盘古 Bot

API 参考

文档版本 01

发布日期 2026-02-05

华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2026。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

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。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

 版权所有 © 华为云计算技术有限公司

1 使用前必读

概述

欢迎使用盘古Bot服务（PanguBot）。盘古Bot服务是针对大模型场景全面升级的智能对话中枢，提供包括意图识别、纠错改写、知识增强问答等关键对话能力，可与大模型结合使用，打造端到端开箱即用的对话平台。

本文档提供了盘古Bot服务API的描述、语法、参数说明及样例等内容。通过配合使用这些接口，您可以轻松地使用盘古Bot服务。

终端节点

终端节点（Endpoint）即调用API的**请求地址**，不同服务不同区域的终端节点不同，您可以从**地区和终端节点**中查询所有服务的终端节点。

基本概念

- 账号

  用户注册时的账号，账号对其所拥有的资源及云服务具有完全的访问权限，可以重置用户密码、分配用户权限等。由于账号是付费主体，为了确保账号安全，建议您不要直接使用账号进行日常管理工作，而是创建用户并使用用户进行日常管理工作。

- 用户

  由账号在IAM中创建的用户，是云服务的使用人员，具有身份凭证（密码和访问密钥）。

  在**API凭证**下，您可以查看账号ID和用户ID。通常在调用API的鉴权过程中，您需要用到账号、用户和密码等信息。

- 区域

  指云资源所在的物理位置，同一区域内可用区间内网互通，不同区域间内网不互通。通过在不同地区创建云资源，可以将应用程序设计的更接近特定客户的要求，或满足不同地区的法律或其他要求。

- 可用区

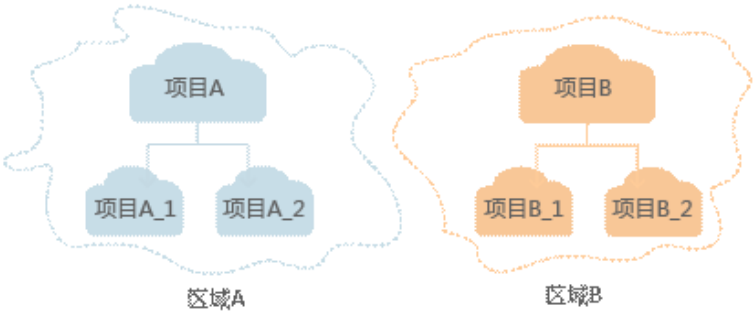
  一个可用区是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

- 项目

版权所有 © 华为云计算技术有限公司

区域默认对应一个项目，这个项目由系统预置，用来隔离物理区域间的资源（计算资源、存储资源和网络资源），以默认项目为单位进行授权，用户可以访问您账号中该区域的所有资源。如果您希望进行更加精细的权限控制，可以在区域默认的项目中创建子项目，并在子项目中购买资源，然后以子项目为单位进行授权，使得用户仅能访问特定子项目中资源，使得资源的权限控制更加精确。

图 1-1 项目隔离模型



- Checkpoint：消费检查点。应用程序消费数据时，记录已消费数据的最新序列号作为检查点。当重新消费数据时，可根据此检查点继续消费。
- APP：应用程序标识符。当多个应用程序分别消费同一通道的数据时，为区分不同应用程序的消费检查点，使用APP作为标识。

版权所有 © 华为云计算技术有限公司

2 API 概览

盘古Bot提供的接口为符合RESTful API设计规范的自研接口。通过盘古Bot的自研接口，您可以使用盘古Bot的如表2-1所示的功能。

表 2-1 接口说明

类型	说明
对话接口	问答 SSE（Server-sent events）接口的主要功能是指定助手进行问答，可通过传递conversation_id实现对话内多轮问答。

版权所有 © 华为云计算技术有限公司

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IAM服务的**管理员创建IAM用户**接口说明如何调用API。

请求 URI

请求URI由如下部分组成。

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

表 3-1 请求 URL

参数	说明
URI-scheme	传输请求的协议，当前所有API均采用**HTTPS**协议。
Endpoint	承载REST服务端点的服务器域名或IP，不同服务在不同区域的Endpoint不同，可以参考**终端节点**获取。例如IAM服务在“西南-贵阳一”区域的Endpoint为“iam.cn-southwest-2.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“管理员创建IAM用户”接口的resource-path为“/v3.0/OS-USER/users”。
query-string	查询参数，可选，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。

例如您需要创建IAM用户，由于IAM为全局服务，则使用任一区域的Endpoint（比如“西南-贵阳一”区域的Endpoint：“iam.cn-southwest-2.myhuaweicloud.com”），并在**管理员创建IAM用户**的URI部分找到resource-path（/v3.0/OS-USER/users），拼接起来如下所示。

```
https://iam.cn-southwest-2.myhuaweicloud.com/v3.0/OS-USER/users
```


版权所有 © 华为云计算技术有限公司

图 3-1 URI 示意图

说明

为查看方便，每个具体API的URI，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。

- **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**管理员创建IAM用户**的URI部分，您可以看到其请求方法为“POST”，则其请求为：

```
POST https://iam.cn-southwest-2.myhuaweicloud.com/v3.0/OS-USER/users
```

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3-2**。

表 3-2 公共请求消息头

参数	是否必选	描述
Content-Type	是	消息体的类型（格式）。推荐用户使用默认值“application/json”。
X-Auth-Token	使用Token认证时必选	用户Token。是调用“获取用户Token”接口的响应值，该接口是唯一不需要认证的接口。请求响应成功后在响应消息头（Header）中包含的“X-Subject-Token”的值即为Token值。

版权所有 © 华为云计算技术有限公司

参数	是否必选	描述
X-Project-Id	否	子项目ID，在多项目场景中使用。如果云服务资源创建在子项目中，AK/SK认证方式下，操作该资源的接口调用需要在请求消息头中携带X-Project-ID。
X-Sdk-Date	使用AK/SK认证时必选	请求的发生时间，当使用AK/SK方式认证时，使用SDK对请求进行签名的过程中会自动填充该字段。 AK/SK认证的详细说明请参见**认证鉴权**。 格式为(YYYYMMDD'T'HHMMSS'Z')，取值为当前系统的GMT时间。
Authorization	使用AK/SK认证时必选	签名认证信息。当使用AK/SK方式认证时，使用SDK对请求进行签名的过程中会自动填充该字段。 AK/SK认证的详细说明请参见**认证鉴权**。
X-Language	否	请求语言。

对于**管理员创建IAM用户**接口，使用AK/SK方式认证时，添加消息头后的请求如下所示。

```
POST https://iam.cn-north-4.myhuaweicloud.com/v3.0/OS-USER/users
Content-Type: application/json
X-Sdk-Date: 20240416T095341Z
Authorization: SDK-HMAC-SHA256 Access=**********, SignedHeaders=content-type;host;x-sdk-date,
Signature=**********
```

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。若请求消息体中参数支持中文，则中文字符必须为UTF-8编码，并在Content-type中声明字符编码方式，例如：Content-Type: application/json; charset=utf-8。

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**管理员创建IAM用户**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，其中加粗的字段需要根据实际值填写。

- **accountid**为IAM用户所属的账号ID。
- **username**为要创建的IAM用户名。
- **************为IAM用户的登录密码。

```
POST https://iam.cn-north-4.myhuaweicloud.com/v3.0/OS-USER/users（中国站）
POST https://iam.ap-southeast-1.myhuaweicloud.com/v3.0/OS-USER/users（国际站）
```


版权所有 © 华为云计算技术有限公司

```
Content-Type: application/json
X-Sdk-Date: 20240416T095341Z
Authorization: SDK-HMAC-SHA256 Access=**********, SignedHeaders=content-type;host;x-sdk-date,
Signature=**********

{
    "user": {
        "domain_id": "accountid",
        "name": "username",
        "password": "**********",
        "description": "IAM User Description"
    }
}
```

到此为止，一个API请求所需要的内容已经准备完成，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

调用接口有如下两种认证方式，您可以选择其中一种进行认证鉴权。

- AK/SK认证：通过AK（Access Key ID）/SK（Secret Access Key)加密调用请求。
- Token认证：通过Token认证调用请求。

AK/SK 认证

说明

- AK/SK签名认证方式仅支持消息体大小12MB以内，12MB以上的请求请使用Token认证。
- AK/SK既可以使用永久访问密钥中的AK/SK，也可以使用临时访问密钥中的AK/SK，但使用临时访问密钥的AK/SK时需要额外携带“X-Security-Token”字段，字段值为临时访问密钥的security_token。

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，也可以使用专门的签名SDK对请求进行签名。详细的签名方法和SDK使用方法请参见**API签名指南**。

须知

签名SDK只提供签名功能，与服务提供的SDK不同，使用时请注意。

您也可以通过这个视频教程了解AK/SK认证的使用：**https://bbs.huaweicloud.com/videos/100697** 。

版权所有 © 华为云计算技术有限公司

Token 认证

说明

- Token的有效期为24小时，需要使用同一个Token鉴权时，可以缓存起来，避免频繁调用。
- 使用Token前请确保Token离过期有足够的时间，防止调用API的过程中Token过期导致调用API失败。

Token在计算机系统中代表令牌（临时）的意思，拥有Token就代表拥有某种权限。Token认证就是在调用API的时候将Token加到请求消息头，从而通过身份认证，获得操作API的权限。

Token可通过调用**获取用户Token**接口获取，调用本服务API需要project级别的Token，即调用获取用户Token接口时，请求body中“auth.scope”的取值需要选择“project”，如下所示。

```
{
    "auth": {
        "identity": {
            "methods": [
                "password"
            ],
            "password": {
                "user": {
                    "name": "username",    //用户名
                    "password": "********",    //登录密码
                    "domain": {
                        "name": "domainname"    //用户所属的账号名称
                    }
                }
            }
        },
        "scope": {
            "project": {
                "name": "xxxxxxxx"    //项目名称
            }
        }
    }
}
```

获取Token后，再调用其他接口时，您需要在请求消息头中添加“X-Auth-Token”，其值即为Token。例如Token值为“ABCDEFJ....”，则调用接口时将“X-Auth-Token: ABCDEFJ....”加到请求消息头即可，如下所示。

```
POST https://iam.cn-north-4.myhuaweicloud.com/v3.0/OS-USER/users
Content-Type: application/json
X-Auth-Token: ABCDEFJ....
```

您还可以通过这个视频教程了解如何使用Token认证：**https://bbs.huaweicloud.com/videos/101333** 。

3.3 返回结果

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

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**管理员创建IAM用户**接口，如果调用后返回状态码为“201”，则表示请求成功。

版权所有 © 华为云计算技术有限公司

响应消息头

对应请求消息头，响应同样也有消息头，如“Content-type”。

对于**管理员创建IAM用户**接口，返回如**图3-2**所示的消息头。

图 3-2 管理员创建 IAM 用户响应消息头

```
"X-Frame-Options": "SAMEORIGIN",
"X-IAM-ETag-id": "2562365939-d8f6f12921974cb097338ac11fceac8a",
"Transfer-Encoding": "chunked",
"Strict-Transport-Security": "max-age=31536000; includeSubdomains;",
"Server": "api-gateway",
"X-Request-Id": "af2953f2bcc67a42325a69a19e6c32a2",
"X-Content-Type-Options": "nosniff",
"Connection": "keep-alive",
"X-Download-Options": "noopen",
"X-XSS-Protection": "1; mode=block;",
"X-IAM-Trace-Id": "token_[redacted]_null_af2953f2bcc67a42325a69a19e6c32a2",
"Date": "Tue, 21 May 2024 09:03:40 GMT",
"Content-Type": "application/json; charset=utf8"
```

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**管理员创建IAM用户**接口，返回如下消息体。篇幅有限，这里只展示部分内容。

```
{
    "user": {
        "id": "c131886aec...",
        "name": "IAMUser",
        "description": "IAM User Description",
        "areacode": "",
        "phone": "",
        "email": "***@***.com",
        "status": null,
        "enabled": true,
        "pwd_status": false,
        "access_mode": "default",
        "is_domain_owner": false,
        "xuser_id": "",
        "xuser_type": "",
        "password_expires_at": null,
        "create_time": "2024-05-21T09:03:41.000000",
        "domain_id": "d78cbac1.....",
        "xdomain_id": "30086000.....",
        "xdomain_type": "",
        "default_project_id": null
    }
}
```

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": "Request body is invalid.",
    "error_code": "IAM.0011"
}
```

其中，error_code表示错误码，error_msg表示错误描述信息。

版权所有 © 华为云计算技术有限公司

4 API

4.1 问答 SSE - GenerateChatSseAnswer

功能介绍

问答SSE（Server-Sent Events）接口主要用于实现实时或近实时的数据推送。SSE是一种允许服务器向浏览器推送更新的技术，它使用HTTP协议，但与传统的请求-响应模式不同，SSE允许服务器主动向客户端发送数据，而不需要客户端频繁地发起请求。在问答系统中，SSE接口可以用来实现实时的问题回答更新。例如，当用户提交一个问题后，服务器可以通过SSE接口实时推送新的回答或更新，而不需要用户不断刷新页面来查看是否有新的回答。

说明

Web原生SSE不支持POST请求，需使用JavaScript库sse.js。

URI

POST /v1/{project_id}/koochat/assistants/{assistant_id}/chat

表 4-1 路径参数

参数	是否必选	参数类型	描述
project_id	是	String	**参数解释**: 项目ID。获取方法请参见获取项目ID和名称。 **约束限制**: 不涉及 **取值范围**: 账户的项目ID。 **默认取值**: 不涉及

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
assistant_id	是	String	**参数解释**: 盘古Bot页面创建的助手ID。获取方法请参见**获取助手ID**。 **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 4-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	否	String	**参数解释**: 用户鉴权信息。获取方法请参见**认证鉴权**。 **约束限制**: 不涉及 **取值范围**: 不涉及。 **默认取值**: 不涉及

表 4-3 请求 Body 参数

参数	是否必选	参数类型	描述
question	是	String	**参数解释:** 用户问题。 **约束限制:** 不涉及 **取值范围:** 1~4096个字符。 **默认取值:** 不涉及

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
conversation_id	否	String	**参数解释：** 会话ID，不携带或携带错误会自动生成新的ID。 **约束限制：** 不涉及 **取值范围：** 只能由英文字母、数字及“-”、“_”组成，且长度为[1~36]个字符，建议使用UUID。 **默认取值：** 不涉及
conversation_conf	否	**Conversation Conf** object	**参数解释：** 对话配置， 创建新会话时需要。 **约束限制：** 不涉及
source	否	String	**参数解释：** 问答来源。 **约束限制：** 不涉及 **取值范围：** ● API：接口调用来源的对话。 ● CONSOLE：盘古Bot控制台来源的对话。 **默认取值：** 不涉及
debug	否	Boolean	**参数解释：** 是否需要进行调试。 **约束限制：** 不涉及 **取值范围：** ● true：启用debug模式。 ● false：不启用debug模式。 **默认取值：** false

版权所有 © 华为云计算技术有限公司

参数	是否必选	参数类型	描述
extra_info	否	String	**参数解释：** 对话额外信息，通过技能透传给FunctionGraph。 **约束限制：** 不涉及 **取值范围：** 不涉及 **默认取值：** 不涉及
user_id	否	String	**参数解释：** 记录用户侧的用户ID，支持用户自定义传入。 **约束限制：** 不涉及 **取值范围：** 0-36字符。 **默认取值：** 不涉及
stream	否	Boolean	**参数解释：** 大模型返回的结果是否通过流式方式输出，以实现打字机效果。 **约束限制：** 不涉及 **取值范围：** ● true：流式输出。 ● false：非流式输出。 **默认取值：** false

版权所有 © 华为云计算技术有限公司

表 4-4 ConversationConf

参数	是否必选	参数类型	描述
ref_enable	否	Boolean	**参数解释：** 是否引用溯源开关。 **约束限制：** 不涉及 **取值范围：** true：开启引用溯源。返回结果中会包含References事件，返回RAG和联网搜索的文件关联信息。 false：关闭引用溯源。 **默认取值：** false

响应参数

状态码：200

表 4-5 响应 Body 参数

参数	参数类型	描述
event-message	**AnswerMessage** object	**参数解释：** 流式对话回复，需要拼接获取最终答案。
event-resp	**RespMessage** object	**参数解释：** 对话的辅助信息，如意图、槽位等信息。
event-references	**ReferencesMessage** object	**参数解释：** 引用溯源。
event-faq_log	**FaqLogMessage** object	**参数解释：** FAQ日志。
event-intention_retrieve_log	**IntentionRetrieveLog** object	**参数解释：** 意图检索日志。
event-intention_distribute_log	**IntentionDistributeLog** object	**参数解释：** 意图分发日志。
event-rag_log	**RagLog** object	**参数解释：** RAG日志。

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
event-llm_request_log	LlmRequestLog object	**参数解释：** 大模型日志。
event-moderation	String	**参数解释：** 内容审核。 **取值范围：** 不涉及

表 4-6 AnswerMessage

参数	参数类型	描述
answer	String	**参数解释：** 流式返回的答案片段。 **取值范围：** 不涉及
created	Long	**参数解释：** 创建事件。 **取值范围：** 不涉及

表 4-7 RespMessage

参数	参数类型	描述
conversation_id	String	**参数解释：** 会话ID。 **取值范围：** 不涉及
created	Long	**参数解释：** 响应时间。 **取值范围：** 不涉及
answer	String	**参数解释：** 完整回复内容。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
request_id	String	**参数解释：** 请求ID。 **取值范围：** 不涉及
reply_type	Integer	**参数解释：** 回复类型。 **取值范围：** ● 0：技能。 ● 1：问答。 ● 2：闲聊。 ● 3：无答案。
detail	**ChatResp** object	**参数解释：** 返回的详情。
rewritten_question	String	**参数解释：** Query改写的结果。 **取值范围：** 不涉及

表 4-8 ChatResp

参数	参数类型	描述
log_id	String	**参数解释：** 对话ID。 **取值范围：** 不涉及
answer	String	**参数解释：** 引擎回复。 **取值范围：** 不涉及
cost	Long	**参数解释：** 耗时。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
resp_type	Integer	**参数解释：** 响应类型。 **取值范围：** ● 0：技能。 ● 1：问答。
score	Float	**参数解释：** 分数。 **取值范围：** 0~1
fusion_result	Integer	**参数解释：** 问答类型。仅当“resp_type”为“1”时才有值。 **取值范围：** ● 0：FAQ。 ● 1：知识。 ● 2：LLM直接回复。
skill_id	String	**参数解释：** 使用的技能ID。仅当“resp_type”为“1”时才有值。 **取值范围：** 不涉及
frame	**FrameDTO** object	**参数解释：** 技能Frame数据。仅当“resp_type”为“1”时才有值。

表 4-9 FrameDTO

参数	参数类型	描述
history_global_variables	Map<String,String>	**参数解释：** 当前会话（可包含多轮对话）中，所有收集到的变量值，包含历史轮对话。Key为变量名称，Value为变量值。
current_variables	Map<String,String>	**参数解释：** 当前对话轮次中，收集到的变量值。Key为变量名称，Value为变量值。

版权所有 © 华为云计算技术有限公司

表 4-10 ReferencesMessage

参数	参数类型	描述
references	**references** object	**参数解释**: 引用数据。

表 4-11 references

参数	参数类型	描述
document_name	Map<String,Object>	**参数解释**: 具体文档名称。
type	String	**参数解释**: 具体文档类型。 **取值范围**: document:文档。
reference	Boolean	**参数解释**: 是否被引用。 **取值范围**: ● true：被引用。 ● false：未被引用。
referenceIndex	Integer	**参数解释**: 索引。 **取值范围**: 不涉及
page_content	String	**参数解释**: 引用原文内容。 **取值范围**: 不涉及
metadata	**metadata** object	**参数解释**: 原始数据。

版权所有 © 华为云计算技术有限公司

表 4-12 metadata

参数	参数类型	描述
title	String	**参数解释**: 文档标题。 **取值范围**: 不涉及
content	String	**参数解释**: 引用原文内容。 **取值范围**: 不涉及
doc_name	String	**参数解释**: 文档分片名称。 **取值范围**: 不涉及
_id	String	**参数解释**: 文档分片Id。 **取值范围**: 不涉及
_score	String	**参数解释**: 分数。 **取值范围**: 0~1

表 4-13 FaqLogMessage

参数	参数类型	描述
directory_ids	Array of strings	**参数解释:** 目录ID。
faqs	Array of **faqs** objects	**参数解释:** FAQ信息。
cost	Integer	**参数解释:** 耗时。 **取值范围:** 不涉及

版权所有 © 华为云计算技术有限公司

表 4-14 faqs

参数	参数类型	描述
id	String	**参数解释：** FAQ ID。 **取值范围：** 不涉及
questions	Array of **questions** objects	**参数解释：** 问题列表。
top_score	Float	**参数解释：** 最大匹配分数。 **取值范围：** 0~1
answers	Array of **answers** objects	**参数解释：** 答案列表。

表 4-15 questions

参数	参数类型	描述
question	String	**参数解释：** 命中的问题。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1

表 4-16 answers

参数	参数类型	描述
content	String	**参数解释：** 答案内容。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
type	Integer	**参数解释：** 答案类型。 **取值范围：** ● 0：纯文本。 ● 1：富文本。

表 4-17 IntentionRetrieveLog

参数	参数类型	描述
candidate_intentions	Array of **candidate_intentions** objects	**参数解释：** 候选意图。
cost	Integer	**参数解释：** 耗时。 **取值范围：** 不涉及

表 4-18 candidate_intentions

参数	参数类型	描述
intention_id	String	**参数解释：** 意图ID。 **取值范围：** 不涉及
intention_name	String	**参数解释：** 意图名称。 **取值范围：** 不涉及
questions	Array of **questions** objects	**参数解释：** 问题列表。
assistant_id	String	**参数解释：** 助手ID。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
version	String	**参数解释：** 版本。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1
threshold	Float	**参数解释：** 采纳阈值。 **取值范围：** 不涉及

表 4-19 questions

参数	参数类型	描述
embedding	Array of floats	**参数解释：** Embedding列表。
question	String	**参数解释：** 问题。 **取值范围：** 不涉及
question_id	String	**参数解释：** 问题ID。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1

版权所有 © 华为云计算技术有限公司

表 4-20 IntentionDistributeLog

参数	参数类型	描述
distributed_skill_id	String	**参数解释：** 分发技能ID。 **取值范围：** 不涉及
distributed_skill_name	String	**参数解释：** 分发技能名称。 **取值范围：** 不涉及
intention_recognized_result	**intention_recognized_result** object	**参数解释：** 意图识别结果。
score	Float	**参数解释：** 分数。 **取值范围：** 0~1

表 4-21 intention_recognized_result

参数	参数类型	描述
intention_id	String	**参数解释：** 意图ID。 **取值范围：** 不涉及
intention_name	String	**参数解释：** 意图名称。 **取值范围：** 不涉及
questions	Array of **questions** objects	**参数解释：** 问题列表。
assistant_id	String	**参数解释：** 助手ID。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
version	String	**参数解释：** 版本。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1
threshold	Float	**参数解释：** 采纳阈值。 **取值范围：** 不涉及

表 4-22 questions

参数	参数类型	描述
embedding	Array of floats	**参数解释：** Embedding列表。
question	String	**参数解释：** 问题内容。 **取值范围：** 不涉及
question_id	String	**参数解释：** 问题ID。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1

版权所有 © 华为云计算技术有限公司

表 4-23 RagLog

参数	参数类型	描述
directory_ids	Array of strings	**参数解释：** 文档目录ID。
debug_info	**debug_info** object	**参数解释：** 调用RAG引擎关键信息。
cost	Integer	**参数解释：** 调用RAG引擎接口耗时。 **取值范围：** 不涉及

表 4-24 debug_info

参数	参数类型	描述
doc_search_response	**doc_search_response** object	**参数解释：** 搜索知识管理的结果。
websearch_response	Array of strings	**参数解释：** 公网搜索结果。
rag_knowledge	Array of strings	**参数解释：** 最终选取作为RAG的知识。

表 4-25 doc_search_response

参数	参数类型	描述
docs	Array of **docs** objects	**参数解释：** 文档描述。

表 4-26 docs

参数	参数类型	描述
file_id	String	**参数解释：** 文件ID。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

参数	参数类型	描述
title	String	**参数解释：** 文件Title。 **取值范围：** 不涉及
subtitle	String	**参数解释：** 文件子Title。 **取值范围：** 不涉及
content	String	**参数解释：** 内容。 **取值范围：** 不涉及
doc_type	String	**参数解释：** 文件类型。 **取值范围：** 不涉及
file_path	String	**参数解释：** 文件路径。 **取值范围：** 不涉及
update_date_time	String	**参数解释：** 文件更新时间。 **取值范围：** 不涉及
score	Float	**参数解释：** 分数。 **取值范围：** 0~1
extra_info	String	**参数解释：** 文档扩展信息。 **取值范围：** 不涉及

版权所有 © 华为云计算技术有限公司

表 4-27 LlmRequestLog

参数	参数类型	描述
llm_chat_req	**llm_chat_req** object	**参数解释：** 大模型调用请求。

表 4-28 llm_chat_req

参数	参数类型	描述
messages	Array of **messages** objects	**参数解释：** 大模型调用请求中的Massages列表。
conversationId	String	**参数解释：** 对话ID。 **取值范围：** 不涉及

表 4-29 messages

参数	参数类型	描述
role	String	**参数解释：** 角色。 **取值范围：** • user：用户。 • system：系统。
content	String	**参数解释：** Content问题。 **取值范围：** 不涉及

请求示例

流式调用对话问答接口。

```
POST https://{endpoint}/v1/{project_id}/koochat/assistants/{assistant_id}/chat

{
  "stream" : true,
  "source" : "API",
  "question" : "查询天气"
}
```


版权所有 © 华为云计算技术有限公司

响应示例

状态码：200

请求成功。

对话接口 SSE 响应消息。

属性名 "event:{type}"，“{type}” 为消息事件类型；属性值"data" 为 SSE 消息数据。

```
event:faq_log
retry:0
data:{"cost":125,"directory_ids":[]}

event:intention_retrieve_log
retry:0
data:{"cost":138,"candidate_intentions":[]}

event:intention_distribute_log
retry:0
data:{"score":null,"distributed_skill_id":null,"distributed_skill_name":null,"intention_recognized_result":null}

event:qaflow_log
retry:0
data:{"cost":561,"directories":[]}

event:rag_log
retry:0
data:{"cost":444,"directory_ids":["123"],"debug_info":{"intention":"无意图","doc_search_response":{"docs":
[]},"rag_knowledge":[],"websearch_response":[],"error_info":[],"planning_queries":["你好"]}}

event:llm_request_log
retry:0
data:{"llm_chat_req":{"conversationId":"1769568381849004","customizedLlmConfig":{"region":"cn-
southwest-2","url":"http://127.0.0.1:80/modelarts-maas/deepseek-v3/v1/chat/
completions","service_type":2,"maas_model_name":"DeepSeek-V3","cloud_system_enable":true},"messages":
[{"role":"system","content":"你是智能对话助手，用户消息为：\n你好"}],"stream":true}}

event:llm_response_log
retry:0
data:{"first_token_cost":1555}

event:e2e_log
retry:0
data:{"first_token_cost":3571}

event:message
retry:0
data:{"created":1769590001702,"answer":"你好
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001707,"answer":"！
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001714,"answer":"今天是
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:
{"created":1769590001715,"answer":"2026","conversation_id":"1769568381849004","request_id":"176956838
1849004"}

event:message
```


版权所有 © 华为云计算技术有限公司

```
retry:0
data:{"created":1769590001715,"answer":"年
1","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001715,"answer":"月
28","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001716,"answer":"日，
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001716,"answer":"星期三下午
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001717,"answer":"好
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001717,"answer":"。
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001718,"answer":"有什么
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001718,"answer":"可以
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001719,"answer":"帮您的
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001719,"answer":"吗?
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001719,"answer":"无论是问题
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001720,"answer":"解答、
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001720,"answer":"内容
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001720,"answer":"创作还是
","conversation_id":"1769568381849004","request_id":"1769568381849004"}
```

 版权所有 © 华为云计算技术有限公司

```
event:message
retry:0
data:{"created":1769590001721,"answer":"技术支持
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001721,"answer":"，
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001721,"answer":"我
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001721,"answer":"随时
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001721,"answer":"为您服务
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:message
retry:0
data:{"created":1769590001722,"answer":"。
","conversation_id":"1769568381849004","request_id":"1769568381849004"}

event:reference
retry:0
data:[]

event:resp
retry:0
data:{"created":1769590001723,"answer":"你好！今天是2026年1月28日，星期三下午好。有什么可以帮您的吗？无论是问题解答、内容创作还是技术支持，我随时为您服务。
","conversation_id":"1769568381849004","request_id":"1769568381849004","detail":{"answer":"你好！今天是2026年1月28日，星期三下午好。有什么可以帮您的吗？无论是问题解答、内容创作还是技术支持，我随时为您服务。","resp_type":1,"llm_generated":false}}

event:message
retry:0
data:[DONE]
```

状态码

状态码	描述
200	请求成功。 对话接口 SSE 响应消息。 属性名 "event:{type}"，“{type}” 为消息事件类型；属性值"data" 为 SSE 消息数据。

错误码

请参见**错误码**。

版权所有 © 华为云计算技术有限公司

5 附录

5.1 状态码

状态码如表5-1所示。

表 5-1 状态码说明

状态码	编码	状态码说明
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。只能切换到更高版本的协议。 例如，切换到HTTPS的高版本协议。
200	OK	请求已成功。
201	Created	创建类的请求已成功。
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	请求已成功，同时HTTPS响应不包含响应体。 在响应OPTIONS方法的HTTPS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。

版权所有 © 华为云计算技术有限公司

状态码	编码	状态码说明
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。
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HTTPS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	服务器等待请求时发生超时。 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删除。

版权所有 © 华为云计算技术有限公司

状态码	编码	状态码说明
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。
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HTTPS协议的版本，无法完成处理。

5.2 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。

版权所有 © 华为云计算技术有限公司

错误码	错误信息	描述	处理措施
CBS.0001	Request parameter error	请求参数错误。 示例：请求参数取值不合法、意图名称重复。	根据错误提示修改请求体。
CBS.0002	Authentication failure error	认证失败错误。 示例：资源非法或被冻结，没有访问权限、鉴权失败。	重新登录。
CBS.0003	Resource not found error	资源不存在错误。 示例：资源directory：directory_id不存在。	确认资源的关联关系是否正确。
CBS.0004	Internal error	内部错误。 示例：http请求执行失败、http请求构建错误。	请联系技术支持。

5.3 获取项目 ID 和名称

在调用接口的时候，部分请求中需要填入项目ID（project_id）或项目名称（project name），所以需要获取到项目ID和名称。

从控制台获取项目 ID 和名称

1. 登录盘古Bot控制台。
2. 在页面右上角单击用户名，在下拉列表中单击“我的凭证”，进入“API凭证”页面。
3. 在项目列表中查看“项目ID”和“项目”（“项目”即项目名称）。

图 5-1 查看项目 ID 和名称

版权所有 © 华为云计算技术有限公司

调用 API 获取项目 ID

项目ID还可通过调用**查询指定条件下的项目信息**API获取。

获取项目ID的接口为“GET https://{iam-endpoint}/v3/projects”，其中{iam-endpoint}为IAM的终端节点，可以参考**终端节点**获取，接口的认证鉴权请参见**认证鉴权**。

响应示例如下，例如CSS部署的区域为“xxx”，在响应消息体中搜索“name”为“xxx”，其对应的“projects”下的“id”即为项目ID。

```
{
   "projects": [
      {
         "domain_id": "65382450e8f64ac0870cd180d14exxxx",
         "is_domain": false,
         "parent_id": "65382450e8f64ac0870cd180d14exxxx",
         "name": "xxx",    //项目名称，即部署区域名称
         "description": "",
         "links": {
            "next": null,
            "previous": null,
            "self": "https://www.example.com/v3/projects/a4a5d4098fb4474fa22cd05f897dxxxx"
         },
         "id": "a4a5d4098fb4474fa22cd05f897dxxxx",    //项目ID
         "enabled": true
      }
   ],
   "links": {
      "next": null,
      "previous": null,
      "self": "https://www.example.com/v3/projects"
   }
}
```

5.4 获取助手 ID

在调用接口的时候，部分请求中需要填入助手ID（assistant_id），所以需要获取到助手ID。

从盘古大脑页面获取助手 ID

1. **登录盘古Bot控制台**。
2. 单击“助手管理”，找到想要调用的助手，单击“设置”按钮。
3. 查看助手ID字段。

图 5-2 查看 ID

版权所有 © 华为云计算技术有限公司