTF-8）、.xlsx、.docx、.pptx、.html、.json、.xml、.md格式，单个文件最大为10M，总上传大小最大为500M。 ● 图片：支持.png、.jpg、.jpeg、.gif、.webp、.bmp格式，单张图片最大为10M，总上传大小最大为200M。 ● 图片-摘要：支持本地文件上传.png、.jpg、.jpeg、.gif、.webp、.bmp格式，需对图片填写摘要信息，单张图片最大为10M，总上传大小最大为300M。 ● 视频-摘要：支持本地文件上传.mp4、.webm、.wmv、.mov、.avi格式，需对视频填写摘要信息，单个视频最大为100M，总上传大小最大为300M。
数据接入	接入方式	选择数据集的接入方式。支持以下两种方式： ● 本地上传：数据文件在本地，从本地选择文件进行上传。 ● OBS接入：数据文件存放在华为云OBS桶，从OBS桶接入数据。 仅支持使用区域位置为北京四的OBS桶接入数据。
	数据文件	当接入方式选择“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，支持上传的文件类型请参考“数据类型”参数说明。
	文件类型	● 当接入方式选择“本地上传”时，无需配置，根据上传的文件自动识别文件类型。 ● 当接入方式选择“OBS接入”时，选择接入的文档、图片的格式。
	OBS桶名	当接入方式选择“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	当接入方式选择“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	当接入方式选择“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

参数名称		参数说明
	版本更新模式	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：当OBS桶内内容发生变化时，调度成功后会生成一个新版本。
	执行周期	当“调度类型”选择“定时调度”时，需配置此参数。 设置执行周期，支持选择为天、周。
	执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 ● 当执行周期为“天”时，设置每日开始执行的时间。 ● 当执行周期为“周”时，指定每周周几，并设置当日开始执行的时间。
	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。
数据加工	数据清洗（可选）	数据类型选择为“文档”、“图片-摘要”、“视频-摘要”时，显示此参数。 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
	PDF预处理	数据类型选择为“文档”时，显示此参数。 ● 提取PDF富媒体：提取PDF文件中的富媒体（图片、表格等）。 ● 无处理
	智能匹配图表	当“PDF预处理”选择为“提取PDF富媒体”时，显示此参数。 数据切分时，有助于对PDF文件中的图片、表格进行完整提取。

参数名称		参数说明
	数据切分	数据类型选择为“文档”时，显示此参数。 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。 ● 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。
	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。输入的分隔符不允许包含以下特殊字符*./$^?+以及带有\的分隔符，仅支持输入\n，用于表示按行分隔，其它诸如\\n，\n\n等均不支持。
	包含分隔符	切分后的分段内容中是否包含分隔符，选择“包含”或“不包含”。
	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。如果当前分片长度小于该值，则会和其他分片进行合并直到接近该值，所以如果不想合并，请将分段长度设置为1。

参数名称		参数说明
	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。
切片提取配置	切片提取配置	开启“切片提取配置”开关，可以对已切分好的分片，提取相关内容作为检索字段、文本过滤字段或向量化字段。
	切片提取模式	数据类型选择为“文档”时，显示此参数。 ● 智能提取：通过大语言模型提取分片内的问题与回答，效果由大模型决定。目前仅支持提取问题和答案，且耗时较长。 ● 规则提取：按照用户配置的提取规则对切片内容进行提取。
	切片内容示例	数据类型选择为“文档”时，显示此参数。 根据输入的切片内容示例，提取并展示。
	切片提取配置	切片提取模式为“规则提取”时，配置此参数。 单击“添加提取字段”，配置如下参数： ● 切片片段名称：自定义切片片段名称。 添加的提取字段数量不超过10个，切片片段名称长度最多为20个字符，不允许重复，不允许为以下名称（大小写不敏感）：file_name、file_id、path、order、document、base64、chunk，不能以ki_、ko_开头，仅可包含字母、数字、下划线，并且以字母开头。 ● 提取规则：分隔符提取、按规则提取。 ● 提取设置 – 当提取规则为“分隔符提取”时，例如：以分隔符**；**对切片进行分段，选择第**2**个分段，片段中**不包含**分隔符。 输入的分隔符不允许包含以下特殊字符*./$^?+以及带有\的分隔符，仅支持输入\n，用于表示按行分隔，其它诸如\\n，\n\n等均不支持。 在填写“选择第*{{片段序号}}*个片段”时，如果此处的片段序号超出片段数量，则提取内容为空。 – 当提取规则为“按规则提取”时，例如：从切片提取以**项目名称：**开头和以**（完）**结尾的片段，选择第**3**个片段，片段中**包含**开头，**不包含**结尾。
	切片提取效果	数据类型选择为“文档”时，显示此参数。 展示对切片内容提取后的效果。

步骤4 单击“创建数据集”。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。

如果需要为数据集创建索引配置，单击“下一步”，参考**创建索引配置**。

----结束

创建索引配置

步骤1 知识数据集创建完成后，单击“下一步”，进入索引配置页面。

步骤2 在索引配置页面，参照**表8-2**进行相关参数的配置。

表 8-2 索引配置参数说明

参数		说明
基础配置	索引配置名称	自定义索引配置名称。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。
	索引描述	索引配置的描述信息。
	RAG类型	GraphRAG类型仅支持用于“文档”类型的数据。 • VectorRAG：向量RAG，是一种结合了向量化和大语言模型的RAG技术。VectorRAG将非结构化的数据转化为结构化的向量空间，利用向量库实现高效的信息检索。 • GraphRAG：知识图谱RAG，是一种结合了知识图谱和大语言模型的RAG技术。GraphRAG能够处理各种类型的文档，从中提取实体（文档中具体的对象或概念）、关系以及文本内容构建知识图谱（一种结构化的知识表示方式），从而增强大语言模型对复杂信息的理解和推理能力。
向量化配置	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。 支持选择模型服务商API、预置模型API、我的模型API（我部署的、我接入的）。 当前向量化模型支持的最大长度为512 token，对应的中文约为512个字，英文与符号约900个字符，请注意分片长度。 模型服务商API使用前需要先设置鉴权，具体操作请参见**如何对模型供应商提供的模型服务设置鉴权**。

参数		说明
	长文本策略	• 截断模式：如果待向量化分片字段token长度超过向量化模型限制的token总数，则进行截断，取前top k个token。 • 智能模式：如果待向量化分片字段token超过向量化模型限制的token总数，首先利用对话大模型对超长分片进行重写，如果仍然超长则进入截断模式。 • 默认模式：如果待向量化分片字段token长度超过了向量化模型限制的token总数，则创建知识库失败。
知识图谱配置	实体抽取模型	选择实体抽取模型。从文本中提取实体（文档中具体的对象或概念）、关系以及文本内容，用于构建知识图谱。
	实体抽取提示语	• 默认：模型按照默认的提示语要求，对文本内容进行抽取。 • 自定义：对默认的提示语进行编辑，自定义实体抽取方式。
索引字段配置		• 当RAG类型为“VectorRAG”时，配置如下参数： – 检索字段：选择对哪个字段进行检索，并作为检索命中内容返回。 – 文本过滤字段：选择哪些字段支持在检索时同步进行文本过滤，并返回该字段的内容。 – 附加返回字段：选择哪些字段在检索时附加返回。 • 当RAG类型为“GraphRAG”时，配置如下参数： – 实体抽取字段：选择对哪个字段进行实体抽取。选择从chunk抽取实体时，chunk是一个完整切片，是指从这个切片的内容里面去抽取实体。 – 附加返回字段：选择哪些字段在检索时附加返回。

步骤3 单击“创建数据集和索引配置”，创建的数据集和索引配置显示在“知识数据集”页面的数据集列表中。

----结束

更多操作

创建数据集完成后，可根据需要执行如表8-3所示的操作。

表 8-3 更多操作

操作	步骤
查看数据集详情	在数据集列表中单击数据集名称，在知识数据集详情页面查看数据概况、索引配置、调度历史以及溯源。
修改数据集	在数据集列表中单击“操作”列的“修改”，支持修改数据集的描述和标签。
删除数据集	• 单个删除数据集 1. 在数据集列表中，选择单击“操作”列的“更多 > 删除”。 2. 单击“确定”。 • 批量删除数据集 1. 在数据集列表勾选多个数据集，再单击列表上方“批量删除”。 2. 在“批量删除”对话框，单击“确认”。
创建索引配置	在数据集列表中单击“操作”列的“创建索引配置”，参考**表8-2**进行配置。
编辑切片	数据集的数据类型为“文档”时，支持编辑切片。编辑切片不会直接更新原数据集，需要生成新的数据集版本使编辑内容生效。 1. 在数据集列表中，选择“操作”列的“更多 > 编辑切片”。 2. 在“选择切片来源”弹窗中，选择数据集版本以及待编辑的文档，单击“确认”。 3. 在切片详情页面，可以编辑、新增及删除切片，操作完成后单击“更新”。 • 选中切片，在页面右侧切片预览区域，编辑当前切片内容。 • 单击切片右下方的，在当前切片前增加新的切片。 • 单击切片右下方的，在当前切片后增加新的切片。 • 单击切片右下方的，删除当前切片。
生成新的数据集版本	切片编辑完成后，必须生成新的数据集版本，才能使编辑内容生效。 1. 切片编辑完成后，在数据集列表中单击列表前的 > 。 2. 选择“数据集版本”页签，单击数据集版本操作列的“生成新版本”，生成新的数据集版本。
修改索引配置	1. 在数据集列表中单击列表前的 > 。 2. 选择“索引配置”页签，单击索引配置操作列的“修改”，修改索引描述。
复制索引配置	1. 在数据集列表中单击列表前的 > 。 2. 选择“索引配置”页签，单击索引配置操作列的“复制”，复制当前索引配置，进行修改后，单击“保存”。

操作	步骤
删除索引配置	1. 在数据集列表中单击列表前的 > 。 2. 选择“索引配置”页签，单击索引配置操作列的“删除”，删除索引配置。

8.2 创建知识库

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

- 通过OBS接入数据时，操作账号需获得OBS（对象存储服务）只读权限，具体操作请参见**对其他账号授予桶的读写权限**。
- 需具备充足的知识库容量包资源（包含OBS存储配额和向量库存储配额，两者比例为5:1），每个租户默认1G的向量库存储配额，默认配额用完后，请参考**购买AppStage**购买知识库容量包。
- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建知识库

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 知识库”，单击“创建知识库”。

步骤2 在“创建知识库”页面，参照**表8-4**进行基础配置，并根据选择的数据来源进行源数据接入配置或添加知识数据集。

表 8-4 知识库基础配置参数说明

参数名称	参数说明
知识库名称	自定义知识库的名称，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。
知识库描述	知识库的相关信息描述。只能包含英文、中文、数字、下划线、中划线、空格及,.?:;"'：； “”’ ‘，。？ 、()（）/
RAG类型	GraphRAG类型仅支持用于“文档”类型的数据。 • VectorRAG：向量RAG，是一种结合了向量化和大语言模型的RAG技术。VectorRAG将非结构化的数据转化为结构化的向量空间，利用向量库实现高效的信息检索。 • GraphRAG：知识图谱RAG，是一种结合了知识图谱和大语言模型的RAG技术。GraphRAG能够处理各种类型的文档，从中提取实体（文档中具体的对象或概念）、关系以及文本内容构建知识图谱（一种结构化的知识表示方式），从而增强大语言模型对复杂信息的理解和推理能力。

参数名称	参数说明
数据来源	知识库的数据来源。 • 接入源数据，参考表8-5接入源数据。 • 选择知识数据集，参考表8-6添加知识数据集。
数据类型	当数据来源为“接入源数据”时，配置此参数。 支持接入的源数据类型： • 文档：支持.pdf、.txt（只支持UTF-8）、.csv（只支持UTF-8）、.xlsx、.docx、.pptx、.html、.json、.xml、.md格式，单个文件最大为10M，总上传大小最大为500M。 • 图片：支持.png、.jpg、.jpeg、.gif、.webp、.bmp格式，单张图片最大为10M，总上传大小最大为200M。 • 图片-摘要：支持本地文件上传.png、.jpg、.jpeg、.gif、.webp、.bmp格式，需对图片填写摘要信息，单张图片最大为10M，总上传大小最大为300M。 • 视频-摘要：支持本地文件上传.mp4、.webm、.wmv、.mov、.avi格式，需对视频填写摘要信息，单个视频最大为100M，总上传大小最大为300M。

表 8-5 接入源数据参数说明

参数名称	参数说明
数据接入	
接入方式	选择数据集的接入方式。支持以下两种方式： • 本地上传 • OBS接入 仅支持使用区域位置为北京四的OBS桶接入数据。
数据文件	当接入方式选择“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，支持上传的文件类型请参考表8-4的“数据类型”。
文件类型	• 当接入方式选择“本地上传”时，无需配置，根据上传的文件自动识别文件类型。 • 当接入方式选择“OBS接入”时，选择接入的文档、图片的格式。
OBS桶名	当接入方式选择“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	当接入方式选择“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
调度类型	当接入方式选择“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
版本更新模式	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：当OBS桶内内容发生变化时，调度成功后会生成一个新版本。
执行周期	当“调度类型”选择“定时调度”时，需配置此参数。 设置执行周期，支持选择为天、周。
执行时间	当“调度类型”选择“定时调度”时，需配置此参数。 ● 当执行周期为“天”时，设置每日开始执行的时间。 ● 当执行周期为“周”时，指定每周周几，并设置当日开始执行的时间。
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。
数据加工	
数据清洗（可选）	数据类型选择为“文档”、“图片-摘要”、“视频-摘要”时，显示此参数。 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
PDF预处理	当接入的文档类型为.pdf时，显示此参数。 ● 提取PDF富媒体：提取PDF文件中的富媒体（图片、表格等）。 ● 无处理
智能匹配图表	当“PDF预处理”选择为“提取PDF富媒体”时，显示此参数。 数据切分时，有助于对PDF文件中的图片、表格进行完整提取。

参数名称	参数说明
数据切分	数据类型选择为“文档”时，显示此参数。 为各文档格式选择一种切分方式，默认为自动切分。 ● 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（仅docx格式的文档支持标题切分）。 ● 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。
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。输入的分隔符不允许包含以下特殊字符*./$^?+以及带有\的分隔符，仅支持输入\n，用于表示按行分隔，其它诸如\\n，\n\n等均不支持。
包含分隔符	切分后的分段内容中是否包含分隔符，选择“包含”或“不包含”。
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。
索引配置	

参数名称	参数说明
向量化模型	选择向量化模型，向量化模型可以将文本数据转换为数值向量，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。 当前模型仅支持向量化512 token的内容，对应的中文约为512个字，英文与符号约900个字符，请注意分片长度。
长文本策略	● 默认模式：如果待向量化分片字段token长度超过了向量化模型限制的token总数，则创建知识库失败。 ● 截断模式：如果待向量化分片字段token长度超过向量化模型限制的token总数，则进行截断，取前top k个token。 ● 智能模式：如果待向量化分片字段token超过向量化模型限制的token总数，首先利用对话大模型对超长分片进行重写，如果仍然超长则进入截断模式。
检索配置	当RAG类型为“VectorRAG”时，支持配置此参数；RAG类型为“GraphRAG”时，默认为“语义检索”方式。
检索方式	当数据类型为“图片”时，不支持选择混合检索和全文检索。 ● 混合检索：同时使用语义检索与全文检索，并对结果进行综合排序。 ● 语义检索：使用向量进行文本语义查询，即调用向量数据库根据向量的相似性检索。 ● 全文检索：使用关键字进行文本匹配，适合查找一些关键词和主题语的数据。
知识图谱配置	当RAG类型为“GraphRAG”时，需要配置此参数。
实体抽取模型	选择实体抽取模型。从文本中提取实体（文档中具体的对象或概念）、关系以及文本内容，用于构建知识图谱。
实体抽取提示语	● 默认：模型按照默认的提示语要求，对文本内容进行抽取。 ● 自定义：对默认的提示语进行编辑，自定义实体抽取方式。
高级配置	开启“高级配置”开关，可以对已切分好的分片，提取相关内容作为检索字段、文本过滤字段或向量化字段。
切片提取模式	数据类型选择为“文档”时，显示此参数。 ● 智能提取：通过大语言模型提取分片内的问题与回答，效果由大模型决定。目前仅支持提取问题和答案，且耗时较长。 ● 规则提取：按照用户配置的提取规则对切片内容进行提取。
切片内容示例	数据类型选择为“文档”时，显示此参数。 切片内容示例，用以测试切片提取效果。

参数名称	参数说明
切片提取配置	切片提取模式为“规则提取”时，配置此参数。 单击“添加提取字段”，配置如下参数： ● 切片片段名称：自定义切片片段名称。 添加的提取字段数量不超过10个，切片片段名称长度最多为20个字符，不允许重复，不允许为以下名称（大小写不敏感）：file_name、file_id、path、order、document、base64、chunk，不能以ki_、ko_开头，仅可包含字母、数字、下划线，并且以字母开头。 ● 提取规则：分隔符提取、按规则提取。 ● 提取设置 – 当提取规则为“分隔符提取”时，例如：以分隔符**;**对切片进行分段，选择第**2**个分段，片段中**不包含**分隔符。输入的分隔符不允许包含以下特殊字符*./$^?+以及带有\的分隔符，仅支持输入\n，用于表示按行分隔，其它诸如\\n，\n\n等均不支持。 在填写“选择第*{{片段序号}}*个片段”时，如果此处的片段序号超出片段数量，则提取内容为空。 – 当提取规则为“按规则提取”时，例如：从切片提取以**项目名称：**开头和以（**完**）结尾的片段，选择第**3**个片段，片段中**包含**开头，**不包含**结尾。
切片提取效果	数据类型选择为“文档”时，显示此参数。 展示对切片内容提取后的效果。
索引字段配置	● 当RAG类型为“VectorRAG”时，配置如下参数： – 检索字段：选择对哪个字段进行检索，并作为检索命中内容返回。 – 文本过滤字段：选择哪些字段支持在检索时同步进行文本过滤，并返回该字段的内容。 – 附加返回字段：选择哪些字段在检索时附加返回。 ● 当RAG类型为“GraphRAG”时，配置如下参数： – 实体抽取字段：选择对哪个字段进行实体抽取。选择从chunk抽取实体时，chunk是一个完整切片，是指从这个切片的内容里面去抽取实体。 – 附加返回字段：选择哪些字段在检索时附加返回。

表 8-6 选择知识数据集参数说明

参数名称	参数说明
数据集选择	单击“选择知识数据集”，在“选择知识数据集”面板，勾选目标数据集，并选择数据集版本及索引配置。如果当前数据集未创建索引配置，可以单击索引配置下拉框中的“创建索引配置”，参考**创建索引配置**进行创建。 您也可以单击“创建知识数据集”，参考**创建知识数据集**创建新的数据集。 说明： ● 每个VectorRAG知识库最多只能添加5个数据集，且添加的数据集必须为同一类型。 ● 每个GraphRAG知识库只能添加1个数据集。

步骤3 单击“保存”，保存知识库的参数配置。

单击“保存并启用”，创建知识库完成并启用该知识库。

----结束

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，单击操作列的“命中测试”。
2. 在命中测试页面，配置测试输入参数，参数说明如**表8-7**所示，配置完单击“测试”。

表 8-7 测试输入参数配置

参数	说明
检索内容	输入测试文本。
选择数据集	选择待测试的数据集。
相似度阈值	取值范围为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	查询最大返回数量。

3. 在“测试结果”区域查看测试效果，测试结果根据相似度从大到小进行排序。
4. 在“测试历史”区域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。

更多操作

创建知识库完成后，可执行如下**表8-8**所示的管理知识库相关操作。

表 8-8 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基础信息、调度历史，并支持进行知识库溯源。
修改知识库	不支持修改已启用的知识库，如需删除，请先进行停用。 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、知识数据集及检索配置等。
删除知识库	不支持删除已启用的知识库，如需删除，请先进行停用。 1. 在知识库列表中，单击“操作”列的“删除”。 2. 在弹出的提示框中，选择删除范围。 • 删除知识库和数据集：删除知识库，同时删除知识库添加的数据集。 • 删除知识库：仅删除知识库。
启用知识库	在知识库列表中，对于已停用的知识库，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Agent**时引用。
停用知识库	在知识库列表中，对于已启用的知识库，可在“操作”列单击“停用”将其暂停使用。
执行知识库调度	当自动调度失败或知识库配置更改时，可以手动执行调度，更新知识库。当数据集更新时，系统会自动执行知识库调度，无需手动执行调度。 在知识库列表中，选择“操作”列的“更多 > 执行”，手动执行知识库调度。

8.3 创建知识检索流

知识检索流是一种特殊的工作流，除工作流基础节点外，还具备检索规划、召回、重排序三个节点。

通常可以使用检索规划节点对原始查询内容进行意图识别、拆解或改写，提升查询的准确性，然后使用召回节点从知识库中检索并召回所有与查询相关的信息，最后通过重排序节点对召回结果进行重排序，确保最相关的信息能够排在前面，从而优化知识检索过程。

在Agent中调用知识检索流，可以提升用户体验和Agent响应质量。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建知识检索流

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 知识检索流”，单击“创建知识检索流”。

步骤2 在“基本信息”弹窗，设置检索流名称、描述，单击“确认”。

步骤3 单击“起始节点”，配置表8-9所示参数。

当接收到匹配的请求时，起始节点会解析请求中的参数，并根据这些参数来初始化检索流实例。

起始节点是流的入口，可以接受和存储初始化参数，当Agent调用检索流时，会提取起始节点入参，并初始化检索流实例，随后按照检索流定义的逻辑顺序执行各个节点。

表 8-9 起始节点配置参数说明

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：用于从服务器获取数据，通常使用URL参数传递数据。 • post：用于向服务器提交数据，通常将数据放在请求体中。 • delete：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：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。 • patch：请求服务器更新资源的部分内容。当资源不存在的时候，patch可能会去创建一个新的资源。
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。 知识检索流默认在起始节点的请求体中引入了WISEAGENT_USER_INPUT参数，表示在Agent调用知识检索流时，以用户在问答对话中输入的内容作为知识检索流的请求参数。 您也可以单击请求体操作列的，新增参数。支持数据类型为：string、number、boolean、integer、array、object。
节点备注		输入节点备注信息，方便后续查阅节点功能。

步骤4 添加其他节点，设置执行动作。

单击“添加执行动作”，选择节点和执行动作，根据**知识检索流相关节点说明**和**工作流基础节点说明**配置节点参数。

步骤5 （可选）单击起始节点，在界面参数配置面板中单击“设置参数”，输入参数，用于调测知识检索流。

步骤6 （可选）单击其他后续节点，在界面参数配置面板中单击“调测节点”，对当前节点进行正确性测试。调测成功后，会将测试的输出数据（即样本数据）及输入数据进行展示，并会在该条节点的左上角标记图标。如果提示“调测失败，请检查接口参数配置是否准确”，请检查并重新配置参数后重试。

步骤7 执行动作设置完成后，单击“保存”。

步骤8 在“流保存成功”弹框中选择是否开启检索流。

- 单击“确定”，立即开启流，启用后的检索流才可在创建Agent时引用。
- 单击“取消”，暂不开启，如有需要可参考**更多操作**开启检索流。

----结束

知识检索流相关节点说明

本节仅介绍知识检索流特有的检索规划、召回及重排序三个节点，工作流其他基础节点说明请参见**工作流基础节点说明**。

- 检索规划

  检索规划包含“Query拆解”、“Query改写”、“意图识别”三个执行动作，执行动作参数配置说明如**表8-10**所示。

  - Query拆解：配置适当的模型将原始查询内容拆解为更简单、易理解的请求。
  - Query改写：配置适当的模型对原始查询内容进行改写、优化，使得原始请求更准确。
  - 意图识别：配置适当的模型对原始内容进行意图判断。

表 8-10 检索规划执行动作参数说明

参数		说明
输入	模型服务调用ID	执行检索规划所调用的模型。 • 对于资产中心预置的模型，在资产中心选择“大模型”页签，单击模型卡片进入模型详情页面，查看模型服务调用ID。 • 对于我的模型（我部署的、我接入的）和我的路由策略，需要填写模型服务调用ID，请单击“获取模型服务调用ID”，进入“我的模型服务”页面，在模型服务列表中单击复制。
	原始查询内容	当执行动作为“拆解/改写”时，需要配置此参数。 表示待处理（拆解/改写）的原始内容。
	原始文本内容	当执行动作为“意图识别”时，需要配置此参数。 表示待进行意图识别的原始内容。

参数		说明
	意图类别	当执行动作为“意图识别”时，需要配置此参数。 定义意图类别，大模型会按照定义的意图类别对用户问题进行归类。 支持自定义数组或选择数组类型的节点输出，例如：["儿科医疗问题","消化科医疗问题"]。
输出	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	实例是节点的鉴权方式，如果未新增实例，节点就无法调通。 ● 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。 ● API Key：必填项，具体介绍请参见**创建API Key**。	
节点备注	输入节点备注信息，方便后续查阅节点功能。	

- 召回

  召回节点包含“召回”一个执行动作，用于从知识库中检索并召回所有与查询相关的信息，参数配置说明如**表8-11**所示。

  表 8-11 召回执行动作参数说明

参数		说明
输入	知识库ID	请单击“获取知识库ID”，进入知识库列表，单击 复制。
	向量化检索内容	向量化检索的内容。
	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。
	召回数量	从检索结果中返回的内容片段数量，取值范围：0~10。

参数		说明

	文本过滤	过滤条件。默认为空，支持填入SearchSqlFilter类对象，SearchSqlFilter参数说明如**表8-12**所示，样例如下：
		{
		"group_type": "OR",
		"expressions": [
		{
		"field": "metadata.file_name",
		"field_type": "STRING",
		"operator": "EQUAL",
		"values": [
		"四大名著介绍.txt"
		]
		},
		{
		"field": "metadata.path",
		"field_type": "STRING",
		"operator": "EQUAL",
		"values": [
		"四大名著介绍.txt"
		]
		}
		]
		}

	召回排序	排序规则。默认为空，支持填入SqlOrder类对象，SqlOrder参数说明如**表8-14**所示，样例如下：
		{
		"order_items": [
		{
		"field": "metadata.order",
		"field_type": "INT",
		"order_type": "DESC"
		}
		]
		}

输出	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	实例是节点的鉴权方式，如果未新增实例，节点就无法调通。 • 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。 • API Key：必填项，具体介绍请参见**创建API Key**。	
节点备注		输入节点备注信息，方便后续查阅节点功能。

表 8-12 SearchSqlFilter

参数	是否必选	参数类型	描述
group_type	否	String	**参数解释：** 过滤条件运算符。 **约束限制：** 只有一个expression时，不需要group_type，group_type可以为null。 **取值范围：** 可以为null，如果不为null，枚举值AND和OR。 **默认取值：** 不涉及。
expressions	否	Array of **Expression** objects	**参数解释：** 过滤条件。 **约束限制：** 不涉及。 **取值范围：** 非空，条件数量介于1到10之间。 **默认取值：** 不涉及。

表 8-13 Expression

参数	是否必选	参数类型	描述
field	否	String	**参数解释：** 过滤字段。 **约束限制：** 不涉及。 **取值范围：** 非空，字符串长度介于1到100之间。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
field_type	否	String	**参数解释：** 过滤字段类型。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：INT、FLOAT、BOOLEAN和STRING。 **默认取值：** 不涉及。
operator	否	String	**参数解释：** 过滤操作符。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：EQUAL、NOT_EQUAL、GREAT_THAN、GREAT_EQUAL、LESS_THAN、LESS_EQUAL、IN、NOTIN和STARTS_WITH。 **默认取值：** 不涉及。
values	否	Array of strings	**参数解释：** 过滤值。 **约束限制：** 不涉及。 **取值范围：** 非空，数量介于1到100之间，每个字符串长度最大不超过2000。 **默认取值：** 不涉及。

表 8-14 SqlOrder

参数	是否必选	参数类型	描述
order_items	否	Array of **OrderItem** objects	**参数解释：** 排序规则。 **约束限制：** 不涉及。 **取值范围：** 非空，数量介于1到10之间。 **默认取值：** 不涉及。

表 8-15 OrderItem

参数	是否必选	参数类型	描述
field	否	String	**参数解释：** 排序字段。 **约束限制：** 不涉及。 **取值范围：** 非空，字符串长度介于1到100之间。 **默认取值：** 不涉及。
field_type	否	String	**参数解释：** 排序字段类型。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：INT、FLOAT、BOOLEAN和STRING。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
order_type	否	String	**参数解释：** 排序类型。 **约束限制：** 不涉及。 **取值范围：** 不为null，枚举值：ASC（升序）和DESC（降序）。 **默认取值：** 不涉及。

- 重排序

  重排序节点包含“重排序”一个执行动作，用于对召回结果进行重排序，参数配置说明如表8-16所示。

  表 8-16 重排序执行动作参数说明

参数		说明
输入	模型服务调用ID	执行重排序所调用的模型，例如平台预置的bge-reranker-large。 • 对于资产中心预置的模型，在资产中心选择“大模型”页签，单击模型卡片进入模型详情页面，查看模型服务调用ID。 • 对于我的模型（我部署的、我接入的）和我的路由策略，需要填写模型服务调用ID，请单击“获取模型服务调用ID”，进入“我的模型服务”页面，在模型服务列表中单击 复制。
	原始查询内容	用户的请求内容。
	排序返回数量	返回排名前n个文档。
	召回结果	召回的结果，提供给模型进行重排序。
输出	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	实例是节点的鉴权方式，如果未新增实例，节点就无法调通。 • 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。 • API Key：必填项，具体介绍请参见创建API Key。	
节点备注	输入节点备注信息，方便后续查阅节点功能。	

测试检索流

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 在检索流列表中，单击检索流名称，进入检索流查看页面。

步骤2 单击“测试”，在测试面板，输入参数，单击“提交测试”。

步骤3 查看测试结果。

步骤4 单击“查看运行历史详情”，在运行详情页面，查看本次测试过程中检索流的运行总次数、成功次数、失败次数，以及各节点的执行时长、状态、参数信息等，方便定位问题。

----结束

更多操作

检索流创建完成后，可执行如下表8-17所示的相关操作。

表 8-17 管理检索流

操作	说明
启用检索流	在检索流列表中，对于“已停用”状态的检索流，在操作列单击“启用”，启用后的检索流才可在创建Agent时引用。
停用检索流	在检索流列表中，对于“已启用”状态的检索流，可在操作列单击“停用”。
查看检索流详情	在检索流列表中单击检索流名称，查看检索流最近一次运行预览图、基本信息、运行历史及历史版本。 ● 最近一次运行预览图：在检索流详情页面左侧区域，展示最近一次流运行预览图，单击页面右上角“编辑”，可进入检索流编辑页面。 ● 基本信息：显示检索流的基本信息，包括名称、状态、调用地址等。 ● 运行历史：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：展示检索流的历史版本，最多20条。 – 选择历史版本操作列的“更多 > 保存为最新版本”，会产生一条新的记录并应用于当前的流。如果流已经被其他应用调用，请谨慎操作。 – 选择历史版本操作列的“更多 > 删除”，删除历史版本。 – 单击历史版本操作列的“编辑”，编辑工作流，保存后会产生一条新的记录并应用于当前的流。如果流已经被其他应用调用，请谨慎操作。
修改检索流	在检索流列表中，单击操作列的“修改”，单击检索流名称后的，可修改检索流名称、描述，并支持增加、删除节点以及修改执行动作参数等。
删除检索流	已启用的检索流需要先停用，才可删除。 在检索流列表中，选择操作列的“更多 > 删除”，在弹出的确认框中单击“确认”。

操作	说明
复制检索流	在检索流列表中，选择操作列的“更多 > 复制”，在弹出的复制流提示框中单击“确认”。

9 管理工具

9.1 创建工具

工具是一组相关的API集合，一个工具通常包含多个执行动作，每个执行动作用于实现特定功能。在创建Agent时调用工具，可以对Agent进行能力扩展。

平台在资产中心预置了部分工具，同时也支持用户根据需求自定义创建工具。在创建工具时，需要先将选定的API服务注册为一个工具，然后再添加该服务下的API作为工具的执行动作。

前提条件

- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。
- 已获取API服务的连接信息以及调用API所需的访问权限和密钥。

创建工具

步骤1 在AI原生应用引擎的左侧导航栏选择“Agent编排中心 > 我的工具”，单击“创建工具”。

步骤2 在“创建工具”页面，配置如**表9-1**所示参数。

表 9-1 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。名称长度不能超过32个字符，可包含中文、大小写字母、数字、下划线、中划线、英文小括号，不能以下划线、中划线、英文小括号开头。
描述	填写工具功能或作用等描述。
图标	支持选择系统图标。

参数名称	参数说明
协议	API服务接口通信协议。 • https • http
主机地址	提供API服务的服务器地址。以https://aiae.appstage.myhuaweicloud.com/v1/chat/completions为例，主机地址为aiae.appstage.myhuaweicloud.com。
基准URL	即Base URL，域名的根路径，默认为/。必须以/开头，且不能有连续多个/，不包含除/._-以外的特殊字符和空格。 以https://aiae.appstage.myhuaweicloud.com/v1/chat/completions为例，基准URL为可以填写为/、/v1、/v1/chat或/v1/chat/completions。原则上基准URL+执行动作中填写的接口路径拼接起来为完整的/v1/chat/completions即可。
验证方式	API的验证方式。 • 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3 配置完单击“创建”。

工具创建成功后，进入工具详情页面，请参考**创建执行动作**添加执行动作。

----结束

创建执行动作

步骤1 在工具详情页面，单击“创建执行动作”，配置执行动作基础信息，参数如**表9-2**所示。

表 9-2 执行动作基本信息参数说明

参数	说明
名称	执行动作是需要完成的特定任务，自定义执行动作的名称，比如，“发送电子邮件”。“更新行”。 长度不能超过64个字符，可包含中文、大小写字母、数字及下划线、中划线、英文小括号，不能以下划线、中划线、英文小括号开头。
英文名称	执行动作的英文名称。
类型	默认为API，表示通过调用API的方式创建执行动作。
可见性	• 可见 • 隐藏 设置为隐藏的执行动作，在流编排中将不可见。
描述	执行动作的描述信息。

步骤2 单击“下一步”，配置输入，参数如表9-2所示，配置完成后单击“下一步”。

表 9-3 输入参数说明

参数	说明
接口路径	API的请求路径。必须以/开头，且不能有连续多个/，不包含除/_-:@%+.~#?&={}[]()、$以外的特殊字符和空格。 以https://aiae.appstage.myhuaweicloud.com/v1/chat/completions为例，接口路径可以填写为v1/chat/completions、/chat/completions、/completions或不填。原则上基准URL+执行动作中填写的接口路径拼接起来为完整的/v1/chat/completions即可。

参数	说明
输入参数	API的请求参数，如果被调用API没有请求参数可不填。 • 请求头（Header）：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。 单击参数列表“操作”列的可以新增参数，参数配置说明请参见表9-4。 • 请求体（Body）：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。 – JSON/XML：JSON、XML格式的数据。 参数列表“操作”列的：可选择是否开启“是否支持根节点输入”参数，开启并发布对应工具后，在创建流中添加该执行动作时可以自定义请求体参数的值。 参数列表“操作”列的：新增参数，参数配置说明请参见表9-4。 导入：可直接粘贴被调用API的JSON、JSON Schema或XML数据，减少逐个配置参数的工作量。导入文件示例请参见JSON Schema/JSON/XML文件示例。 如果请求体使用XML格式，XML header参数必须配置。 复制：复制请求体参数的JSON或XML数据。 预览：可以预览参数的JSON或XML结构。 – form-data：文件格式数据。 参数列表“操作”列的：新增参数，参数配置说明请参见表9-4。 – Binary：文件格式数据。输入、输出仅支持配置一处。例如，输出入参数选择了“Binary”，则输出不显示，反之亦成立。 • 查询参数（Query）：查询参数会追加到URL。例如，在 /items?id=#### 中，查询参数为ID。 单击参数列表“操作”列的可以新增参数，参数配置说明请参见表9-4。 • 路径参数（Path）：路径与路径模板一起使用，其中参数值实际上是操作URL的一部分。

表 9-4 参数配置说明

参数	说明
参数名称	输入参数的名称。
显示字段	用户在表单中看到的参数项标签。

参数	说明
必填	勾选该参数是否是用户必填项。
参数类型	选择参数类型，支持string、number、boolean、integer类型。 此外，当请求体（Body）为JSON或XML时，还支持array、object类型参数；当请求体（Body）为form-data时，还支持file类型参数。
说明	关于输入值的介绍说明。
操作	• 单击  ：配置参数在界面的显示样式。配置完成并发布对应工具后，在创建流中添加该执行动作时可以查看参数的界面显示效果。 – 参数类型：选择参数类型。 – 格式：设置用户输入该参数时界面显示的样式，可选择文本框、日期时间、富文本、下拉列表。不同的参数类型可选择的格式不同。 当格式选择为下拉列表时，需设置标签和下拉列表的值。单击“新增”可进行添加。 设置完成后，如果勾选了“支持多选”，则可在创建流中添加该执行动作时选择多个值，否则，只能选择一个值。需要配置分隔符，分隔符只能输入单个特殊字符，如“,”“$”、“%”、“^”、“&”等，不设置时，默认为“,”。 – 可见性：设置用户输入该参数时界面显示的可见性。 无：在流中正常显示。 高级：默认隐藏在高级设置菜单里。 隐藏：该参数向用户隐藏。 重要：优先显示在表单的最开始。 – 默认值：当格式选择为文本框、日期时间、富文本时，支持设置该参数的预设值。 • 单击  ：新增节点。 • 单击  ：删除该节点。

步骤3 单击“下一步”，配置输出参数，参数如表9-2所示。

表 9-5 输出参数说明

参数	说明
添加响应	单击“添加响应”，根据被调用API的响应码信息添加响应码。

参数	说明
输出参数	配置输出参数，如被调用API没有响应参数可不填。 ● 响应体：HTTP响应消息的组成部分之一，响应体呈现发送给服务器的数据。 – JSON/XML：JSON、XML格式的数据。 参数列表“操作”列的：新增参数，参数配置说明请参见表9-4。 当选择XML格式时，单击参数列表“操作”列的，配置XML标签名， 导入：可直接粘贴被调用API的JSON Schema、JSON或XML数据，减少逐个配置参数的工作量。导入文件示例请参见JSON Schema/JSON/XML文件示例。 如果响应体使用XML格式，XML header参数必须配置。 复制：复制请求体参数的JSON或XML数据。 预览：可以预览参数的JSON或XML结构。 – Binary：文件格式数据。输入、输出仅支持配置一处。例如，输出入参数选择了“Binary”，则输出不显示，反之亦成立。 ● 响应头：HTTP响应消息的组成部分之一，响应头负责通知服务器有关于客户端请求的信息。 单击参数列表“操作”列的可以新增参数，参数配置说明请参见表9-4。

步骤4 单击“下一步”，调试校验工具，验证工具是否可用。

1. 配置用例设置参数，参数说明如表9-6所示。配置完成后，单击“提交测试”。

表 9-6 配置用例设置参数说明

参数	说明
实例	选择已创建好的实例。 也支持新建实例，单击“新建实例”，配置实例名称、描述及验证信息，验证信息填写工具创建时所配置的鉴权信息，单击“保存”。 验证信息与步骤3所选的验证方式相关，如果验证方式为“无验证”，则无需配置实例。
定义参数	配置输入参数。

2. 在测试结果预览区域，查看测试结果。

步骤5 工具调试完成后，单击“保存”。

新创建的工具显示在“我的工具”列表中，任务状态为“待上架”，请参考**表9-9**上架工具。

----结束

JSON Schema/JSON/XML 文件示例

- JSON Schema

```
{
  "properties": {
    "str": {
      "description": "",
      "default": "",
      "x-hw-default": "",
      "type": "string",
      "x-hw-label": "",
      "x-hw-visibility": "none",
      "format": "input",
      "x-hw-format": "input",
      "x-hw-select-options": []
    },
    "obj": {
      "description": "",
      "default": "",
      "x-hw-default": "",
      "type": "object",
      "x-hw-label": "",
      "x-hw-visibility": "none",
      "format": "input",
      "x-hw-format": "input",
      "x-hw-select-options": [],
      "properties": {
        "obj_str1": {
          "description": "",
          "default": "",
          "x-hw-default": "",
          "type": "string",
          "x-hw-label": "",
          "x-hw-visibility": "none",
          "format": "input",
          "x-hw-format": "input",
          "x-hw-select-options": []
        },
        "obj_str2": {
          "description": "",
          "default": "",
          "x-hw-default": "",
          "type": "string",
          "x-hw-label": "",
          "x-hw-visibility": "none",
          "format": "input",
          "x-hw-format": "input",
          "x-hw-select-options": []
        }
      },
      "required": [
        "obj_str1",
        "obj_str2"
      ]
    },
```



```
        "arr": {
            "description": "",
            "default": "",
            "x-hw-default": "",
            "type": "array",
            "x-hw-label": "",
            "x-hw-visibility": "none",
            "format": "input",
            "x-hw-format": "input",
            "x-hw-select-options": [],
            "items": {
                "description": "",
                "default": "",
                "x-hw-default": "",
                "type": "string",
                "x-hw-label": "",
                "x-hw-visibility": "none",
                "format": "input",
                "x-hw-format": "input",
                "x-hw-select-options": []
            }
        }
    },
    "required": [
        "str",
        "obj",
        "arr"
    ],
    "type": "object"
}
```

- JSON

```
{
    "str": "string",
    "obj": {
        "obj_str1": "string",
        "obj_str2": "string"
    },
    "arr": [
        "string", "string"
    ]
}
```

- XML

```
<root>
    <str>string</str>
    <arr>string</arr>
    <arr>string</arr>
    <arr>string</arr>
    <obj>
        <obj_str1>string</obj_str1>
        <obj_str2>string</obj_str2>
    </obj>
</root>
```

更多操作

工具创建完成后，您可以执行如**表9-7**的操作。

表 9-7 相关操作

操作	说明
设置工具鉴权	在工具列表中，单击操作列的“设置鉴权”，设置鉴权信息，单击“保存”。只有经过身份验证和授权的用户才能使用工具。
申请上架工具	将工具上架至资产中心，具体操作请参见**将创建的工具上架到资产中心**。
创建执行动作	在工具列表中，单击工具名称，在工具详情页面创建执行动作，具体操作请参见**创建执行动作**。
下载工具	在工具列表中，选择操作列的“更多 > 下载”，下载工具的json格式文件。
导入更新工具	在工具列表中，选择操作列的“更多 > 导入更新”，以导入json格式API文件的方式更新工具。
编辑工具	已上架的工具编辑后，需要重新上架，资产中心的工具才能更新生效。 1. 在工具列表中，单击操作列的“修改”，支持编辑工具的名称、描述、图标、协议、主机地址、基准URL以及验证方式。 2. 单击“更新”。
删除工具	1. 在工具列表中，选择操作列的“更多 > 删除”。 2. 在弹出的提示框中单击“确认”。
编辑工具的执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“编辑”。
测试执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“测试”。
删除执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“删除”。

9.2 导入工具

创建工具时，对于复杂的API服务，每个API都需要手动添加和配置，会导致操作量大且容易出错。平台提供了导入OpenAPI文件的功能，以减少手动创建工具的工作量。

前提条件

- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。
- 已获取API服务的json文件。

创建工具

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

步骤2 在我的工具页面，单击“导入工具”。

步骤3 选择本地json文件进行上传。

上传完成后，新创建的工具显示在“我的工具”列表中。

----结束

更多操作

工具创建完成后，您可以执行如**表9-8**的操作。

表 9-8 相关操作

操作	说明
设置工具鉴权	在工具列表中，单击操作列的“设置鉴权”，设置鉴权信息，单击“保存”。只有经过身份验证和授权的用户才能使用工具。
申请上架工具	将工具上架至资产中心，具体操作请参见**将创建的工具上架到资产中心**。
创建执行动作	在工具列表中，单击工具名称，在工具详情页面创建执行动作，具体操作请参见**创建执行动作**。
下载工具	在工具列表中，选择操作列的“更多 > 下载”，下载工具的json格式文件。
导入更新工具	在工具列表中，选择操作列的“更多 > 导入更新”，以导入json格式API文件的方式更新工具。
编辑工具	已上架的工具编辑后，需要重新上架，资产中心的工具才能更新生效。 1. 在工具列表中，单击操作列的“修改”，支持编辑工具的名称、描述、图标、协议、主机地址、基准URL以及验证方式。 2. 单击“更新”。
删除工具	1. 在工具列表中，选择操作列的“更多 > 删除”。 2. 在弹出的提示框中单击“确认”。
编辑工具的执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“编辑”。
测试执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“测试”。
删除执行动作	在工具列表中，单击 > ，展开执行动作列表，单击执行动作列表操作列的“删除”。

9.3 将创建的工具上架到资产中心

工具创建完成后，可以将工具上架至资产中心，其他租户或资源相互隔离的部门收藏工具后，可便捷地进行使用。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

上架工具

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

步骤2 在我的工具列表中，选择操作列的“更多 > 申请上架”。

步骤3 配置上架信息，参数如**表9-9**所示。

表 9-9 工具上架参数说明

参数	说明
工具名称	自定义工具上架后的名称。默认显示工具在创建时定义的名称，可根据需要修改。 名称长度不能超过32个字符，可包含中文、大小写字母、数字、下划线、中划线及英文小括号，开头不能是下划线、中划线、英文小括号。
工具描述	输入工具描述信息。
工具鉴权获取地址	鉴权信息获取地址。三方API请到对应官网获取，个人API请留联系方式，方便用户联系获取。
执行动作	显示工具的执行动作及动作描述信息。

步骤4 单击“确定”。

上架申请提交后，工具上架状态为“审批中”，请等待平台运营者审批。您可以在“我的工具”列表中，选择操作列的“更多 > 审核详情”，查看审批日志。

- 审批通过后，工具上架状态为“已上架”，在AI原生应用引擎“资产中心”的“工具”页签下可以查看到该工具。
- 审批未通过时，工具上架状态为“已驳回”。

----结束

下架工具

步骤1 在工具列表中，选择已上架工具操作列的“更多 > 申请下架”，即可提交工具下架申请。

步骤2 待平台运营者审核通过后，工具会从资产中心下架。

----结束

9.4 收藏上架的工具

其他租户或资源相互隔离的部门，如需使用资产中心上架的工具，需要先收藏到“我的工具”列表中，方可便捷地进行使用。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

收藏其他租户上架的工具

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“工具”页签。

步骤3 在左侧“筛选”区域，可以按照行业、类型、其他三个维度快速筛选查找工具。

步骤4 鼠标光标移至工具卡片上，单击卡片右上角  。

单击工具卡片右上角的  ，可以取消收藏。

收藏后的工具在“Agent编排中心 > 我的工具”页面的工具列表中展示。

----结束

9.5 调用资产中心工具前设置认证鉴权

资产中心展示了平台预置的第三方厂商工具以及租户上架的工具，这些三方工具可以在创建Agent时进行便捷调用。

如果三方工具在创建时设置了鉴权信息，在调用前还需要配置认证鉴权。

前提条件

- 需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。
- 已获取三方工具的鉴权信息。
  - 对于第三方厂商工具，需要在该厂商的官网进行购买或注册，以获取鉴权信息。
  - 对于租户上架的工具，可以在设置鉴权过程中，根据界面提示进行获取。

设置鉴权信息

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“工具”页签。

步骤3 鼠标光标移至工具卡片上，单击“设置鉴权”。

步骤4 在“设置鉴权信息”弹框中，输入鉴权信息，单击“保存”。

对于租户上架的工具，设置鉴权信息弹框中通常会展示工具鉴权获取地址，请根据界面提示进行获取。

鉴权信息与工具创建时所采用的验证方式有关，一般常见的有appkey、X-API-Key等。

步骤5 对于已设置鉴权的工具，在“设置鉴权信息”对话框，单击“移除”。

移除鉴权信息后将影响该工具的调用，需重新设置才能进行调用。

----结束

10 管理工作流

10.1 创建工作流

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，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。工作流可在用户创建Agent时调用，Agent使用过程中，当起始节点触发，后续动作即可自动执行，完成一系列复杂的任务。

- 创建Agent（LLM模式）时，由大模型根据用户问题与工作流的关联性决策是否调用工作流。
- 创建的Agent（工作流模式）时，用户与工作流进行对话，每次对话都会调用工作流。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建工作流

步骤1 在AI原生应用引擎的左侧导航栏，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”。

步骤3 在“基本信息”对话框，设置工作流名称、描述，单击“确认”，进入工作流构建页面。

构建页面的画布中默认包含起始节点 ，起始节点用于启动工作流。

步骤4 配置起始节点，具体配置说明请参见**起始节点**。

步骤5 添加其他节点和执行动作。

1. 在画布中单击“  > 添加执行动作”或“添加执行动作”。
2. 在弹框中选择基础节点或工具节点，选择节点的执行动作，各节点的详细介绍请参见**工作流基础节点说明**和**工作流工具节点说明**。
   - 基础节点：工作流系统通常会提供一系列基础节点，如开始节点、结束节点、任务节点等。这些节点构成了工作流的基本框架。

- 工具节点：工具节点用于执行特定任务，可以是自定义的，也可以是系统提供的，用于实现特定的业务逻辑或功能。

图 10-1 添加节点

步骤6 配置其他节点输入、输出等参数，各节点参数说明请参见**工作流基础节点说明**和**工作流工具节点说明**。

步骤7 （可选）单击起始节点，在界面参数配置面板中单击“设置参数”，输入参数，用于调测工作流。

步骤8 （可选）单击其他后续节点，在界面参数配置面板中单击“调测节点”，对当前节点进行正确性测试。调测成功后，会将测试的输出数据（即样本数据）及输入数据进行展示，并会在该条节点的左上角标记图标。如果提示“调测失败，请检查接口参数配置是否准确”，请检查并重新配置参数后重试。

步骤9 配置完成后，单击“保存”。

步骤10 在“流保存成功”弹框中选择是否开启工作流。

- 单击“确定”，立即开启流，启用后的工作流才能进行调测以及在创建Agent时调用。
- 单击“取消”，暂不开启，如有需要可参考**更多操作**开启工作流。

----结束

测试工作流

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 在工作流列表中，单击工作流名称，进入工作流查看页面。

步骤2 单击“测试”，在测试面板，输入参数，单击“提交测试”。

步骤3 查看测试结果。

步骤4 单击“查看运行历史详情”，在运行详情页面，查看本次测试过程中工作流的运行总次数、成功次数、失败次数，以及各节点的执行时长、输入参数及输出参数等，方便定位问题。

----结束

画布操作说明

工作流构建过程中，画布中可以执行的操作如表10-1所示。

表 10-1 画布操作说明

操作	说明
删除节点	不支持删除起始节点。 鼠标光标移至节点上，单击，删除节点。
复制节点	鼠标光标移至节点上，单击“ ··· > 复制节点”，选择待粘贴的节点位置，单击“ > 粘贴节点”。
剪切节点	鼠标光标移至节点上，单击“ ··· > 剪切节点”，选择待粘贴的节点位置，单击“ > 粘贴节点”。 如果将当前流中剪切的节点进行跨流粘贴时，该节点在原流中不会被删除，而是相当于执行了一个跨流的复制操作。
修改基本信息	鼠标光标移至节点上，单击“ ··· > 修改基本信息”，支持编辑节点名称及备注。
撤销操作	单击画布中的 ，撤销操作，最多可撤销10步。
撤销恢复	单击画布中的 ，恢复撤销操作，最多可恢复10步。

更多操作

工作流创建完成后，可执行如下表10-2所示的相关操作。

表 10-2 管理工作流

操作	说明
启用工作流	在工作流列表中，对于“已停用”状态的工作流，在操作列单击“启用”，启用后的工作流才可在**创建Agent**时引用。
停用工作流	在工作流列表中，对于“已启用”状态的工作流，可在操作列单击“停用”。
查看工作流详情	在工作流列表中单击工作流名称，查看工作流最近一次运行预览图、基本信息、运行历史及历史版本。 • 最近一次运行预览图：在工作流详情页面左侧区域，展示最近一次流运行预览图，单击页面右上角“编辑”，可进入工作流编辑页面。 • 基本信息：展示工作流的基本信息，包括名称、启用状态、最近运行状态及调用地址等。 • 运行历史：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。在运行历史列表中，单击触发时间，进入运行详情页面，查看本次测试过程中工作流的运行总次数、成功次数、失败次数，以及各节点的执行时长、输入参数及输出参数等。 • 历史版本：展示工作流的历史版本，最多20条。 – 选择历史版本操作列的“更多 > 保存为最新版本”，会产生一条新的记录并应用于当前的流。如果流已经被其他应用调用，请谨慎操作。 – 选择历史版本操作列的“更多 > 删除”，删除历史版本。 – 单击历史版本操作列的“编辑”，参考**创建工作流**编辑工作流，保存后会产生一条新的记录并应用于当前的流。如果流已经被其他应用调用，请谨慎操作。
修改工作流	在工作流列表中，单击操作列的“修改”，单击检索流名称后的，可修改工作流名称、描述，并支持增加、删除节点以及修改执行动作参数等。
删除工作流	已启用的工作流需要先停用，才可删除。 在工作流列表中，选择操作列的“更多 > 删除”，在弹出的确认框中单击“确认”。
复制工作流	在工作流列表中，选择操作列的“更多 > 复制”，在弹出的复制流提示框中单击“确认”。

10.2 工作流基础节点说明

10.2.1 起始节点

起始节点是工作流的入口，可以接受和存储初始化参数，当Agent调用工作流时，会提取起始节点入参，并初始化工作流实例，随后按照工作流定义的逻辑顺序执行各个节点。

请参考表10-3配置起始节点参数，配置完成后可以单击“设置参数”，对当前节点进行正确性测试。调测成功后，会将测试的输出数据（即样本数据）及输入数据进行展示，并会在该条节点的左上角标记图标。如果提示“调测失败，请检查接口参数配置是否准确”，请检查并重新配置参数后重试。

表 10-3 起始节点配置参数说明

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。 ● patch：请求服务器更新资源的部分内容。当资源不存在的时候，patch可能会去创建一个新的资源。
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。

参数		说明
	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。 • 选择“引入更多 > 引入用户对话输入”，新增默认的WISEAGENT_USER_INPUT参数，表示在Agent调用工作流时，以用户在问答对话中输入的内容作为工作流的请求参数。在**创建Agent（工作流模式）**时添加的工作流，必须引入用户对话输入。 • 选择“引入更多 > 引入历史对话”，新增默认的WISEAGENT_CONVERSATION参数，表示在Agent调用工作流时，引入用户与Agent的历史对话内容作为工作流的请求参数。 • 选择“引入更多 > 引入变量”，新增默认的WISEAGENT_VARIABLES参数，可以添加一般变量和敏感变量，同时，需要在**创建Agent（LLM模式）**时添加名称相同的变量，以便在Agent调用工作流时，将名称相同的变量值作为工作流的参数输入。 – 一般变量可用于存储Agent记忆信息，有助于Agent生成个性化回答，同时也可以作为工作流的输入参数。 – 敏感变量仅支持作为工作流的输入参数，不会用于个性化回答，平台不会存储敏感变量的内容，可用于传输密码、密钥等敏感数据。 • 选择“引入更多 > 引入图片”，新增默认的WISEAGENT_IMAGE参数，表示在Agent调用工作流时，引入图片作为工作流的请求参数，支持上传的图片大小为20M。图片参数的描述是大模型识别图片的关键信息，删除描述可能会导致大模型无法识别。
节点备注		输入节点备注信息，方便后续查阅节点功能。

10.2.2 调用子工作流

调用子工作流是工作流的基础节点之一，仅包含“调用子工作流”一个执行动作。

您可以使用该节点调用或触发另一个工作流（即子工作流）。输入子工作流的ID和输入参数，即可调用子工作流。

调用子工作流节点的输入和输出通常以JSON格式传递，因此，一般在调用子工作流节点前添加一个JSON构造器节点，将对象转换为JSON格式字符串，在调用子工作流节点后添加一个JSON解析动作，用于解析调用子工作流的输出，提供给后续节点使用，如**图10-2**所示。

图 10-2 调用子工作流

- 输入

  表 10-4 运行动作属性配置输入参数说明

参数	是否必填项	说明
子流Id	是	子工作流的ID。 单击“获取工作流ID”，进入“我的工作流”列表，单击工作流ID列的 复制。
子流输入参数	是	子工作流的输入参数。

- 输出

  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

  在工作流中首次调用LLM节点需要新增实例，实例是节点的鉴权方式，如果未新增实例，节点就无法调通。

  a. 单击“新增实例”，在“创建实例”面板，配置表10-5参数信息。

  表 10-5 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“凭证管理 > 我的凭证”页面，在“平台API Key”页签获取，具体介绍请参见创建API Key。

  b. 单击“保存”，创建实例成功。

- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

10.2.3 数据连接器

数据连接器是工作流的基础节点之一，包含“json解析”和“cdm解析”两个执行动作。

数据解析连接器用于解析接收到的一个对象或者数组，以获取到用户想要的数据。

json 解析

- 输入参数

  json解析执行动作，输入参数说明如表10-6所示。

  表 10-6 json 解析输入参数说明

参数	必须	说明
基本输入	是	待解析的对象或者数组。
JSON对象	是	根据对象定义模式设置参数。

- 输出参数

  用户可以在之后的执行动作中调用该输出参数，输出参数说明如表10-7所示。

  表 10-7 json 解析输出参数说明

参数	说明
响应体	json解析输出参数对象。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

cdm 解析

- 输入参数
  cdm解析执行动作，输入参数说明如表10-8所示。

  表 10-8 cdm 解析输入参数说明

参数	必须	说明
基本输入	是	待解析的对象或者数组。
CDM对象	是	根据对象定义模式设置参数。

- 输出参数
  用户可以在之后的执行动作中调用该输出参数，输出参数说明如表10-9所示。

  表 10-9 cdm 解析输出参数说明

参数	说明
响应体	cdm解析输出参数对象。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

10.2.4 LLM

LLM（Large Language Model，大语言模型）即大模型，是工作流的基础节点之一，仅包含“chat”一个执行动作。

在工作流中添加LLM节点，可以使用大语言模型推理服务实现智能问答，在输入参数中引入前置节点的输出或自定义文本作为输入问题，大语言模型根据问题生成回答。

chat 配置说明

- 输入
  用户配置运行动作执行动作，相关参数说明如表10-10所示。

表 10-10 输入参数说明

参数	是否必填项	说明
模型服务调用ID	是	需要调用的大模型。 • 对于资产中心预置的模型，在资产中心选择“大模型”页签，单击模型卡片进入模型详情页面，查看模型服务调用ID。 • 对于我的模型（我部署的、我接入的）和我的路由策略，需要填写模型服务调用ID，请单击“获取模型服务调用ID”，进入“我的模型服务”页面，在模型服务列表中单击 复制。
高级配置		
频率惩罚	否	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。
输入加输出最大token数	否	表示模型输入加输出的最大长度。
存在惩罚	否	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。
温度	否	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”只设置1个。
多样性	否	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”只设置1个。
消息	是	选择数组类型的节点输出。 单击 可切换为数组样式，配置“角色”和“对话内容（输入）”。 • 角色：对话内容对应的角色，支持user或system。 user表示用户向大模型提问。 system表示给大模型输入对话背景及设定。 • 对话内容：支持自定义输入文本，也可将前置节点的输出作为输入。 当角色为user时，输入发送给大模型的问题。例如：请帮我分析一下这个股票的潜在价值。 当角色为system时，输入大模型的对话背景，即对大模型的设定。例如输入给大模型：你是一个理财专家，请在后续的回答中，结合理财技巧给出答复。

- 输出

  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

在工作流中首次调用LLM节点需要新增实例，实例是节点的鉴权方式，如果未新增实例，节点就无法调通。

a. 单击“新增实例”，在“创建实例”面板，配置表10-11参数信息。

表 10-11 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“凭证管理 > 我的凭证”页面，在“平台API Key”页签获取，具体介绍请参见创建API Key。

b. 单击“保存”，创建实例成功。

- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

10.2.5 知识库

知识库是工作流的基础节点之一，仅包含“查询知识库”一个执行动作。

在工作流中添加知识库节点，可以根据输入参数从指定知识库内召回匹配的信息。

查询知识库配置说明

- 输入参数

  用户配置运行动作执行动作，相关参数说明如表10-12所示。

表 10-12 输入参数说明

参数	是否必填项	说明
知识库ID	是	需要使用的知识库。 单击“获取知识库ID”，进入“我的知识库”列表，单击 复制。
最小相似度	否	搜索的关键字和返回内容的相似度阈值，取值范围是0~1。 示例：如果输入0.5，则返回相似度大于等于0.5的结果。
限量	否	检索返回切片限制数量，默认为10条。

参数	是否必填项	说明

过滤项	否	过滤条件。默认为null，如果不为null，则为SearchSqlFilter类对象，SearchSqlFilter参数说明如表10-13所示，样例如下：
		{
		"group_type": "OR",
		"expressions": [
		{
		"field": "metadata.file_name",
		"field_type": "STRING",
		"operator": "EQUAL",
		"values": [
		"四大名著介绍.txt"
		]
		},
		{
		"field": "metadata.path",
		"field_type": "STRING",
		"operator": "EQUAL",
		"values": [
		"四大名著介绍.txt"
		]
		}
		]
		}

排序项	否	排序规则。默认为null，如果不为null，则为SqlOrder类对象，SqlOrder参数说明如表10-13所示，样例如下：
		{
		"order_items": [
		{
		"field": "metadata.order",
		"field_type": "INT",
		"order_type": "DESC"
		}
		]
		}

关键字	否	检索的关键字。 根据输入的关键字从知识库内召回匹配的信息，支持自定义输入文本，也可将前序节点的输出作为输入。

表 10-13 SearchSqlFilter

参数	是否必选	参数类型	描述
group_type	否	String	**参数解释：** 过滤条件运算符。 **约束限制：** 只有一个expression时，不需要group_type，group_type可以为null。 **取值范围：** 可以为null，如果不为null，枚举值AND和OR。 **默认取值：** 不涉及。
expressions	否	Array of **Expression** objects	**参数解释：** 过滤条件。 **约束限制：** 不涉及。 **取值范围：** 非空，条件数量介于1到10之间。 **默认取值：** 不涉及。

表 10-14 Expression

参数	是否必选	参数类型	描述
field	否	String	**参数解释：** 过滤字段。 **约束限制：** 不涉及。 **取值范围：** 非空，字符串长度介于1到100之间。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
field_type	否	String	**参数解释：** 过滤字段类型。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：INT、FLOAT、BOOLEAN和STRING。 **默认取值：** 不涉及。
operator	否	String	**参数解释：** 过滤操作符。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：EQUAL、NOT_EQUAL、GREAT_THAN、GREAT_EQUAL、LESS_THAN、LESS_EQUAL、IN、NOTIN和STARTS_WITH。 **默认取值：** 不涉及。
values	否	Array of strings	**参数解释：** 过滤值。 **约束限制：** 不涉及。 **取值范围：** 非空，数量介于1到100之间，每个字符串长度最大不超过2000。 **默认取值：** 不涉及。

表 10-15 SqlOrder

参数	是否必选	参数类型	描述
order_items	否	Array of **OrderItem** objects	**参数解释：** 排序规则。 **约束限制：** 不涉及。 **取值范围：** 非空，数量介于1到10之间。 **默认取值：** 不涉及。

表 10-16 OrderItem

参数	是否必选	参数类型	描述
field	否	String	**参数解释：** 排序字段。 **约束限制：** 不涉及。 **取值范围：** 非空，字符串长度介于1到100之间。 **默认取值：** 不涉及。
field_type	否	String	**参数解释：** 排序字段类型。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，枚举值：INT、FLOAT、BOOLEAN和STRING。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
order_type	否	String	**参数解释：** 排序类型。 **约束限制：** 不涉及。 **取值范围：** 不为null，枚举值：ASC（升序）和DESC（降序）。 **默认取值：** 不涉及。

- 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
  在工作流中首次调用知识库节点需要新增实例，实例是节点的鉴权方式，如果未新增实例，节点就无法调通。
  a. 单击“新增实例”，在“创建实例”面板，配置表10-17参数信息。

  表 10-17 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“凭证管理 > 我的凭证”页面，在“平台API Key”页签获取，具体介绍请参见创建API Key。

  b. 单击“保存”，创建实例成功。
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

10.2.6 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10-18。

  表 10-18 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。
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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10-19。

  表 10-19 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。
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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表10-20。

表 10-20 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。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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**表10-21**。

表 10-21 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。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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**表10-22**。

表 10-22 变量定义参数说明

参数	是否必填项	说明	示例
变量名	是	用于指定将要命名的变量的名称。	re

参数	是否必填项	说明	示例
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。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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**表10-23**。

  表 10-23 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。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

10.2.7 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。

图 10-3 中断

- 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。
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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：

图 10-4 条件判断

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。

图 10-5 条件判断

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10-24所示。

  表 10-24 条件判断输入参数说明

参数	是否必填项	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。

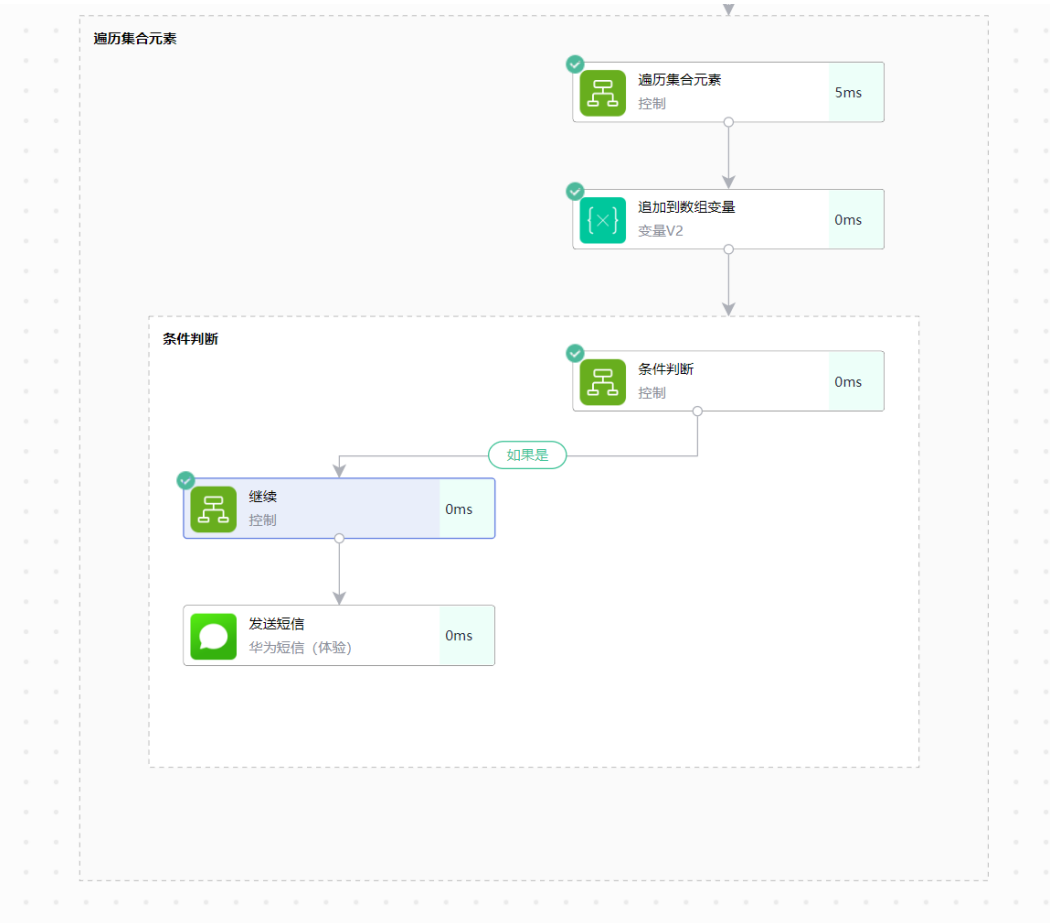
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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。

图 10-6 继续



- 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。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

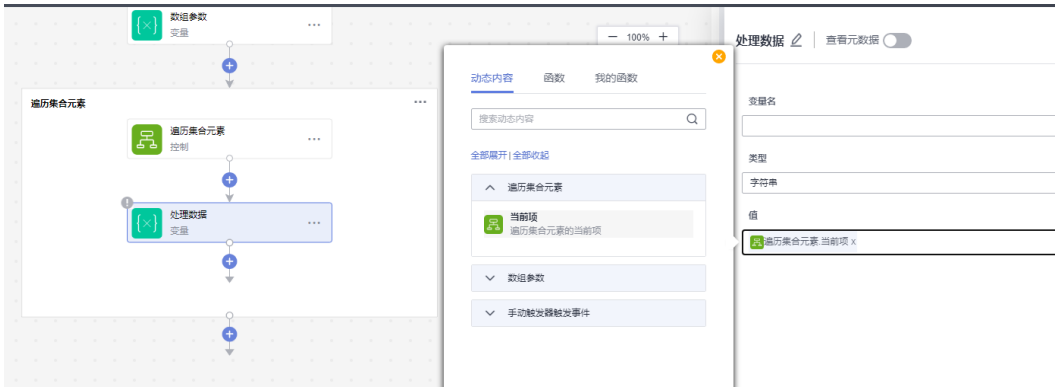
遍历集合元素

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：

图 10-7 遍历集合元素

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10-25**所示。

  表 10-25 遍历集合元素输入参数说明

参数	是否必填项	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10-26**。

  表 10-26 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。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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：

图 10-8 分支

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10-27**。

  表 10-27 分支输出参数说明

参数	说明
当前项	分支中的当前项。

- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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10-28**所示。

表 10-28 数据分片输入参数说明

参数	是否必填项	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10-29**。

表 10-29 数据分片输出参数说明

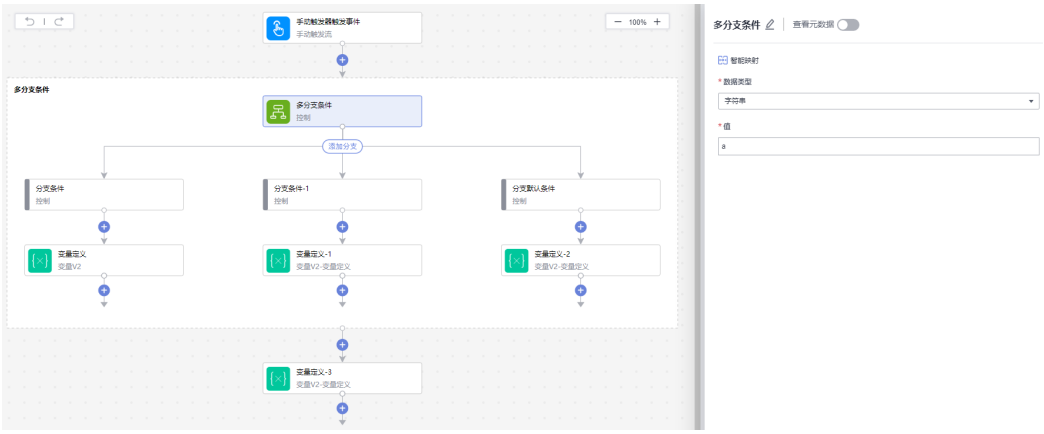
参数	说明
执行结果	数据分片是否运行成功。 • true：表示成功。 • false：表示失败。
数据分片结果	数据分片后的结果，返回值为二维数组。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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：

图 10-9 多分支条件



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10-30所示。

  表 10-30 多分支条件输入参数说明

参数	是否必填项	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10-31。

  表 10-31 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。

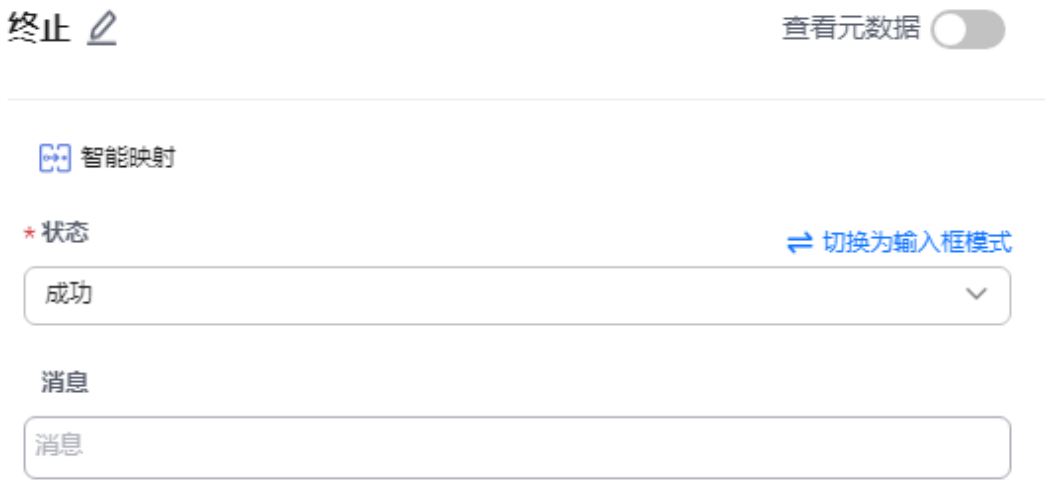
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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）。

图 10-10 终止



- 输入参数
  用户配置终止执行动作，相关参数说明如表10-32所示。

表 10-32 终止输入参数说明

参数	是否必填项	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

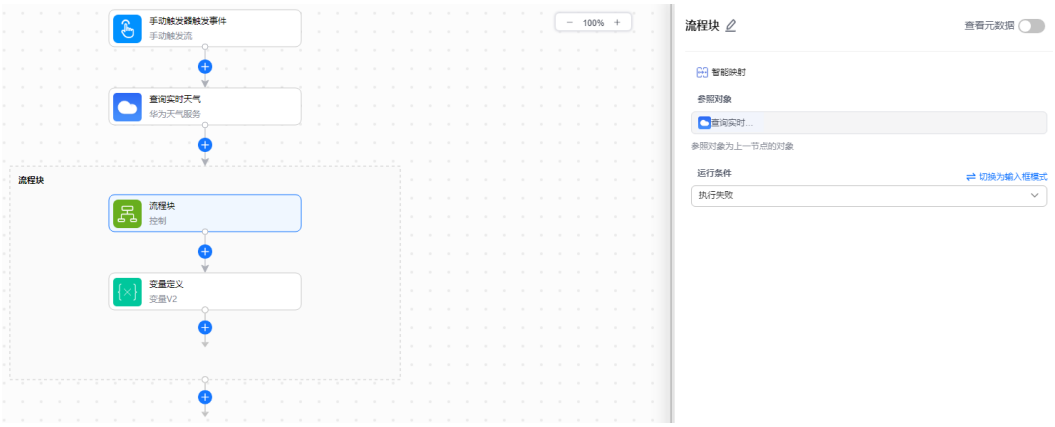
  该执行动作无输出参数。
- 节点备注

  输入节点备注信息，方便后续查阅节点功能。

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。

图 10-11 流程块



- 输入参数

  用户配置终止执行动作，相关参数说明如**表10-33**所示。

表 10-33 流程块输入参数说明

参数	是否必填项	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**表10-34**。

表 10-34 流程块输出参数说明

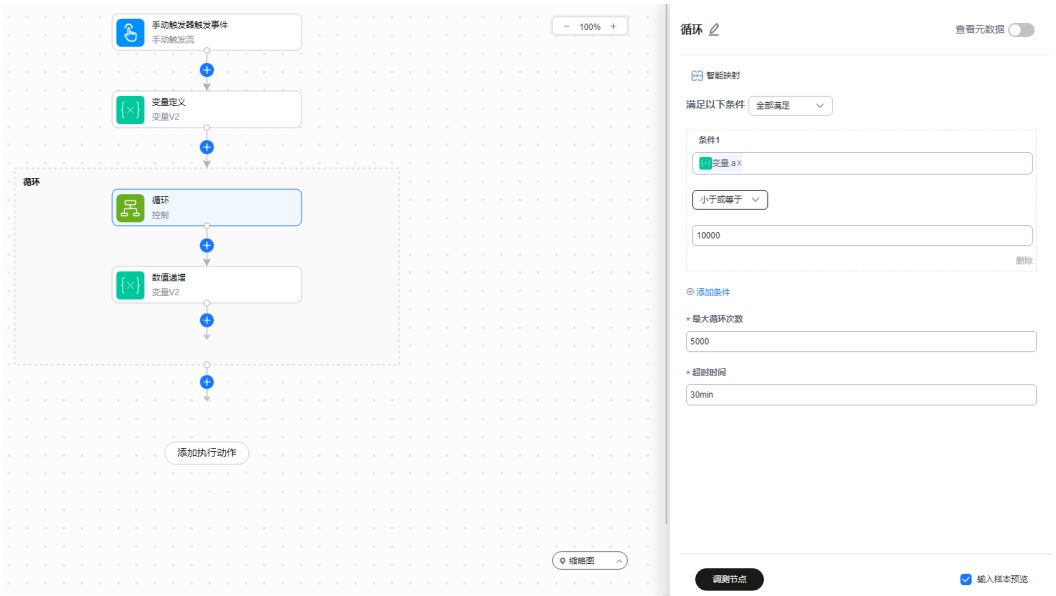
参数	说明
是否执行	流程块内步骤是否执行。
异常响应内容	异常响应内容。
异常连接器	异常连接器。
异常执行动作	异常执行动作。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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。

图 10-12 循环



- 输入参数
  用户配置终止执行动作，相关参数说明如**表10-35**所示。

表 10-35 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表10-36。

  表 10-36 循环输出参数说明

参数	说明
循环次数	循环次数。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

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。

图 10-13 异常监控和处理

- 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表10-37。

  表 10-37 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。

参数	说明
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。

- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

10.2.8 JSON 构造器

JSON构造器为用户提供构造JSON对象的能力，用户通过界面化操作可以构造出复杂的JSON结构，也可以通过“切换为源码模式”，将复杂的JSON格式数据转换到界面显示，包含“构造JSON对象”执行动作。

构造 JSON 对象

- 输入参数
  用户配置构造JSON对象源码模式执行动作，相关参数说明如表10-38所示。

表 10-38 构造 JSON 对象源码模式输入参数说明

参数	必须	说明
JSON对象	是	根据对象定义模式设置参数。

表 10-39 构造 JSON 对象表单模式输入参数说明

参数	必须	说明
JSON对象	是	根据对象定义模式设置参数。
root	否	根据下拉框选择数据类型，有“object”、“array”可以选择。两种类型都可以添加子节点。
key0	否	是root d的子节点，有“string”、“number”、“boolean”、“object”、“array”可以选择。其中只有“object”和“array”类型可以添加子节点。

- 输出参数
  该执行动作无输出参数。
- 节点备注
  输入节点备注信息，方便后续查阅节点功能。

10.2.9 Code 代码

Code代码是工作流的基础节点之一，仅包含“运行代码”一个执行动作。

支持在工作流中编写Python代码，可以将前置节点的输出作为函数的输入参数，函数的返回结果则作为输出参数供后置节点引用，从而提高工作流的灵活性和智能性。

运行代码配置说明

表 10-40 运行代码配置参数说明

参数	是否必填项	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，参数说明如表10-41所示，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。
输入参数	是	按照函数定义中指定的参数列表配置入参，即传递给函数的实际值。 输入参数或选择前序节点的输出作为输入。
节点备注	否	输入节点备注信息，方便后续查阅节点功能。

表 10-41 创建函数参数说明

参数	说明
名称	函数名，用于调用函数。
描述	函数功能描述。
入参	输入参数。
出参	输出参数。每个变量都可在后置节点中引用。
执行语言	当前仅支持Python3.9，即运行函数的环境，请查看Python函数开发指南。

参数	说明
编辑源码	在源码编辑区，编写函数内部的代码运行逻辑，如图10-14所示，图中各模块说明如下： ①：导入模块，是Python标准库中的模块，无需修改。 ②：用户自定义导入模块。 ③：公共函数使用方法示例，提供了如何使用公共函数和mssiAuthData参数的示例，无需修改。 ④：函数定义和注释，extractRequestParam函数和handler函数是系统预置的模板代码，无需修改。 ⑤：系统方法，无需修改。 ⑥：用户自定义函数中的逻辑。

图 10-14 源码编辑区

10.2.10 结束

结束节点是工作流的基础节点之一，仅包含“结束节点”一个执行动作。

结束节点作为整条工作流的输出返回，需配置响应体、状态码、响应头参数。

表 10-42 结束节点参数说明

参数	说明
选择回答模式	• 由Agent生成回答：Agent绑定了大模型时，由大模型对工作流的输出进行总结，生成自然语言回答。 • 使用设定内容直接回答（对象或数组类型）：该模式仅单Agent工作流模式或工作流选择精确模式时生效。Agent不会对工作流的输出进行处理，直接将“回答内容”参数中配置的对象或数组作为回答。 • 使用设定内容直接回答（字符串类型）：该模式仅单Agent工作流模式或工作流选择精确模式时生效。Agent不会对工作流的输出进行处理，将“回答内容”参数中配置的对象或数组类型转换为字符串作为回答。
响应体	当回答模式为“由Agent生成回答”时，配置此参数。 工作流的输出，支持自定义，也可选择前序节点的输出参数，只接受对象或数组类型，基本类型请使用JSON构造器组装成对象。
回答内容	当回答模式为“使用设定内容直接回答”时，配置此参数。 工作流的输出，支持自定义，也可选择前序节点的输出参数，只接受对象或数组类型，基本类型请使用JSON构造器组装成对象。
状态码	选择状态码。
响应头	选择数组类型的节点输出。
节点备注	输入节点备注信息，方便后续查阅节点功能。

10.3 工作流工具节点说明

工作流的工具节点可以是系统提供的，也可以是用户自定义的工具，用于实现特定的业务逻辑或功能。，包含以下三种类型：

- 华为类：为用户提供各种华为类的工具节点，如华为会议、华为天气服务等。具体介绍请参见**华为类**。
- 生活服务类：为用户提供各种用途全面，功能丰富的API资产，如银行网点查询、生活小窍门等。具体介绍请参见**生活服务类**。
- 我的工具类：包含AI原生应用引擎资产中心预置的三方工具以及自创建的工具。
  - 资产中心预置的三方工具：在AI原生应用引擎的左侧导航栏选择“资产中心”，选择“工具”页签，单击工具卡片，在工具详情页面可以查看工具描述、执行动作、参数配置等信息。

- 自创建的工具：工具是API的代理或容器，用户可以将常用API封装为工具。在创建工具时，需要先将选定的API服务注册为一个工具，然后再添加该服务下的API作为工具的执行动作。具体介绍请参见**创建工具**。

在工作流中首次调用我的工具类的节点需要新增实例，实例是工具的鉴权方式，如果未新增实例，工具就无法调通。

a. 在工具类节点配置时，单击“新增实例”，此处以“历史上的今天”节点配置为例，如**图10-15**所示。

图 10-15 历史上的今天节点配置

b. 在“创建实例”面板，配置表10-43参数信息。

表 10-43 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。 ● 资产中心预置的三方工具 – 对于第三方厂商工具，需要在该厂商的官网进行购买或注册，以获取鉴权信息。 – 对于其他租户上架的工具，在AI原生应用引擎的左侧导航栏选择“资产中心”，选择“工具”页签，鼠标光标移至工具卡片上，单击“设置鉴权”，设置鉴权信息弹框中通常会展示工具鉴权获取地址，请根据界面提示进行获取。 ● 租户自创建的工具，请填写工具创建时设置的鉴权信息。

11 管理提示语

11.1 创建提示语

平台在资产中心预置了提示语模板，同时也支持用户根据需求自定义创建提示语。在模型调测时引用创建的提示语模板，可以快速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。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建提示语

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 提示语”，单击“创建提示语”。

步骤2 在“创建提示语”页面，参照**表11-1**进行基础配置后，单击“下一步”。

表 11-1 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，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。
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

参数名称	参数说明
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(())
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3 在“在线优化”页面，参照表11-2进行参数配置。

表 11-2 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	显示创建时填写的提示语内容。
推理模型	将提示语应用于我创建的、平台预置的或第三方模型服务中，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11-3所示。

表 11-3 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4 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5 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6 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11-4所示的操作。

表 11-4 更多操作

操作	说明
修改提示语	1. 在“我创建的”的提示语列表中，单击“操作”列“修改”。 2. 参照表11-1，修改提示语的基础配置参数。
优化提示语	1. 在“我创建的”的提示语列表中，单击“操作”列“优化”。 2. 参照表11-2，配置参数，优化提示语。
删除提示语	• 删除：在“我创建的”的提示语列表中，单击提示语所在行的“操作”列的“删除”。 • 批量删除：在“我创建的”的提示语列表中，勾选需要删除的提示语，单击“批量删除”。

11.2 对创建的提示语进行优化

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。

前提条件

已创建提示语。

优化提示语

步骤1 在AI原生应用引擎的左侧导航栏选择“知识中心 > 提示语”，选择“我创建的”页签。

步骤2 在提示语列表中，单击操作列的“优化”，参照表11-5进行参数配置。

表 11-5 在线优化提示语参数说明

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	显示创建时填写的提示语内容。
推理模型	将提示语应用于我创建的、平台预置的或第三方模型服务中，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11-6所示。

表 11-6 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

11.3 管理资产中心预置提示语

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，也可以是带有一系列参数的文字描述。

AI原生应用引擎资产中心预置了多款提示语模板，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，支持测试、一键快速复制及收藏等。在模型调测时引用提示语模板，可以快速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。

测试提示语

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“提示语模板”页签。

步骤3 将鼠标光标移至提示语模板卡片上，单击“测试”，进入模型调测页面。

步骤4 在调测文本对话类型模型时，将提示语模板内容作为输入问题，按Enter键或单击 预览效果。

----结束

收藏提示语

将自己关注的提示语收藏后，可便捷地在收藏列表中查看提示语详情，并且在模型调测引用提示语模板时，可以在“我收藏的”页签下快速选择使用。

前提条件

需要具备AI原生应用引擎管理员或开发者权限。

操作步骤

步骤1 将鼠标光标移至提示语模板卡片上，单击卡片右上角 。

单击提示语模板卡片右上角的 ，可以取消收藏。

步骤2 收藏成功后，您可以在“知识中心 > 提示语”页面“我收藏的”页签下查看收藏结果。

单击提示语列表操作列的“取消收藏”，可以取消收藏。

步骤3 单击收藏列表中的提示语名称，可便捷地查看提示语详情，在详情页面可以测试提示语、基于提示语模板去创建新的提示语。

----结束

基于提示语模板创建提示语

以资产中心预置提示语模板为基础，对模板内容和配置信息稍作修改，便能快速创建生成新的提示语。

前提条件

需要具备AI原生应用引擎管理员或开发者权限。

操作步骤

步骤1 将鼠标光标移至提示语模板卡片上，单击“去创建”。

步骤2 在创建提示语页面，参考**创建提示语**进行创建。

----结束

12 管理 Agent

12.1 Agent 编排使用指引

操作指引

图 12-1 AI 原生应用引擎使用流程

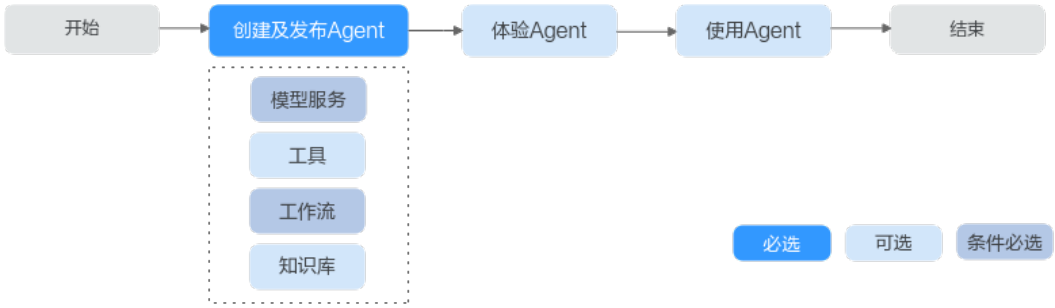


表 12-1 编排 Agent 的流程详解

序号	流程环节	说明
1	创建及发布Agent	一站式创建专属Agent，并将应用程序及相关组件进行发布，使其能够正常运行。当前支持创建LLM模式和工作流模式两种类型的Agent。 • LLM模式下，将准备好的模型服务（必选）、工具、工作流及知识库等编排成Agent。 • 工作流模式下，用户与工作流进行对话，因此必须添加工作流，不支持添加模型、工具、知识库等配置。
2	体验Agent	以对话的形式，对创建的Agent或平台资产中心预置的AI应用进行体验调测，以发现并解决Agent接口上的问题和错误。
3	使用Agent	支持通过API接口调用或Web界面访问两种方式使用Agent。

12.2 创建并发布 Agent

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。

平台在资产中心预置了部分AI应用，同时也支持用户创建Agent，当前支持创建LLM模式和工作流模式两种类型的Agent。

- LLM模式下，将准备好的模型服务、工具、工作流、知识库等编排成Agent，用户与大模型进行对话，由大模型决策并灵活调用工作流、知识库等，同时，该模式还支持使用平台自带的智能创建功能快速搭建应用。
- 工作流模式下，不支持添加工具、知识库等配置，用户与工作流进行对话，每次对话都会调用该工作流，这种模式一般适用于相对固定的场景，例如客户服务热线，Agent在接收到用户输入后，按照既定的流程响应处理，不需要进行复杂的分析和决策。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建 Agent（LLM 模式）

步骤1 在AI原生应用引擎的左侧导航栏选择“Agent编排中心 > 我的Agent”，单击“创建Agent”。

步骤2 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“Agent生成”弹框中，在“Agent名称”输入框输入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。
     您也可以仅输入“想要的Agent”信息，系统会根据Agent的功能、用途等描述智能生成Agent名称。
  b. 单击“生成”，系统将智能生成Agent配置及Agent。
- 方式二：配置Agent相关参数信息，生成Agent，具体操作如下：
  a. 关闭“Agent生成”对话框，在“创建Agent”页面左上角选择“LLM模式（智能创建）”，参照**表12-2**配置基础信息、选择模型及设定角色。

表 12-2 创建 Agent 参数说明

参数名称	参数说明
基础信息	设置Agent名称、描述信息。Agent名称不能以数字、下划线开头，不能包含特殊字符，长度2-19个字符。 您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：选择思考模型和问答模型。 思考模型用于任务规划和选择组件，主要用于工作流、知识库、工具的调用，以及入参的识别传递等。 问答模型用于问答及总结。您可以结合资产中心大模型详情介绍进行模型选择和使用。 ○ 输出最大token数：简称max_tokens，表示模型输出最大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。 ○ 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。 说明： ■ 模型服务商API在调用前需要配置认证鉴权，具体介绍请参见**如何对平台接入的第三方模型服务设置鉴权**。 ■ 如果要选择“我接入的”模型API作为思考模型，需要在模型服务描述中填写“SupportFunctionCall，AdaptFunctionCall”进行适配，具体介绍请参见**接入模型服务**。

参数名称	参数说明
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。 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. 参照**表12-3**可拓展性地添加技能、知识库、开场白以及输入推荐问题。

表 12-3 拓展性能力参数说明

参数名称	参数说明
技能	
工具	用于实现特定功能的模块或组件，添加工具可以使Agent具备更多能力，工具具体介绍请参见**管理工具**。 单击＋，选择我创建的工具或系统预置的通用工具；单击 由系统智能生成工具。 资产中心预置的三方工具在调用前需要配置认证鉴权，具体介绍请参见**调用资产中心工具前设置认证鉴权**。

参数名称	参数说明
工作流	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，在创建Agent时调用，可以有效提高Agent开发的效率，工作流具体介绍请参见**管理工作流**。 单击+，选择自创建的工作流，然后配置工作流模式： ■ 总结模式：通过匹配模型对工作流的输出内容做进一步总结。 ■ 精确模式：直接返回工作流的输出内容，保证回答的准确度。 调用工作流时，如果提示所选的工作流中存在当前Agent没有的变量，是由于在工作流节点中添加了变量，需要以Agent中名称相同的变量值作为工作流的参数输入，但Agent中没有设置这些变量。请单击“查看详情”，在补充变量弹窗中，单击“补充所选变量”，在Agent中添加变量。
知识	

参数名称	参数说明
知识库设置	单击 自动调用 ，设置如下参数： ■ 自动调用：每一轮对话自动调用知识库，利用知识库召回内容辅助大模型生成回复内容。 ■ 按需调用：由大模型决策是否调用知识库，利用知识库召回内容辅助大模型生成回复内容。 ■ 最大召回数量：从检索结果中返回的内容片段数量，取值范围：0~10。 ■ 最小匹配度：知识库召回内容与检索需求匹配程度的最低阈值，用于确保召回内容具有一定的相关性，取值范围：0~1。 ■ 提示语内容：当调用方式为“自动调用”时，支持配置此参数。 将用户问题和知识库检索的内容通过提示语形式进行组装，提供给大模型，有助于大模型提供准确回答。 如果需要在提示语内容中对数据信息进行调用，需要在输入字段中包含**{{query}}**和**{{context}}**，以这段输入为例： A问B**{{query}}**，B回答了**{{context}}**。 假如知识库数据中：context为月亮，query为李白的静夜思主题是什么？那提示语就能对该数据进行调用，得出： A问B **李白的静夜思主题是什么?** B回答了 **月亮**
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”；单击 由系统智能生成知识库。
知识检索流	知识检索流作为知识库检索工具，基于意图识别、Query改写、Query拆解、召回和重排序等，支持可视化RAG检索编排，可以优化知识检索过程，提升用户体验和Agent响应质量。 单击＋，在“知识检索流”弹框中选择知识检索流。
记忆	

参数名称	参数说明
变量	变量用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并存储在变量中。 1. 单击“变量”参数后面的 +，弹出“编辑变量”页面。 2. 单击“添加一般变量”，输入字段名、默认值、描述。例如：字段名为“职业”，默认值为“医生”。 一般变量可用于存储Agent记忆信息，有助于Agent生成个性化回答，同时也可以作为工作流的输入参数。 3. 单击“添加敏感变量”，输入字段名、描述。 敏感变量仅支持作为工作流的输入参数，不会用于个性化回答，平台不会存储敏感变量的内容，可用于传输密码、密钥等敏感数据。 4. 单击“保存”。 单击“变量”参数后面的，可以查看、编辑变量最新值。
片段记忆	开启片段记忆开关，Agent可以形成对用户的个人记忆，提供个性化回复。 勾选“支持自动整合更新”，可以对片段记忆存储的信息进行自动整合和优化更新。
文件盒子	开启文件盒子开关，在Agent体验或使用过程中，可以上传pdf、docx、txt等纯文本文件，利用模型能力对文件进行解读、总结，并基于文件内容进行问答。
对话设置	
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。
语音设置	■ 支持语音输入：开启后，支持使用语音输入问题。 ■ 支持语音输出：开启后，Agent以语音形式输出回答。

步骤3 在“Agent预览”区域单击“开始体验”，在对话输入框输入问题，按Enter键或单击预览Agent效果。

- 单击对话输入框中的，上传.wav、.mp4、.mp3格式文件或图片，可以对上传的音频文件或图片提问。
- 如果Agent开启了“支持语音输入”开关，您也可以单击对话输入框中的，通过语音输入问题。
- 如果Agent使用配置变量的方式实现了记忆能力，在预览时，会自动识别对话与变量匹配的内容，自动更新变量取值，单击“Agent预览”区域右上角的“记忆 > 变量”，可以查看变量使用效果，修改变量取值可以手动更新Agent记忆信息。
- 如果Agent开启了“片段记忆”开关，在预览时，会自动识别并保留用户个性化信息，单击“Agent预览”区域右上角的“记忆 > 片段记忆”，可以查看片段记忆使用效果，单击记忆内容后面的，修改记忆内容，可以手动更新Agent记忆信息。
- 如果Agent开启了文件盒子，执行如下操作进行体验：
  a. 单击“Agent预览”区域右上角的“记忆 > 文件盒子”，单击“上传文件”。
     只支持pdf、docx、txt等纯文本文件，文件大小小于10M。
  b. 文件上传完成后，单击文件盒子列表操作列的，引用文件。
  c. 在对话输入框对引用文件提问，Agent会根据输入问题对文件进行回答、总结。

步骤4 单击“保存”，完成Agent创建。

“我创建的”列表中生成一条Agent记录，Agent状态为“草稿”。

----结束

创建 Agent（工作流模式）

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

步骤2 单击“创建Agent”，在页面左上角选择“工作流模式”。

步骤3 在“创建Agent”页面，参照表12-4配置参数。

表 12-4 创建 Agent 参数说明

参数名称	参数说明
基础信息	设置Agent名称、描述信息。Agent名称不能以数字、下划线开头，不能包含特殊字符，长度2-19个字符。

参数名称	参数说明
工作流配置	单击+，选择自创建的工作流。您也可以单击“创建工作流”，参考**创建工作流**创建新的工作流。 工作流模式的Agent只能绑定一个工作流，且此工作流的起始节点必须引入用户对话输入，即包含默认的WISEAGENT_USER_INPUT参数，表示以用户在问答对话中输入的内容作为工作流的请求参数，与Agent的每次对话都会对该工作流进行调用，具体介绍请参见**起始节点**。 WISEAGENT_VARIABLES和WISEAGENT_CONVERSATION为可选参数。
变量	变量用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并存储在变量中。 1. 单击“变量”参数后面的+，弹出“编辑变量”页面。 2. 单击“添加一般变量”，输入字段名、默认值、描述。例如：字段名为“职业”，默认值为“医生”。 一般变量可作为记忆信息存储，有助于Agent生成个性化回答，同时也可以作为工作流的输入参数。 3. 单击“添加敏感变量”，输入字段名、描述。 敏感变量仅支持作为工作流的输入参数，不会用于个性化回答。 4. 单击“保存”。 单击“变量”参数后面的，可以查看、编辑变量最新值。
开场白	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?
推荐问题	单击+，在输入框输入推荐的问题语句。
语音设置	• 支持语音输入：开启后，支持使用语音输入问题。 • 支持语音输出：开启后，Agent以语音形式输出回答。

步骤4 在“Agent预览”区域单击“开始体验”，在对话输入框输入问题，按Enter键或单击预览Agent效果。

- 单击对话输入框中的，上传.wav、.mp4、.mp3格式的文件或图片，可以对上传的音频文件或图片提问。
- 如果Agent开启了“支持语音输入”开关，您也可以单击对话输入框中的，通过语音输入问题。
- 如果Agent使用配置变量的方式实现了记忆能力，在预览时，会自动识别对话与变量匹配的内容，自动更新变量取值，单击“Agent预览”区域右上角的“记忆 > 变量”，可以查看变量使用效果，修改变量取值可以手动更新Agent记忆信息。

步骤5 单击“保存”，完成Agent创建。

“我创建的”列表中生成一条Agent记录，Agent状态为“草稿”。

----结束

发布 Agent

Agent发布后，用户即可通过API接口调用或Web界面进行访问。

步骤1 Agent创建完成后，在创建页面单击“发布”。

也可以在“我创建的”列表中，单击Agent列表操作列的“发布Agent”，进入发布页面。

步骤2 选择发布渠道并设置发布密钥，单击“发布”。

表 12-5 发布参数说明

参数	说明
选择发布渠道	• API：以API的方式发布Agent，发布成功后可复制API地址进行分享。 • Web Url：以Web Url的方式发布Agent，发布成功后可复制Web链接进行分享。
设置发布密钥	设置该密钥是确保发布分享Agent后，用户能正常调用Agent相关联的模型、工具、工作流和知识库。 API Key：输入AI原生应用引擎平台API Key，获取方式请参见**创建API Key**。
部署资源	• 选择发布方式 – 免费额度：每个租户有3个免费额度，使用免费额度发布Agent运行速度相对缓慢。 – 运行时引擎SKU额度：请参见**购买AI原生应用引擎**进行订购，一个额度代表一个节点数量。 • 节点数量 – 使用免费额度发布，默认为单节点部署。 – 使用运行时引擎SKU额度发布，可选择多节点集群部署。

步骤3 发布后，在Agent列表的“发布地址”列，单击  或  ，复制发布地址进行分享。

----结束

更多操作

Agent创建完成后，可执行如**表12-6**所示操作。

表 12-6 更多操作

操作	说明
查看Agent详情	在“我创建的”Agent列表中单击Agent名称，进入Agent详情页面，可查看Agent的基础信息、Agent组成、接口信息以及对话日志等。您还可以导出用户反馈数据： 1. 在“对话日志”页签中，单击“导出用户反馈数据”。 2. 在弹框中选择导出范围，默认导出最近30天的数据。 3. 单击“确定”。
修改Agent参数	支持修改状态为“草稿”的Agent。 在“我创建的”Agent列表中，在“操作”列选择“更多 > 修改”，修改Agent配置参数。
取消发布Agent	1. 在“我创建的”Agent列表中，单击“操作”列的“取消发布”。 2. 单击“确认”。
删除Agent	对于已发布的Agent，可先取消发布再删除。 1. 在“我创建的”Agent列表中，在“操作”列选择“更多 > 删除”。 2. 单击“确认”。
体验Agent	我创建的Agent发布后，可以进行体验，具体介绍请参见**体验Agent**。

12.3 体验 Agent

Agent体验是指以对话的形式，对自创建的Agent或平台资产中心预置的AI应用进行体验调测，以发现并解决Agent接口上的问题和错误。

体验我的 Agent

1. 在AI原生应用引擎的左侧导航栏选择“Agent编排中心 > 我的Agent”。
2. 选择“我创建的”页签，单击Agent列表“操作”列的“体验”。
3. 如果创建Agent时，使用配置变量的方式实现了记忆能力，体验时输入变量值，作为Agent记忆信息存储，单击“保存”。
4. 在Agent体验页面的对话输入框输入问题，按Enter键或单击体验Agent。
   - 单击对话输入框中的，上传.wav、.mp4、.mp3格式文件或图片，可以对上传的音频文件或图片提问。
     资产中心预置Agent不支持对音频文件提问功能。
   - 如果创建Agent时开启了“支持语音输入”开关，您也可以单击对话输入框中的，通过语音输入问题。
   - 如果创建Agent时，使用配置变量的方式实现了记忆能力，在Agent体验时会自动识别对话与变量匹配的内容，自动更新变量取值，选择页面右上角的“记忆 > 变量”，可以查看变量使用效果。

- 如果创建Agent时开启了“片段记忆”，在Agent体验时会自动识别并保留用户个性化信息，选择页面右上角的“记忆 > 片段记忆”，可以查看片段记忆使用效果。
- 如果创建Agent时开启了“文件盒子”，根据以下操作体验文件盒子：
  i. 选择对话框右上角的“记忆 > 文件盒子”，上传文件。
  只支持上传pdf、docx、txt等纯文本文件，文件大小小于10M。
  ii. 文件上传完成后，单击文件盒子列表操作列的，引用文件。
  iii. 在对话输入框对引用文件提问，Agent会根据输入问题对文件进行回答、总结。

5. 对于Agent生成的答案可以进行复制、点赞、点踩等。
   - ：如果创建Agent时开启了“支持语音输出”开关，可以语音播放答案。
   - ：重新生成答案。
   - ：复制答案。
   - ：对答案点赞。
   - ：对答案点踩。
6. 在“我创建的”Agent列表中，单击Agent名称，进入Agent详情页面，在“对话日志”页签中可以查看所有问答的对话日志。

体验资产中心预置的 AI 应用

1. 在AI原生应用引擎的左侧导航栏选择“资产中心”。
2. 在资产中心页面，选择“AI应用”页签。
3. 将鼠标光标移至应用卡片上，单击“体验”。
4. 在Agent体验页面的对话输入框输入问题，按Enter键或单击进行体验。

12.4 使用 Agent

支持通过API接口调用或Web界面访问两种方式使用Agent。

前提条件

需要获取Agent的发布地址。

使用 Agent

- Web Url
  a. 直接打开Web链接访问应用。
  b. 如果创建Agent时，使用配置变量的方式实现了记忆能力，体验时输入变量值，作为Agent记忆信息存储，单击“保存”。
  c. 在Agent的对话输入框输入问题，按Enter键或单击使用Agent。
     - 单击对话输入框中的，上传.wav、.mp4、.mp3格式文件或图片，可以对上传的音频文件或图片提问。

- 如果创建Agent时开启了“支持语音输入”开关，您也可以单击对话输入框中的，通过语音输入问题。
- 如果创建Agent时，使用配置变量的方式实现了记忆能力，在使用Agent时会自动识别对话与变量匹配的内容，自动更新变量取值，选择页面右上角的“记忆 > 变量”，可以查看变量使用效果。
- 如果创建Agent时，开启了“片段记忆”，在使用Agent时会自动识别并保留用户个性化信息，选择页面右上角的“记忆 > 片段记忆”，可以查看片段记忆使用效果。
- 如果创建Agent时，开启了“文件盒子”，根据以下操作体验文件盒子：
  1) 选择对话框右上角的“记忆 > 文件盒子”，上传文件。
     只支持上传pdf、docx、txt等纯文本文件，文件大小小于10M。
  2) 文件上传完成后，单击文件盒子列表操作列的，引用文件。
  3) 在对话输入框对引用文件提问，Agent会根据输入问题对文件进行回答、总结。

d. 对于Agent生成的答案可以进行复制、点赞、点踩等。

- ：如果创建Agent时开启了“支持语音输出”开关，可以语音播放答案。
- ：重新生成答案。
- ：复制答案。
- ：对答案点赞。
- ：对答案点踩。

- API地址
  可参考**调用Agent**进行使用。

12.5 收藏资产中心预置的 AI 应用

支持收藏平台资产中心预置的AI应用。将自己关注的AI应用收藏后，可便捷地在收藏列表中查看应用详情及体验Agent。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

收藏资产中心预置 AI 应用

步骤1 在AI原生应用引擎的左侧导航栏选择“资产中心”。

步骤2 在资产中心页面，选择“AI应用”页签。

步骤3 将鼠标光标移至应用卡片上，单击卡片右上角。

单击模型卡片右上角的，可以取消收藏。

步骤4 收藏成功后，您可以在“Agent编排中心 > 我的Agent”页面“我收藏的”页签下，查看收藏结果。

单击Agent列表操作列的“取消收藏”，可以取消收藏。

步骤5 单击Agent列表中的Agent名称，可以便捷地查看Agent详情及体验Agent。

----结束

13 管理我的凭证

13.1 创建 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。用户创建Agent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。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

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。
- 如果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。
- 每个用户最多只能拥有两个AK/SK访问密钥。

创建 AK/SK 访问密钥

步骤1 在AI原生应用引擎的左侧导航栏选择“凭证管理 > 我的凭证”。

步骤2 在“我的凭证”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”。

为了保证历史兼容性，系统会使用访问密钥创建时间作为初始值。

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 访问密钥

密钥删除后无法恢复，请谨慎删除。

步骤1 在“我的凭证”页面，选择“AK/SK访问密钥”页签。

步骤2 在AK/SK访问密钥列表中，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

13.2 创建 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。用户开发的Agent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。

操作须知

每个用户最多可添加两个平台API Key。

前提条件

需要具备AI原生应用引擎管理员或开发者权限，权限申请操作请参见**AppStage组织成员申请权限**。

创建 API Key

步骤1 在AI原生应用引擎的左侧导航栏选择“凭证管理 > 我的凭证”。

步骤2 在“我的凭证”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”。

API Key名称最多为32个字符。

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。

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“凭证管理 > 我的凭证”页面，选择“平台API Key”页签。

步骤2 在API Key列表中，单击“操作”列“删除”。

步骤3 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

14 下载 AI 原生应用引擎 SDK

AI原生应用引擎面向开发者提供了一套搭建原生应用的Python SDK，包含了模型调用，知识获取，工具调用等功能。开发者可以使用SDK调用AI原生应用引擎的各种能力，快速构建大模型应用。

用户可以通过AI原生应用引擎平台下载SDK，同时对SDK完整性进行校验以确保获取的SDK为原始文件。本文介绍如何下载SDK以及完整性校验方法，SDK的具体使用方法和功能介绍请参考**SDK参考**。

AI原生应用引擎SDK面向开发者开放下载，无需登录AI原生应用引擎，直接访问下载地址：**wiseagent-dev-sdk-python**，也可以获取SDK。

下载 SDK 并校验完整性

步骤1 登录AI原生应用引擎，鼠标光标移至右上角登录的用户名。

步骤2 单击“下载SDK”，进入下载SDK页面。

步骤3 在“操作流程”区域，单击“下载SDK”，可获取完整的AI原生应用开发套件。

步骤4 生成SDK包的SHA256哈希值，用于校验SDK完整性。生成方法如下：

- **Windows系统SHA256哈希值生成方法**
  a. 执行**Windows+R**，唤起任务调用，输入cmd打开命令行调用窗口。
  b. 执行**certutil -hashfile** *绝对路径下文件 校验值*。
     例如：certutil -hashfile C:\Users\xxxx\tcp.xml sha256
- **Linux系统SHA256哈希值生成方法**
  a. 打开终端，进入到SDK文件所在目录。
  b. 执行**sha256sum** *文件名*，生成该文件的SHA256。
     例如：sha256sum tcp.xml
- **Mac系统SHA256哈希值生成方法**
  a. 打开终端，进入到SDK文件所在目录。
  b. 执行**shasum -a 256** *文件名*，生成该文件的SHA256。
     例如：shasum -a 256 tcp.xml

步骤5 将生成的SHA256哈希值与“操作流程”区域展示的原始SHA256进行对比，验证文件的完整性。

如果两者一致，说明SDK包在下载过程中未被篡改或损坏；如果不一致，则说明存在问题，建议重新下载。

----结束

15 管理账号信息

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

查看账户信息

登录AI原生应用引擎，将鼠标移至右上角登录的用户名，弹出“账户信息”页面，可查看当前登录用户的账户信息：账号名、岗位。

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 登录AI原生应用引擎，鼠标光标移至右上角登录的用户名，弹出“账户信息”页面。

步骤2 在“账户信息”页面，单击“修改密码”。

步骤3 为确认是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。

- 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件验证码一种方式。
- 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”。

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”。
   - 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“账户信息”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”。

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”。
   - 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