应用平台

API 参考

文档版本	01
发布日期	2024-02-08

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

1.1 概述

欢迎使用应用平台（AppStage），华为云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》**。

1.2 调用说明

AppStage提供了REST（Representational State Transfer）风格API，支持您通过HTTPS请求调用，调用方法请参见**如何调用API**。

1.3 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“appstage.huaweicloud.com/wiseagent”。

1.4 约束与限制

无。

1.5 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。

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
模型集市	调用预置大预言/向量化模型服务接口。
在线测试	私有部署的对话/向量化推理服务的在线测试接口。
知识库	知识库数据查询接口。

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**调用预置大语言模型流式模型服务**接口说明如何调用API，通过该API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。

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。

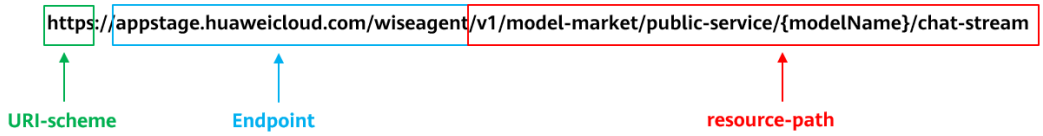
表 3-1 URI 中的参数说明

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“appstage.huaweicloud.com/wiseagent”。
resource-path	资源路径，也即API访问路径。从具体API的URI模块获取，例如“**调用预置大语言模型流式模型服务**”API的resource-path为“/v1/model-market/public-service/{modelName}/chat-stream”，其中*{modelName}*为模型名称。
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**调用预置大语言模型流式模型服务**接口，则需使用“华北-北京四”区域的Endpoint（appstage.huaweicloud.com/wiseagent），并在**调用预置大语言模型流式模型服务**的URI部分找到resource-path（/v1/model-market/public-service/{modelName}/chat-stream），拼接起来如下所示。

```
https://appstage.huaweicloud.com/wiseagent/v1/model-market/public-service/{modelName}/chat-stream
```

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**调用预置大语言模型流式模型服务**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://appstage.huaweicloud.com/wiseagent/v1/model-market/public-service/{modelName}/chat-stream
```

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**。

表 3-3 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3-4**。	是	modelmarket.chat

表 3-4 Resource-code

Resource-code	接口
modelmarket.chat	**调用预置大语言模型非流式模型服务**
modelmarket.chat.stream	**调用预置大语言模型流式模型服务**
modelmarket.embedding.batch	**调用预置向量化模型批量服务**
onlinetest.chat.test	**我部署的对话推理服务API在线测试**
onlinetest.embedding.test.batch	**我部署的向量化推理服务API在线测试**
dataset.query.embeddata	**知识库数据查询**

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**调用预置大语言模型流式模型服务**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://appstage.huaweicloud.com/wiseagent/v1/model-market/public-service/{modelName}/chat-stream

{
  "query" : "请介绍一下你自己",
  "history" : [ ],
  "system" : "你是一名程序员",
  "do_sample" : true,
  "max_length" : 2048,
  "max_new_tokens" : 1024,
  "temperature" : 0.8,
  "top_p" : 0.1,
  "repetition_penalty" : 1.1
}
```

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。

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) {
    }
}
```

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**调用预置大语言模型流式模型服务**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**调用预置大语言模型流式模型服务**接口，返回如下消息体。为篇幅起见，这里只展示部分内容。

```
data:我


data:是一名


data:人工智能


data:助手
```

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_msg": "The format of message is error",
   "error_code": "AS.0001"
}
```

其中，error_code表示错误码，error_msg表示错误描述信息。

4 API

4.1 模型集市

4.1.1 调用预置大语言模型非流式模型服务

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，大语言模型完整生成回答后一次性返回。

URI

POST /v1/model-market/public-service/{modelName}/chat

表 4-1 路径参数

参数	是否必选	参数类型	描述
modelName	是	String	模型名称，目前支持baichuan-13b-chat、chatglm3-6b。

请求参数

表 4-2 请求 Body 参数

参数	是否必选	参数类型	描述
history	否	Array of Array of objects	历史对话信息。
max_length	否	Integer	输入加输出最大token数。

参数	是否必选	参数类型	描述
max_new_tokens	否	Integer	输出最大token数。
query	是	String	对话输入。
repetition_penalty	否	Float	重复惩罚。
temperature	否	Float	温度。
system	否	String	角色。
do_sample	否	Boolean	是否概率采样token得到结果。
top_p	否	Float	多样性。

响应参数

状态码： 200

表 4-3 响应 Body 参数

参数	参数类型	描述
history	Array of objects	历史对话信息。
query	String	对话输入。
input_token_length	Integer	输入token数。
output_token_length	Integer	输出token数。
response	String	响应信息。
request_id	String	请求ID。

请求示例

```
https://{endpoint}/v1/model-market/public-service/{modelName}/chat

{
  "query" : "请介绍一下你自己",
  "history" : [ ],
  "system" : "你是一名程序员",
  "do_sample" : true,
  "max_length" : 2048,
  "max_new_tokens" : 1024,
  "temperature" : 0.8,
  "top_p" : 0.1,
  "repetition_penalty" : 1.1
}
```

响应示例

状态码： 200

OK


```
{
  "input_token_length" : 10,
  "response" : "我是一名人工智能助手，擅长处理各种问题，帮助用户解答疑问、提供建议和执行任务。我的知识库不断更新，可以为用户提供最新的信息和最专业的建议。我可以帮助用户编写代码、优化算法、分析数据以及其他各种编程需求。此外，我还具备自然语言处理能力，可以与用户进行流畅的对话，提供实时的帮助和支持。",
  "query" : "请介绍一下你自己",
  "history" : [ [ "请介绍一下你自己", "我是一名人工智能助手，擅长处理各种问题，帮助用户解答疑问、提供建议和执行任务。我的知识库不断更新，可以为用户提供最新的信息和最专业的建议。我可以帮助用户编写代码、优化算法、分析数据以及其他各种编程需求。此外，我还具备自然语言处理能力，可以与用户进行流畅的对话，提供实时的帮助和支持。" ] ],
  "output_token_length" : 82,
  "request_id" : "7f340105-7243-45c6-9388-2d32603c24ea-1706237665137234"
}
```


状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

4.1.2 调用预置大语言模型流式模型服务

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，逐个token的快速返回模式，不用等待大语言模型生成完成。

URI

POST /v1/model-market/public-service/{modelName}/chat-stream

表 4-4 路径参数

参数	是否必选	参数类型	描述
modelName	是	String	模型名称，目前支持baichuan-13b-chat、chatglm3-6b。

请求参数

表 4-5 请求 Body 参数

参数	是否必选	参数类型	描述
history	否	Array of Array of objects	历史对话信息。
max_length	否	Integer	输入加输出最大token数。
max_new_tokens	否	Integer	输出最大token数。
query	是	String	对话输入。
repetition_penalty	否	Float	重复惩罚。
temperature	否	Float	温度。
system	否	String	角色。
do_sample	否	Boolean	是否概率采样token得到结果。
top_p	否	Float	多样性。

响应参数

状态码： 200

表 4-6 响应 Body 参数

参数	参数类型	描述
timeout	Long	超时时长。

请求示例

```
https://{endpoint}/v1/model-market/public-service/{modelName}/chat-stream

{
  "query" : "请介绍一下你自己",
  "history" : [ ],
```

```
  "system" : "你是一名程序员",
  "do_sample" : true,
  "max_length" : 2048,
  "max_new_tokens" : 1024,
  "temperature" : 0.8,
  "top_p" : 0.1,
  "repetition_penalty" : 1.1
}
```

响应示例

状态码： 200

OK

```
data:我

data:是一名

data:人工智能

data:助手
```

状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

4.1.3 调用预置向量化模型批量服务

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。

URI

POST /v1/model-market/public-service/{modelName}/embedding-batch

表 4-7 路径参数

参数	是否必选	参数类型	描述
modelName	是	String	模型名称，目前支持bge-large-zh-v1.5。

请求参数

表 4-8 请求 Body 参数

参数	是否必选	参数类型	描述
text	是	Array of strings	输入的多条句子列表。

响应参数

状态码： 200

表 4-9 响应 Body 参数

参数	参数类型	描述
vectors	Array of objects	输入的多条句子转换的向量表示。
input_token_length	Integer	输入token数。

请求示例

```
https://{endpoint}/v1/model-market/public-service/{modelName}/embedding-batch

{
  "text" : [ "你好，你是哪个模型", "那是一个快乐的人", "那是一个快乐的狗" ]
}
```

响应示例

状态码： 200

OK

```
{
  "vectors" : [ [ 0.017777875065803528, -0.027557365596294403, -0.03859279677271843,
0.023178197443485260, "....." ], [ 0.00554633280262351, -0.04635364189743996, -0.07506467401981354,
0.03592068701982498, "....." ], [ 0.036464523524045944, -0.05596702918410301, 0.028902683407068253,
0.007492372300475836, "....." ] ],
  "input_token_length" : 13
}
```

状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

4.2 在线测试

4.2.1 我部署的对话推理服务 API 在线测试

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。非流式接口提供大语言模型完整生成回答后一次性返回。

URI

POST /v1/model-online-test/inference-service/test/{serviceId}/{modelName}/chat

表 4-10 路径参数

参数	是否必选	参数类型	描述
modelName	是	String	模型名称，目前支持 baichuan-13b-chat、chatglm3-6b。
serviceId	是	String	私人部署的模型服务ID。

请求参数

表 4-11 请求 Body 参数

参数	是否必选	参数类型	描述
history	否	Array of Array of objects	历史对话信息。
max_length	否	Integer	输入加输出最大token数。
max_new_tokens	否	Integer	输出最大token数。
query	是	String	对话输入。
repetition_penalty	否	Float	重复惩罚。
temperature	否	Float	温度。
system	否	String	角色。
do_sample	否	Boolean	是否概率采样token得到结果。
top_p	否	Float	多样性。

响应参数

状态码： 200

表 4-12 响应 Body 参数

参数	参数类型	描述
history	Array of Array of objects	历史对话信息。
query	String	对话输入。
input_token_length	Integer	输入token数。
output_token_length	Integer	输出token数。
response	String	响应信息。
request_id	String	请求ID。

请求示例

```
https://{endpoint}/v1/model-online-test/inference-service/test/{serviceId}/{modelName}/chat

{
```

```
  "query" : "请介绍一下你自己",
  "history" : [ ],
  "system" : "你是一名程序员",
  "do_sample" : true,
  "max_length" : 2048,
  "max_new_tokens" : 1024,
  "temperature" : 0.8,
  "top_p" : 0.1,
  "repetition_penalty" : 1.1
}
```

响应示例

状态码： 200

OK

```
{
  "input_token_length" : 10,
  "response" : "我是一名人工智能助手，擅长处理各种问题，帮助用户解答疑问、提供建议和执行任务。我的知识库不断更新，可以为用户提供最新的信息和最专业的建议。我可以帮助用户编写代码、优化算法、分析数据以及其他各种编程需求。此外，我还具备自然语言处理能力，可以与用户进行流畅的对话，提供实时的帮助和支持。",
  "query" : "请介绍一下你自己",
  "history" : [ [ "请介绍一下你自己", "我是一名人工智能助手，擅长处理各种问题，帮助用户解答疑问、提供建议和执行任务。我的知识库不断更新，可以为用户提供最新的信息和最专业的建议。我可以帮助用户编写代码、优化算法、分析数据以及其他各种编程需求。此外，我还具备自然语言处理能力，可以与用户进行流畅的对话，提供实时的帮助和支持。" ] ],
  "output_token_length" : 82,
  "request_id" : "7f340105-7243-45c6-9388-2d32603c24ea-1706237665137234"
}
```

状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

4.2.2 我部署的向量化推理服务 API 在线测试

功能介绍

向量化模型服务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。

URI

POST /v1/model-online-test/inference-service/test/{serviceId}/{modelName}/embedding-batch

表 4-13 路径参数

参数	是否必选	参数类型	描述
modelName	是	String	模型名称，目前支持bge-large-zh-v1.5。
serviceId	是	String	私人部署的模型服务ID。

请求参数

表 4-14 请求 Body 参数

参数	是否必选	参数类型	描述
text	是	Array of strings	输入的多条句子列表。

响应参数

状态码： 200

表 4-15 响应 Body 参数

参数	参数类型	描述
vectors	Array of objects	输入的多条句子转换的向量表示。
input_token_length	Integer	输入token数。

请求示例

```
https://{endpoint}/v1/model-online-test/inference-service/test/{serviceId}/{modelName}/embedding-batch

{
  "text" : [ "你好，你是哪个模型", "那是一个快乐的人", "那是一个快乐的狗" ]
}
```

响应示例

状态码： 200

OK

```
{
  "vectors" : [ [ 0.017777875065803528, -0.027557365596294403, -0.03859279677271843,
0.02317819744348526, "......" ], [ 0.00554633280262351, -0.04635364189743996, -0.07506467401981354,
0.03592068701982498, "......" ], [ 0.036464523524045944, -0.05596702918410301, 0.028902683407068253,
0.007492372300475836, "......" ] ],
  "input_token_length" : 13
}
```


状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

4.3 知识库

4.3.1 知识库数据查询

功能介绍

知识库数据查询接口。

URI

POST /v1/embedDataSet/queryEmbedData/{uuid}

表 4-16 路径参数

参数	是否必选	参数类型	描述
uuid	是	String	知识库页面复制的Embedding API中的uuid。

请求参数

表 4-17 请求 Body 参数

参数	是否必选	参数类型	描述
filter	否	Object	过滤条件。
is_cal_L2Distance	否	Boolean	是否计算向量距离，默认取false。取true时会默认按照向量距离降序排序。
limit	否	String	数据条目限制，默认为100条。
order_by	否	Object	排序条件。
query	是	String	查询内容。

响应参数

状态码： 200

表 4-18 响应 Body 参数

参数	参数类型	描述
body	Array of **results** objects	返回体body参数，列表类型。
code	String	响应状态id
msg	String	响应状态

表 4-19 results

参数	参数类型	描述
id	String	数据唯一id。
document	String	文档知识库：切分后的文档内容；图片知识库。
metadata	Object	**metadata用于过滤。**
uuid	String	数据唯一id。
dist	Number	向量距离。

表 4-20 metadata

参数	参数类型	描述
file_id	String	文件id。(默认)。
file_name	String	完整文件名。（默认）。
order	Integer	段落在全文中切分后的顺序。
source	String	文件路径。文件上传时仅为文件名，数据接入时为除去’obs路径’后的路径值。
source_path	String	文件路径。

请求示例

```
https://{endpoint}/v1/embedDataSet/queryEmbedData/{uuid}

{
  "query" : "介绍",
  "limit" : "10",
  "is_cal_L2Distance" : true,
  "filter" : {
    "and" : [ {
      "expression" : {
        "field" : "metadata.file_id",
        "operation" : "in",
        "value" : [ "tes-t_1", "test_2" ]
      }
    }, {
      "expression" : {
        "field" : "dist",
        "operation" : "<=",
        "value" : [ "40", "30" ]
      }
    } ]
  }
}
```

响应示例

状态码： 200

OK

```
{
  "code" : "0",
  "msg" : "success!",
  "body" : [ {
    "results" : "{\"id\":null,\"document\"           :\"AppStage平台提供运营中心、运维中心、开发中心、AI中心等服务，其中AI中心（WisepilotStack）目前支持数据接           入以及管理，提示语管理，模型微调以及应用编排等功能\",\"metadata\":{\"file_id\":\"AppStage_常见问题\"           ,\"file_name\":\"AppStage_常见问题.docx\",\"order\":42,\"source\":\"AppStage_常见问题.docx\"           ,\"source_path\":\"AppStage_常见问题.docx\"},\"uuid\":\"d30ad89f-8162-462a-9486-c4e90a552f27\",\"dist\":0           .24643265061367536}"
  }, {
    "results" : "{\"id\":null,\"document\"           :\"AppStage平台是什么？\\nAppStage平台由华为云研发，提供运营中心、运维中心、开发中心、AI中心服务的平台，其           中AI中心（WisepilotStack）是由\",\"metadata\":{\"file_id\":\"AppStage_常见问题\",\"file_name\"           :\"AppStage_常见问题.docx\",\"order\":40,\"source\":\"AppStage_常见问题.docx\",\"source_path\"           :\"AppStage_常见问题.docx\"},\"uuid\":\"023149c8-d109-4bab-acef-a17c65f8d455\",\"dist\":0           .2554842085986647}"
  }, {
    "results" : "{\"id\":null,\"document\"           :\"华为云SmartStage团队AI部门研发，包含了数据接入，提示语管理，模型微调以及管理，应用编排等功能。\\nAppSta           ge平台目前有哪些产品功能？\",\"metadata
```

\":{\"file_id\":\"AppStage_常见问题\",\"file_name\" :\"AppStage_常见问题.docx\",\"order\":41,\"source
\":\"AppStage_常见问题.docx\",\"source_path\" :\"AppStage_常见问题.docx\"},\"uuid
\":\"4df4528c-92de-460a-bfa1-515bfe7472bd\",\"dist\":0 .32361903803934944}"
}]
}

状态码

状态码	描述
200	OK
201	Created
401	Unauthorized
403	Forbidden
404	Not Found

错误码

请参见**错误码**。

5 权限和授权项

AppStage调用接口时使用AK/SK进行鉴权，具体AK/SK的创建方法请参考**认证鉴权**。创建AK/SK时需要使用管理员权限，具体权限介绍请参考AppStage《产品介绍》中的**权限管理**章节，具体权限申请操作请参考AppStage《用户指南》中的**权限管理**章节。

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
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

7 修订记录

发布日期	修订记录
2024-02-08	第一次正式发布。