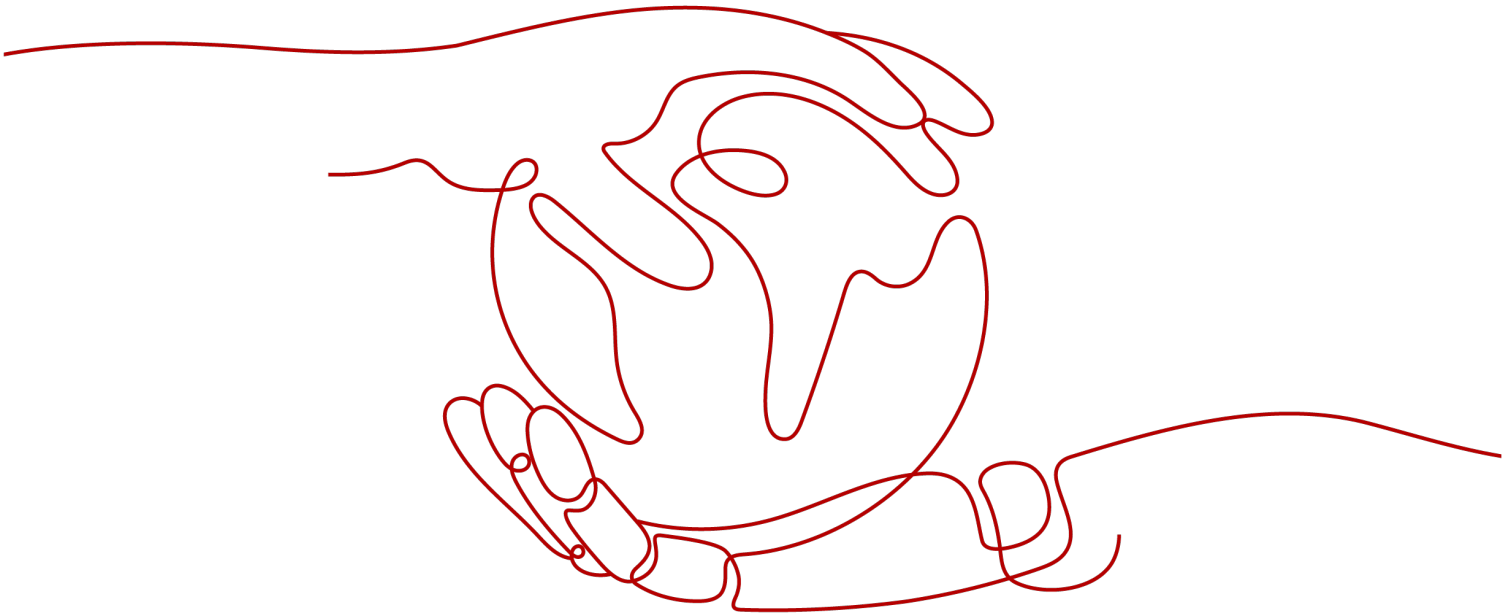
应用平台

API 参考

文档版本 06

发布日期 2025-01-20



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 使用前必读

欢迎使用应用平台（AppStage），AppStage是基于平台工程（Platform Engineering）理念打造的下一代应用全生命周期管理和AI原生应用生命周期管理平台，帮助客户快速高效地实现传统应用及AI原生应用全生命周期管理，为应用构建、运维和运营等生命周期管理活动提供自助式服务能力。

目前AppStage的AI原生应用引擎提供API供您调用。在调用AppStage的AI原生应用引擎API之前，请确保已经充分了解AppStage的相关概念，详细信息请参见AppStage服务的**产品介绍**。

终端节点

终端节点即调用API的**请求地址**，不同服务不同区域的终端节点不同，AppStage目前仅部署在“**华北-北京四**”区域，Endpoint为“aiae.appstage.myhuaweicloud.com”。

基本概念

- 大模型推理服务

  直接调用预置大模型提供API完成推理过程。

- 私有模型部署

  针对已经微调训练好的模型，如需评测此模型效果，或通过应用调用此模型，则需将模型部署为线上服务。

- 向量知识库

  通过引入多种类型和格式的企业知识，将数据转化为向量，并利用高效的存储和索引方式进行查询，实现基于检索增强的大模型能力。

- 工作流

  任务流程的细化分解是一种有效策略，能够简化系统架构，并降低对大语言模型能力的过度依赖。通过将繁复的工作拆解为一系列独立节点，不仅增强了复杂任务处理的效率，还在很大程度上提升了整个系统的透明度、鲁棒性和错误容忍度。这种方法使得LLM的应用范围得以扩大，即便面对高度复杂的任务也能表现出色。

2 API 概览

AppStage接口的分类与说明如表2-1所示。

表 2-1 API 概览

类型	说明
模型调用	包含文本对话类、文本向量化类模型服务调用接口。
应用中心	包含Agent调用、用户配置（工作流、技能）调用、文件盒子等接口。
知识中心	包含知识库和知识数据集的创建、删除、执行及查询等接口。

3 如何调用 API

3.1 构造请求

本节介绍REST API请求的组成，并以调用AppStage服务的文本对话接口说明如何调用API。

您还可以通过这个视频教程了解如何构造请求调用API：**https://bbs.huaweicloud.com/videos/102987** 。

请求 URI

请求URI由如下部分组成。

{URI-scheme} :// {Endpoint} / {resource-path} ? {query-string}

尽管请求URI包含在请求消息头中，但大多数语言或框架都要求您从请求消息中单独传递它，所以在此单独强调。

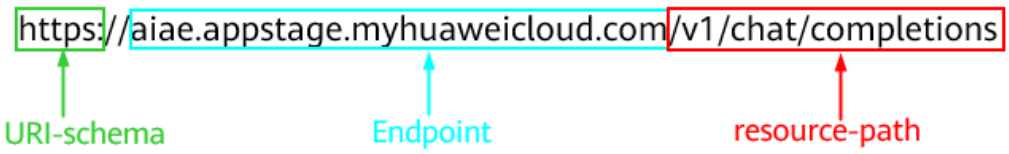
表 3-1 URI 中的参数说明

参数	描述
URI-scheme	表示用于传输请求的协议，当前所有API均采用**HTTPS**协议。
Endpoint	指定承载REST服务端点的服务器域名或IP，不同服务不同区域的Endpoint不同，当前AppStage服务只在“华北-北京四”部署，Endpoint为“aiae.appstage.myhuaweicloud.com”。
resource-path	资源路径，即API访问路径。从具体API的URI模块获取，例如文本对话API的resource-path为“/v1/chat/completions”。
query-string	查询参数，是可选部分，并不是每个API都有查询参数。查询参数前面需要带一个**“?”**，形式为**“参数名=参数取值”**，例如**“?limit=10”**，表示查询不超过10条数据。

例如，您需要调用AppStage在“华北-北京四”区域的文本对话接口，则需使用“华北-北京四”区域的Endpoint（aiae.appstage.myhuaweicloud.com），并在文本对话的URI部分找到resource-path（/v1/chat/completions），拼接起来如下所示。

```
https://aiae.appstage.myhuaweicloud.com/v1/chat/completions
```

图 3-1 URI 示意图



说明

为查看方便，在每个具体API的URI部分，只给出resource-path部分，并将请求方法写在一起。这是因为URI-scheme都是HTTPS，而Endpoint在同一个区域也相同，所以简洁起见将这两部分省略。

请求方法

HTTP请求方法（也称为操作或动词），它告诉服务你正在请求什么类型的操作。

表 3-2 HTTP 方法

方法	说明
GET	请求服务器返回指定资源。
PUT	请求服务器更新指定资源。
POST	请求服务器新增资源或执行特殊操作。
DELETE	请求服务器删除指定资源，如删除对象等。
HEAD	请求服务器资源头部。
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。

在调用文本对话接口的URI部分，您可以看到其请求方法为“POST”，则其请求为：

```
POST https://aiae.appstage.myhuaweicloud.com/v1/chat/completions
```

请求消息头

附加请求头字段，如指定的URI和HTTP方法所要求的字段。例如定义消息体类型的请求头“Content-Type”，请求鉴权信息等。

详细的公共请求消息头字段请参见表3-3和表3-4。

表 3-3 AK/SK 认证公共请求消息头

名称	描述	是否必选	示例
Content-Type	消息体的类型（格式），当前只支持application/json。	是	application/json
ts	毫秒时间戳。	是	1707101222000
nonce	请求唯一标识（UUID）。从**AK/SK认证**中获取。	是	-
ak	为AK/SK凭证文件中的AK明文。从**AK/SK认证**中获取。	是	-
sign	签名字符串。从**AK/SK认证**中获取。	是	-
resource-code	WiseAgent对外开放接口对应的唯一编码，每个接口唯一。请参考**表3-5**。	是	modelmarket.chat

表 3-4 API Key 认证公共请求消息头

名称	描述	是否必选	示例
Content-Type	消息体的类型（格式），当前只支持application/json。	是	application/json
Authorization	认证信息。格式为：Bearer ${API Key}	是	Bearer sk-5db9*********dd58

表 3-5 Resource-code

Resource-code	接口
modelrouter.chat	**调用文本对话模型服务**
modelrouter.embeddings	**调用文本向量化模型服务**

Resource-code	接口
knowledgeBases.query.embeddata	检索知识库数据

请求消息体

请求消息体通常以结构化格式发出，与请求消息头中Content-type对应，传递除请求消息头之外的内容。如果请求消息体中参数支持中文，则中文字符必须为UTF-8编码。

每个接口的请求消息体内容不同，也并不是每个接口都需要有请求消息体（或者说消息体为空），GET、DELETE操作类型的接口就不需要消息体，消息体具体内容需要根据具体接口而定。

对于文本对话接口，您可以从接口的请求部分看到所需的请求参数及参数说明。将消息体加入后的请求如下所示。

```
POST https://aiae.appstage.myhuaweicloud.com/v1/chat/completions

{
   "model": "platform:chatglm3-6b",
   "messages": [
      {
         "role": "user",
         "content": "你好!"
      }
   ],
   "stream": false
}
```

到这里为止这个请求需要的内容就具备齐全了，您可以使用curl、Postman或直接编写代码等方式发送请求调用API。

3.2 认证鉴权

AppStage调用接口支持AK/SK和API Key认证鉴权。

AK/SK认证：通过AK（Access Key ID）/SK（Secret Access Key）进行API调用时的认证。

API Key：通过API密钥进行API调用时的认证。

AK/SK 认证

AK/SK认证就是使用AK/SK对请求进行签名，在请求时将签名信息添加到消息头，从而通过身份认证。

- AK(Access Key ID)：访问密钥ID。与私有访问密钥关联的唯一标识符；访问密钥ID和私有访问密钥一起使用，对请求进行加密签名。
- SK(Secret Access Key)：与访问密钥ID结合使用的密钥，对请求进行加密签名，可标识发送方，并防止请求被修改。

使用AK/SK认证时，您可以基于签名算法使用AK/SK对请求进行签名。详细的签名认证操作流程如下。

1. AK/SK申请

   使用具有管理员权限（admin）账号登录到WiseAgent主页，从右上角凭证管理进入到AK/SK管理页面，新建AK/SK。

   每个用户只能同时拥有两个AK/SK凭证。

2. AK/SK下载

   成功创建AK/SK后，会立刻弹出AK/SK凭证下载弹窗，下载后得到凭证文件。

   每个凭证仅能下载一次，且无法找回，请妥善保管凭证文件。

3. 使用AK/SK鉴权

   在请求头里添加如下header：

   ts: 毫秒时间戳

   nonce: 请求唯一标识（UUID）

   ak: 凭证文件中的AK明文

   resource-code: WiseAgent对外开放接口对应的唯一编码，每个接口唯一

   sign: 按如下规则拼接字符串"ts={变量名}&nonce={nonce}&ak={ak}"，对拼接得到的字符串plain进行SHA256散列后得到散列值hash，再使用凭证中的SK明文对刚才生产的hash进行再散列，最后进行Base64转码，得到签名字符串。

   签名样例代码（JAVA）：

```
public String sha256(String plain) {
    try {
        MessageDigest messageDigest = MessageDigest.getInstance("SHA-256");
        messageDigest.update(plain.getBytes(StandardCharsets.UTF_8));
        byte[] bytes = messageDigest.digest();
        StringBuffer hexBuffer = new StringBuffer();
        for (byte aByte : bytes) {
            String hex = Integer.toHexString(0xff & aByte);
            if (hex.length() == 1) {
                hexBuffer.append('0');
            }
            hexBuffer.append(hex);
        }
        return hexBuffer.toString();
    } catch (NoSuchAlgorithmException ignore) {
    }
}

public String hmacSha256(String hash, String sk) {
    try {
        Mac hmacSHA256 = Mac.getInstance("HmacSHA256");
        SecretKeySpec secretKeySpec = new SecretKeySpec(sk.getBytes(StandardCharsets.UTF_8),
"HmacSHA256");
        hmacSHA256.init(secretKeySpec);
        byte[] bytes = hmacSHA256.doFinal(hash.getBytes(StandardCharsets.UTF_8));
        return Base64.encodeBase64String(bytes);
    } catch (NoSuchAlgorithmException | InvalidKeyException ignore) {
    }
}
```

API Key 认证

API Key全称为应用程序接口密钥，是一种用于验证和授权API请求的代码。它通常是一串字符，用于识别调用API的应用程序和开发者。

1. 获取API Key

   以管理员身份登录AI原生应用引擎工作台，参考**创建API Key**获取。

2. 使用API Key鉴权

调用时，在请求头里新增字段Authorization，值填写为Bearer ${API Key}，拼接起来如下所示。

```
Authorization:Bearer sk-5db9*********dd58
```

3.3 返回结果

状态码

请求发送以后，您会收到响应，包含状态码、响应消息头和消息体。

状态码是一组从1xx到5xx的数字代码，状态码表示了请求响应的状态，完整的状态码列表请参见**状态码**。

对于文本对话接口，如果调用后返回状态码为“200”，则表示请求成功。

响应消息头

对应请求消息头，响应同样也有消息头。例如，Content-Type=application/json

响应消息体

响应消息体通常以结构化格式返回，与响应消息头中Content-type对应，传递除响应消息头之外的内容。

对于文本对话接口，返回如下消息体。为篇幅起见，这里只展示部分内容。

```
{
    "created": 1718772336,
    "usage": {
        "completion_tokens": 23,
        "prompt_tokens": 45,
        "total_tokens": 68
    },
    "model": "chatglm3-6b",
    "id": "chatcmpl-xxx",
    "choices": [{
        "finish_reason": "stop",
        "index": 0,
        "message": {
            "role": "assistant",
            "content": "你好，有什么我可以帮助你的吗？"
        },
        "logprobs": null
    }],
    "object": "chat.completion"
}
```

当接口调用出错时，会返回错误码及错误信息说明，错误响应的Body体格式如下所示。

```
{
    "error_code": "AIAE.31001702",
    "error_msg": "Model not exists, please check and try again later!"
}
```

其中，error_code表示错误码，error_msg表示错误描述信息。

4 API

4.1 模型调用

4.1.1 调用文本对话模型服务

功能介绍

调用大语言模型推理服务，根据用户问题，获取大语言模型的回答。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/chat/completions

请求参数

表 4-1 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-2 请求 Body 参数

参数	是否必选	参数类型	描述
messages	是	Array of **ChatCompletionRequestMessage** objects	**参数解释：** 文本对话消息体类。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
model	是	String	**参数解释：** 模型服务调用唯一id字段。平台定义了4种模型服务： 1.平台预置模型服务 登录**AI原生应用引擎**，在左侧导航栏选择”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。 2.租户部署模型服务 登录**AI原生应用引擎**，在左侧导航栏选择“模型中心 > 我的模型服务”，选择“我部署的”页签，在模型服务列表中复制模型服务调用ID。 3.租户接入模型服务 登录**AI原生应用引擎**，在左侧导航栏选择“模型中心 > 我的模型服务”，选择“我接入的”页签，在模型服务列表中复制模型服务调用ID。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
frequency_penalty	否	Number	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 介于-2.0和2.0之间。 **默认取值：** 0。

参数	是否必选	参数类型	描述
logit_bias	否	Map<String,Integer>	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
max_tokens	否	Integer	**参数解释：** 返回体允许的最大token数。 **约束限制：** 不涉及。 **取值范围：** 介于1和4096之间。 **默认取值：** 4096。
n	否	Integer	**参数解释：** 返回体中包含的chatCompletionChoice数量，建议默认设置为1，最大限度地降低成本。 **约束限制：** 不涉及。 **取值范围：** 介于1和128之间。 **默认取值：** 1。

参数	是否必选	参数类型	描述
presence_penalty	否	Number	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 介于-2.0和2.0之间。 **默认取值：** 0。
stream	否	Boolean	**参数解释：** 是否流式返回。 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。 **约束限制：** 不涉及。 **取值范围：** true或者false。 **默认取值：** false。
temperature	否	Number	**参数解释：** 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。建议该参数设置为1。 **约束限制：** 不涉及。 **取值范围：** 介于0.0和2.0之间。 **默认取值：** 1。

参数	是否必选	参数类型	描述
top_p	否	Number	**参数解释：** 影响输出文本的多样性，取值越大，生成文本的多样性越强。 **约束限制：** 不涉及。 **取值范围：** 介于0.0和1.0之间。 **默认取值：** 0.5。
tools	否	**FunctionCall Tool** object	**参数解释：** 可供模型调用的工具。 **约束限制：** 目前仅如下模型支持此功能： glm-4 glm-3-turbo moonshot-v1-8k moonshot-v1-32k moonshot-v1-128k **取值范围：** 参考约束限制。 **默认取值：** 不涉及。
tool_choice	否	String	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 当前仅支持auto。 **默认取值：** auto。

参数	是否必选	参数类型	描述
content_security_verify	否	ContentSecurityVerify object	**参数解释：** 控制是否开启内容审核。如果开启内容审核，AI原生应用引擎会对模型返回结果进行审查。 **约束限制：** 不涉及。 **取值范围：** true或者false。 **默认取值：** false。

表 4-3 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
content	是	String	**参数解释：** 消息具体内容。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
role	是	String	**参数解释：** 消息体对应的角色。 **约束限制：** 不涉及。 **取值范围：** 如果是系统则为system。 如果是用户则为user。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
name	否	String	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-4 FunctionCallTool

参数	是否必选	参数类型	描述
type	否	String	**参数解释：** 调用工具类型。 **约束限制：** 目前仅支持function。 **取值范围：** 不涉及。 **默认取值：** function。
function	否	**function** object	**参数解释：** 仅当工具类型为function时补充。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-5 function

参数	是否必选	参数类型	描述
name	否	String	**参数解释：** 函数名称。 **约束限制：** 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。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
description	否	String	**参数解释：** 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。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
parameters	否	Object	**参数解释：** Json Schema对象，用于定义函数所接受的参数。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-6 ContentSecurityVerify

参数	是否必选	参数类型	描述
is_response_verify	否	Boolean	**参数解释：** 是否开启返回体内容审核（默认不开启）。 有文本内容，则对文本进行内容审核； 有图片内容，则会对图片进行内容审核。 **约束限制：** 不涉及。 **取值范围：** true或false **默认取值：** false。

响应参数

状态码： 200

表 4-7 响应 Body 参数

参数	参数类型	描述
id	String	**参数解释：** 文本对话唯一标识符。 **取值范围：** 不涉及。
choices	Array of **choices** objects	**参数解释：** 返回体列表。 如果‘n’大于1，则结果为多个。 **取值范围：** 不涉及。
created	Integer	**参数解释：** 问答发生的时间（格式为时间戳）。 **取值范围：** 不涉及。

参数	参数类型	描述
model	String	**参数解释：** 实际转发后调用的模型名称，与请求体中model可能不同。 **取值范围：** 不涉及。异常详情
object	String	**参数解释：** 固定值。 **取值范围：** ‘chat.completion’。
usage	**CompletionUsage** object	**参数解释：** 每次请求的用量统计。 **取值范围：** 不涉及。

表 4-8 choices

参数	参数类型	描述
finish_reason	String	**参数解释：** 返回结束的原因。 1.stop：模型达到自然停止点或提供的停止序列； 2.length：达到请求中指定的最大令牌数； 3.content_filter：由于内容过滤器的标志而省略了内容。 **取值范围：** 不涉及。
index	Integer	**参数解释：** 返回多个choices时，每个choice对应的顺序。 **取值范围：** 不涉及。
message	**ChatCompletionResponseMessage** object	**参数解释：** 模型服务返回的具体消息体内容。 **取值范围：** 不涉及。

表 4-9 ChatCompletionResponseMessage

参数	参数类型	描述
content	String	**参数解释：** 返回消息体的内容。 **取值范围：** 不涉及。
role	String	**参数解释：** 返回消息体的角色。 **取值范围：** 不涉及。

表 4-10 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	**参数解释：** 回答包含的token数。 **取值范围：** 不涉及。
prompt_tokens	Integer	**参数解释：** 提问包含的token数。 **取值范围：** 不涉及。
total_tokens	Integer	**参数解释：** 提问+回答token总数。 **取值范围：** 不涉及。

状态码： 500

表 4-11 响应 Body 参数

参数	参数类型	描述
error	**Error** object	**参数解释：** 异常详情。 **取值范围：** 不涉及。

参数	参数类型	描述
error_code	String	**参数解释：** 平台异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常信息。 **取值范围：** 不涉及。

表 4-12 Error

参数	参数类型	描述
code	String	**参数解释：** 异常码。 **取值范围：** 不涉及。
message	String	**参数解释：** 异常信息。 **取值范围：** 不涉及。
param	String	**参数解释：** 异常参数，暂未使用。 **取值范围：** 不涉及。
type	String	**参数解释：** 异常类型，同code。 **取值范围：** 不涉及。

请求示例

```
{
  "model" : "publisher:baichuan:Baichuan2-Turbo",
  "messages" : [ {
    "role" : "system",
    "content" : "You are a helpful assistant."
  }, {
    "role" : "user",
    "content" : "你好!"
  } ]
}
```

响应示例

状态码： 200

OK

```
{
  "created" : 1718772336,
  "usage" : {
    "completion_tokens" : 23,
    "prompt_tokens" : 45,
    "total_tokens" : 68
  },
  "model" : "Baichuan2-Turbo",
  "id" : "chatcmpl-xxx",
  "choices" : [ {
    "finish_reason" : "stop",
    "index" : 0,
    "message" : {
      "role" : "assistant",
      "content" : "你好，有什么我可以帮助你的吗？"
    },
    "logprobs" : null
  } ],
  "object" : "chat.completion"
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error" : {
    "message" : "Internal server error, please try again later!",
    "type" : "internal_error",
    "param" : null,
    "code" : "internal_error"
  },
  "error_code" : "AIAE.31001001",
  "error_msg" : "Internal server error, please try again later!"
}
```

状态码

状态码	描述
200	OK
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.1.2 调用文本向量化模型服务

功能介绍

将用户输入的文本转化成数字向量，多用于从向量化知识库中查询相似的文本。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/embeddings

请求参数

表 4-13 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-14 请求 Body 参数

参数	是否必选	参数类型	描述
input	是	Array of strings	**参数解释：** 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"]。 **约束限制：** 输入长度小于25M，且列表元素数量小于1000。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
model	是	String	**参数解释：** 模型服务调用唯一id字段。平台定义了4种模型服务： 1.平台预置模型服务 登录**AI原生应用引擎**，在左侧导航栏选择”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。 2.租户部署模型服务 登录**AI原生应用引擎**，在左侧导航栏选择“模型中心 > 我的模型服务”，选择“我部署的”页签，在模型服务列表中复制模型服务调用ID。 3.租户接入模型服务 登录**AI原生应用引擎**，在左侧导航栏选择“模型中心 > 我的模型服务”，选择“我接入的”页签，在模型服务列表中复制模型服务调用ID。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-15 响应 Body 参数

参数	参数类型	描述
data	Array of **Embedding** objects	**参数解释：** 向量化结果。 **取值范围：** 不涉及。

参数	参数类型	描述
model	String	**参数解释：** 实际转发后调用的模型名称，与请求体中model可能不同。 **取值范围：** 不涉及。
object	String	**参数解释：** 固定值。 **取值范围：** ‘list’
usage	**usage** object	**参数解释：** 每次请求的用量统计。 **取值范围：** 不涉及。

表 4-16 Embedding

参数	参数类型	描述
index	Integer	**参数解释：** 向量在向量列表中的排序。 **取值范围：** 不涉及。
embedding	Array of numbers	**参数解释：** 向量数组（Float类型）。 **取值范围：** 不涉及。
object	String	**参数解释：** 固定值。 **取值范围：** ‘embedding’

表 4-17 usage

参数	参数类型	描述
prompt_tokens	Integer	**参数解释：** 提问包含的token数。 **取值范围：** 不涉及。
total_tokens	Integer	**参数解释：** 提问包含的token数。 **取值范围：** 不涉及。

状态码： 500

表 4-18 响应 Body 参数

参数	参数类型	描述
error	**Error** object	**参数解释：** 异常详情。 **取值范围：** 不涉及。
error_code	String	**参数解释：** 平台异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常信息。 **取值范围：** 不涉及。

表 4-19 Error

参数	参数类型	描述
code	String	**参数解释：** 异常码。 **取值范围：** 不涉及。

参数	参数类型	描述
message	String	**参数解释：** 异常信息。 **取值范围：** 不涉及。
param	String	**参数解释：** 异常参数，暂未使用。 **取值范围：** 不涉及。
type	String	**参数解释：** 异常类型，同code。 **取值范围：** 不涉及。

请求示例

```
{
  "model" : "publisher:zhipu:embedding-2",
  "input" : "你好啊"
}
```

响应示例

状态码： 200

OK

```
{
   "data": [
      {
         "index": 0,
         "embedding": [
            0.02513289265334606,
            -0.017512470483779907,
            -0.029955564066767693,
            ...
         ],
   "object": "embedding"
         }
   ],
   "usage": {
      "prompt_tokens": 5,
      "total_tokens": 5
   },
   "model": "embedding-2",
   "object": "list"
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error" : {
```

```
  "message" : "Internal server error, please try again later!",
  "type" : "internal_error",
  "param" : null,
  "code" : "internal_error"
 },
 "error_code" : "AIAE.31001001",
 "error_msg" : "Internal server error, please try again later!"
}
```

状态码

状态码	描述
200	OK
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2 应用中心

4.2.1 调用知识检索流

功能介绍

该接口用于调用用户配置的知识检索流。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/workflow-adapter-open/rag-flows/{flow_id}

表 4-20 路径参数

参数	是否必选	参数类型	描述
flow_id	是	String	**参数解释：** 知识检索流ID。进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识检索流”，在流列表中复制检索流ID。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-21 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-22 请求 Body 参数

参数	是否必选	参数类型	描述
body	是	Object	**参数解释：** 结构与知识检索流的起始节点配置相关，如果为GET请求则为非必填，如果为POST请求则为必填。 比如检索流配置了query_param作为查询参数，header_param作为请求头参数，body_param_1与body_param_2作为请求体参数，此时调用本接口只需要将这些参数依次传入，AI原生应用引擎自动按照名称进行分配，并完成检索流的调用。 具体结构请参照本接口的请求实例。 **约束限制：** 不涉及

响应参数

状态码： 200

表 4-23 响应 Body 参数

参数	参数类型	描述
responseBody	String	**参数解释：** 流执行结果的内容。 **取值范围：** 不涉及。
responseHeaders	Object	**参数解释：** 流执行结果的响应头。 **取值范围：** 不涉及。

状态码： 500

表 4-24 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "query_param" : "query_example",
  "header_param" : "header_example",
  "body_param_1" : "body_example_1",
  "body_param_2" : "body_example_2"
}
```

响应示例

状态码： 200

成功。

```
{
  "data" : {
    "responseBody" : "something in response body",
    "responseHeaders" : {
      "Server" : "api-gateway",
      "X-Request-Id" : "787b7740f42e75b007ac3bfb599fcef4",
      "X-Content-Type-Options" : "nosniff",
      "Connection" : "keep-alive",
      "lubanops-nspan-id" : "1",
      "X-Download-Options" : "noopen",
      "Date" : "Tue, 23 Jul 2024 11:38:29 GMT",
      "lubanops-ntrace-id" : "2748112-1721734708992-1130609",
      "Referrer-Policy" : "no-referrer",
      "X-Frame-Options" : "SAMEORIGIN",
      "Strict-Transport-Security" : "max-age=31536000; includeSubdomains;",
      "lubanops-nenv-id" : "28164",
      "Content-Length" : "0",
      "X-XSS-Protection" : "1; mode=block;",
      "Content-Type" : "application/json"
    },
    "statusCode" : 200
  }
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.22009001",
```

```
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.2 调用流

功能介绍

该接口用于调用用户配置的流。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/workflow-adapter-open/flows/{flow_id}

表 4-25 路径参数

参数	是否必选	参数类型	描述
flow_id	是	String	**参数解释：** 流ID，进入**AI原生应用引擎**，在左侧导航栏选择“Agent编排中心 > 我的工作流”，在流列表中复制流ID。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-26 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-27 请求 Body 参数

参数	是否必选	参数类型	描述
body	是	Object	**参数解释：** 结构与流的起始节点配置相关，如果为GET请求则为非必填，如果为POST请求则为必填。 比如工作流配置了query_param作为查询参数，header_param作为请求头参数，body_param_1与body_param_2作为请求体参数，此时调用本接口只需要将这些参数依次传入，AI原生应用引擎自动按照名称进行分配，并完成工作流的调用。 具体结构请参照本接口的请求实例。 **约束限制：** 不涉及。

响应参数

状态码： 200

表 4-28 响应 Body 参数

参数	参数类型	描述
responseBody	String	**参数解释：** 流执行结果的内容。 **取值范围：** 不涉及。
responseHeaders	Object	**参数解释：** 流执行结果的响应头。 **取值范围：** 不涉及。

状态码： 500

表 4-29 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "query_param" : "query_example",
  "header_param" : "header_example",
  "body_param_1" : "body_example_1",
  "body_param_2" : "body_example_2"
}
```

响应示例

状态码： 200

成功。

```
{
  "data" : {
    "responseBody" : "something in response body",
    "responseHeaders" : {
      "Server" : "api-gateway",
      "X-Request-Id" : "787b7740f42e75b007ac3bfb599fcef4",
```

```
    "X-Content-Type-Options" : "nosniff",
    "Connection" : "keep-alive",
    "lubanops-nspan-id" : "1",
    "X-Download-Options" : "noopen",
    "Date" : "Tue, 23 Jul 2024 11:38:29 GMT",
    "lubanops-ntrace-id" : "2748112-1721734708992-1130609",
    "Referrer-Policy" : "no-referrer",
    "X-Frame-Options" : "SAMEORIGIN",
    "Strict-Transport-Security" : "max-age=31536000; includeSubdomains;",
    "lubanops-nenv-id" : "28164",
    "Content-Length" : "0",
    "X-XSS-Protection" : "1; mode=block;",
    "Content-Type" : "application/json"
   },
   "statusCode" : 200
  }
}
```


状态码： 500

服务器内部错误或三方服务器内部错误。


```
{
  "error_code" : "AIAE.22009001",
  "error_msg" : "Internal Server Error."
}
```


状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.3 调用工具的执行动作

功能介绍

该接口用于调用用户配置的工具的执行动作。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/workflow-adapter-open/skills/{skill_id}

表 4-30 路径参数

参数	是否必选	参数类型	描述
skill_id	是	String	**参数解释：** 工具的执行动作ID，进入**AI原生应用引擎**，在左侧导航栏选择“Agent编排中心 > 我的工具”，在工具列表中复制执行动作ID。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-31 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-32 请求 Body 参数

参数	是否必选	参数类型	描述
body	是	Object	**参数解释：** 结构与工具的执行动作的配置相关，并且所有请求头中的入参与请求参数均添加至请求体中，由AI原生应用引擎自动完成分配。如果为GET请求则为非必填，如果为POST请求则为必填。 比如工作流配置了query_param作为查询参数，header_param作为请求头参数，body_param_1与body_param_2作为请求体参数，此时调用本接口只需要将这些参数依次传入，AI原生应用引擎自动按照名称进行分配，并完成工具的执行动作的调用。 具体结构请参照本接口的请求实例。 **约束限制：** 不涉及。

响应参数

状态码： 200

表 4-33 响应 Body 参数

参数	参数类型	描述
data	**data** object	响应的body参数。

表 4-34 data

参数	参数类型	描述
id	String	**参数解释：** 调用记录ID。 **取值范围：** 只由英文字母、数字以及“-”组成，且长度为36个字符。

参数	参数类型	描述
version	Number	**参数解释：** 工具的版本号。 **取值范围：** 正整数。
connector_id	String	**参数解释：** 工具ID。 **取值范围：** 只由英文字母、数字以及“-”组成，且长度为36个字符。
action_id	String	**参数解释：** 工具的执行动作ID。 **取值范围：** 只由英文字母、数字以及“-”组成，且长度为36个字符。
start_time	String	**参数解释：** 本次调用的开始时间。 **取值范围：** UTC格式的日期。
end_time	String	**参数解释：** 本次调用的结束时间。 **取值范围：** UTC格式的日期。
cost	Number	**参数解释：** 本次调用的总耗时，单位为毫秒。 **取值范围：** 正整数。
status	String	**参数解释：** 本次调用的结果。 **取值范围：** ● success ● failure
status_code	Number	**参数解释：** 本次调用的状态码。 **取值范围：** 不涉及。

参数	参数类型	描述
method	String	**参数解释：** 本次调用的方法。 **取值范围：** • GET • POST • PUT • DELETE
path	String	**参数解释：** 本次调用的url。 **取值范围：** 不涉及。
invoke_output	**invoke_output** object	**参数解释：** 本次调用的输出结果，即返回体。 **取值范围：** 不涉及。
invoke_input	**invoke_input** object	**参数解释：** 本次调用的输入内容，即请求体。 **取值范围：** 不涉及。

表 4-35 invoke_output

参数	参数类型	描述
body	String	**参数解释：** 本次调用的输出结果具体内容。 **取值范围：** 不涉及。

表 4-36 invoke_input

参数	参数类型	描述
body	String	**参数解释：** 本次调用的请求体。 **取值范围：** 不涉及。

参数	参数类型	描述
header	String	**参数解释：** 本次调用的请求头。 **取值范围：** 不涉及。

状态码： 500

表 4-37 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "query_param" : "query_example",
  "header_param" : "header_example",
  "body_param_1" : "body_example_1",
  "body_param_2" : "body_example_2"
}
```

响应示例

状态码： 200

成功。

```
{
  "data" : {
    "id" : "6f46e379-9adf-4395-af0d-4549e09c4048",
    "version" : 3,
    "connector_id" : "d5a2b8fd-ad02-437d-9234-2225eb992fd6",
    "action_id" : "a224ce98-07b5-479a-b75e-560029399312",
    "start_time" : "2024-12-28T01:17:31.146Z",
    "end_time" : "2024-12-28T01:17:33.582Z",
    "cost" : 2436,
    "status" : "success",
    "status_code" : 200,
    "method" : "POST",
    "path" : "some path",
    "invoke_output" : {
      "body" : "something in response body"
```

```
  },
  "invoke_input" : {
    "body" : "something in request body",
    "header" : "something in request header"
  }
 }
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
 "error_code" : "AIAE.22009001",
 "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.4 上传文件用于测试流

功能介绍

该接口用于测试流前向AI原生引用引擎上传文件，目前仅支持图片格式（jpg、png、jpeg），为上传的文件提供临时访问路径，后续可以使用该访问路径调用测试流接口完成图片流的测试。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/workflow-adapter-open/common/file-upload

表 4-38 Query 参数

参数	是否必选	参数类型	描述
file	是	String	**参数解释：** 文件内容，目前仅支持jpg、png、jpeg格式的图片。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
type	是	String	**参数解释：** 文件类型，目前仅支持image图片类型。 **约束限制：** 不涉及。 **取值范围：** image **默认取值：** 不涉及。

请求参数

表 4-39 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
Content-Type	是	String	**参数解释：** 内容类型，调用此接口时固定为multipart/form-data。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-40 响应 Body 参数

参数	参数类型	描述
id	String	**参数解释：** 文件ID。 **取值范围：** 只由英文字母、数字以及“-”组成，且长度为36个字符。
temp_url	String	**参数解释：** 文件的临时访问外链。 **取值范围：** 只由英文字母、数字以及“/”、“_”、“-”组成。
expire_time	String	**参数解释：** 文件的临时访问外链的过期时间。 **取值范围：** 不涉及。

状态码： 500

表 4-41 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "file" : "file content",
  "type" : "image"
}
```

响应示例

状态码： 200

成功。

```
{
  "data" : {
    "id" : "file_id",
    "temp_url" : "temporary url of file which can access for some time",
    "expire_time" : "current temporary url will expire at this time"
  }
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.22009001",
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.5 上传文件用于调用 Agent

功能介绍

该接口用于调用agent前向AI原生应用引擎上传文件，目前仅支持图片格式（jpg、png、jpeg），为上传的文件提供访问路径，后续可以使用该访问路径调用agent对话接口。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/routes/open/file/upload

表 4-42 Query 参数

参数	是否必选	参数类型	描述
file	是	String	**参数解释：** 文件内容，目前仅支持jpg、png、jpeg格式的图片。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
type	是	String	**参数解释：** 文件类型，目前仅支持image图片类型。 **约束限制：** 不涉及。 **取值范围：** image **默认取值：** 不涉及。

请求参数

表 4-43 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
Content-Type	是	String	**参数解释：** 内容类型，调用此接口时固定为multipart/form-data。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-44 响应 Body 参数

参数	参数类型	描述
url	String	**参数解释：** 文件路径，可用于**调用Agent**。 **取值范围：** 不涉及。

状态码： 500

表 4-45 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "file" : "file content",
  "type" : "image"
}
```

响应示例

状态码： 200

成功。

```
{
  "url" : "url of file"
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.22009001",
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.6 上传文件至文件盒子

功能介绍

在Agent的文件盒子中上传文件。在完成文件上传后，可以在**调用Agent**时引用上传的文件进行对话。只支持上传pdf、txt、docx等纯文本文件，且文件大小不超过10MB。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/routes/open/fileBox/upload

表 4-46 Query 参数

参数	是否必选	参数类型	描述
agent-id	是	String	**参数解释：** Agent的唯一id。进入**AI原生应用引擎**，在左侧导航栏选择“Agent编排中心 > 我的Agent”，选择“我创建的”页签，选择列表操作列的“更多 > 修改”，在浏览器地址栏查看id。 **约束限制：** 不涉及。 **取值范围：** 由英文字母和数字组成，长度为32个字符。 **默认取值：** 不涉及。
attachment-code	是	String	**参数解释：** 附件码。此处请填写固定值：ai-file-box。 **约束限制：** 不涉及。 **取值范围：** • ai-file-box **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
file	是	String	**参数解释：** 待上传的文件。 **约束限制：** 文件需为pdf、txt、docx等文本文件，且大小不超过10MB。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求参数

表 4-47 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
Content-Type	是	String	**参数解释：** 消息体的类型。 **约束限制：** 由于需要上传文件，所以必须为multipart/form-data。 **取值范围：** • application/json • application/json;charset=utf-8 • multipart/form-data **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-48 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 生成的文件ID。 **取值范围：** 由英文字母和数字组成，长度为32个字符。

状态码： 500

表 4-49 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

向Agent的文件盒子中上传上海旅游攻略文件。

```
/v1/routes/open/fileBox/upload

{
  "agent-id" : "1eb7f2f6f105496c8065be77dc038b63",
  "attachment-code" : "ai-file-box",
  "file" : "上海旅游攻略.docx"
}
```

响应示例

状态码： 200

成功。

```
{
  "data" : "053f5dda365345a9a80cc63895df1647"
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.00001500",
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.7 删除文件盒子中的文件

功能介绍

调用本接口删除文件盒子中已有的文件。

调用方法

请参见**如何调用API**。

URI

DELETE https://aiae.appstage.myhuaweicloud.com/v1/routes/open/file/{file-id}

表 4-50 路径参数

参数	是否必选	参数类型	描述
file-id	是	String	**参数解释：** 文件id。调用**上传文件至文件盒子**接口时，返回体中的id即为上传成功的文件id。如果是在AI原生应用引擎页面中上传的文件，则可查看开发者工具，在页面中单击Agent名称查看详情，在detail接口的响应中查看"file_infos"字段中的"id"，即为文件id。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-51 Query 参数

参数	是否必选	参数类型	描述
agent-id	是	String	**参数解释：** Agent的唯一ID。进入**AI原生应用引擎**，在左侧导航栏选择“Agent编排中心 > 我的Agent”，选择“我创建的”页签，选择列表操作列的“更多 > 修改”，在浏览器地址栏查看id。 **约束限制：** 不涉及。 **取值范围：** 由英文字母和数字组成，长度为32个字符。 **默认取值：** 不涉及。

请求参数

表 4-52 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-53 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 删除成功。 **取值范围：** ● success

状态码： 500

表 4-54 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。

参数	参数类型	描述
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

- 删除Agent文件盒子中的文件。

```
DELETE https://aiae.appstage.myhuaweicloud.com/v1/routes/open/file/
053f5dda365345a9a80cc63895df1647?agent-id=1eb7f2f6f105496c8065be77dc038b63 \
--header "Authorization: sk-162xxxxxxxxxxx"
```

响应示例

状态码： 200

删除成功。

```
{
  "data" : "success"
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.00001500",
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	删除成功。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.2.8 调用 Agent

功能介绍

调用本接口，向已发布的Agent发起一次对话请求。若无已发布的Agent，请先在**AI原生应用引擎**中创建Agent并进行发布。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/routes/open/{id}/execute

表 4-55 路径参数

参数	是否必选	参数类型	描述
id	是	String	**参数解释：** Agent的唯一id。进入**AI原生应用引擎**，在左侧导航栏选择“Agent编排中心 > 我的Agent”，选择“我创建的”页签，选择列表操作列的“更多 > 修改”，在浏览器地址栏查看id。 **约束限制：** 不涉及。 **取值范围：** 由英文字母和数字组成，长度为32个字符。 **默认取值：** 不涉及。

请求参数

表 4-56 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-57 请求 Body 参数

参数	是否必选	参数类型	描述
query	是	String	**参数解释：** 发起对话的内容。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
quote	否	String	**参数解释：** 对话中引用文件盒子中文件的完整名称，包含后缀。 **约束限制：** 如果发布Agent时，用户在该Agent的文件盒子中没有文件，需要先**上传文件**。 **取值范围：** 只支持pdf、txt、docx等文本文件格式。 **默认取值：** 不涉及。
memory	否	Array of **memory** objects	**参数解释：** 在本次对话请求中，让大模型提前记住的部分。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
user_id	否	String	**参数解释：** 用户id，允许自定义，与conversation_id共同使用可以使Agent自动获取该用户相同对话id下的前几轮对话内容。 **约束限制：** 不与memory同时生效，memory存在时memory优先生效。 **取值范围：** 不涉及。 **默认取值：** AI引擎用户的真实user_id，orgid的不可逆加密值。
conversation_id	否	String	**参数解释：** 对话id，允许自定义，与user_id共同使用可以使Agent自动获取该用户相同对话id下的前几轮对话内容。 **约束限制：** 不与memory同时生效，memory存在时memory优先生效。 **取值范围：** 不涉及。 **默认取值：** 随机生成的仅包含英文字母与数字的32位字符串。

参数	是否必选	参数类型	描述
variables	否	Object	**参数解释：** 变量用于用户个人信息，例如语言偏好等，并让Agent记住这些特征，使回复更加个性化。 **约束限制：** 变量包括一般变量和敏感变量，一般变量可用于对话和工作流，敏感变量只用于工作流。如果Agent中设置了敏感变量，敏感变量为必传，一般变量非必需。 **取值范围：** 一般变量common_variables和敏感变量sensitive_variables。 **默认取值：** 不涉及。

表 4-58 memory

参数	是否必选	参数类型	描述
role	是	String	**参数解释：** 对话角色。 **约束限制：** 只支持user，assistant或tool三种取值。 **取值范围：** user，assistant或tool。 **默认取值：** 不涉及。
content	是	String	**参数解释：** 对话内容。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-59 响应 Body 参数

参数	参数类型	描述
request_id	String	**参数解释：** 唯一请求id。 **取值范围：** 只由英文字母及数字组成，长度为32个字符。
agent_id	String	**参数解释：** Agent的唯一id。 **取值范围：** 只由英文字母及数字组成，长度为32个字符。
user_id	String	**参数解释：** 本轮对话的用户唯一身份标识。 **取值范围：** ● 在发起对话请求时自定义的user_id。 ● AI引擎用户的真实user_id，orgid的不可逆加密值。
conversation_id	String	**参数解释：** 会话ID。 **取值范围：** ● 在发起对话请求时自定义的conversation_id。 ● 随机生成的仅包含英文字母与数字的32位字符串。
type	String	**参数解释：** 返回内容的类型。 **取值范围：** ● hint ● workflow ● tool ● knowledge ● message。
data	Object	**参数解释：** 不同响应类型的响应体中包含不同的参数，见示例。 **取值范围：** 不涉及。

状态码： 400

表 4-60 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

状态码： 500

表 4-61 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 错误信息。 **取值范围：** 不涉及。

请求示例

- 设置memory，向Agent发起一次对话请求。

```
/v1/routes/open/{id}/execute

{
  "query" : "查询北京天气",
  "memory" : [ {
    "role" : "user",
    "content" : "你是谁"
  }, {
    "role" : "assistant",
    "content" : "我是盘古大模型"
  }, {
    "role" : "user",
    "content" : "南京天气"
  }, {
    "role" : "assistant",
```

```
    "tool_calls" : [ {
      "id" : "efd6ff92-422c-4ba4-b531-ac1991af7c1a",
      "type" : "function",
      "function" : {
        "name" : "查询当前天气 查询当前天气",
        "arguments" : "{\"city\":\"320100\",\"extensions\":\"all\"}"
      }
    } ]
  }, {
    "role" : "tool",
    "tool_call_id" : "efd6ff92-422c-4ba4-b531-ac1991af7c1a",
    "content" : "{\"data\":{\"status\":\"1\",\"count\":\"1\",\"info\":\"OK\",\"infocode
\":\"10000\",\"forecasts\":[{\"city\":\"南京市\",\"adcode\":\"320100\",\"province\":\"江苏\",\"reporttime
\":\"2024-08-20 16:32:01\",\"casts\":[{\"date\":\"2024-08-20\",\"week\":\"2\",\"dayweather\":\"中雨
\",\"nightweather\":\"中雨\",\"daytemp\":\"32\",\"nighttemp\":\"26\",\"daywind\":\"西北\",\"nightwind
\":\"西北\",\"daypower\":\"1-3\",\"nightpower\":\"1-3\",\"daytemp_float\":\"32.0\",\"nighttemp_float
\":\"26.0\"}]}]}}"
  } ]
}
```

- 设置引用文件、自定义的用户id和对话id、一般变量，向Agent发起一次对话请求。

```
/v1/routes/open/{id}/execute

{
  "query" : "上海有哪些旅游景点",
  "quote" : "上海旅游攻略.docx",
  "user_id" : 1008600000300604420,
  "conversation_id" : "70dc6dfd397244edbbf4847acb78bfa9a",
  "variables" : {
    "common_variables" : {
      "上海古典名园" : "豫园"
    }
  }
}
```

响应示例

状态码： 200

- 表示Agent成功接收对话请求，正常响应。
- 响应采用Server-Sent Events（SSE）机制进行流式输出，数据内容用data字段表示，示例如下：data: message；需注意，这并不是一个JSON数据。此外，还可以有冒号开头的行，表示注释，通常可忽略。
- Agent的响应类型为hint，用于提示接下来使用knowledge，tool或workflow进行响应，示例中的提示为workflow。

```
{
  "request_id" : "266e71692aba45ec8b111d847963109d",
  "agent_id" : "50b58e0041e843b0b7bd343dca076443",
  "user_id" : 1008600000300604420,
  "conversation_id" : "c568e962b0004650bc12b63aec96366d",
  "type" : "hint",
  "data" : {
    "id" : "ce7f20a1-a3dd-4249-b7ca-8c5039dd8c74",
    "name" : "儿科问答",
    "tool_type" : "workflow"
  }
}
```

- Agent使用workflow进行响应。

```
{
  "request_id" : "266e71692aba45ec8b111d847963109d",
  "agent_id" : "50b58e0041e843b0b7bd343dca076443",
  "user_id" : 1008600000300604420,
  "conversation_id" : "c568e962b0004650bc12b63aec96366d",
```

```
 "type" : "workflow",
 "data" : {
  "id" : "ce7f20a1-a3dd-4249-b7ca-8c5039dd8c74",
  "name" : "儿科问答",
  "status" : "SUCCESS",
  "request" : {
   "query" : "婴儿肥胖怎么办"
  },
  "response" : {
   "data" : {
    "responseBody" : "{\"result\":\"问题分析:主要控制儿童饮食，合理饮食不喝酒，不吃油炸食物意见和建议:建议孩子们多锻炼一点，每天至少锻炼一到两个小时，而且他们必须坚持锻炼。他们也应该少吃油和脂肪，多吃水果和蔬菜。我认为我们应该在一段时间后恢复正常。就食疗而言，父母必须参与其中，并被要求掌握一些相关知识，如不允许孩子吃得太多或太多，不给予高糖、高脂肪、高热量的饮食。治疗节食中的儿童并让他们挨饿也很困难。因此，在进行饮食控制之前，有必要耐心而详细地告诉儿童肥胖的危害、\"}",
    "responseHeaders" : {
     "Server" : "api-gateway",
     "X-Request-Id" : "6701c75b8f23102a659e63a3cc5a20d6",
     "X-Content-Type-Options" : "nosniff",
     "Connection" : "keep-alive",
     "X-Download-Options" : "noopen",
     "Date" : "Tue, 20 Aug 2024 08:37:27 GMT",
     "Referrer-Policy" : "no-referrer",
     "X-Frame-Options" : "SAMEORIGIN",
     "Strict-Transport-Security" : "max-age=31536000; includeSubdomains;",
     "lubanops-nenv-id" : "28164",
     "Content-Length" : "660",
     "X-XSS-Protection" : "1; mode=block;",
     "Content-Type" : "application/json"
    },
    "statusCode" : 200
   }
  }
 }
}
```


- Agent使用tool进行响应。


```
{
 "request_id" : "266e71692aba45ec8b111d847963109d",
 "agent_id" : "50b58e0041e843b0b7bd343dca076443",
 "user_id" : 1008600000300604420,
 "conversation_id" : "c568e962b0004650bc12b63aec96366d",
 "type" : "tool",
 "data" : {
  "id" : "b6dbf1a6-f374-4d44-96fb-45726f7fa7f0",
  "name" : "航班信息 航班信息",
  "status" : "SUCCESS",
  "request" : {
   "city" : "南京",
   "endcity" : "大理",
   "date" : "2024-08-24"
  },
  "response" : {
   "data" : {
    "status" : 0,
    "msg" : "ok",
    "result" : {
     "city" : "NKG",
     "endcity" : "DLU",
     "date" : "2024-08-24",
     "list" : [ {
      "flightno" : "ZH2010",
      "airline" : "深圳航空",
      "realflightno" : "TV6026",
      "departportcode" : "NKG",
      "departport" : "禄口国际机场",
      "arrivalportcode" : "DLU",
      "arrivalport" : "大理荒草坝机场",
      "departterminal" : "T1",
```

```
          "arrivalterminal" : "",
          "departdate" : "2024-08-24",
          "arrivaldate" : "2024-08-24",
          "departtime" : "16:35",
          "arrivaltime" : "19:40",
          "departdateadd" : 0,
          "arrivaldateadd" : 0,
          "craft" : "19N",
          "stopnum" : "0",
          "costtime" : "03:05",
          "punctualrate" : "95",
          "pricelist" : [ ],
          "minprice" : "0",
          "airporttax" : "50",
          "fueltax" : "50",
          "food" : "1,",
          "isasr" : "1,",
          "iseticket" : "1,",
          "iscodeshare" : 1
        } ]
      }
    }
  }
 }
}
```


- Agent使用knowledge进行响应。


```
{
 "request_id" : "266e71692aba45ec8b111d847963109d",
 "agent_id" : "50b58e0041e843b0b7bd343dca076443",
 "user_id" : 1008600000300604420,
 "conversation_id" : "c568e962b0004650bc12b63aec96366d",
 "type" : "knowledge",
 "data" : {
  "id" : "7e4cf06bd8404ec594c621c8f47c44f1",
  "name" : "华西医院",
  "status" : "SUCCESS",
  "request" : {
   "query" : "肠息肉怎么办"
  },
  "response" : "结直肠息肉应该怎么办？\n结直肠息肉是什么\n结肠息肉需不需要切除"
 }
}
```


- Agent使用大模型进行响应。


```
{
 "request_id" : "266e71692aba45ec8b111d847963109d",
 "agent_id" : "50b58e0041e843b0b7bd343dca076443",
 "user_id" : 1008600000300604420,
 "conversation_id" : "c568e962b0004650bc12b63aec96366d",
 "type" : "knowledge",
 "data" : {
  "id" : "202412302111448bd332d627ed4c5f",
  "content" : "著名的",
  "url" : null,
  "raw" : {
   "role" : "assistant",
   "content" : "著名的"
  }
 }
}
```


状态码： 400

缺少请求体。


```
{
 "error_code" : "AIAE.00001400",
 "error_msg" : "Request body is missing"
}
```


状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.00001500",
  "error_msg" : "Internal Server Error."
}
```

状态码

状态码	描述
200	• 表示Agent成功接收对话请求，正常响应。 • 响应采用Server-Sent Events（SSE）机制进行流式输出，数据内容用data字段表示，示例如下：data: message；需注意，这并不是一个JSON数据。此外，还可以有冒号开头的行，表示注释，通常可忽略。
400	缺少请求体。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.3 知识中心

4.3.1 检索知识库数据

功能介绍

检索知识库数据，根据用户提供的检索信息，返回命中的信息数据。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases/{knowledge_base_id}/embed-datas

表 4-62 路径参数

参数	是否必选	参数类型	描述
knowledge_base_id	是	String	**参数解释：** 知识库id。获取方式： 1.可从创建知识库接口返回获取； 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，可从页面知识库ID栏获取。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-63 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-64 请求 Body 参数

参数	是否必选	参数类型	描述
keyword	否	String	**参数解释：** 搜索关键字，用于检索匹配结果。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，字符串长度介于0到2048之间。 **默认取值：** 不涉及。
similarity_min	否	Float	**参数解释：** 相似度最小值，数值越大表示相似度越高。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，取值[0, 1]。 **默认取值：** 不涉及。
limit	是	Integer	**参数解释：** 检索返回切片限制数量。 **约束限制：** 不涉及。 **取值范围：** 非空，取值[1, 100]，默认上限100。 **默认取值：** 不涉及。
filter	否	Object	**参数解释：** 过滤条件。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，则为SearchSqlFilter类对象。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
order_by	否	Object	**参数解释：** 排序规则。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，则为SqlOrder对象。 **默认取值：** 不涉及。
data_sets	否	Array of **DataSetSearchInfo** objects	**参数解释：** 检索的数据集信息，用于指定检索知识库中部分数据集。 **约束限制：** 不涉及。 **取值范围：** 可以为null，如果不为null，则为DataSetSearchInfo对象列表，且数量介于1到10之间。 **默认取值：** 不涉及。

表 4-65 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。

参数	是否必选	参数类型	描述
expressions	是	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 4-66 Expression

参数	是否必选	参数类型	描述
field	是	String	**参数解释：** 过滤字段。 **约束限制：** 需以“metadata.”开头，后续可接上path、order、file_name以及索引配置时配置的文本过滤字段，如metadata.path、metadata.answer（假设answer为配置的文本过滤字段）。 **取值范围：** 非空，字段名称长度介于1到100之间。 **默认取值：** 不涉及。
field_type	是	String	**参数解释：** 过滤字段类型。 **约束限制：** 不涉及。 **取值范围：** 非空，枚举值：INT、FLOAT、BOOLEAN和STRING，取值视field的类型而定。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
operator	是	String	**参数解释：** 过滤操作符。 **约束限制：** 不涉及。 **取值范围：** 非空，枚举值：EQUAL（等于）、NOT_EQUAL（不等于）、GREAT_THAN（大于）、GREAT_EQUAL（大于等于）、LESS_THAN（小于）、LESS_EQUAL（小于等于）、IN（在xxx之间）、NOTIN（不在xxx之间）和STARTS_WITH（以xxx开头）。 **默认取值：** 不涉及。
values	是	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 4-67 SqlOrder

参数	是否必选	参数类型	描述
order_items	是	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 4-68 OrderItem

参数	是否必选	参数类型	描述
field	是	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：** STRING。
order_type	是	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。

表 4-69 DataSetSearchInfo

参数	是否必选	参数类型	描述
data_set_id	否	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集ID。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-70 响应 Body 参数

参数	参数类型	描述
data	Array of **ChunkData** objects	**参数解释：** 知识库检索响应数据。 **取值范围：** 不涉及。

表 4-71 ChunkData

参数	参数类型	描述
id	String	**参数解释：** 切片id。 **取值范围：** 由数字、字母和中划线组成，长度36。

参数	参数类型	描述
document	String	**参数解释：** 向量化内容。 **取值范围：** 不涉及。
chunk	String	**参数解释：** 完整切片内容。若用户在创建索引时配置了chunk作为检索附加字段，则检索时在此返回。 **取值范围：** 不涉及。
chunk_fragments	Map<String,String>	**参数解释：** 附加返回字段及内容（map类型，key为附加返回字段名称，value为附加返回字段内容），若用户在创建索引时配置了除chunk之外的附加返回字段，则字段名称及内容检索时在此返回。 **取值范围：** 不涉及。
similarity	Float	**参数解释：** 向量化内容（document）和检索关键字（keyword）的向量相似度门槛，只有相似度大于等于similarity才会返回。 **取值范围：** 大小为[0,1]。
metadata	**metadata** object	**参数解释：** 切片元数据信息，若用户在创建索引时选择文本过滤字段，则该字段及其内容会作为元数据一部分返回。 **取值范围：** 不涉及。
download_addresses	Map<String,String>	**参数解释：** 图片或视频临时下载地址（map类型，key为文件路径，value为下载地址）。 **取值范围：** 不涉及。
download_address	String	**参数解释：** 废弃字段，请使用download_addresses。 **取值范围：** 不涉及。

参数	参数类型	描述
data_set_id	String	**参数解释：** 切片所属知识数据集id。 **取值范围：** 由数字、字母和中划线组成，长度36。

表 4-72 metadata

参数	参数类型	描述
order	Integer	**参数解释：** 切片序号。 **取值范围：** 不涉及。
file_name	String	**参数解释：** 文件名称。 **取值范围：** 不涉及。
path	String	**参数解释：** 文件相对路径。 **取值范围：** 不涉及。

状态码： 500

表 4-73 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "keyword" : "户外",
  "similarity_min" : "0.78",
  "limit" : 10,
  "filter" : {
    "group_type" : "AND",
    "expressions" : [ {
      "field" : "metadata.answer",
      "field_type" : "STRING",
      "operator" : "EQUAL",
      "values" : [ "我是xxx。" ]
    } ]
  },
  "order_by" : {
    "order_items" : [ {
      "field" : "metadata.order",
      "field_type" : "INT",
      "order_type" : "DESC"
    } ]
  },
  "data_sets" : [ {
    "data_set_id" : "a31ed909-xxxx-xxxx-xxxx-10958c90b3f7"
  } ]
}
```

响应示例

状态码： 200

OK。

```
{
  "data" : [ {
    "id" : "812857ef-xxxx-xxxx-xxxx-24ba9fd5e95c",
    "document" : "问题：你是谁。回答：我是xxx。回答用户id：000。",
    "chunk" : "问题：你是谁。回答：我是xxx。回答用户id：000。",
    "chunk_fragments" : {
      "question" : "你是谁。",
      "answer" : "我是xxx。"
    },
    "similarity" : 0.87,
    "metadata" : {
      "order" : 10,
      "file_name" : "户外运动热度大大带动各相关产业发展.docx",
      "path" : "户外运动热度大大带动各相关产业发展.docx",
      "question" : "你是谁。",
      "answer" : "我是xxx。"
    },
    "download_addresses" : {
      "xxx.png" : "https://xxxx"
    },
    "download_address" : null,
    "data_set_id" : "3967c49d-xxxx-xxxx-xxxx-5eda056a1f1b"
  } ]
}
```

状态码： 500

服务器内部错误或三方服务器内部错误。

```
{
  "error_code" : "AIAE.00001500",
  "error_msg" : "系统内部错误。"
}
```

状态码

状态码	描述
200	OK。
500	服务器内部错误或三方服务器内部错误。

错误码

请参见**错误码**。

4.3.2 创建知识库

功能介绍

该接口用于创建知识库，创建的知识库启用后可在创建Agent时引用。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases

请求参数

表 4-74 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-75 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	**参数解释：** 知识库名称。 **约束限制：** 不涉及。 **取值范围：** 支持中英文、数字、“_”，长度为[2-50]，以中英文、数字开头。 **默认取值：** 不涉及。
description	否	String	**参数解释：** 知识库描述。 **约束限制：** 不涉及。 **取值范围：** 只能包含英文，中文，数字，下划线，中划线，空格及,.?:;"'：；“”’ ‘，。？、()（）/等符号，最长255个字符。 **默认取值：** 不涉及。
retrieval_status	是	String	**参数解释：** 知识库召回状态。 **约束限制：** 不涉及。 **取值范围：** 枚举值：ENABLE（启用召回）、DISABLE（禁用召回）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
rag_type	否	String	**参数解释：** 知识库RAG类型。 **约束限制：** 不涉及。 **取值范围：** 可以为空，为空则使用默认值。枚举值：VECTOR_RAG（向量RAG，是一种结合了向量化和大语言模型的RAG技术）、GRAPH_RAG（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）。 **默认取值：** VECTOR_RAG。
retrieval_config	否	**RetrievalConfig** object	**参数解释：** 知识库检索召回配置。 **约束限制：** 不涉及。 **取值范围：** 可以为空，为空则使用默认值。 **默认取值：** 若不传，则检索配置（retrieval_config）中，检索模式（retrieval_modes）默认为语义检索（SEMANTIC_RETRIEVAL）。
knowledge_data_sets	是	Array of **KnowledgeDataSet** objects	**参数解释：** 知识数据集信息列表，选择知识数据集创建知识库。 **约束限制：** 不涉及。 **取值范围：** 不允许为空，数量最小为1，最大为5。 **默认取值：** 不涉及。

表 4-76 RetrievalConfig

参数	是否必选	参数类型	描述
retrieval_modes	是	Array of strings	**参数解释：** 检索模式，用于设置知识库检索召回时的检索方式。 **约束限制：** 不涉及。 **取值范围：** 可传多个值。枚举值：SEMANTIC_RETRIEVAL（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），FULL_TEXT_RETRIEVAL（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）。 **默认取值：** 不涉及。
retrieval_hybrid_mode	否	String	**参数解释：** 检索模式选择SEMANTIC_RETRIEVAL和FULL_TEXT_RETRIEVAL时，为混合检索。此参数用于指定混合检索的模式。 **约束限制：** 检索模式需选择SEMANTIC_RETRIEVAL和FULL_TEXT_RETRIEVAL。 **取值范围：** 当前仅支持RRF，枚举值：RRF。 **默认取值：** 不涉及。

表 4-77 KnowledgeDataSet

参数	是否必选	参数类型	描述
data_set_id	是	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。 **约束限制：** 需要先调用创建知识数据集接口，接口返回即为知识数据集id。 **取值范围：** 仅支持数字、字母和中划线。 **默认取值：** 不涉及。
data_set_version	是	String	**参数解释：** 数据集版本。 **约束限制：** 需根据知识数据集id，调用查询数据集详情接口，获取版本。 **取值范围：** 格式为：v2024-11-21T11:36:55Z。 **默认取值：** 不涉及。
index_config_id	是	String	**参数解释：** 索引配置id。 **约束限制：** 需根据知识数据集id，调用查询数据集详情接口，获取索引配置id。 **取值范围：** 仅支持数字、字母和中划线。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-78 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 正常返回的结果。 **取值范围：** 不涉及。

状态码： 400

表 4-79 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "name" : "知识库名称",
  "description" : "知识库描述",
  "retrieval_status" : "ENABLE",
  "rag_type" : "VECTOR_RAG",
  "retrieval_config" : {
    "retrieval_modes" : [ "SEMANTIC_RETRIEVAL", "FULL_TEXT_RETRIEVAL" ],
    "retrieval_hybrid_mode" : "RRF"
  },
  "knowledge_data_sets" : [ {
    "data_set_id" : "djh28e62-xxxxxxxxx-a15be0d63812",
    "data_set_version" : "v2024-11-21T11:36:55Z",
    "index_config_id" : "d3f28e62-xxxxxxx-a15be0d638a2"
  } ]
}
```

响应示例

状态码： 200

操作成功，返回知识库id。

```
{
  "data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回知识库id。
400	请求错误。

错误码

请参见**错误码**。

4.3.3 删除知识库

功能介绍

该接口用于删除知识库。

调用方法

请参见**如何调用API**。

URI

DELETE https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases/{knowledge_base_id}

表 4-80 路径参数

参数	是否必选	参数类型	描述
knowledge_base_id	是	String	**参数解释：** 知识库id。获取方式： 1.从创建知识库接口返回获取。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，可从页面知识库id栏获取。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求参数

表 4-81 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-82 响应 Body 参数

参数	参数类型	描述
data	DeleteResult object	**参数解释：** 知识数据集删除结果。 **取值范围：** 不涉及。

表 4-83 DeleteResult

参数	参数类型	描述
id	String	**参数解释：** 知识库id。 **取值范围：** 由数字、字母和中划线组成，长度36。
name	String	**参数解释：** 知识库名称。 **取值范围：** 由中英文、数字、“_”组成，长度为[2-50]。
result	Boolean	**参数解释：** 知识库删除成功或失败。 **取值范围：** true或false。
reason	String	**参数解释：** 知识库删除失败原因。 **取值范围：** 不涉及.

状态码： 400

表 4-84 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。

参数	参数类型	描述
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回删除结果。

```
{
  "data" : {
    "id" : "djh28e62-xxxxxxxxx-a15be0d63812",
    "name" : "知识库名称",
    "result" : false,
    "reason" : "知识库已启用"
  }
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回删除结果。
400	请求错误。

错误码

请参见**错误码**。

4.3.4 执行知识库

功能介绍

该接口用于执行知识库，将知识数据集的更新同步到知识库中。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases/{knowledge_base_id}/execute

表 4-85 路径参数

参数	是否必选	参数类型	描述
knowledge_base_id	是	String	**参数解释：** 知识库id。获取方式： 1.从创建知识库接口返回获取。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，可从页面知识库id栏获取。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求参数

表 4-86 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-87 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 正常返回的结果。 **取值范围：** 不涉及。

状态码： 400

表 4-88 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回执行记录id。

```
{
  "data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回执行记录id。
400	请求错误。

错误码

请参见**错误码**。

4.3.5 查询知识库最新执行记录

功能介绍

该接口用于查询知识库最新执行记录，通过该接口可获取知识库最新执行的结果，开始时间，进度，耗时等信息。

调用方法

请参见**如何调用API**。

URI

GET https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases/{knowledge_base_id}/latest-execution-record

表 4-89 路径参数

参数	是否必选	参数类型	描述
knowledge_base_id	是	String	**参数解释：** 知识库id。获取方式： 1.从创建知识库接口返回获取。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，可从页面知识库id栏获取。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求参数

表 4-90 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-91 响应 Body 参数

参数	参数类型	描述
data	**ExecutionRecord** object	**参数解释：** 执行记录与日志。 **取值范围：** 不涉及。

表 4-92 ExecutionRecord

参数	参数类型	描述
id	String	**参数解释：** 任务执行记录id。 **取值范围：** 由数字、字母和中划线组成，长度36。

参数	参数类型	描述
result	String	**参数解释：** 任务执行结果。 **取值范围：** 枚举值：SUCCESS（任务执行成功）、FAILURE（任务执行失败）、SKIP（跳过）、RUNNING（任务正在执行）。
progress	Float	**参数解释：** 任务执行进度。 **取值范围：** 范围[0,100]。
run_time	Long	**参数解释：** 任务执行时长，单位毫秒。 **取值范围：** 不涉及。
start_time	String	**参数解释：** 任务开始时间。 **取值范围：** 格式为：yyyy-mm-ddThh:mm:ss.000+00:00，如2024-11-21T11:36:55.000+00:00。
end_time	String	**参数解释：** 任务结束时间。 **取值范围：** 格式为：yyyy-mm-ddThh:mm:ss.000+00:00，如2024-11-21T11:36:55.000+00:00。
log_detail	String	**参数解释：** 任务日志详细信息。 **取值范围：** 不涉及。

状态码： 400

表 4-93 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回执行记录信息。

```
{
  "data" : {
    "id" : "djh28e62-3a81-4018-a48f-a15be0d63812",
    "result" : "SUCCESS",
    "progress" : 100.0,
    "run_time" : 27684,
    "start_time" : "2024-11-22T03:15:49.000+00:00",
    "end_time" : "2024-11-22T03:16:17.000+00:00",
    "log_detail" : "开始任务.....结束任务"
  }
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回执行记录信息。
400	请求错误。

错误码

请参见**错误码**。

4.3.6 修改知识库召回状态

功能介绍

该接口用于修改知识库召回状态，启用或禁用知识库召回功能。若知识库被禁用，将无法被agent使用。

调用方法

请参见**如何调用API**。

URI

PUT https://aiae.appstage.myhuaweicloud.com/v1/knowledge-bases/{knowledge_base_id}/retrieval-status

表 4-94 路径参数

参数	是否必选	参数类型	描述
knowledge_base_id	是	String	**参数解释：** 知识库id。获取方式： 1.从创建知识库接口返回获取。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，可从页面知识库id栏获取。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

请求参数

表 4-95 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 4-96 请求 Body 参数

参数	是否必选	参数类型	描述
-	否	String	**参数解释：** 知识库召回状态。 **约束限制：** 不涉及。 **取值范围：** 枚举值：ENABLE（启用召回）与DISABLE（禁用召回）。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-97 响应 Body 参数

参数	参数类型	描述
data	Boolean	**参数解释：** 正常返回的结果。 **取值范围：** 不涉及。

状态码： 400

表 4-98 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

```
ENABLE
```

响应示例

状态码： 200

操作成功，返回操作结果。

```
{
  "data" : true
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回操作结果。
400	请求错误。

错误码

请参见**错误码**。

4.3.7 创建知识数据集

功能介绍

该接口用于创建知识数据集，可将原始文档按照一定规则进行处理，用于后续生成知识库。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/knowledge-datasets

表 4-99 Query 参数

参数	是否必选	参数类型	描述
data_set	是	String	**参数解释：** 创建知识数据集请求体，参数结构请参见“附录 > **知识数据集请求参数说明**”。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
file	否	Array of strings	**参数解释：** 上传的文件。 **取值范围：** 数量不超过10个。 **默认取值：** 不涉及。 **约束限制：** ● 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、wov、.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。

请求参数

表 4-100 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-101 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 正常返回的结果。 **取值范围：** 不涉及。

状态码： 400

表 4-102 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。

参数	参数类型	描述
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

```
{
  "name" : "测试",
  "description" : "测试",
  "data_type" : "TEXT",
  "tags" : [ "对话问答", "文案生成" ],
  "ingestion_config" : {
    "data_source" : "FILE_UPLOAD",
    "file_types" : [ "txt" ]
  },
  "schedule_config" : {
    "schedule_type" : "ONCE"
  },
  "preprocess_config" : {
    "cleaning_methods" : [ "invisible" ],
    "pdf_preprocess_type" : "NO_PREPROCESS"
  },
  "chunk_config" : {
    "slicing_configs" : {
      "txt" : {
        "slicing_method" : "autoSlicing"
      }
    }
  },
  "extraction_config" : {
    "extraction_example" : "今天天气如何？答：还不错哦。",
    "extraction_mode" : "RULE_EXTRACTION",
    "rule_extraction_configs" : [ {
      "extraction_rule" : "SEPARATOR",
      "field_name" : "question",
      "separator_extraction" : {
        "contain_separator" : false,
        "extraction_code" : 1,
        "separator" : "？ "
      }
    }, {
      "extraction_rule" : "TEMPLATE",
      "field_name" : "answer",
      "template_extraction" : {
        "contain_end" : true,
        "contain_start" : false,
        "end_with" : "。",
        "extraction_code" : 1,
        "start_with" : "答"
      }
    } ]
  },
  "index_config" : {
    "description" : "索引配置",
    "long_text_solution" : "TRUNCATE_MODE",
    "name" : "索引配置",
    "rag_type" : "VECTOR_RAG",
    "retrieval_configs" : [ {
      "category" : "FULL_CHUNK",
      "name" : "chunk",
      "retrieval_return" : false,
```

```
    "text_filter" : false,
    "vector_retrieval" : false
  }, {
    "category" : "CHUNK_FRAGMENT",
    "name" : "question",
    "retrieval_return" : true,
    "text_filter" : true,
    "vector_retrieval" : true
  }, {
    "category" : "CHUNK_FRAGMENT",
    "name" : "answer",
    "retrieval_return" : true,
    "text_filter" : true,
    "vector_retrieval" : false
  } ],
  "vector_model_service_key" : "GPT-4"
 }
}
```

响应示例

状态码： 200

操作成功，返回数据集id。

```
{
 "data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

状态码： 400

请求错误。

```
{
 "error_code" : "AIAE.40001001",
 "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回数据集id。
400	请求错误。

错误码

请参见**错误码**。

4.3.8 查询知识数据集详情

功能介绍

该接口用于查询知识数据集详情。

调用方法

请参见**如何调用API**。

URI

GET https://aiae.appstage.myhuaweicloud.com/v1/knowledge-datasets/{data_set_id}

表 4-103 路径参数

参数	是否必选	参数类型	描述
data_set_id	是	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-104 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-105 响应 Body 参数

参数	参数类型	描述
data	**KnowledgeDataSetDetail** object	**参数解释：** 知识数据集详情。 **取值范围：** 不涉及。

表 4-106 KnowledgeDataSetDetail

参数	参数类型	描述
data_set_versions	Array of **KdsVersionInfo** objects	**参数解释：** 知识数据集版本配置信息列表。 **取值范围：** 列表最大长度100。
index_configs	Array of **KdsIndexConfigInfo** objects	**参数解释：** 知识数据集索引配置信息列表。 **取值范围：** 列表最大长度100。

表 4-107 KdsVersionInfo

参数	参数类型	描述
id	String	**参数解释：** 数据集版本id。 **取值范围：** 由数字、字母和中划线组成，长度36。
version	String	**参数解释：** 数据集版本号。 **取值范围：** 格式为：vyyyy-mm-ddThh:mm:ssZ，如v2024-11-21T11:36:55Z。

参数	参数类型	描述
created_date	String	**参数解释：** 数据集版本创建时间。 **取值范围：** 格式为：yyyy-mm-dd hh:mm:ss，如2024-11-21 11:36:55。
last_updated_date	String	**参数解释：** 数据集版本最近更新时间。 **取值范围：** 格式为：yyyy-mm-dd hh:mm:ss，如2024-11-21 11:36:55。

表 4-108 KdsIndexConfigInfo

参数	参数类型	描述
id	String	**参数解释：** 索引id。 **取值范围：** 由数字、字母和中划线组成，长度36。
name	String	**参数解释：** 索引名称。 **取值范围：** 由中英文、数字、“_”组成，长度为[2-50]。
description	String	**参数解释：** 索引描述。 **取值范围：** 不涉及。
data_set_id	String	**参数解释：** 知识数据集id。 **取值范围：** 由数字、字母和中划线组成，长度36。
vector_model_service_key	String	**参数解释：** 向量化模型的service_key。 **取值范围：** 不涉及。

参数	参数类型	描述
index_vector_config	IndexVectorConfig object	**参数解释：** 索引向量化配置。 **取值范围：** 不涉及。

表 4-109 IndexVectorConfig

参数	参数类型	描述
long_text_solution	String	**参数解释：** 知识数据集切片长文本处理方式。 **取值范围：** 枚举值： TRUNCATE_MODE（截断模式：如果分片的token长度超过向量化模型的token数，则自动对超长部分进行截断处理）。 SMART_MODE（智能模式：如果分片的token长度超过向量化模型的token数，则知识库向量化失败）。 DEFAULT_MODE（默认模式：如果分片的token长度超过向量化模型的token数，则知识库向量化失败）。

状态码： 400

表 4-110 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回数据集详情。

```
{
  "data" : {
    "data_set_versions" : [ {
      "id" : "askdjh28e62-xxxxxxxxx-a15be0d63812",
      "version" : "v2024-11-21T11:36:55Z",
      "created_date" : "2024-11-11 19:36:57",
      "last_updated_date" : "2024-11-21 19:36:57"
    } ],
    "index_configs" : [ {
      "id" : "d3f28e62-xxxxxxx-a15be0d638a2",
      "name" : "索引配置名称",
      "description" : "索引配置",
      "data_set_id" : "d3f28e62-3a81-4018-a48f-a15be0d638a2",
      "vector_model_service_key" : "service_key",
      "index_vector_config" : {
        "long_text_solution" : "TRUNCATE_MODE"
      }
    } ]
  }
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回数据集详情。
400	请求错误。

错误码

请参见**错误码**。

4.3.9 删除知识数据集

功能介绍

该接口用于根据数据集id删除知识数据集。

调用方法

请参见**如何调用API**。

URI

DELETE https://aiae.appstage.myhuaweicloud.com/v1/knowledge-datasets/{data_set_id}

表 4-111 路径参数

参数	是否必选	参数类型	描述
data_set_id	是	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-112 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-113 响应 Body 参数

参数	参数类型	描述
data	**DeleteResult** object	**参数解释：** 知识数据集删除结果。 **取值范围：** 不涉及。

表 4-114 DeleteResult

参数	参数类型	描述
id	String	**参数解释：** 知识数据集id。 **取值范围：** 由数字、字母和中划线组成，长度36。

参数	参数类型	描述
name	String	**参数解释：** 知识数据集名称。 **取值范围：** 由中英文、数字、“_”组成，长度为[2-50]。
result	Boolean	**参数解释：** 知识数据集删除成功或失败。 **取值范围：** true或false。
reason	String	**参数解释：** 知识数据集删除失败原因。 **取值范围：** 不涉及。

状态码： 400

表 4-115 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回删除结果。

```
{
  "data" : {
    "id" : "djh28e62-xxxxxxxxx-a15be0d63812",
```

```
    "name" : "知识数据集名称",
    "result" : false,
    "reason" : "知识数据集已被知识库引用"
  }
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回删除结果。
400	请求错误。

错误码

请参见**错误码**。

4.3.10 执行知识数据集

功能介绍

该接口用于根据知识数据集id，触发知识数据集的调度执行，调度执行完毕，数据集的内容将被更新。该接口仅在数据来源为OBS接入有效。

调用方法

请参见**如何调用API**。

URI

POST https://aiae.appstage.myhuaweicloud.com/v1/knowledge-datasets/{data_set_id}/execute

表 4-116 路径参数

参数	是否必选	参数类型	描述
data_set_id	是	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-117 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-118 响应 Body 参数

参数	参数类型	描述
data	String	**参数解释：** 正常返回的结果。 **取值范围：** 不涉及。

状态码： 400

表 4-119 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，执行记录id。

```
{
  "data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，执行记录id。
400	请求错误。

错误码

请参见**错误码**。

4.3.11 查询知识数据集最新执行记录

功能介绍

该接口用于查询知识数据集最新执行记录，可以获取知识数据集调度执行的结果，执行时间，耗时等信息。

调用方法

请参见**如何调用API**。

URI

GET https://aiae.appstage.myhuaweicloud.com/v1/knowledge-datasets/{data_set_id}/latest-execution-record

表 4-120 路径参数

参数	是否必选	参数类型	描述
data_set_id	是	String	**参数解释：** 知识数据集id，获取方式： 1.创建知识数据集接口返回值即为知识数据集id。 2.进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。 **约束限制：** 不涉及。 **取值范围：** 只能由英文字母、数字以及“-”组成，且长度为36个字符。 **默认取值：** 不涉及。

请求参数

表 4-121 请求 Header 参数

参数	是否必选	参数类型	描述
Authorization	是	String	**参数解释：** 鉴权信息。获取平台API Key，并为API Key添加前缀Bearer，得到标准鉴权信息，例如Bearer sk-74e4157***。API Key获取方法请参见**创建API Key**。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

响应参数

状态码： 200

表 4-122 响应 Body 参数

参数	参数类型	描述
data	**ExecutionRecord** object	**参数解释：** 执行记录与日志。 **取值范围：** 不涉及。

表 4-123 ExecutionRecord

参数	参数类型	描述
id	String	**参数解释：** 任务执行记录id。 **取值范围：** 由数字、字母和中划线组成，长度36。
result	String	**参数解释：** 任务执行结果。 **取值范围：** 枚举值：SUCCESS（任务执行成功）、FAILURE（任务执行失败）、SKIP（跳过）、RUNNING（任务正在执行）。
progress	Float	**参数解释：** 任务执行进度。 **取值范围：** 范围[0,100]。
run_time	Long	**参数解释：** 任务执行时长，单位毫秒。 **取值范围：** 不涉及。
start_time	String	**参数解释：** 任务开始时间。 **取值范围：** 格式为：yyyy-mm-ddThh:mm:ss.000+00:00，如2024-11-21T11:36:55.000+00:00。

参数	参数类型	描述
end_time	String	**参数解释：** 任务结束时间。 **取值范围：** 格式为：yyyy-mm-ddThh:mm:ss.000+00:00，如2024-11-21T11:36:55.000+00:00。
log_detail	String	**参数解释：** 任务日志详细信息。 **取值范围：** 不涉及。

状态码： 400

表 4-124 响应 Body 参数

参数	参数类型	描述
error_code	String	**参数解释：** 异常错误码。 **取值范围：** 不涉及。
error_msg	String	**参数解释：** 异常错误信息。 **取值范围：** 不涉及。

请求示例

无

响应示例

状态码： 200

操作成功，返回执行记录信息。

```
{
  "data" : {
    "id" : "djh28e62-3a81-4018-a48f-a15be0d63812",
    "result" : "SUCCESS",
    "progress" : 100.0,
    "run_time" : 27684,
    "start_time" : "2024-11-22T03:15:49.000+00:00",
    "end_time" : "2024-11-22T03:16:17.000+00:00",
    "log_detail" : "开始任务.....结束任务"
```

```
  }
}
```

状态码： 400

请求错误。

```
{
  "error_code" : "AIAE.40001001",
  "error_msg" : "参数xxxx不合法。"
}
```

状态码

状态码	描述
200	操作成功，返回执行记录信息。
400	请求错误。

错误码

请参见**错误码**。

5 应用示例

5.1 创建知识库并进行检索

操作场景

本文通过调用一系列知识中心的API，介绍从零开始创建知识库并进行检索的基本流程。API调用方法请参见**如何调用API**。

前提条件

准备一篇名为“问题.txt”的文档，文档内容为：

“什么是量子计算？它是一种利用量子力学原理进行信息处理的计算方式。====

什么是RAG？它是一种通过整合检索系统和生成模型的优势，来提升模型生成文本的质量和上下文相关性。====”

操作流程

步骤一：创建数据集。

步骤二：查询知识数据集创建结果。

步骤三：查询数据集版本和索引信息。

步骤四：创建知识库。

步骤五：查询知识库创建结果。

步骤六：启用或停用知识库召回功能。

步骤七：检索知识库数据。

步骤一：创建数据集

调用**创建知识数据集**接口创建数据集，示例如下：

```
{
"name": "测试",
"description": "测试",
"data_type": "TEXT",
```

```
"tags": [
"对话问答",
"文案生成"
],
"ingestion_config": {
"data_source": "FILE_UPLOAD",
"file_types": [
"txt"
]
},
"schedule_config": {
"schedule_type": "ONCE"
},
"preprocess_config": {
"cleaning_methods": [
"invisible"
],
"pdf_preprocess_type": "NO_PREPROCESS"
},
"chunk_config": {
"slicing_configs": {
"txt": {
"slicing_method": "sentence",
"sentence_slicing_config": {
"slicing_strategy": "equivalent",
"spec_symbols": [
"===="
],
"contain_separator": false,
"chunk_size": 26,
"chunk_overlap": 0
}
}
}
},
"extraction_config": {
"extraction_example": "今天天气如何？答：还不错哦。",
"extraction_mode": "RULE_EXTRACTION",
"rule_extraction_configs": [
{
"extraction_rule": "SEPARATOR",
"field_name": "question",
"separator_extraction": {
"contain_separator": false,
"extraction_code": 2,
"separator": "？ "
}
},
{
"extraction_rule": "TEMPLATE",
"field_name": "answer",
"template_extraction": {
"contain_end": true,
"contain_start": false,
"end_with": "。",
"extraction_code": 1,
"start_with": "它是"
}
}
]
},
"index_config": {
"description": "索引配置",
"long_text_solution": "TRUNCATE_MODE",
"name": "索引配置",
"rag_type": "VECTOR_RAG",
"retrieval_configs": [
{
"category": "FULL_CHUNK",
```



```
"name": "chunk",
"retrieval_return": false,
"text_filter": false,
"vector_retrieval": false
},
{
"category": "CHUNK_FRAGMENT",
"name": "question",
"retrieval_return": true,
"text_filter": true,
"vector_retrieval": true
},
{
"category": "CHUNK_FRAGMENT",
"name": "answer",
"retrieval_return": true,
"text_filter": true,
"vector_retrieval": false
}
],
"vector_model_service_key": "GPT-4"
}
}
```

在创建知识数据集中，有几个参数需要注意：

- ingestion_config内的data_source：表明数据来源，FILE_UPLOAD为文件上传，您需上传文件；OBS_INGESTTION表示OBS接入，您需添加OBS信息的配置。
- chunk_config：表示如何对文本进行切分。
- extraction_config：表明要对切片做提取操作，切片提取出的字段可以在索引配置中使用。
- index_config：索引配置，其中retrieval_configs配置完整切片内容，以及提取出的字段内容，在知识库检索时如何使用。

记录下接口返回的内容，该内容为知识数据集id。

```
{
"data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

步骤二：查询知识数据集创建结果

调用**查询知识数据集最新执行记录**接口查询创建结果，该接口所需的知识数据集id为**步骤一：创建数据集**返回的内容。

根据返回结果响应：

```
{
"data" : {
"id" : "djh28e62-3a81-4018-a48f-a15be0d63812",
"result" : "SUCCESS",
"progress" : 100.0,
"run_time" : 27684,
"start_time" : "2024-11-22T03:15:49.000+00:00",
"end_time" : "2024-11-22T03:16:17.000+00:00",
"log_detail" : "开始任务.....结束任务"
}
}
```

- 如果result字段值为SUCCESS，则表明数据集创建成功。
- 如果result字段值为RUNNING，则表明数据集正在创建，请稍候。
- 如果result字段值为FALIURE，则表明数据集创建失败，请检查文件是否符合要求。

步骤三：查询数据集版本和索引信息

知识数据集创建成功后，调用**查询知识数据集详情**接口，查询数据集版本和索引信息，响应示例如下：

```
{
"data" : {
"data_set_versions" : [ {
"id" : "askdjh28e62-xxxxxxxxx-a15be0d63812",
"version" : "v2024-11-21T11:36:55Z",
"created_date" : "2024-11-11 19:36:57",
"last_updated_date" : "2024-11-21 19:36:57"
} ],
"index_configs" : [ {
"id" : "d3f28e62-xxxxxxx-a15be0d638a2",
"name" : "索引配置名称",
"description" : "索引配置",
"data_set_id" : "d3f28e62-3a81-4018-a48f-a15be0d638a2",
"vector_model_service_key" : "service_key",
"index_vector_config" : {
"long_text_solution" : "TRUNCATE_MODE"
}
} ]
}
}
```

保存创建知识库对应的数据集版本id、版本号、索引配置id。数据集版本和索引可能有多个，可以任意组合。

步骤四：创建知识库

调用**创建知识库**接口创建知识库。根据保存的数据集id，数据集版本号，索引配置id，构建请求体，创建知识库即可。请求体如下：

```
{
"name": "知识库名称",
"description": "知识库描述",
"retrieval_status": "ENABLE",
"rag_type": "VECTOR_RAG",
"retrieval_config": {
"retrieval_modes": [
"SEMANTIC_RETRIEVAL",
"FULL_TEXT_RETRIEVAL"
],
"retrieval_hybrid_mode": "RRF"
},
"knowledge_data_sets": [
{
"data_set_id": "djh28e62-xxxxxxxxx-a15be0d63812",
"data_set_version": "v2024-11-21T11:36:55Z",
"index_config_id": "d3f28e62-xxxxxxx-a15be0d638a2"
}
]
}
```

注意：retrieval_status字段设置知识库是否启用检索，如果启用，则检索前不需要再调用修改知识库召回状态接口启用知识库。

将创建接口响应内容保存起来，该内容为知识库id：

```
{
"data" : "3f28e62-xxxxxxxx-a15be0d638a2"
}
```

步骤五：查询知识库创建结果

调用**查询知识库最新执行记录**接口查询创建结果。该接口所需的知识库id为**步骤四：创建知识库**返回的内容。

返回结果响应如下：

```
{
"data" : {
"id" : "djh28e62-3a81-4018-a48f-a15be0d63812",
"result" : "SUCCESS",
"progress" : 100.0,
"run_time" : 27684,
"start_time" : "2024-11-22T03:15:49.000+00:00",
"end_time" : "2024-11-22T03:16:17.000+00:00",
"log_detail" : "开始任务.....结束任务"
}
}
```

- 如果result字段值为SUCCESS，则表明知识库创建成功。
- 如果result字段值为RUNNING，则表明知识库正在创建，请稍候。
- 如果result字段值为FALIURE，则表明知识库创建失败，请检查文件是否符合要求。

步骤六：启用或停用知识库召回功能

调用**修改知识库召回状态**接口启用或停用知识库召回功能。

若创建知识库时已经启用知识库，则不需要执行此步骤。否则需启用知识库，请求体为：

```
ENABLE
```

若响应如下，则启用成功。

```
{
"data" : true
}
```

步骤七：检索知识库数据

调用**检索知识库数据**接口检索知识库（知识库召回状态需为启用），请求体示例如下：

```
{
"keyword" : "什么是",
"similarity_min" : "0.78",
"limit" : 10,
"filter" : {
"group_type" : "AND",
"expressions" : [ {
"field" : "metadata.answer",
"field_type" : "STRING",
"operator" : "EQUAL",
"values" : [ "一种利用量子力学原理进行信息处理的计算方式。" ]
} ]
},
"order_by" : {
"order_items" : [ {
"field" : "metadata.order",
"field_type" : "INT",
"order_type" : "DESC"
} ]
},
```



```
"data_sets" : [ {
"data_set_id" : "a31ed909-xxxx-xxxx-xxxx-10958c90b3f7"
} ]
}
```

根据索引配置，返回结果为：

```
{
"data" : [ {
"id" : "812857ef-xxxx-xxxx-xxxx-24ba9fd5e95c",
"document" : "什么是量子计算？它是一种利用量子力学原理进行信息处理的计算方式。",
"chunk" : "什么是量子计算？它是一种利用量子力学原理进行信息处理的计算方式。",
"chunk_fragments" : {
"question" : "什么是量子计算。",
"answer" : "一种利用量子力学原理进行信息处理的计算方式"
},
"similarity" : 0.87,
"metadata" : {
"order" : 10,
"file_name" : "问题.txt",
"path" : "问题.txt ",
"question" : "什么是量子计算。",
"answer" : "一种利用量子力学原理进行信息处理的计算方式。"
},
"download_addresses" : {
"xxx.png" : "https://xxxx"
},
"download_address" : null,
"data_set_id" : "3967c49d-xxxx-xxxx-xxxx-5eda056a1f1b"
} ]
}
```

响应参数解释：

- document：表示向量化检索内容，索引配置时选择某个字段为向量化字段，检索命中时返回该字段内容，即vector_retrieval为true。
- chunk：表示完整切片，索引配置时配置chunk作为附加字段返回时，该字段有内容，即索引配置中category为FULL_CHUNK，retrieval_return为true。
- chunk_fragments：表示切片提取字段，索引配置时配置所提取的字段作为附加字段返回时，该字段有内容，即索引配置中category为CHUNK_FRAGMENT，retrieval_return为true。
- metadata：默认包含order、file_name、path三个字段，若索引配置时配置所提取的字段作为文本过滤字段时，即索引配置中category为CHUNK_FRAGMENT，text_filter为true时，metadata下会新增该字段及其内容。

5.2 更新知识库

操作场景

本文通过调用一系列知识中心的API介绍知识库的更新流程，适用于当数据源为OBS接入时，在OBS上进行文件增删改后，将改动同步到知识库的场景。API调用方法请参见**如何调用API**。

前提条件

- 用户接入的OBS目录下，存在文件的增删改其中一种情况。
- 需获取待更新数据集的id，支持通过如下两种方式获取：
  - **创建知识数据集**接口返回值即为知识数据集id。

- 进入**AI原生应用引擎**，在左侧导航栏选择“知识中心 > 知识库”，选择页面右上角的“... > 知识数据集”，在数据集列表中，单击数据集名称，进入详情页即可获取数据集id。

更新流程

1. 执行知识数据集调度，更新数据集内容。

   调用**执行知识数据集**接口，根据知识数据集id，触发知识数据集的调度执行，调度执行完毕，数据集的内容将被更新。

   接口返回内容如下，该内容为执行记录id。

   ```
   {
     "data" : "3f28e62-xxxxxxxx-a15be0d638a2"
   }
   ```

2. 查询知识数据集调度执行的结果。

   调用**查询知识数据集最新执行记录**接口，根据知识数据集id，查询调度执行结果。根据返回结果响应：

   ```
   {
   "data" : {
   "id" : "djh28e62-3a81-4018-a48f-a15be0d63812",
   "result" : "SUCCESS",
   "progress" : 100.0,
   "run_time" : 27684,
   "start_time" : "2024-11-22T03:15:49.000+00:00",
   "end_time" : "2024-11-22T03:16:17.000+00:00",
   "log_detail" : "开始任务.....结束任务"
   }
   }
   ```

   - 如果result字段值为SUCCESS，则表明数据集更新成功。
   - 如果result字段值为RUNNING，则表明数据集正在更新，请稍候。
   - 如果result字段值为FALIURE，则表明数据集更新失败，请检查文件是否符合要求。

6 附录

6.1 状态码

状态码如表6-1所示

表 6-1 状态码

状态码	编码	错误码说明
100	Continue	继续请求。 这个临时响应用来通知客户端，它的部分请求已经被服务器接收，且仍未被拒绝。
101	Switching Protocols	切换协议。只能切换到更高级的协议。 例如，切换到HTTP的新版本协议。
201	Created	创建类的请求完全成功。
202	Accepted	已经接受请求，但未处理完成。
203	Non-Authoritative Information	非授权信息，请求成功。
204	NoContent	请求完全成功，同时HTTP响应不包含响应体。 在响应OPTIONS方法的HTTP请求时返回此状态码。
205	Reset Content	重置内容，服务器处理成功。
206	Partial Content	服务器成功处理了部分GET请求。
300	Multiple Choices	多种选择。请求的资源可包括多个位置，相应可返回一个资源特征与地址的列表用于用户终端（例如：浏览器）选择。
301	Moved Permanently	永久移动，请求的资源已被永久的移动到新的URI，返回信息会包括新的URI。

状态码	编码	错误码说明
302	Found	资源被临时移动。
303	See Other	查看其它地址。 使用GET和POST请求查看。
304	Not Modified	所请求的资源未修改，服务器返回此状态码时，不会返回任何资源。
305	Use Proxy	所请求的资源必须通过代理访问。
306	Unused	已经被废弃的HTTP状态码。
400	BadRequest	非法请求。 建议直接修改该请求，不要重试该请求。
401	Unauthorized	在客户端提供认证信息后，返回该状态码，表明服务端指出客户端所提供的认证信息不正确或非法。
402	Payment Required	保留请求。
403	Forbidden	请求被拒绝访问。 返回该状态码，表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为该请求被设置为拒绝访问，建议直接修改该请求，不要重试该请求。
404	NotFound	所请求的资源不存在。 建议直接修改该请求，不要重试该请求。
405	MethodNotAllowed	请求中带有该资源不支持的方法。 建议直接修改该请求，不要重试该请求。
406	Not Acceptable	服务器无法根据客户端请求的内容特性完成请求。
407	Proxy Authentication Required	请求要求代理的身份认证，与401类似，但请求者应当使用代理进行授权。
408	Request Time-out	服务器等候请求时发生超时。 客户端可以随时再次提交该请求而无需进行任何更改。
409	Conflict	服务器在完成请求时发生冲突。 返回该状态码，表明客户端尝试创建的资源已经存在，或者由于冲突请求的更新操作不能被完成。
410	Gone	客户端请求的资源已经不存在。 返回该状态码，表明请求的资源已被永久删除。
411	Length Required	服务器无法处理客户端发送的不带Content-Length的请求信息。

状态码	编码	错误码说明
412	Precondition Failed	未满足前提条件，服务器未满足请求者在请求中设置的其中一个前提条件。
413	Request Entity Too Large	由于请求的实体过大，服务器无法处理，因此拒绝请求。为防止客户端的连续请求，服务器可能会关闭连接。如果只是服务器暂时无法处理，则会包含一个Retry-After的响应信息。
414	Request-URI Too Large	请求的URI过长（URI通常为网址），服务器无法处理。
415	Unsupported Media Type	服务器无法处理请求附带的媒体格式。
416	Requested range not satisfiable	客户端请求的范围无效。
417	Expectation Failed	服务器无法满足Expect的请求头信息。
422	UnprocessableEntity	请求格式正确，但是由于含有语义错误，无法响应。
429	TooManyRequests	表明请求超出了客户端访问频率的限制或者服务端接收到多于它能处理的请求。建议客户端读取相应的Retry-After首部，然后等待该首部指出的时间后再重试。
500	InternalServerError	表明服务端能被请求访问到，但是不能理解用户的请求。
501	Not Implemented	服务器不支持请求的功能，无法完成请求。
502	Bad Gateway	充当网关或代理的服务器，从远端服务器接收到了一个无效的请求。
503	ServiceUnavailable	被请求的服务无效。 建议直接修改该请求，不要重试该请求。
504	ServerTimeout	请求在给定的时间内无法完成。客户端仅在为请求指定超时（Timeout）参数时会得到该响应。
505	HTTP Version not supported	服务器不支持请求的HTTP协议的版本，无法完成处理。

6.2 错误码

当您调用API时，如果遇到“APIGW”开头的错误码，请参见**API网关错误码**进行处理。

状态码	错误码	错误信息	描述	处理措施
200	AIAE.22001001	API调用异常	API调用异常	调用接口url、请求方式错误或出现访问其他用户资源的越权问题，请检查后重试
200	AIAE.22001002	IAM认证异常	IAM认证异常	后端服务错误，请联系技术支持
200	AIAE.22001003	认证失败:{reason}	认证失败:{reason}	请参考返回的error message，或联系技术支持
200	AIAE.22001004	参数{parameterName}异常	参数{parameterName}异常	输入的参数有误，请参考返回的error message进行修改后重试
200	AIAE.22001005	{type}类型远程调用失败，错误信息为{error_msg}	{type}类型远程调用失败，错误信息为{error_msg}	请参考返回的error message处理后重试
200	AIAE.22001006	文件上传失败:{reason}	文件上传失败:{reason}	文件上传失败，请参考返回的error message处理后重试
200	AIAE.22001007	技能未设置鉴权信息	技能未设置鉴权信息	在wiseAgent页面为技能设置鉴权信息。
200	AIAE.22001008	用户无权限访问当前资源	用户无权限访问当前资源	请更换用户后访问
200	AIAE.22001009	开启的流无法被删除	开启的流无法被删除	关闭流后重试
200	AIAE.22009001	系统内部错误，请联系管理员	系统内部错误，请联系管理员	系统内部错误，请联系技术支持
400	AIAE.31001106	AK/SK signature verify failed, please check and try again later!	很抱歉，AK/SK签名验证失败!	很抱歉，AK/SK签名验证失败!
400	AIAE.31001601	Sensitive request error, please try again later!	很抱歉，请求内容中包含敏感信息，请重试!	很抱歉，请求内容中包含敏感信息，请重试!

状态码	错误码	错误信息	描述	处理措施
400	AIAE.31001602	Sensitive response error, please try again later!	很抱歉，返回内容中包含敏感信息，请重试！	很抱歉，返回内容中包含敏感信息，请重试！
400	AIAE.31001603	Sensitive content error, please try again later!	很抱歉，请求或返回内容中包含敏感信息，请重试！	很抱歉，请求或返回内容中包含敏感信息，请重试！
400	AIAE.31001701	Bad request parameter error, please check and try again later!	很抱歉，请求参数异常，请检查后重试！	很抱歉，请求参数异常，请检查后重试！
400	AIAE.40001003	Authentication failed	X-Auth-Token鉴权失败	很抱歉，X-Auth-Token 鉴权失败
400	AIAE.40002605	knowledgeBase status is not ENABLE	很抱歉，知识库未启用，没有权限查询	很抱歉，知识库未启用，没有权限查询
401	AIAE.31001101	User not login, please check and try again later!	很抱歉，用户未登录，请登录后重试！	很抱歉，用户未登录，请登录后重试！
401	AIAE.31001102	AK/SK verify failed, please check and try again later!	很抱歉，AK/SK校验失败！	很抱歉，AK/SK校验失败！
401	AIAE.31001103	Authentication verify failed, please check and try again later!	很抱歉，鉴权失败！	很抱歉，鉴权失败！
401	AIAE.31001104	API Key verify failed, please check and try again later!	很抱歉，API Key校验失败！	很抱歉，API Key校验失败！
401	AIAE.31001201	Tenant id is empty, please check and try again later!	很抱歉，空的租户id，请检查后重试！	很抱歉，空的租户id，请检查后重试！

状态码	错误码	错误信息	描述	处理措施
401	AIAE.31005001	The third model service authentication is abnormal, please check and try again later!	很抱歉，三方模型服务鉴权异常，请检查您的鉴权信息！	很抱歉，三方模型服务鉴权异常，请检查您的鉴权信息！
401	AIAE.31005002	Invalid third api key, please check and try again later!	很抱歉，三方模型服务鉴权API Key异常，请检查您的鉴权信息！	很抱歉，三方模型服务鉴权API Key异常，请检查您的鉴权信息！
401	AIAE.31005007	The third model service authentication is empty, please set and try again later!	很抱歉，三方模型服务鉴权未设置，请设置后重试！	很抱歉，三方模型服务鉴权未设置，请设置后重试！
402	AIAE.31005005	The third model service exceeded current quota error, please check and try again later!	很抱歉，账户异常，请检查您的账户余额！	很抱歉，账户异常，请检查您的账户余额！
403	AIAE.31001105	Role permission verify failed, please check and try again later!	很抱歉，当前用户不允许该操作！	很抱歉，当前用户不允许该操作！
403	AIAE.31001501	SKU not subscribed to model service, please try again after subscribed!	很抱歉，您未订阅当前模型服务的SKU，请联系管理员订阅！	很抱歉，您未订阅当前模型服务的SKU，请联系管理员订阅！
403	AIAE.31001502	SKU verify failed, please check and try again later!	很抱歉，SKU校验异常，请检查您的SKU！	很抱歉，SKU校验异常，请检查您的SKU！
403	AIAE.40001004	User does not have permission	当前用户没有权限	很抱歉，当前用户没有权限

状态码	错误码	错误信息	描述	处理措施
404	AIAE.31001202	Model not published, please try again after model published!	很抱歉，模型服务未发布，请发布后重试！	很抱歉，模型服务未发布，请发布后重试！
404	AIAE.31001702	Model not exists, please check and try again later!	很抱歉，模型服务不存在，请检查您输入的模型服务名称！	很抱歉，模型服务不存在，请检查您输入的模型服务名称！
408	AIAE.31001003	Connection timeout, please try again later!	很抱歉，网络连接超时，请稍后重试！	很抱歉，网络连接超时，请稍后重试！
408	AIAE.31005006	The third model service connect timeout, please try again later!	很抱歉，三方服务连接超时，请稍后重试！	很抱歉，三方服务连接超时，请稍后重试！
429	AIAE.31001002	Request too frequent error, please try again later!	很抱歉，您的请求过于频繁，请稍后重试！	很抱歉，您的请求过于频繁，请稍后重试！
429	AIAE.31005003	The third model service rate limit exceeded, please try again later!	很抱歉，您的请求当前已达最大并发数，请稍后重试！	很抱歉，您的请求当前已达最大并发数，请稍后重试！
429	AIAE.31005004	The third model service overload error, please try again later!	很抱歉，服务当前超载，请稍后重试！	很抱歉，服务当前超载，请稍后重试！
500	AIAE.31001001	Internal server error, please try again later!	很抱歉，服务内部出现了问题，请稍后重试！	很抱歉，服务内部出现了问题，请稍后重试！

状态码	错误码	错误信息	描述	处理措施
500	AIAE.31001004	Ai security governance service error, please try again later or disable ai security governance service!	很抱歉，内容审核服务出现了问题，请稍后重试！	很抱歉，内容审核服务出现了问题，请稍后重试！
500	AIAE.31005000	Invalid third response, please try again later!	很抱歉，调用三方模型服务异常，请稍后重试！	很抱歉，调用三方模型服务异常，请稍后重试！
400	UniModel.Request.0001	请求参数错误	模型服务的请求参数错误	检查模型的请求参数
500	UniDataEmbed.Internal.0001	请求失败	请求向量化服务失败	检查向量知识库配置
500	UniModel.Internal.0001	模型访问失败	无法访问选择的模型	检查模型是否已经正常部署
500	UniModel.Internal.0002	模型返回超时	模型服务返回超时	检查网络情况，或者减少模型返回内容
500	WS.00100001	AUTHENTICATION_ERROR	鉴权错误	检查访问权限
500	WS.00100002	SHA_ERROR	SHA算法错误	检查签名所用的算法
500	WS.00100003	SIGN_ERROR	请求签名错误	检查请求签名
500	WS.00100005	NO_ACCESS_ERROR	无访问权限错误	检查接口访问权限

6.3 知识数据集请求参数说明

CreateKnowledgeDataSetReq

创建知识数据集的data_set参数具有特定结构，需要按照以下实体进行构造后转成json。

表 6-2 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	**参数解释：** 数据集名称。 **约束限制：** 不涉及。 **取值范围：** 长度2-50个字符，支持中英文、数字、下划线(_)，以中英文、数字开头。 **默认取值：** 不涉及。
description	否	String	**参数解释：** 数据集描述。 **约束限制：** 不涉及。 **取值范围：** 长度0-255个字符，只能包含英文、中文、数字、下划线、中划线、空格及,.?:;"'：；“”‘，。？、()（）/等符号。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
tags	否	Array of strings	**参数解释：** 数据集标签。 **约束限制：** 不涉及。 **取值范围：** 传入数量0~5个，需为以下标签：航空、语音转文本、电力、文本、城市数字化、文案生成、水运、1M-10M、NL2SQL、全功能、公路交通、银行业务、制造、数字基础设施、高质量数据（训练）、英文、流媒体、图像理解、托管服务、政府、医疗、>100M、文本向量化、文本生图、城市交通、对话问答、多模生成、功能调用、语音合成、城轨、图文向量化、证券业务、大语言模型、铁路、互联网交换中心、企业基础设施与运营、通用、口岸海关和特殊监管区、10M-100M、代码生成、0-1M、中文、矿业、教育、油气、大企业、种子数据（数据膨胀）、任务规划、保险业务、政务/政党数字化。 **默认取值：** 不涉及。
data_type	是	String	**参数解释：** 数据集类型。 **约束限制：** 不涉及。 **取值范围：** 枚举值：TEXT（文档）、IMAGE_TO_TEXT（图片摘要）、VIDEO_TO_TEXT（视频摘要）、IMAGE（图片）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
preprocess_config	否	PreprocessConfig object	**参数解释：** 数据集预处理配置。 **约束限制：** data_type为IMAGE时不传。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
ingestion_config	是	IngestionConfig object	数据集数据接入配置。 **参数解释：** 数据集数据接入配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
chunk_config	否	ChunkConfig object	**参数解释：** 数据集切分配置。 **约束限制：** 不涉及。 **取值范围：** data_type不为TEXT不传。 **默认取值：** 不涉及。
schedule_config	是	ScheduleConfig object	**参数解释：** 调度配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
extraction_config	否	ExtractionConfig object	**参数解释：** 切片提取配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
index_config	否	IndexConfig object	**参数解释：** 知识数据集索引配置。 **约束限制：** 创建知识库需要索引，若需创建知识库则需传入。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-3 PreprocessConfig

参数	是否必选	参数类型	描述
cleaning_methods	否	Array of strings	**参数解释：** 数据集清洗方法。 **约束限制：** 不涉及。 **取值范围：** 枚举值：URL_AND_EMAIL（删除所有的URL和电子邮件地址）、CONTINUOUS_SYMBOL（清除连续的空格，换行符和制表符）、INVISIBLE（清除不可见字符）、WHITESPACE（规范化空格）、GARBLE（清除乱码）、WEB_SYMBOL（清除网页标识符）、EMOJI（清除表情）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
pdf_preprocess_type	否	String	**参数解释：** 数据集pdf文件预处理类型。 **约束限制：** 仅data_type为TEXT时支持传入。 **取值范围：** 枚举值：EXTRACT_RICH_MEDIA（提取富媒体，如表、图）、NO_PREPROCESS（不做处理）。 **默认取值：** 不涉及。
rich_media_intelligent_match	否	String	数据集pdf预处理后，富媒体提取类型，仅data_type为TEXT时支持传入，枚举值：SMART_MATCH_IMAGE_TABLE（智能提取，仅预处理为EXTRACT_RICH_MEDIA支持）、NO_MATCH（不提取）。 **参数解释：** 数据集pdf预处理后，富媒体提取类型。 **约束限制：** 仅data_type为TEXT时支持传入。 **取值范围：** 枚举值：SMART_MATCH_IMAGE_TABLE（智能提取，仅预处理为EXTRACT_RICH_MEDIA支持）、NO_MATCH（不提取）。 **默认取值：** 不涉及。

表 6-4 IngestionConfig

参数	是否必选	参数类型	描述
data_source	是	String	**参数解释：** 数据来源。 **约束限制：** 不涉及。 **取值范围：** 枚举值：FILE_UPLOAD（文件上传）、OBS_INGESTION（OBS接入）。 **默认取值：** 不涉及。
obs_ingestion	否	**ObsIngestion** object	**参数解释：** OBS接入配置。 **约束限制：** data_source为FILE_UPLOAD（文件上传）则不传OBS接入配置，否则需传入。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
file_types	是	Array of strings	**参数解释：** 数据集支持的文件类型。 **约束限制：** 不涉及。 **取值范围：** 枚举值：PDF、TXT、CSV、XLSX、DOCX、PPTX、HTML、JSON、XML、JPG、JPEG、PNG、MP4、WEBM。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
summary_configs	否	Array of SummaryConfig objects	**参数解释：** 摘要类型数据集摘要配置。 **约束限制：** 在data_type为IMAGE_TO_TXT或VIDEO_TO_TEXT时需传入，其它类型则不传。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-5 ObsIngestion

参数	是否必选	参数类型	描述
obs_bucket_name	是	String	**参数解释：** OBS桶名。 **约束限制：** 不涉及。 **取值范围：** 长度3~63个字符。 **默认取值：** 不涉及。
obs_input_directory	是	String	**参数解释：** OBS接入目录路径。 **约束限制：** 不涉及。 **取值范围：** OBS接入路径与目录下文件名组合成的路径，最长不超过200。 **默认取值：** 不涉及。

表 6-6 SummaryConfig

参数	是否必选	参数类型	描述
file_name	是	String	**参数解释：** 文件名。 **约束限制：** 需与上传文件名称一致。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
summary	是	String	**参数解释：** 摘要。 **约束限制：** 不涉及。 **取值范围：** 长度1~600。 **默认取值：** 不涉及。

表 6-7 ChunkConfig

参数	是否必选	参数类型	描述
slicing_configs	否	Map<String,SlicingConfig>	**参数解释：** 数据集切分配置列表。 **约束限制：** 切分配置数量需要与文件类型数量保持一致。 **取值范围：** 范围1~30。 **默认取值：** 不涉及。

表 6-8 SlicingConfig

参数	是否必选	参数类型	描述
slicing_method	是	String	**参数解释：** 数据集切分方法。 **约束限制：** 不涉及。 **取值范围：** 枚举值：AUTO_SLICING（自动切分）、TITLE（标题切分）、SENTENCE（自定义切分）、JSON（Json切分）、XML（XML切分），除自动切分外，其它类型切分需传入对应切分配置。 **默认取值：** 不涉及。
sentence_slicing_config	否	**SentenceSlicingConfig** object	**参数解释：** 自定义切分配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
title_slicing_config	否	**TitleSlicingConfig** object	**参数解释：** 标题切分配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
json_slicing_config	否	**JsonSlicingConfig** object	**参数解释：** json切分配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
xml_slicing_config	否	XmlSlicingConfig object	**参数解释：** xml切分配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-9 SentenceSlicingConfig

参数	是否必选	参数类型	描述
slicing_strategy	是	String	**参数解释：** 文本切分策略。 **约束限制：** 不涉及。 **取值范围：** 枚举值：RECURSIVE（递归切分）、EQUIVALENT（等价切分）。 **默认取值：** 不涉及。
spec_symbols	是	Array of strings	**参数解释：** 分段分隔符。 **约束限制：** 不涉及。 **取值范围：** 长度1~20，除\n外，不允许包含以下字符 *./$^?+ 且不允许为<!I@M#A%G&E!>、<!T@A#B%L&E!>。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
chunk_size	是	Integer	**参数解释：** 分段长度。 **约束限制：** 不涉及。 **取值范围：** 取值1~500。 **默认取值：** 不涉及。
chunk_overlap	是	Integer	**参数解释：** 分段重叠长度。 **约束限制：** 不涉及。 **取值范围：** 取值0~50。 **默认取值：** 不涉及。
contain_separator	是	Boolean	**参数解释：** 切片是否包含分隔符。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。

表 6-10 TitleSlicingConfig

参数	是否必选	参数类型	描述
slicing_strategy	是	String	**参数解释：** 文本切分策略。 **约束限制：** 不涉及。 **取值范围：** 枚举值：RECURSIVE（递归切分）、EQUIVALENT（等价切分）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
title_level	是	String	**参数解释：** 标题层级深度。 **约束限制：** 不涉及。 **取值范围：** 枚举值：H1、H2、H3、H4、H5。 **默认取值：** 不涉及。
title_saved_method	是	String	**参数解释：** 标题保存方式。 **约束限制：** 不涉及。 **取值范围：** 枚举值：COMBINATION（多标题组合）、LAST（最后一级标题）。 **默认取值：** 不涉及。
spec_symbols	是	Array of strings	**参数解释：** 分段分隔符。 **约束限制：** 不涉及。 **取值范围：** 长度1~20，除\n外，不允许包含以下字符 *./$^?+ 且不允许为<!I@M#A%G&E!>、<!T@A#B%L&E!>。 **默认取值：** 不涉及。
chunk_size	是	Integer	**参数解释：** 分段长度。 **约束限制：** 不涉及。 **取值范围：** 取值1~500。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
chunk_overlap	是	Integer	**参数解释：** 分段重叠长度。 **约束限制：** 不涉及。 **取值范围：** 取值0~50。 **默认取值：** 不涉及。
contain_separator	是	Boolean	**参数解释：** 切片是否包含分隔符。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。

表 6-11 JsonSlicingConfig

参数	是否必选	参数类型	描述
levels_back	是	Integer	**参数解释：** 输出层级。 **约束限制：** 不涉及。 **取值范围：** 取值0~20。 **默认取值：** 不涉及。
collapse_length	是	Integer	**参数解释：** 递归最小长度。 **约束限制：** 不涉及。 **取值范围：** 取值0~1000。 **默认取值：** 不涉及。

表 6-12 XmlSlicingConfig

参数	是否必选	参数类型	描述
tree_level_split	是	Integer	**参数解释：** 遍历层级。 **约束限制：** 不涉及。 **取值范围：** 取值0~20。 **默认取值：** 不涉及。

表 6-13 ScheduleConfig

参数	是否必选	参数类型	描述
schedule_type	是	String	**参数解释：** 调度类型。 **约束限制：** data_source为FILE_UPLOAD时仅支持ONCE（一次性调度）。 **取值范围：** 枚举值：ONCE（一次性调度）、SCHEDULE（周期性调度）。 **默认取值：** 不涉及。
scheduled_task_config	否	**ScheduledTaskConfig** object	**参数解释：** 定时调度配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-14 ScheduledTaskConfig

参数	是否必选	参数类型	描述
cycle_type	是	String	**参数解释：** 定时任务周期类型。 **约束限制：** 不涉及。 **取值范围：** 枚举值：DAY（按日更新）、WEEK（按周更新）。 **默认取值：** 不涉及。
run_time	是	String	**参数解释：** 定时任务执行时间。 **约束限制：** 不涉及。 **取值范围：** 格式为：HH:mm:ss，如：18:30:00。 **默认取值：** 不涉及。
week_day	否	String	**参数解释：** 定时任务执行日期（星期）。 **约束限制：** cycle_type为WEEK时需传入。 **取值范围：** 枚举值：SUNDAY（星期天）、MONDAY（星期一）、TUESDAY（星期二）、WEDNESDAY（星期三）、THURSDAY（星期四）、FRIDAY（星期五）、SATURDAY（星期六）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
execute_immediately	是	Boolean	**参数解释：** 定时任务是否立即执行一次。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。
version_refresh_mode	是	String	**参数解释：** 版本刷新模式。 **约束限制：** 不涉及。 **取值范围：** 枚举值：ONE_VERSION（每次执行覆盖原版本）、MULTI_VERSION（每次执行生成新版本）。 **默认取值：** 不涉及。

表 6-15 ExtractionConfig

参数	是否必选	参数类型	描述
extraction_example	否	String	**参数解释：** 切片提取样例。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
extraction_mode	是	String	**参数解释：** 切片提取模式。 **约束限制：** 不涉及。 **取值范围：** 枚举：RULE_EXTRACTION（规则提取）、SMART_EXTRACTION（智能提取）。 **默认取值：** 不涉及。
rule_extraction_configs	否	Array of **RuleExtractionConfig** objects	**参数解释：** 规则提取配置列表。 **约束限制：** extraction_mode为RULE_EXTRACTION时需传入，为SMART_EXTRACTION时则不传。 **取值范围：** 规则提取配置数量不超过10个，提取字段名称长度1~20，不允许重复。 **默认取值：** 不涉及。

表 6-16 RuleExtractionConfig

参数	是否必选	参数类型	描述
field_name	是	String	**参数解释：** 提取字段名称。 **约束限制：** 不涉及。 **取值范围：** 字段数量不超过10个，其中名称长度1~20，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。 **默认取值：** 不涉及。
extraction_rule	是	String	**参数解释：** 提取规则。 **约束限制：** 不涉及。 **取值范围：** 枚举：SEPARATOR（分隔符提取）、TEMPLATE（模板提取）。 **默认取值：** 不涉及。
separator_extraction	是	SeparatorExtractionConfig object	**参数解释：** 分隔符提取配置。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
template_extraction	是	**HeadAndTailExtractionTemplate** object	**参数解释：** 首尾匹配提取模板。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-17 SeparatorExtractionConfig

参数	是否必选	参数类型	描述
separator	是	String	**参数解释：** 分隔符。 **约束限制：** 不涉及。 **取值范围：** 长度1~20，除\n外，不允许包含以下字符 *./$^?+ 且不允许为<!I@M#A%G&E!>、<!T@A#B%L&E!>。 **默认取值：** 不涉及。
extraction_code	是	Integer	**参数解释：** 提取分段序号。 **约束限制：** 不涉及。 **取值范围：** 范围1~100，提取序号大于可提取分段数量时字段内容为空串。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
contain_separator	是	Boolean	**参数解释：** 提取分段是否包含分隔符。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。

表 6-18 HeadAndTailExtractionTemplate

参数	是否必选	参数类型	描述
start_with	是	String	**参数解释：** 提取分段开头。 **约束限制：** 不涉及。 **取值范围：** 长度1~20，除\n外，不允许包含以下字符 *./$^?+，且不允许为<!I@M#A%G&E!>、<!T@A#B%L&E!>。 **默认取值：** 不涉及。
contain_start	是	Boolean	**参数解释：** 提取分段是否包含开头。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
end_with	是	String	**参数解释：** 提取分段结尾。 **约束限制：** 不涉及。 **取值范围：** 长度1~20，除\n外，不允许包含以下字符 *./$^?+，且不允许为<!I@M#A%G&E!>、<!T@A#B%L&E!>。 **默认取值：** 不涉及。
contain_end	是	Boolean	**参数解释：** 提取分段是否包含结尾。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。
extraction_code	是	Integer	**参数解释：** 提取分段序号。 **约束限制：** 不涉及。 **取值范围：** 范围1~100，提取序号大于可提取分段数量时字段内容为空串。 **默认取值：** 不涉及。

表 6-19 IndexConfig

参数	是否必选	参数类型	描述
name	是	String	**参数解释：** 索引配置名称。 **约束限制：** 不涉及。 **取值范围：** 支持中英文、数字、“_”，长度为2~50个字符，以中英文、数字开头。 **默认取值：** 不涉及。
description	否	String	**参数解释：** 索引配置描述。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。
rag_type	否	String	**参数解释：** 知识库RAG类型。 **约束限制：** 不涉及。 **取值范围：** 可以为空，为空则使用默认值。 枚举值：VECTOR_RAG（向量RAG，是一种结合了向量化和大语言模型的RAG技术）、GRAPH_RAG（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）。 **默认取值：** VECTOR_RAG

参数	是否必选	参数类型	描述
vector_model_service_key	是	String	**参数解释：** 向量化模型的service_key。 **约束限制：** 不涉及。 **取值范围：** 长度1~128，不能为空白字符，如空格。 **默认取值：** 不涉及。
long_text_solution	是	String	**参数解释：** 知识数据集切片长文本处理方式。 **约束限制：** 不涉及。 **取值范围：** 枚举值： • TRUNCATE_MODE（如果分片的token长度超过向量化模型的token数，则知识库向量化失败）。 • SMART_MODE（如果分片的token长度超过向量化模型的token数，则自动对超长部分进行截断处理）。 • DEFAULT_MODE（如果分片的token长度超过向量化模型的token数，则大模型对超长部分进行重写；如果重写后仍然超长，则进入截断模式。此模式较为耗时）。 **默认取值：** 不涉及。
index_graph_config	否	Object **IndexGraphConfig** objects	**参数解释：** 知识图谱相关配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
retrieval_configs	是	Array of **IndexConfigField** objects	**参数解释：** 知识库召回配置。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。

表 6-20 IndexConfigField

参数	是否必选	参数类型	描述
name	是	String	**参数解释：** 索引字段名称。 **约束限制：** 不涉及。 **取值范围：** 名称长度1~20，仅可包含字母、数字、下划线，并且以字母开头，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。 **默认取值：** 不涉及。
category	是	String	**参数解释：** 索引字段类型。 **约束限制：** 不涉及。 **取值范围：** 枚举值：FULL_CHUNK（完整切片）、CHUNK_FRAGMENT（切片提取片段）。 **默认取值：** 不涉及。

参数	是否必选	参数类型	描述
vector_retrieval	是	Boolean	**参数解释：** 是否为向量化字段。 **约束限制：** 整个索引配置中，必须有且只有一个向量化字段。 **取值范围：** true或false。 **默认取值：** 不涉及。
graph_extract	否	Boolean	**参数解释：** 是否为知识图谱抽取字段。 **约束限制：** 索引配置适配RAG类型为GRAPH_RAG时有效，整个索引配置中，最多有一个字段为true。 **取值范围：** true或false。 **默认取值：** false。
text_filter	是	Boolean	**参数解释：** 是否为文本过滤字段。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** false。
retrieval_return	是	Boolean	**参数解释：** 是否为附加返回字段。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** false。

表 6-21 IndexGraphConfig

参数	是否必选	参数类型	描述
entity_extract_method	是	String	**参数解释：** 实体抽取方式。 **约束限制：** 不涉及。 **取值范围：** 枚举值：TRIPLET（三元组抽取）。 **默认取值：** 不涉及。
extract_model_service_key	是	String	**参数解释：** 实体抽取模型服务key。 **约束限制：** 不涉及。 **取值范围：** 长度1~128，不能为空白字符，如空格。 **默认取值：** 不涉及。
customize_extract_prompt	是	Boolean	**参数解释：** 是否自定义实体抽取提示语。 **约束限制：** 不涉及。 **取值范围：** true或false。 **默认取值：** 不涉及。
extract_prompt	否	String	**参数解释：** 用户自定义实体抽取Prompt。 **约束限制：** 不涉及。 **取值范围：** 不涉及。 **默认取值：** 不涉及。