

Flexus 企业搜索服务

API 参考

文档版本 01
发布日期 2024-08-27



版权所有 © 华为技术有限公司 2024。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



HUAWEI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。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://www.huawei.com>

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：

<https://www.huawei.com/cn/psirt/vul-response-process>

如企业客户须获取漏洞信息，请参见如下网址：

<https://securitybulletin.huawei.com/enterprise/cn/security-advisory>

目录

1 使用前必读	1
1.1 概述	1
1.2 调用说明	1
1.3 终端节点	1
1.4 基本概念	2
2 API 概览	4
3 如何调用 API	8
3.1 构造请求	8
3.2 认证鉴权	11
3.3 接口获取	11
3.4 返回结果	12
4 知识库管理	14
4.1 创建知识库	14
4.2 修改知识库配置	25
4.3 设置知识库搜索问答 prompt	37
4.4 设置知识库通用无搜索的 prompt	40
4.5 开启知识库	42
4.6 关闭知识库	44
4.7 获取知识库列表	46
4.8 查看知识库列表	58
4.9 删除知识库	70
5 结构化数据	73
5.1 上传结构化数据	73
5.2 下载失败的条目	75
5.3 搜索结构化数据文件	78
6 文件管理	88
6.1 上传文档接口	88
6.2 查询文档接口	92
6.3 下载文件接口	101
6.4 删除文档接口	103
7 FAQ 管理	106

7.1 创建 FAQ.....	106
7.2 更新 FAQ.....	109
7.3 获取 FAQ.....	112
7.4 查询 FAQ 列表.....	116
7.5 删除 FAQ.....	120
7.6 批量删除 FAQ.....	122
8 FAQ 批量管理.....	125
8.1 批量 FAQ 上传.....	125
8.2 列举批量上传 FAQ 文件.....	128
8.3 下载上传失败的 FAQ.....	137
8.4 下载批量导入 FAQ 原文件.....	139
8.5 删除批量导入 FAQ 文件.....	141
9 搜索与问答.....	144
9.1 搜索知识库.....	144
9.2 生成答案.....	148
10 对话历史.....	156
10.1 查询对话历史列表.....	156
10.2 获取对话历史.....	160
10.3 删除对话历史.....	164
10.4 批量删除对话历史.....	166
11 图片管理.....	168
11.1 上传图片.....	168
11.2 获取图片内容.....	170
12 模型管理.....	174
12.1 新增模型.....	174
12.2 修改模型配置.....	178
12.3 列举模型.....	182
12.4 删除模型.....	187
13 用户的文档解析规则定义.....	189
13.1 创建用户解析规则.....	189
13.2 修改用户解析规则.....	191
13.3 列举用户规则列表.....	194
13.4 删除用户定义规则解析接口.....	196
14 公共参数.....	199
14.1 状态码.....	199
14.2 错误码.....	202

1 使用前必读

[概述](#)

[调用说明](#)

[终端节点](#)

[基本概念](#)

1.1 概述

欢迎使用Koosearch服务。Koosearch服务为您提供托管的分布式搜索引擎服务+大模型生成能力。本文档提供了Koosearch服务API的描述、语法、参数说明及样例等内容。

1.2 调用说明

KooSearch服务提供了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（Endpoint）即调用API的**请求地址**，KooSearch局点在贵阳一，终端节点见[表1-1](#)。

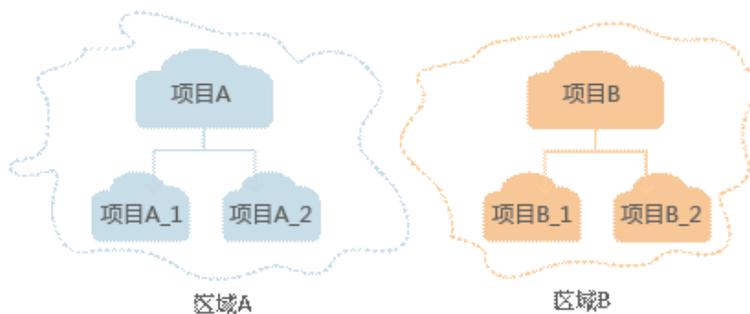
表 1-1 KooSearch 终端节点

区域名称	区域	终端节点（Endpoint）	协议类型
西南-贵阳一	cn-southwest-2	css.cn-southwest-2.myhuaweicloud.com	HTTPS

1.4 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他们进行日常管理工作
- 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
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。
- 企业项目

企业项目是项目的升级版，针对企业不同项目间资源的分组和管理，是逻辑隔离。企业项目中可以包含多个区域的资源，且项目中的资源可以迁入迁出。关于企业项目ID的获取及企业项目特性的详细信息，请参见《[企业管理服务用户指南](#)》。

2 API 概览

Koosearch提供的接口为符合RESTful API设计规范的自研接口。通过Koosearch的自研接口，您可以使用Koosearch的[表1 接口说明](#)所示的功能。

表 2-1 接口说明

类型	说明
知识库管理接口	知识库管理等相关接口。
结构化数据接口	将结构化数据文件纳入系统进行管理接口。
文件管理接口	用户纳入系统的文档管理接口。
FAQ管理接口	用户自定义的FAQ相关管理接口。
FAQ批量管理接口	用户自定义的批量FAQ相关管理接口。
搜索与问答接口	Koosearch的搜索与模型问答接口。
对话历史接口	用户对话历史接口。
图片管理接口	图片管理相关接口。
模型管理接口	模型管理的相关接口。
用户的文档解析规则定义接口	用户文档解析自定义规则配置接口。

表 2-2 知识库管理接口

API	说明
创建知识库	创建名称为xxx的知识库。

API	说明
修改知识库配置	修改知识库配置。包括： 1. 解析设置：是否使用OCR增强、是否需要解析图片、是否需要解析页眉页脚、是否需要解析目录页。 2. 文档拆分设置： <ul style="list-style-type: none">自动分段长度分段（按文本长度分段）层级分段（按小标题分段，可自定义小标题解析规则） 3. 搜索模型设置：rerank模型配置。 4. NLP模型设置。 5. 其他设置：召回数量、参考文档数量、精排开关、Query改写开关、意图分类。
设置知识库搜索问答prompt	设置知识库搜索问答prompt，该prompt包括参考文档占位符{0}和搜索query占位符{1}。可以根据不同的使用场景，配置不同的prompt，例如增加人设、限定条件、答案格式要求等，帮助模型生成更优的结果。
设置通用问答prompt	设置通用问答prompt。该prompt包含问题占位符{0}，和对回答的要求。
开启知识库	根据ID开启知识库，知识库开启后为可检索状态。
关闭知识库	根据ID关闭知识库，知识库关闭后为不可检索状态。
获取知识库列表	获取当前账号下所有知识库列表。包括知识库ID、知识库名称、知识库状态、创建人、创建时间、更新时间等。
查询知识库详情	根据知识库ID获取当前知识库下的所有文件list，包括文档名称、文档类型、文档大小、文档状态、创建人、创建时间等。
删除知识库	根据ID删除知识库。

表 2-3 结构化数据接口

API	说明
结构化数据上传	将结构化数据文件上传到对象存储服务OBS中进行结构化数据批处理功能。
下载失败的条目	下载一个JSON文件，该文件包含给定上传Field的失败条目。
结构化数据文件搜索	根据文件名搜索目标文件。

表 2-4 文件管理接口

API	说明
上传文件	上传文件至某个知识库。
查询文件	在当前知识库查询名称为xx的文件。
下载文件	下载指定ID的文件。
删除文件	删除指定ID的文件。

表 2-5 FAQ 管理接口

API	说明
创建FAQ	输入问题、相似问题、答案，创建单条FAQ。
更新FAQ	更新指定ID的FAQ的问题、答案、相似问题。
获取FAQ	根据ID获取FAQ详情，包括问题、答案。
查询FAQ列表	获取FAQ列表，包括ID、问题、答案、创建人、创建时间等。可以根据问题关键字过滤出需要查看的FAQ。
删除FAQ	根据知识库ID和FAQ的ID，删除FAQ。
批量删除FAQ	选择多个需要删除的FAQ，根据ID批量删除。

表 2-6 FAQ 批量管理接口

API	说明
批量FAQ上传	上传xlsx/xls文件，用于FAQ的批量导入。
下载上传失败的FAQ	下载给定上传的批量FAQ文件中，上传失败的条目。
列举批量上传faq文件	列举批量上传faq文件。
下载批量导入FAQ原文件	下载批量导入FAQ的原始文件。
删除批量导入FAQ文件	删除指定的FAQ文件。

表 2-7 搜索与问答接口

API	说明
搜索知识库	根据query，在指定知识库中搜索得到topN相关的结果。
生成答案	将content输入给模型，使用模型能力生成答案。

表 2-8 对话历史接口

API	说明
查询对话历史列表	查询历史问题列表。包含对话ID、对话问题、用户名、对话开始时间。
获取对话历史	根据chat_id获取对话历史，返回历史的question和answer数据。
删除对话历史	根据指定ID，删除对应的对话历史。
批量删除对话历史	根据多个ID，批量删除对话历史信息。

表 2-9 图片管理接口

API	说明
上传图片	上传图片入库。
获取图片内容	根据ID获取图片内容。

表 2-10 模型管理接口

API	说明
新增模型	新增配置模型，包括模型名称、模型描述、模型endpoint、模型配置。
修改模型配置	修改模型。
列举模型	列举模型。
删除模型	删除模型。

表 2-11 用户的文档解析规则定义接口

API	说明
创建用户解析规则	列用于创建用户自定义的解析规则，解析规则用于解析需要作为subtitle的文本，规则由上而下依次校验。
修改用户解析规则	用于修改已配置的解析规则。
列举用户规则列表	列举用户配置的所有解析规则列表。
删除用户定义规则解析接口	用于删除用户定义的某种规则。

3 如何调用 API

- 构造请求
- 认证鉴权
- 接口获取
- 返回结果

3.1 构造请求

本节介绍REST API请求的组成，以调用**获取用户Token**接口说明如何调用API，该API获取用户的Token，Token是用户的访问令牌，承载身份与权限信息，Token可以用于调用其他API时鉴权。

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}

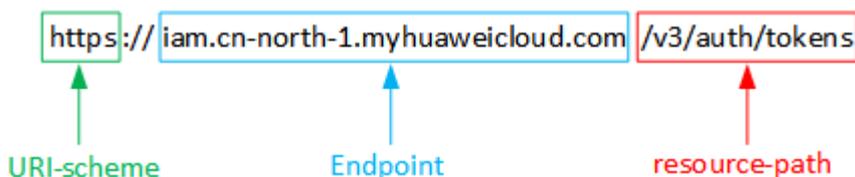
表 3-1 请求 URL

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，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“获取用户Token”API的resource-path为“/v3/auth/tokens”。

例如您需要获取“西南-贵阳一”区域的token，则需使用“西南-贵阳一”区域的Endpoint（iam.cn-southwest-2.myhuaweicloud.com），并在[获取用户Token](#)的URI部分找到resource-path（v3/auth/toknes），拼接起来如下所示。

```
https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens
```

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。

在获取用户Token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/auth/tokens
```

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 公共请求消息头](#)。

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。

参数	是否必选	描述
X-Auth-Token	否（使用Token认证时必选）	用户Token。 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。

对于“获取用户Token”接口，由于不需要认证，所以只添加“Content-Type”即可，添加消息头后的请求如下所示。

```
POST https://iam.xxx.myhuaweicloud.com/v3/auth/tokens
Content-Type: application/json
```

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。

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（即消息体为空），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“获取用户Token”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，加粗的斜体字段需要根据实际值填写，其中 ***username*** 为用户名，*********** 为用户的登录密码，***domainname*** 为用户所属的账号名称，如果是账号本身获取token，***username*** 和 ***domainname*** 填为一致，***xxxxxxxxxxxxxxxxxxxx*** 为project的ID。

说明

scope参数定义了Token的作用范围，表示获取的Token仅能访问指定project下的资源，scope参数的详细说明请参见：[获取用户Token](#)。

```
POST https://iam.cn-southwest-2.myhuaweicloud.com/v3/auth/tokens
{
  "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": {
        "id": "xxxxxxxxxxxxxxxxxxxx" //项目ID
      }
    }
  }
}
```

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。对于“获取用户Token”接口，返回的响应消息头中“x-subject-token”就是需要获取的用户Token。有了Token之后，您就可以使用Token认证调用其他API。

3.2 认证鉴权

KooSearch服务支持通过Token认证进行认证鉴权。

📖 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。

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": "xxxxxxx" //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://{endpoint}/v3/auth/projects
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 接口获取

KooSearch服务接口组成：

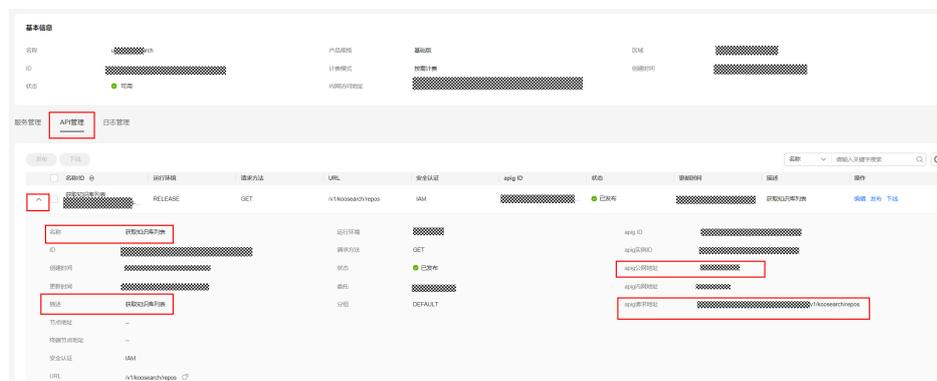
https://{apig公网地址}/{apig请求地址}

服务接口请求地址获取方式:

- 步骤1** 账号登录CSS控制台。
- 步骤2** 左侧导航栏选择：“KooSearch > Flexus 企业搜索服务”。
- 步骤3** 根据“名称/ID”，单击进入已购买的服务。
- 步骤4** 单击“API管理”，即可看到支持的所有API列表。
- 步骤5** 单击展开箭头，即可查看每个API详情，包括“名称”、“描述”、“请求方法”、“apig公网地址”、“apig请求地址”等信息。

----结束

图 3-2 KooSearch-API 管理界面



3.4 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](#)。

对于获取用户Token接口，如果调用后返回状态码为“201”，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，如“Content-type”。

对于获取用户Token接口，返回如[图 获取用户Token响应消息头](#)所示的消息头，其中“x-subject-token”就是需要获取的用户Token。有了Token之后，您就可以使用Token认证调用其他API。

图 3-3 获取用户 Token 响应消息头

```
connection → keep-alive
content-type → application/json
date → Tue, 12 Feb 2019 06:52:13 GMT
server → Web Server
strict-transport-security → max-age=31536000; includeSubdomains;
transfer-encoding → chunked
via → proxy A
x-content-type-options → nosniff
x-download-options → noopen
x-frame-options → SAMEORIGIN
x-iam-trace-id → 218d45ab-d674-4995-af3a-2d0255ba41b5
x-subject-token → MIIYXQVJKoZlhveNAQcCoIIYtJCCEoCAQExDTALBglghkgBZQMEAgEwgharBgkqhkiG9w0BBwGgghacBIIWmHsidG9rZW4iOnsiZmVhcnVzX29udmVhcnVzMTktMDItMTNUMC
fj3KIs6YgKnpVNRbW2eZ5eb78SZOkajACgkIQ1wi4JlGzrpd18LGXK5tdfdq4lqHCYb8P4NaY0NYejcAgzJVeFYtLWT1GSO0zxKZmlQHqJ82HBqHdgZO9fuEbL5dMhdavj+33wEI
xHRCe9I87o+k9-
j+CMZSEB7bUgd5Uj6eRASX11jipPEGA270g1Fruool6jggfFkNPQuF5OU8+uSsttVwRtNfc+qTp22Rkd5McqFGQ8LcuJxC3a+9CM8nOintWW7oeRUVhVpxk8pxiX1wTEboX-
RzT6MUbvpvGw-qPNFYxJECKnoH3HRozv0vN--n5d6Nbxg==
x-xss-protection → 1; mode=block;
```

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。

对于获取用户Token接口，返回如下消息体。为篇幅起见，这里只展示部分内容。

```
{
  "token": {
    "expires_at": "2019-02-13T06:52:13.855000Z",
    "methods": [
      "password"
    ],
    "catalog": [
      {
        "endpoints": [
          {
            "region_id": "xxx",
            .....

```

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 知识库管理

- [创建知识库](#)
- [修改知识库配置](#)
- [设置知识库搜索问答prompt](#)
- [设置知识库通用无搜索的prompt](#)
- [开启知识库](#)
- [关闭知识库](#)
- [获取知识库列表](#)
- [查看知识库列表](#)
- [删除知识库](#)

4.1 创建知识库

功能介绍

创建知识库。

URI

POST /v1/koosearch/repos

请求参数

表 4-1 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 4-2 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	知识库名称 最小长度: 1 最大长度: 64
detail	否	String	知识库描述 最小长度: 1 最大长度: 100
embedding_model	否	String	embedding模型名称 最小长度: 1 最大长度: 32
rerank_model	否	String	rerank模型名称 最小长度: 1 最大长度: 32
pangu_nlp_model	否	String	nlp大模型名称 最小长度: 1 最大长度: 32
file_extract	否	FileExtractConf object	文档解析整体配置

参数	是否必选	参数类型	描述
search_plan_category_ids	否	Array of strings	搜索规划category类型配置 default.category.list[0].id=talk default.category.list[0].category=闲聊类 default.category.list[0].locale=zh default.category.list[1].id=language_task default.category.list[1].category=语言任务类 default.category.list[1].locale=zh default.category.list[2].id=human default.category.list[2].category=人设类 default.category.list[2].locale=zh default.category.list[3].id=common default.category.list[3].category=通用知识类 default.category.list[3].locale=zh default.category.list[4].id=special_knowledge default.category.list[4].category=行业知识类 default.category.list[4].locale=zh 最小长度：0 最大长度：10 数组长度：0 - 10
language_id	否	String	知识库语言ID zh==CHINESE en==ENGLISH th==THAI ar==ARABIC 最小长度：2 最大长度：4
cache_enabled	否	Boolean	是否开启缓存
answer_reference_enabled	否	Boolean	是否开启引用

参数	是否必选	参数类型	描述
answer_image_reference_enabled	否	Boolean	是否开启图文并茂
session_config	否	SessionConfig object	缓存策略
extend_config	否	KnowledgeRepoExtendConfig object	知识库扩展配置
tags	否	Array of TagInfo objects	标签列表 数组长度: 0 - 10
refs	否	String	引用知识库标识列表, 以逗号分隔 最小长度: 0 最大长度: 1024
search_plan_model	否	String	search_plan模型名称 最小长度: 1 最大长度: 32

表 4-3 FileExtractConf

参数	是否必选	参数类型	描述
parse_conf	否	ParseConf object	文档解析配置, 包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	否	SplitConf object	拆分配置, 包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等
id	否	String	文档解析ID 最小长度: 0 最大长度: 64

表 4-4 ParseConf

参数	是否必选	参数类型	描述
ocr_enabled	否	Boolean	ocr增强 缺省值: false
image_enabled	否	Boolean	图片解析 缺省值: false
header_footer_enabled	否	Boolean	解析页眉页脚 缺省值: false
catalog_enabled	否	Boolean	解析目录页 缺省值: false
image_conf	否	String	图片解析开启后 (TEXT 提取图片文本、IMAGE 保留原图) 缺省值: TEXT 枚举值: <ul style="list-style-type: none">• TEXT• IMAGE

表 4-5 SplitConf

参数	是否必选	参数类型	描述
split_mode	否	String	分段设置/层级解析模型 <ul style="list-style-type: none">• LENGTH-长度拆分, 即为字数拆分• CATALOG-层级分段下的自动解析• RULE- 层级分段下的规则解析• AUTO- 自动拆分, 自动识别文档格式匹配适合的拆解析方式 缺省值: AUTO 枚举值: <ul style="list-style-type: none">• LENGTH• CATALOG• RULE• AUTO

参数	是否必选	参数类型	描述
separator_ids	否	Array of strings	<p>分段标识符ID列表</p> <p>文件解析分类</p> <p>default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en</p> <p>default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh</p> <p>default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en</p> <p>default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh</p> <p>default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en</p> <p>default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh</p> <p>default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en</p> <p>default.separator.list[7].id=comma_zh default.separator.list[7].separator=,</p>

参数	是否必选	参数类型	描述
			default.separator.list[7].locale=zh default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en 最小长度：1 最大长度：128 数组长度：1 - 100
rule_regex_id	否	String	用户定义解析规则ID 最小长度：1 最大长度：256
chunk_size	否	Integer	分段预计长度 最小值：0 最大值：6000 缺省值：500
title_level	否	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	否	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	否	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的结果

表 4-6 SessionConfig

参数	是否必选	参数类型	描述
similarity_threshold	是	Float	命中相似度阈值 最小值：0.1 最大值：1.0

参数	是否必选	参数类型	描述
answer_select_policy	是	String	缓存命中选择策略 FIRST: 选择第一个命中 RANDOM: 随机选择 枚举值: <ul style="list-style-type: none">• FIRST• RANDOM
eviction	是	Eviction object	缓存过期策略
model_name	是	String	query2query模型名称 最小长度: 1 最大长度: 32

表 4-7 Eviction

参数	是否必选	参数类型	描述
policy	是	String	缓存过期策略 LRU: (Least Recently Used) now - accessTime > ttl 清除 FIFO: (First In First Out) now - createTime > ttl 清除 LFU: (Least Frequency Used) hit_count < 阈值的清除 最小长度: 1 最大长度: 100 枚举值: <ul style="list-style-type: none">• LRU• FIFO• LFU
ttl	否	Long	缓存过期时间 毫秒 最小值: 0 最大值: 3153600000
hit_count_threshold	否	Long	缓存命中次数阈值 最小值: 1 最大值: 10000

表 4-8 KnowledgeRepoExtendConfig

参数	是否必选	参数类型	描述
extend_context	否	Boolean	是否扩展长上下文。提供更广泛的上下文，以生成完整回答。 缺省值: false
effective_input_length	否	Integer	效果最好的上下文长度，与模型相关，确保输入令牌的有效长度以保证最佳输出 考虑到多轮对话，建议控制在模型上下文长度的60%（向上取整）。 最小值: 2 最大值: 128
top_p	否	Float	通过限制词汇的选择来控制生成文本的多样性。值越高，候选单词越多，文本多样性越高。 最小值: 0.1 最大值: 1 缺省值: 0.1
max_tokens	否	Integer	模型生成最大新词数 最小值: 1 最大值: 131072 缺省值: 131072
chat_temperature	否	Float	非搜索增强模型生成多样性 最小值: 0 最大值: 1
search_temperature	否	Float	搜索增强模型生成多样性 最小值: 0 最大值: 1 缺省值: 0.3
presence_penalty	否	Float	文本重复度 最小值: -2 最大值: 2 缺省值: 0
use_system_prompt	否	Boolean	是否使用系统Prompt，与盘古RAG场景Prompt标准拼接方案保持一致 缺省值: false

参数	是否必选	参数类型	描述
system_prompt	否	String	系统Prompt, 注意: 1.当 use_system_prompt为true时必填 2.不需要拼接query 最小长度: 0 最大长度: 8192

表 4-9 TagInfo

参数	是否必选	参数类型	描述
tag_key	是	String	tag key 最小长度: 1 最大长度: 128
tag_value	是	String	tag value 最小长度: 1 最大长度: 128

响应参数

状态码: 200

表 4-10 响应 Body 参数

参数	参数类型	描述
repo_id	String	知识库ID 最小长度: 1 最大长度: 64

状态码: 400

表 4-11 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-12 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos
{
  "name": "知识库-name",
  "language_id": "zh",
  "detail": "",
  "tags": [],
  "file_extract": {
    "parse_conf": {
      "ocr_enabled": true,
      "image_enabled": false,
      "header_footer_enabled": false,
      "catalog_enabled": false
    },
    "split_conf": {
      "split_mode": "AUTO"
    }
  },
  "extend_config": {
    "extend_context": false,
    "effective_input_length": 3
  },
  "embedding_model": "pangu_embedding_dev",
  "rerank_model": "pangu_rerank_dev",
  "search_plan_model": "search_plan_model",
  "pangu_nlp_model": "pangu-7b",
  "cache_enabled": false,
  "answer_reference_enabled": false,
  "answer_image_reference_enabled": false
}
```

响应示例

状态码：200

知识库ID

```
{  
  "repo_id" : "1235abc"  
}
```

状态码

状态码	描述
200	知识库ID
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.2 修改知识库配置

功能介绍

修改知识库配置。

包括：

- 解析设置：是否使用OCR增强、是否需要解析图片、是否需要解析页眉页脚、是否需要解析目录页
- 文档拆分设置：自动分段、长度分段（按文本长度分段）、层级分段（按小标题分段，可自定义小标题解析规则）
- 搜索模型设置：rerank模型配置
- NLP模型设置：生成模型选择
- 其他设置：召回数量、精排开关、参考文档数量、意图分类、Query改写开关

URI

PUT /v1/koosearch/repos/{repo_id}

表 4-13 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-14 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 4-15 请求 Body 参数

参数	是否必选	参数类型	描述
top_k	否	Integer	top_k配置。top_k指召回前k条与query相关的片段 最小值：10 最大值：500
reference_count	否	Integer	参考文档数量。参考文档指最终和query一起输入给NLP模型，用于生成最终答案的文档 最小值：1 最大值：10

参数	是否必选	参数类型	描述
rerank_enabled	否	Boolean	rerank开关, 包含【打开、关闭】两个状态“打开”状态下, 召回的top_k结果会使用rerank模型进行重排序“关闭”状态下, 召回的top_k结果不会进行重排序
query_rewrite_enabled	否	Boolean	是否使用改写结果搜索
search_plan_category_ids	否	Array of strings	搜索规划category类型配置 default.category.list[0].id=talk default.category.list[0].category=闲聊类 default.category.list[0].locale=zh default.category.list[1].id=language_task default.category.list[1].category=语言任务类 default.category.list[1].locale=zh default.category.list[2].id=human default.category.list[2].category=人设类 default.category.list[2].locale=zh default.category.list[3].id=common default.category.list[3].category=通用知识类 default.category.list[3].locale=zh default.category.list[4].id=special_knowledge default.category.list[4].category=行业知识类 default.category.list[4].locale=zh 最小长度: 0 最大长度: 10 数组长度: 0 - 10
file_extract	否	FileExtract object	文档解析整体配置, 包含文档解析使用哪些组件、文档拆分规则

参数	是否必选	参数类型	描述
rerank_model	否	String	rerank模型名称 最小长度：1 最大长度：32
pangu_nlp_model	否	String	nlp大模型名称 最小长度：1 最大长度：32
search_threshold	否	Float	搜索接口过滤的阈值，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 最小值：0 最大值：200
chat_ref_threshold	否	Float	参考文档过滤的阈值，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 最小值：0 最大值：200
faq_threshold	否	Float	FAQ直出阈值：超过阈值的FAQ会作为答案直接输出，不需要经过大模型总结 注意事项：1.参数值小于等于0时，表示FAQ不直出 2.旧版本query2doc，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 3.新版本query2query，阈值为0-1 最小值：0 最大值：200
cache_enabled	否	Boolean	是否开启缓存
session_config	否	SessionConfig object	缓存策略
answer_reference_enabled	否	Boolean	是否开启引用
answer_image_reference_enabled	否	Boolean	是否开启图文并茂
extend_config	否	KnowledgeRepoExtendConfig object	知识库扩展配置

参数	是否必选	参数类型	描述
tags	否	Array of TagInfo objects	标签列表 数组长度：0 - 10
refs	否	String	引用知识库标识列表，以逗号分隔 最小长度：0 最大长度：1024
name	否	String	知识库名称 最小长度：1 最大长度：64
search_plan_model	否	String	search_plan模型名称 最小长度：1 最大长度：32

表 4-16 FileExtract

参数	是否必选	参数类型	描述
parse_conf	否	ParseConf object	文档解析配置，包含是否使用 OCR 增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	否	SplitConf object	拆分配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern 配置等

表 4-17 ParseConf

参数	是否必选	参数类型	描述
ocr_enabled	否	Boolean	ocr 增强 缺省值：false
image_enabled	否	Boolean	图片解析 缺省值：false
header_footer_enabled	否	Boolean	解析页眉页脚 缺省值：false

参数	是否必选	参数类型	描述
catalog_enabled	否	Boolean	解析目录页 缺省值: false
image_conf	否	String	图片解析开启后 (TEXT 提取图片文本、IMAGE 保留原图) 缺省值: TEXT 枚举值: <ul style="list-style-type: none">• TEXT• IMAGE

表 4-18 SplitConf

参数	是否必选	参数类型	描述
split_mode	否	String	分段设置/层级解析模型 <ul style="list-style-type: none">• LENGTH-长度拆分, 即为字数拆分• CATALOG-层级分段下的自动解析• RULE- 层级分段下的规则解析• AUTO- 自动拆分, 自动识别文档格式匹配适合的拆分析方式 缺省值: AUTO 枚举值: <ul style="list-style-type: none">• LENGTH• CATALOG• RULE• AUTO

参数	是否必选	参数类型	描述
separator_ids	否	Array of strings	<p>分段标识符ID列表</p> <p>文件解析分类</p> <p>default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en</p> <p>default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh</p> <p>default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en</p> <p>default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh</p> <p>default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en</p> <p>default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh</p> <p>default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en</p> <p>default.separator.list[7].id=comma_zh default.separator.list[7].separator=,</p>

参数	是否必选	参数类型	描述
			default.separator.list[7].locale=zh default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en 最小长度：1 最大长度：128 数组长度：1 - 100
rule_regex_id	否	String	用户定义解析规则ID 最小长度：1 最大长度：256
chunk_size	否	Integer	分段预计长度 最小值：0 最大值：6000 缺省值：500
title_level	否	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	否	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	否	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的结果

表 4-19 SessionConfig

参数	是否必选	参数类型	描述
similarity_threshold	是	Float	命中相似度阈值 最小值：0.1 最大值：1.0

参数	是否必选	参数类型	描述
answer_select_policy	是	String	缓存命中选择策略 FIRST: 选择第一个命中 RANDOM: 随机选择 枚举值: <ul style="list-style-type: none">• FIRST• RANDOM
eviction	是	Eviction object	缓存过期策略
model_name	是	String	query2query模型名称 最小长度: 1 最大长度: 32

表 4-20 Eviction

参数	是否必选	参数类型	描述
policy	是	String	缓存过期策略 LRU: (Least Recently Used) now - accessTime > ttl 清除 FIFO: (First In First Out) now - createTime > ttl 清除 LFU: (Least Frequency Used) hit_count < 阈值的清除 最小长度: 1 最大长度: 100 枚举值: <ul style="list-style-type: none">• LRU• FIFO• LFU
ttl	否	Long	缓存过期时间 毫秒 最小值: 0 最大值: 3153600000
hit_count_threshold	否	Long	缓存命中次数阈值 最小值: 1 最大值: 10000

表 4-21 KnowledgeRepoExtendConfig

参数	是否必选	参数类型	描述
extend_context	否	Boolean	是否扩展长上下文。提供更广泛的上下文，以生成完整回答。 缺省值: false
effective_input_length	否	Integer	效果最好的上下文长度，与模型相关，确保输入令牌的有效长度以保证最佳输出 考虑到多轮对话，建议控制在模型上下文长度的60%（向上取整）。 最小值: 2 最大值: 128
top_p	否	Float	通过限制词汇的选择来控制生成文本的多样性。值越高，候选单词越多，文本多样性越高。 最小值: 0.1 最大值: 1 缺省值: 0.1
max_tokens	否	Integer	模型生成最大新词数 最小值: 1 最大值: 131072 缺省值: 131072
chat_temperature	否	Float	非搜索增强模型生成多样性 最小值: 0 最大值: 1
search_temperature	否	Float	搜索增强模型生成多样性 最小值: 0 最大值: 1 缺省值: 0.3
presence_penalty	否	Float	文本重复度 最小值: -2 最大值: 2 缺省值: 0
use_system_prompt	否	Boolean	是否使用系统Prompt，与盘古RAG场景Prompt标准拼接方案保持一致 缺省值: false

参数	是否必选	参数类型	描述
system_prompt	否	String	系统Prompt, 注意: 1.当 use_system_prompt为true时必填 2.不需要拼接query 最小长度: 0 最大长度: 8192

表 4-22 TagInfo

参数	是否必选	参数类型	描述
tag_key	是	String	tag key 最小长度: 1 最大长度: 128
tag_value	是	String	tag value 最小长度: 1 最大长度: 128

响应参数

状态码: 200

表 4-23 响应 Body 参数

参数	参数类型	描述
repo_id	String	知识库ID 最小长度: 1 最大长度: 64

状态码: 400

表 4-24 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-25 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos/v22cca08-a83b-4518-8abf-0666ab5d635a
```

```
{
  "rerank_model": "string",
  "pangu_nlp_model": "string",
  "file_extract": {
    "parse_conf": {
      "ocr_enabled": false,
      "image_enabled": false,
      "header_footer_enabled": false,
      "catalog_enabled": false,
      "image_conf": "TEXT"
    },
    "split_conf": {
      "split_mode": "AUTO",
      "separator_ids": [ ],
      "rule_regex_id": "string",
      "chunk_size": 6000,
      "title_level": 3,
      "combine_title": true
    }
  },
  "search_plan_category_ids": [ "string" ],
  "rerank_enabled": true,
  "top_k": 500,
  "reference_count": 10,
  "query_rewrite_enabled": true
}
```

响应示例

状态码：200

知识库ID

```
{  
  "repo_id" : "1235abc"  
}
```

状态码

状态码	描述
200	知识库ID
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.3 设置知识库搜索问答 prompt

功能介绍

设置知识库搜索问答prompt，该prompt包括参考文档占位符{0}和搜索query占位符{1}。可以根据不同的使用场景，配置不同的prompt，例如增加人设、限定条件、答案格式要求等，帮助模型生成更优的结果。

URI

POST /v1/koosearch/repos/{repo_id}/search-prompt

表 4-26 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-27 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 4-28 请求 Body 参数

参数	是否必选	参数类型	描述
prompt	是	String	具体的值 最小长度: 1 最大长度: 8192

响应参数

状态码: 200

表 4-29 响应 Body 参数

参数	参数类型	描述
repo_id	String	响应字段 最小长度: 1 最大长度: 64

状态码: 400

表 4-30 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-31 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

配置示例

```
/v1/koosearch/repos/919e8b10-b0c4-4bcd-9d5a-8ebb69594514/search-prompt
{
  "prompt": "你是一名问题回答助手，请参考以下给出的文章回答问题。在回答问题之前，请你先仔细判给出的文章是否可以回答问题。如果给出的文章与问题不相关，则直接输出“抱歉，我无法回答此次问题”；如果给出的文章与问题相关，则参考给出的文章回答问题。\\n给出的文章： \\n{0}\\n问题： {1}\\n请你按照要求进行仔细判断后，给出你的回答：}"
}
```

响应示例

状态码：200

设置的搜索增强prompt的知识库ID

```
{
  "repo_id": "919e8b10-b0c4-4bcd-9d5a-8ebb69594514"
}
```

状态码

状态码	描述
200	设置的搜索增强prompt的知识库ID
400	请求体参数错误

状态码	描述
500	内部异常

错误码

请参见[错误码](#)。

4.4 设置知识库通用无搜索的 prompt

功能介绍

设置通用问答prompt。该prompt包含问题占位符{0}，和对回答的要求。

URI

POST /v1/koosearch/repos/{repo_id}/common-prompt

表 4-32 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-33 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 4-34 请求 Body 参数

参数	是否必选	参数类型	描述
prompt	是	String	无搜索的prompt 最小长度：1 最大长度：8192

响应参数

状态码：200

表 4-35 响应 Body 参数

参数	参数类型	描述
repo_id	String	响应字段 最小长度：1 最大长度：64

状态码：400

表 4-36 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-37 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

配置示例

```
/v1/koosearch/repos/919e8b10-b0c4-4bcd-9d5a-8ebb69594514/common-prompt
```

```
{  
  "prompt": "你是一名问题回答助手，请参考以下给出的文章回答问题。在回答问题之前，请你先仔细判给出的文章是否可以回答问题。如果给出的文章与问题不相关，则直接输出“抱歉，我无法回答此次问题”；如果给出的文章与问题相关，则参考给出的文章回答问题。\\n给出的文章： \\n{0}\\n问题： {1}\\n请你按照要求进行仔细判断后，给出你的回答：}"  
}
```

响应示例

状态码： 200

设置通用prompt的知识库ID

```
{  
  "repo_id": "919e8b10-b0c4-4bcd-9d5a-8ebb69594514"  
}
```

状态码

状态码	描述
200	设置通用prompt的知识库ID
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.5 开启知识库

功能介绍

根据ID开启知识库，知识库开启后为可检索状态。

URI

```
PUT /v1/koosearch/repos/{repo_id}/enable
```

表 4-38 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-39 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码：400

表 4-40 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-41 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

无

响应示例

无

状态码

状态码	描述
200	开启知识库成功
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.6 关闭知识库

功能介绍

根据ID关闭知识库，知识库关闭后为不可检索状态。

URI

PUT /v1/koosearch/repos/{repo_id}/disable

表 4-42 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-43 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码：400

表 4-44 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-45 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

无

响应示例

状态码：400

请求体参数错误

```
{  
  "error_code": "string",  
  "error_message": "string"  
}
```

状态码

状态码	描述
200	关闭知识库成功
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.7 获取知识库列表

功能介绍

获取当前账号下所有知识库列表。包括知识库ID、知识库名称、知识库状态、创建人、创建时间、更新时间等。

URI

GET /v1/koosearch/repos

表 4-46 Query 参数

参数	是否必选	参数类型	描述
name	否	String	知识库名称 最小长度：1 最大长度：64
status	否	String	状态(open-开启, close-关闭) 最小长度：1 最大长度：64
page_num	否	Integer	请求页码 最小值：1 最大值：65535 缺省值：1
page_size	是	Integer	请求限定响应结果的分页大小， 例如5条/页, 10条/页 最小值：1 最大值：65535 缺省值：10
tag	否	String	tag信息, 由英文冒号连接key和 value构成, 如: key1:value1 最小长度：0 最大长度：65535
lod	否	String	level of detail, 返回结果详细程 度: simple: 简单 detail: 详细 缺省值: detail 最小长度：1 最大长度：64 枚举值: • simple • detail

请求参数

表 4-47 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

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_list	Array of KnowledgeRepoListInfo objects	知识库列表 数组长度: 1 - 65535
total	Integer	总数 最小值: 1 最大值: 65535
region_ocr_enabled	Boolean	ocr开关
region_rac_enabled	Boolean	rac开关

表 4-49 KnowledgeRepoListInfo

参数	参数类型	描述
id	String	知识库ID 最小长度: 1 最大长度: 100
name	String	知识库名称 最小长度: 1 最大长度: 100

参数	参数类型	描述
detail	String	描述 最小长度: 1 最大长度: 100
status	String	状态 最小长度: 1 最大长度: 100
create_user	String	用户 最小长度: 1 最大长度: 100
create_time	String	创建时间 最小长度: 1 最大长度: 100
update_time	String	更新时间 最小长度: 1 最大长度: 100
top_k	Integer	topK数量 最小值: 10 最大值: 500
prompt	String	prompt 最小长度: 0 最大长度: 8192
common_prompt	String	通用prompt 最小长度: 0 最大长度: 8192
rerank_enabled	Boolean	rerank开关
moderate_enabled	Boolean	内容审核开关
search_plan_enabled	Boolean	搜索规划开关
query_rewrite_enabled	Boolean	改写开关
reference_count	Integer	参考文档数量。参考文档指最终和query一起输入给NLP模型, 用于生成最终答案的文档 最小值: 1 最大值: 10

参数	参数类型	描述
fields	Array of KnowledgeRepoFieldSchema objects	字段描述 数组长度: 1 - 100
search_threshold	Float	搜索接口过滤的阈值, 精排关闭时, 阈值为0-200, 精排打开时, 阈值为0-1 最小值: 0 最大值: 200
chat_ref_threshold	Float	参考文档过滤的阈值, 精排关闭时, 阈值为0-200, 精排打开时, 阈值为0-1 最小值: 0 最大值: 200
faq_threshold	Float	FAQ直出阈值: 超过阈值的FAQ会作为答案直接输出, 不需要经过大模型总结 注意事项: 1.参数值小于等于0时, 表示FAQ不直出 2.旧版本query2doc, 精排关闭时, 阈值为0-200, 精排打开时, 阈值为0-1 3.新版本query2query, 阈值为0-1 最小值: 0 最大值: 200
embedding_model	String	embedding模型名称 最小长度: 1 最大长度: 64
rerank_model	String	rerank模型名称 最小长度: 1 最大长度: 64
nlp_model	String	nlp模型名称 最小长度: 1 最大长度: 64
file_extract	FileExtract object	文档解析详情

参数	参数类型	描述
search_plan_category_ids	Array of strings	搜索规划category类型配置 default.category.list[0].id=talk default.category.list[0].category=闲聊类 default.category.list[0].locale=zh default.category.list[1].id=language_task default.category.list[1].category=语言任务类 default.category.list[1].locale=zh default.category.list[2].id=human default.category.list[2].category=人设类 default.category.list[2].locale=zh default.category.list[3].id=common default.category.list[3].category=通用知识类 default.category.list[3].locale=zh default.category.list[4].id=special_knowledge default.category.list[4].category=行业知识类 default.category.list[4].locale=zh 最小长度：0 最大长度：10 数组长度：0 - 10
language_id	String	知识库语言ID 最小长度：2 最大长度：4
cache_enabled	Boolean	是否开启缓存
session_config	SessionConfig object	缓存策略
answer_reference_enabled	Boolean	是否开启引用
answer_image_reference_enabled	Boolean	是否开启图文并茂
extend_config	KnowledgeRepoExtendConfig object	知识库扩展配置
refs	String	引用知识库标识列表，以逗号分隔 最小长度：0 最大长度：1024

表 4-50 KnowledgeRepoFieldSchema

参数	参数类型	描述
name	String	字段名 最小长度：1 最大长度：100
field_type	String	字段类型 最小长度：1 最大长度：100
name_zh	String	字段中文名 最小长度：1 最大长度：100

表 4-51 FileExtract

参数	参数类型	描述
parse_conf	ParseConf object	文档解析配置，包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	SplitConf object	拆配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等

表 4-52 ParseConf

参数	参数类型	描述
ocr_enabled	Boolean	ocr增强 缺省值： false
image_enabled	Boolean	图片解析 缺省值： false
header_footer_enabled	Boolean	解析页眉页脚 缺省值： false
catalog_enabled	Boolean	解析目录页 缺省值： false

参数	参数类型	描述
image_conf	String	图片解析开启后（TEXT 提取图片文本、IMAGE 保留原图） 缺省值：TEXT 枚举值： <ul style="list-style-type: none">• TEXT• IMAGE

表 4-53 SplitConf

参数	参数类型	描述
split_mode	String	分段设置/层级解析模型 <ul style="list-style-type: none">• LENGTH-长度拆分，即为字数拆分• CATALOG-层级分段下的自动解析• RULE- 层级分段下的规则解析• AUTO- 自动拆分，自动识别文档格式匹配适合的拆分析方式 缺省值：AUTO 枚举值： <ul style="list-style-type: none">• LENGTH• CATALOG• RULE• AUTO

参数	参数类型	描述
separator_ids	Array of strings	分段标识符ID列表 文件解析分类 default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en default.separator.list[7].id=comma_zh default.separator.list[7].separator=, default.separator.list[7].locale=zh default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en 最小长度：1 最大长度：128 数组长度：1 - 100
rule_regex_id	String	用户定义解析规则ID 最小长度：1 最大长度：256
chunk_size	Integer	分段预计长度 最小值：0 最大值：6000 缺省值：500

参数	参数类型	描述
title_level	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的结果

表 4-54 SessionConfig

参数	参数类型	描述
similarity_threshold	Float	命中相似度阈值 最小值：0.1 最大值：1.0
answer_select_policy	String	缓存命中选择策略 FIRST：选择第一个命中 RANDOM：随机选择 枚举值： <ul style="list-style-type: none">• FIRST• RANDOM
eviction	Eviction object	缓存过期策略
model_name	String	query2query模型名称 最小长度：1 最大长度：32

表 4-55 Eviction

参数	参数类型	描述
policy	String	缓存过期策略 LRU: (Least Recently Used) now - accessTime > ttl 清除 FIFO: (First In First Out) now - createTime > ttl 清除 LFU: (Least Frequency Used) hit_count < 阈值的清除 最小长度: 1 最大长度: 100 枚举值: <ul style="list-style-type: none">• LRU• FIFO• LFU
ttl	Long	缓存过期时间 毫秒 最小值: 0 最大值: 31536000000
hit_count_threshold	Long	缓存命中次数阈值 最小值: 1 最大值: 10000

表 4-56 KnowledgeRepoExtendConfig

参数	参数类型	描述
extend_context	Boolean	是否扩展长上下文。提供更广泛的上下文，以生成完整回答。 缺省值: false
effective_input_length	Integer	效果最好的上下文长度，与模型相关，确保输入令牌的有效长度以保证最佳输出 考虑到多轮对话，建议控制在模型上下文长度的60%（向上取整）。 最小值: 2 最大值: 128
top_p	Float	通过限制词汇的选择来控制生成文本的多样性。值越高，候选单词越多，文本多样性越高。 最小值: 0.1 最大值: 1 缺省值: 0.1

参数	参数类型	描述
max_tokens	Integer	模型生成最大新词数 最小值：1 最大值：131072 缺省值：131072
chat_temperature	Float	非搜索增强模型生成多样性 最小值：0 最大值：1
search_temperature	Float	搜索增强模型生成多样性 最小值：0 最大值：1 缺省值：0.3
presence_penalty	Float	文本重复度 最小值：-2 最大值：2 缺省值：0
use_system_prompt	Boolean	是否使用系统Prompt，与盘古RAG场景Prompt标准拼接方案保持一致 缺省值：false
system_prompt	String	系统Prompt，注意：1.当 use_system_prompt 为true时必须填 2.不需要拼接query 最小长度：0 最大长度：8192

请求示例

```
/v1/koosearch/repos?page_num=1&page_size=10&name=知识&status=open
```

响应示例

状态码：200

知识库列表响应体

```
{  
  "data_list": [{  
    "id": "12369797",  
    "name": "知识库1",  
    "detail": "金融知识库",  
    "status": "OPEN",  
    "create_user": "",  
    "create_time": "1235356",  
    "update_time": "1235356",  
    "fields": [{  
      "name": "",  
      "field_type": "",  
      "name_zh": ""  
    }  
  ]  
}
```

```
    }  
  }  
  "total": 10  
}
```

状态码

状态码	描述
200	知识库列表响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.8 查看知识库列表

功能介绍

查看知识库列表，包括知识库ID、知识库名称、知识库状态、创建人、创建时间、更新时间等。

URI

GET /v1/koosearch/repos/{repo_id}

表 4-57 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 4-58 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 4-59 响应 Body 参数

参数	参数类型	描述
id	String	知识库ID 最小长度: 1 最大长度: 100
name	String	知识库名称 最小长度: 1 最大长度: 100
detail	String	描述 最小长度: 1 最大长度: 100
status	String	状态 最小长度: 1 最大长度: 100
create_user	String	用户 最小长度: 1 最大长度: 100
create_time	String	创建时间 最小长度: 1 最大长度: 100
update_time	String	更新时间 最小长度: 1 最大长度: 100

参数	参数类型	描述
top_k	Integer	topK数量 最小值：10 最大值：500
prompt	String	prompt 最小长度：0 最大长度：8192
common_prompt	String	通用prompt 最小长度：0 最大长度：8192
rerank_enabled	Boolean	rerank开关
moderate_enabled	Boolean	内容审核开关
search_plan_enabled	Boolean	搜索规划开关
query_rewrite_enabled	Boolean	改写开关
reference_count	Integer	参考文档数量。参考文档指最终和query一起输入给NLP模型，用于生成最终答案的文档 最小值：1 最大值：10
fields	Array of KnowledgeRepoFieldSchema objects	字段描述 数组长度：1 - 100
search_threshold	Float	搜索接口过滤的阈值，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 最小值：0 最大值：200
chat_ref_threshold	Float	参考文档过滤的阈值，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 最小值：0 最大值：200

参数	参数类型	描述
faq_threshold	Float	FAQ直出阈值：超过阈值的FAQ会作为答案直接输出，不需要经过大模型总结 注意事项：1.参数值小于等于0时，表示FAQ不直出 2.旧版本query2doc，精排关闭时，阈值为0-200，精排打开时，阈值为0-1 3.新版本query2query，阈值为0-1 最小值： 0 最大值： 200
embedding_model	String	embedding模型名称 最小长度： 1 最大长度： 64
rerank_model	String	rerank模型名称 最小长度： 1 最大长度： 64
nlp_model	String	nlp模型名称 最小长度： 1 最大长度： 64
file_extract	FileExtract object	文档解析详情
search_plan_category_ids	Array of strings	搜索规划category类型配置 default.category.list[0].id=talk default.category.list[0].category=闲聊类 default.category.list[0].locale=zh default.category.list[1].id=language_task default.category.list[1].category=语言任务类 default.category.list[1].locale=zh default.category.list[2].id=human default.category.list[2].category=人设类 default.category.list[2].locale=zh default.category.list[3].id=common default.category.list[3].category=通用知识类 default.category.list[3].locale=zh default.category.list[4].id=special_knowledge default.category.list[4].category=行业知识类 default.category.list[4].locale=zh 最小长度： 0 最大长度： 10 数组长度： 0 - 10
language_id	String	知识库语言ID 最小长度： 2 最大长度： 4

参数	参数类型	描述
cache_enabled	Boolean	是否开启缓存
session_config	SessionConfig object	缓存策略
answer_reference_enabled	Boolean	是否开启引用
answer_image_reference_enabled	Boolean	是否开启图文并茂
extend_config	KnowledgeRepoExtendConfig object	知识库扩展配置
tags	Array of TagInfo objects	tag信息 数组长度: 0 - 10
refs	String	引用知识库标识列表, 以逗号分隔 最小长度: 0 最大长度: 1024
search_plan_model	String	search_plan模型名称 最小长度: 1 最大长度: 32

表 4-60 KnowledgeRepoFieldSchema

参数	参数类型	描述
name	String	字段名 最小长度: 1 最大长度: 100
field_type	String	字段类型 最小长度: 1 最大长度: 100
name_zh	String	字段中文名 最小长度: 1 最大长度: 100

表 4-61 FileExtract

参数	参数类型	描述
parse_conf	ParseConf object	文档解析配置，包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	SplitConf object	拆分配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等

表 4-62 ParseConf

参数	参数类型	描述
ocr_enabled	Boolean	ocr增强 缺省值： false
image_enabled	Boolean	图片解析 缺省值： false
header_footer_enabled	Boolean	解析页眉页脚 缺省值： false
catalog_enabled	Boolean	解析目录页 缺省值： false
image_conf	String	图片解析开启后（TEXT 提取图片文本、IMAGE 保留原图） 缺省值： TEXT 枚举值： <ul style="list-style-type: none">• TEXT• IMAGE

表 4-63 SplitConf

参数	参数类型	描述
split_mode	String	分段设置/层级解析模型 <ul style="list-style-type: none">● LENGTH-长度拆分，即为字数拆分● CATALOG-层级分段下的自动解析● RULE- 层级分段下的规则解析● AUTO- 自动拆分，自动识别文档格式匹配适合的拆分解析方式 缺省值： AUTO 枚举值： <ul style="list-style-type: none">● LENGTH● CATALOG● RULE● AUTO

参数	参数类型	描述
separator_ids	Array of strings	分段标识符ID列表 文件解析分类 default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en default.separator.list[7].id=comma_zh default.separator.list[7].separator=, default.separator.list[7].locale=zh default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en 最小长度：1 最大长度：128 数组长度：1 - 100
rule_regex_id	String	用户定义解析规则ID 最小长度：1 最大长度：256
chunk_size	Integer	分段预计长度 最小值：0 最大值：6000 缺省值：500

参数	参数类型	描述
title_level	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的结果

表 4-64 SessionConfig

参数	参数类型	描述
similarity_threshold	Float	命中相似度阈值 最小值：0.1 最大值：1.0
answer_select_policy	String	缓存命中选择策略 FIRST：选择第一个命中 RANDOM：随机选择 枚举值： <ul style="list-style-type: none">• FIRST• RANDOM
eviction	Eviction object	缓存过期策略
model_name	String	query2query模型名称 最小长度：1 最大长度：32

表 4-65 Eviction

参数	参数类型	描述
policy	String	缓存过期策略 LRU: (Least Recently Used) now - accessTime > ttl 清除 FIFO: (First In First Out) now - createTime > ttl 清除 LFU: (Least Frequency Used) hit_count < 阈值的清除 最小长度: 1 最大长度: 100 枚举值: <ul style="list-style-type: none">• LRU• FIFO• LFU
ttl	Long	缓存过期时间 毫秒 最小值: 0 最大值: 31536000000
hit_count_threshold	Long	缓存命中次数阈值 最小值: 1 最大值: 10000

表 4-66 KnowledgeRepoExtendConfig

参数	参数类型	描述
extend_context	Boolean	是否扩展长上下文。提供更广泛的上下文，以生成完整回答。 缺省值: false
effective_input_length	Integer	效果最好的上下文长度，与模型相关，确保输入令牌的有效长度以保证最佳输出 考虑到多轮对话，建议控制在模型上下文长度的60%（向上取整）。 最小值: 2 最大值: 128
top_p	Float	通过限制词汇的选择来控制生成文本的多样性。值越高，候选单词越多，文本多样性越高。 最小值: 0.1 最大值: 1 缺省值: 0.1

参数	参数类型	描述
max_tokens	Integer	模型生成最大新词数 最小值：1 最大值：131072 缺省值：131072
chat_temperature	Float	非搜索增强模型生成多样性 最小值：0 最大值：1
search_temperature	Float	搜索增强模型生成多样性 最小值：0 最大值：1 缺省值：0.3
presence_penalty	Float	文本重复度 最小值：-2 最大值：2 缺省值：0
use_system_prompt	Boolean	是否使用系统Prompt，与盘古RAG场景Prompt标准拼接方案保持一致 缺省值：false
system_prompt	String	系统Prompt，注意：1.当 use_system_prompt 为true时必须填 2.不需要拼接query 最小长度：0 最大长度：8192

表 4-67 TagInfo

参数	参数类型	描述
tag_key	String	tag key 最小长度：1 最大长度：128
tag_value	String	tag value 最小长度：1 最大长度：128

状态码：400

表 4-68 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 4-69 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos/0890bc5b-d57e-4199-82ce-a629c1b5926c
```

响应示例

状态码：200

知识库详情响应体

```
{  
  "id": "1234545",  
  "name": "知识库123",  
  "detail": "金融知识库",  
  "status": "OPEN",  
  "create_user": "",  
  "create_time": "",  
  "update_time": "1693921595857",  
  "fields": [{  
    "name": "id",  
    "field_type": "keyword",  
    "name_zh": "id"  
  }]  
}
```

状态码

状态码	描述
200	知识库详情响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

4.9 删除知识库

功能介绍

根据ID删除知识库。

URI

DELETE /v1/koosearch/repos

请求参数

表 4-70 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 4-71 请求 Body 参数

参数	是否必选	参数类型	描述
repo_ids	是	Array of strings	知识库ID数组 最小长度：1 最大长度：64 数组长度：1 - 65535

响应参数

状态码： 200

表 4-72 响应 Body 参数

参数	参数类型	描述
total_count	Integer	预期删除数量 最小值： 0 最大值： 65535
deleted_count	Integer	实际删除数量 最小值： 0 最大值： 65535

状态码： 400

表 4-73 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

状态码： 500

表 4-74 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

请求示例

```
/v1/koosearch/repos  
{  
  "repo_ids": [ "123235343", "abc" ]  
}
```

响应示例

状态码: 200

删除信息响应体

```
{  
  "total_count": 4,  
  "deleted_count": 0  
}
```

状态码

状态码	描述
200	删除信息响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

5 结构化数据

[上传结构化数据](#)

[下载失败的条目](#)

[搜索结构化数据文件](#)

5.1 上传结构化数据

功能介绍

将结构化数据文件上传到对象存储服务OBS中进行结构化数据批处理功能。

URI

POST /v1/koosearch/repos/{repo_id}/structured-data

表 5-1 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 5-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 5-3 FormData 参数

参数	是否必选	参数类型	描述
file	是	File	仅支持JSON文件上传

响应参数

状态码: 200

表 5-4 响应 Body 参数

参数	参数类型	描述
file_id	String	文件ID 最小长度: 1 最大长度: 64

状态码: 400

表 5-5 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

状态码： 500

表 5-6 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

请求示例

```
/v1/koosearch/repos/36b6d979-7f98-4fda-b8b5-d7d0cc95d296/structured-data  
file
```

响应示例

状态码： 200

上传文档接口响应体

```
{  
  "file_id": "60fa0fbf-a809-4e4e-8d26-ec389a497bf4"  
}
```

状态码

状态码	描述
200	上传文档接口响应体
400	参数错误
500	内部异常

错误码

请参见[错误码](#)。

5.2 下载失败的条目

功能介绍

下载一个JSON文件，该文件包含给定上传Field的失败条目。

URI

GET /v1/koosearch/repos/{repo_id}/structured-data/{file_id}/fails

表 5-7 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64
file_id	是	String	文件ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 最小长度: 1 最大长度: 64

请求参数

表 5-8 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 5-9 响应 Body 参数

参数	参数类型	描述
-	File	执行成功

状态码： 400

表 5-10 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

状态码： 500

表 5-11 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

请求示例

```
/v1/koosearch/repos/043c4673-264a-48fe-b199-a6320cbc393e/structured-data/8756cc16-4f0b-4c33-a97e-c0b34debdbdb/fails
```

响应示例

无

状态码

状态码	描述
200	执行成功
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

5.3 搜索结构化数据文件

功能介绍

根据文件名搜索目标文件。

URI

GET /v1/koosearch/repos/{repo_id}/structured-data/search

表 5-12 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

表 5-13 Query 参数

参数	是否必选	参数类型	描述
file_name	是	String	结构化数据的文件名 最小长度：1 最大长度：1024

参数	是否必选	参数类型	描述
file_status	是	String	文件状态 SUCCESS - 上传成功 ERROR - 上传失败 PENDING - 排队解析中 RUNNING - 正在解析 IMPORT_EXCEPTION - 导入异常 FILE_ENCODING_ERROR - 文章解码异常 最小长度：0 最大长度：128 枚举值： <ul style="list-style-type: none">• SUCCESS• ERROR• PENDING• RUNNING• IMPORT_EXCEPTION• FILE_ENCODING_ERROR
page_num	否	Integer	请求页码 最小值：1 最大值：65535 缺省值：1
page_size	否	Integer	请求限定响应结果的分页大小，例如5条/页，10条/页 最小值：1 最大值：65535 缺省值：10
ids	否	Array	精准查询文件id列表 数组长度：1 - 64

请求参数

表 5-14 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码： 200

表 5-15 响应 Body 参数

参数	参数类型	描述
total	Integer	返回的文档总数 最小值： 1 最大值： 65535
page_num	Integer	页码 最小值： 1 最大值： 65535
page_size	Integer	每页数量 最小值： 1 最大值： 65535
files	Array of FileInfo objects	文件列表 数组长度： 0 - 65535

表 5-16 FileInfo

参数	参数类型	描述
id	String	文件ID 最小长度： 1 最大长度： 64
task_id	String	任务id 最小长度： 0 最大长度： 64
name	String	文件名 最小长度： 1 最大长度： 255
repo_id	String	知识库id 最小长度： 0 最大长度： 64
project_id	String	项目id 最小长度： 1 最大长度： 64

参数	参数类型	描述
application_id	String	应用id 最小长度: 1 最大长度: 64
status	String	文件状态 <ul style="list-style-type: none"> • SUCCESS 执行成功 • ERROR 执行失败 • PENDING 初始状态,未处理 • RUNNING 执行中 • IMPORT_EXCEPTION 导入异常 • FILE_ENCODING_ERROR 编码错误 最小长度: 0 最大长度: 64 枚举值: <ul style="list-style-type: none"> • SUCCESS • ERROR • PENDING • RUNNING • IMPORT_EXCEPTION • FILE_ENCODING_ERROR
type	String	文件类型 最小长度: 1 最大长度: 64
size	Long	文件大小 (bytes) 最小值: 1 最大值: 99999999
category	String	文档目录, 单值, 对应目录树中的叶子节点。建议格式为 叶子节点目录名称(目录 ID), 如: 专利 (3166-1)。 最小长度: 0 最大长度: 255
create_user	String	创建者, 指上传文件的用户, 可能不存在该键值 最小长度: 1 最大长度: 64
create_time	String	创建时间, 指上传文件的时间, 例如: 1692848139119 最小长度: 1 最大长度: 13

参数	参数类型	描述
update_time	String	更新时间, 可能不存在该键值 最小长度: 1 最大长度: 13
file_path	String	文件地址, 可能不存在该键值 最小长度: 0 最大长度: 65535
upload_desc	String	上传说明, 可能不存在该键值 最小长度: 0 最大长度: 65535
file_extract_conf	FileExtractConf object	文件抽取配置项
tags	Array of strings	文档标签, 可以使用标签对文档进行自动分组, 用于筛选过滤。 注意事项: 1.建议不区分大小写, 例如, Approved and approved 是同一个标签 2.文档可以有一个或多个标签 3.建议格式为: 标签名称:标签值, 如果能够保证标签值不冲突, 也可以直接使用 标签值, 例如 《冰箱使用说明书》对应的产品型号: ProductModel:BCD-551WLCTDAFA5U1 《一句顶一万句》的作者是: 刘震云 《外观设计专利交底书模板》的专业领域是: 法务领域 最小长度: 1 最大长度: 255 数组长度: 0 - 250
fail_count	Integer	上传失败的数据条数 最小值: 0 最大值: 65535
fail_records_expire_time	String	上传失败的时间戳 最小长度: 1 最大长度: 13

表 5-17 FileExtractConf

参数	参数类型	描述
parse_conf	ParseConf object	文档解析配置, 包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页

参数	参数类型	描述
split_conf	SplitConf object	拆分配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等
id	String	文档解析ID 最小长度：0 最大长度：64

表 5-18 ParseConf

参数	参数类型	描述
ocr_enabled	Boolean	ocr增强 缺省值： false
image_enabled	Boolean	图片解析 缺省值： false
header_footer_enabled	Boolean	解析页眉页脚 缺省值： false
catalog_enabled	Boolean	解析目录页 缺省值： false
image_conf	String	图片解析开启后（TEXT 提取图片文本、IMAGE 保留原图） 缺省值： TEXT 枚举值： <ul style="list-style-type: none">● TEXT● IMAGE

表 5-19 SplitConf

参数	参数类型	描述
split_mode	String	分段设置/层级解析模型 <ul style="list-style-type: none">• LENGTH-长度拆分，即为字数拆分• CATALOG-层级分段下的自动解析• RULE- 层级分段下的规则解析• AUTO- 自动拆分，自动识别文档格式匹配适合的拆分解析方式 缺省值： AUTO 枚举值： <ul style="list-style-type: none">• LENGTH• CATALOG• RULE• AUTO

参数	参数类型	描述
separator_ids	Array of strings	<p>分段标识符ID列表</p> <p>文件解析分类</p> <p>default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en</p> <p>default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh</p> <p>default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en</p> <p>default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh</p> <p>default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en</p> <p>default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh</p> <p>default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en</p> <p>default.separator.list[7].id=comma_zh default.separator.list[7].separator=, default.separator.list[7].locale=zh</p> <p>default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en</p> <p>最小长度: 1 最大长度: 128 数组长度: 1 - 100</p>
rule_regex_id	String	<p>用户定义解析规则ID</p> <p>最小长度: 1 最大长度: 256</p>
chunk_size	Integer	<p>分段预计长度</p> <p>最小值: 0 最大值: 6000 缺省值: 500</p>

参数	参数类型	描述
title_level	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	Boolean	false 保留最后一级 true 保存多标题组合 缺省值： false
merge_titles	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的结果

状态码：400

表 5-20 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 5-21 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos/36b6d979-7f98-4fda-b8b5-d7d0cc95d296/structured-data/search?
page_num=1&page_size=1
```

响应示例

无

状态码

状态码	描述
200	文档列表响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

6 文件管理

- [上传文档接口](#)
- [查询文档接口](#)
- [下载文件接口](#)
- [删除文档接口](#)

6.1 上传文档接口

功能介绍

上传文件至某个知识库。

URI

POST /v1/koosearch/repos/{repo_id}/files

表 6-1 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64

请求参数

表 6-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 6-3 FormData 参数

参数	是否必选	参数类型	描述
file	是	File	需要上传的文件
category	否	String	文档目录, 单值, 对应目录树中的叶子节点。建议格式为 <i>叶子节点目录名称(目录 ID)</i> , 如: <i>专利 (3166-1)</i> 。 最小长度: 0 最大长度: 255
tags	否	Array of strings	文档标签, 可以使用标签对文档进行自动分组, 用于筛选过滤。 注意事项: 1. 建议不区分大小写, 例如, `Approved` and `approved` 是同一个标签 2. 文档可以有一个或多个标签 3. 不要使用保留字符, 否则查询时需要转义: ``+ - = && > < ! () { } [] ^ " ~ * ? : \ /` `` 最小长度: 1 最大长度: 255 数组长度: 0 - 250

参数	是否必选	参数类型	描述
file_extract_req	否	String	文档级别解析参数 样例 { "mode": 1, "language": "CHINESE", "parse_conf": { "ocr_enabled": true, "image_enabled": true, "image_conf": "IMAGE", "header_footer_enabled": false, "catalog_enabled": false }, "split_conf": { "split_mode": "RULE", "chunk_size": 500, "separators": ["。", "!", ";", "?", ":", "]", "title_level": 3, "combine_title": true, "merge_titles": true, "rule_regexs": ["^第([零〇一二三四五六七八九十百千万1-9]{1,7})章", "^第([零〇一二三四五六七八九十百千万1-9]{1,7})节", "^第([零〇一二三四五六七八九十百千万1-9]{1,7})条"] } } 最小长度：0 最大长度：10000

响应参数

状态码：201

表 6-4 响应 Body 参数

参数	参数类型	描述
file_id	String	文件ID 最小长度：1 最大长度：64

状态码：400

表 6-5 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 6-6 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos/0ac621de-3624-454e-adcf-8c5b4cf7317c/files
```

响应示例

状态码：201

上传文档接口响应体

```
{  
  "file_id": "5c46f7e31c5c4587ac99ef0da0502470"  
}
```

状态码

状态码	描述
201	上传文档接口响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

6.2 查询文档接口

功能介绍

在当前知识库查询名称为xx的文件。

URI

GET /v1/koosearch/repos/{repo_id}/files

表 6-7 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64

表 6-8 Query 参数

参数	是否必选	参数类型	描述
file_name	否	String	文件名 最小长度: 1 最大长度: 1024
page_num	否	Integer	请求页码 最小值: 1 最大值: 65535 缺省值: 1
page_size	否	Integer	请求限定响应结果的分页大小, 例如5条/页, 10条/页 最小值: 1 最大值: 65535 缺省值: 10

参数	是否必选	参数类型	描述
category	否	String	文档目录，单值，对应目录树中的叶子节点 建议格式为 <i>叶子节点目录名称(目录ID)</i> ，如：专利(3166-1)。 最小长度：0 最大长度：255
ids	否	Array	查询文件id列表 最小长度：1 最大长度：64 数组长度：1 - 65535

请求参数

表 6-9 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码：200

表 6-10 响应 Body 参数

参数	参数类型	描述
total	Integer	返回的文档总数 最小值：1 最大值：65535
page_num	Integer	页码 最小值：1 最大值：65535
page_size	Integer	每页数量 最小值：1 最大值：65535

参数	参数类型	描述
files	Array of FileInfo objects	文件列表 数组长度: 0 - 65535

表 6-11 FileInfo

参数	参数类型	描述
id	String	文件ID 最小长度: 1 最大长度: 64
task_id	String	任务id 最小长度: 0 最大长度: 64
name	String	文件名 最小长度: 1 最大长度: 255
repo_id	String	知识库id 最小长度: 0 最大长度: 64
project_id	String	项目id 最小长度: 1 最大长度: 64
application_id	String	应用id 最小长度: 1 最大长度: 64

参数	参数类型	描述
status	String	文件状态 <ul style="list-style-type: none">• SUCCESS 执行成功• ERROR 执行失败• PENDING 初始状态,未处理• RUNNING 执行中• IMPORT_EXCEPTION 导入异常• FILE_ENCODING_ERROR 编码错误 最小长度: 0 最大长度: 64 枚举值: <ul style="list-style-type: none">• SUCCESS• ERROR• PENDING• RUNNING• IMPORT_EXCEPTION• FILE_ENCODING_ERROR
type	String	文件类型 最小长度: 1 最大长度: 64
size	Long	文件大小 (bytes) 最小值: 1 最大值: 99999999
category	String	文档目录, 单值, 对应目录树中的叶子节点。建议格式为 叶子节点目录名称(目录 ID), 如: 专利 (3166-1)。 最小长度: 0 最大长度: 255
create_user	String	创建者, 指上传文件的用户, 可能不存在该键值 最小长度: 1 最大长度: 64
create_time	String	创建时间, 指上传文件的时间, 例如: 1692848139119 最小长度: 1 最大长度: 13
update_time	String	更新时间, 可能不存在该键值 最小长度: 1 最大长度: 13

参数	参数类型	描述
file_path	String	文件地址，可能不存在该键值 最小长度：0 最大长度：65535
upload_desc	String	上传说明，可能不存在该键值 最小长度：0 最大长度：65535
file_extract_conf	FileExtractConf object	文件抽取配置项
tags	Array of strings	文档标签，可以使用标签对文档进行自动分组，用于筛选过滤。 注意事项：1.建议不区分大小写，例如，Approved and approved 是同一个标签 2.文档可以有多个或多个标签 3.建议格式为：标签名称:标签值，如果能够保证标签值不冲突，也可以直接使用 标签值，例如 《冰箱使用说明书》对应的产品型号：ProductModel:BCD-551WLCTDAFA5U1 《一句顶一万句》的作者是：刘震云 《外观设计专利交底书模板》的专业领域是：法务领域 最小长度：1 最大长度：255 数组长度：0 - 250
fail_count	Integer	上传失败的数据条数 最小值：0 最大值：65535
fail_records_expire_time	String	上传失败的时间戳 最小长度：1 最大长度：13

表 6-12 FileExtractConf

参数	参数类型	描述
parse_conf	ParseConf object	文档解析配置，包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	SplitConf object	拆配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等

参数	参数类型	描述
id	String	文档解析ID 最小长度：0 最大长度：64

表 6-13 ParseConf

参数	参数类型	描述
ocr_enabled	Boolean	ocr增强 缺省值：false
image_enabled	Boolean	图片解析 缺省值：false
header_footer_enabled	Boolean	解析页眉页脚 缺省值：false
catalog_enabled	Boolean	解析目录页 缺省值：false
image_conf	String	图片解析开启后（TEXT 提取图片文本、IMAGE 保留原图） 缺省值：TEXT 枚举值： <ul style="list-style-type: none">• TEXT• IMAGE

表 6-14 SplitConf

参数	参数类型	描述
split_mode	String	分段设置/层级解析模型 <ul style="list-style-type: none">● LENGTH-长度拆分，即为字数拆分● CATALOG-层级分段下的自动解析● RULE- 层级分段下的规则解析● AUTO- 自动拆分，自动识别文档格式匹配适合的拆分解析方式 缺省值： AUTO 枚举值： <ul style="list-style-type: none">● LENGTH● CATALOG● RULE● AUTO

参数	参数类型	描述
separator_ids	Array of strings	分段标识符ID列表 文件解析分类 default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en default.separator.list[7].id=comma_zh default.separator.list[7].separator=, default.separator.list[7].locale=zh default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en 最小长度：1 最大长度：128 数组长度：1 - 100
rule_regex_id	String	用户定义解析规则ID 最小长度：1 最大长度：256
chunk_size	Integer	分段预计长度 最小值：0 最大值：6000 缺省值：500

参数	参数类型	描述
title_level	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的 结果

状态码：400

表 6-15 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 6-16 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

无

响应示例

无

状态码

状态码	描述
200	文档列表响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

6.3 下载文件接口

功能介绍

下载指定ID的文件。

URI

GET /v1/koosearch/repos/files/{file_id}

表 6-17 路径参数

参数	是否必选	参数类型	描述
file_id	是	String	文件ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 最小长度: 1 最大长度: 64

请求参数

表 6-18 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 6-19 响应 Body 参数

参数	参数类型	描述
-	File	下载成功的文件

状态码: 400

表 6-20 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

状态码: 500

表 6-21 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/koosearch/repos/files/69692c309a37a00ae746a926a81e5d41
```

响应示例

无

状态码

状态码	描述
200	下载成功的文件
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

6.4 删除文档接口

功能介绍

删除指定ID的文件。

URI

```
DELETE /v1/koosearch/repos/{repo_id}/files/{file_id}
```

表 6-22 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64
file_id	是	String	文件ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 最小长度: 1 最大长度: 64

请求参数

表 6-23 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 400

表 6-24 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 6-25 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

删除指定ID的文件

```
/v1/koosearch/repos/0ac621de-3624-454e-adcf-8c5b4cf7317c/files/69692c309a37a00ae746a926a81e5d41
```

响应示例

无

状态码

状态码	描述
200	删除成功
400	参数错误
500	内部异常

错误码

请参见[错误码](#)。

7 FAQ 管理

- [创建FAQ](#)
- [更新FAQ](#)
- [获取FAQ](#)
- [查询FAQ列表](#)
- [删除FAQ](#)
- [批量删除FAQ](#)

7.1 创建 FAQ

功能介绍

输入问题、相似问题、答案，创建单条FAQ。

URI

POST /v1/koosearch/repos/{repo_id}/faqs

表 7-1 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 7-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 7-3 请求 Body 参数

参数	是否必选	参数类型	描述
question	是	String	问题 最小长度：1 最大长度：1000
answer	是	String	答案 最小长度：1 最大长度：10000

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID 获取方式： 账号登录KooSearch体验平台， 在左侧导航栏选择：“知识库管理”， 在此界面获取“知识库ID”字段的值， 表示该知识库在向量数据库中存储的ID， 具有唯一性 最小长度：1 最大长度：64
question1	否	String	相似问题 最小长度：0 最大长度：1000
question2	否	String	相似问题 最小长度：0 最大长度：1000
question3	否	String	相似问题 最小长度：0 最大长度：1000
question4	否	String	相似问题 最小长度：0 最大长度：1000
category	否	String	目录，单值，对应目录树中的叶子节点 最小长度：0 最大长度：255
tags	否	String	标签列表，多个标签以逗号分隔。 注意事项：1.文档可以有一个或多个标签 2.建议标签值不区分大小写 3.不要使用保留字符，否则查询时需要转义： + - = && // > < ! () { } [] ^ " ~ * ? : / 4.完整替换，不支持部分更新。 最小长度：0 最大长度：4096

响应参数

状态码：200

表 7-4 响应 Body 参数

参数	参数类型	描述
faq_id	String	创建的FAQ的ID 最小长度：1 最大长度：64

请求示例

无

响应示例

状态码：200

成功

```
{  
  "faq_id" : "64b4a1bc-3e60-43dd-981c-d24828ddc8ac"  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

7.2 更新 FAQ

功能介绍

更新指定ID的FAQ的问题、答案、相似问题。

URI

PUT /v1/koosearch/repos/{repo_id}/faqs

表 7-5 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64

请求参数

表 7-6 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 7-7 请求 Body 参数

参数	是否必选	参数类型	描述
id	是	String	编号 最小长度: 1 最大长度: 64
question	是	String	问题 最小长度: 1 最大长度: 1000
answer	是	String	答案 最小长度: 1 最大长度: 10000

参数	是否必选	参数类型	描述
repo_id	否	String	知识库ID 获取方式： 账号登录KooSearch体验平台， 在左侧导航栏选择：“知识库管理”， 在此界面获取“知识库ID”字段的值， 表示该知识库在向量数据库中存储的ID， 具有唯一性 最小长度：1 最大长度：64
question1	否	String	相似问题 最小长度：0 最大长度：1000
question2	否	String	相似问题 最小长度：0 最大长度：1000
question3	否	String	相似问题 最小长度：0 最大长度：1000
question4	否	String	相似问题 最小长度：0 最大长度：1000
file_id	是	String	文档ID 最小长度：36 最大长度：36
category	否	String	目录，单值，对应目录树中的叶子节点 最小长度：0 最大长度：255
tags	否	String	标签列表，多个标签以逗号分隔。 注意事项：1.文档可以有一个或多个标签 2.建议标签值不区分大小写 3.不要使用保留字符，否则查询时需要转义： + - = && // > < ! () { } [] ^ " ~ * ? : / 4.完整替换，不支持部分更新。 最小长度：0 最大长度：4096

响应参数

状态码： 200

表 7-8 响应 Body 参数

参数	参数类型	描述
faq_id	String	更新的FAQ的ID 最小长度： 1 最大长度： 64

请求示例

无

响应示例

状态码： 200

成功

```
{  
  "faq_id" : "64b4a1bc-3e60-43dd-981c-d24828ddc8ac"  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

7.3 获取 FAQ

功能介绍

根据ID获取FAQ详情，包括问题、答案。

URI

GET /v1/koosearch/repos/{repo_id}/faqs/{faq_id}

表 7-9 路径参数

参数	是否必选	参数类型	描述
faq_id	是	String	FAQ ID, 1~64个字符, 只能包含数字、字母、中划线和下划线, 且必须以字母开头。 获取方式: <ol style="list-style-type: none">1. 账号登录KooSearch体验平台2. 左侧导航栏选择: "KooSearch-KooSearch文档问答"3. 根据"名称/ID", 选择购买的服务, 单击"问答", 进入KooSearch文档问答管理页面4. 左侧导航栏选择: "知识库管理", 在知识库列表中, 选择目标知识库, 单击其ID进入5. 单击"FAQ管理", 即可查看FAQ列表, 获取某条FAQ的ID 最小长度: 1 最大长度: 64
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64

请求参数

表 7-10 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 7-11 响应 Body 参数

参数	参数类型	描述
id	String	faq的ID 最小长度: 1 最大长度: 64
repo_id	String	知识库ID 最小长度: 1 最大长度: 64
question	String	问题 最小长度: 1 最大长度: 1000
answer	String	答案 最小长度: 1 最大长度: 10000
status	String	无实际意义, 可忽略 最小长度: 1 最大长度: 64
create_time	String	创建时间 最小长度: 0 最大长度: 64
update_time	String	更新时间 最小长度: 0 最大长度: 64

参数	参数类型	描述
question1	String	相似问1 最小长度: 0 最大长度: 1000
question2	String	相似问2 最小长度: 0 最大长度: 1000
question3	String	相似问3 最小长度: 0 最大长度: 1000
question4	String	相似问4 最小长度: 0 最大长度: 1000
category	String	目录, 单值, 对应目录树中的叶子节点 最小长度: 0 最大长度: 255
tags	Array of strings	标签列表 注意事项: 1.文档可以有一个或多个标签 2.建议标签值不区分大小写 3.不要使用保留字符, 否则查询时需要转义: + - = && // > < ! () { } [] ^ " ~ * ? : / 4.完整替换, 不支持部分更新。 最小长度: 1 最大长度: 255 数组长度: 0 - 250

请求示例

无

响应示例

状态码: 200

成功

```
{
  "id": "5323205cfa9244b385ff5968ece0cd9a",
  "repo_id": "123",
  "question": "问题",
  "answer": "答案",
  "status": 0,
  "create_time": "1716284755573",
  "update_time": "1716284755573"
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

7.4 查询 FAQ 列表

功能介绍

获取FAQ列表，包括ID、问题、答案、创建人、创建时间等。可以根据问题关键字过滤出需要查看的FAQ。

URI

GET /v1/koosearch/repos/{repo_id}/faqs

表 7-12 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度： 1 最大长度： 64

表 7-13 Query 参数

参数	是否必选	参数类型	描述
question	否	String	问题关键字 最小长度： 0 最大长度： 1000

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID 获取方式： 账号登录KooSearch体验平台， 在左侧导航栏选择：“知识库管理”， 在此界面获取“知识库ID”字段的值， 表示该知识库在向量数据库中存储的ID， 具有唯一性 最小长度：1 最大长度：64
answer	否	String	FAQ答案 最小长度：0 最大长度：10000
page_num	否	Integer	请求页码 最小值：1 最大值：65535 缺省值：1
page_size	否	Integer	限定响应体每页返回的数据条数， 例如5页/条，10页/条 最小值：1 最大值：65535 缺省值：10
ids	否	Array	faq id列表 最小长度：1 最大长度：64 数组长度：0 - 65535

请求参数

表 7-14 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码： 200

表 7-15 响应 Body 参数

参数	参数类型	描述
records	Array of FaqInfo objects	Faq信息 数组长度： 0 - 65535
total	Integer	总数量 最小值： 0 最大值： 65535
size	Integer	每页数量 最小值： 0 最大值： 65535
current	Integer	当前页码 最小值： 0 最大值： 65535
optimize_count_sql	Boolean	无实际意义，可忽略
search_count	Boolean	无实际意义，可忽略
pages	Integer	总页数 最小值： 0 最大值： 65535

表 7-16 FaqInfo

参数	参数类型	描述
id	String	faq的ID 最小长度： 1 最大长度： 64
repo_id	String	知识库ID 最小长度： 1 最大长度： 64
question	String	问题 最小长度： 1 最大长度： 1000

参数	参数类型	描述
answer	String	答案 最小长度: 1 最大长度: 10000
status	String	无实际意义, 可忽略 最小长度: 1 最大长度: 64
create_time	String	创建时间 最小长度: 0 最大长度: 64
update_time	String	更新时间 最小长度: 0 最大长度: 64
question1	String	相似问1 最小长度: 0 最大长度: 1000
question2	String	相似问2 最小长度: 0 最大长度: 1000
question3	String	相似问3 最小长度: 0 最大长度: 1000
question4	String	相似问4 最小长度: 0 最大长度: 1000
category	String	目录, 单值, 对应目录树中的叶子节点 最小长度: 0 最大长度: 255
tags	Array of strings	标签列表 注意事项: 1.文档可以有一个或多个标签 2.建议标签值不区分大小写 3.不要使用保留字符, 否则查询时需要转义: + - = && // > < ! () { } [] ^ " ~ * ? : / 4.完整替换, 不支持部分更新。 最小长度: 1 最大长度: 255 数组长度: 0 - 250

请求示例

```
/v1/12345678/applications/12345678/uni-search/12345678aaa/faq??page_num=1&page_size=10&question=你好&repo_id=a12356455
```

响应示例

状态码: 200

成功

```
{
  "records": [ {
    "id": "991a482a-8330-4338-bdda-32852504bb38",
    "repo_id": "b4254b30-f241-476c-b003-63df8f1cc221",
    "question": "无法登录我的账户怎么办?",
    "answer": "请确保你使用的是正确的用户名和密码。如果仍然无法登录, 请尝试重置密码或联系客服。",
    "status": 0,
    "create_time": "1700807345932",
    "update_time": "1700807345932"
  }, {
    "id": "d3477f5c-f7d7-409c-a709-b56df8a7341c",
    "repo_id": "b4254b30-f241-476c-b003-63df8f1cc221",
    "question": "如何申请入学?",
    "answer": "请查看我们的官方网站或联系招生办公室, 了解入学要求、申请截止日期和所需材料。我们还提供入学顾问服务。",
    "status": 0,
    "create_time": "1700807328240",
    "update_time": "1700807328240"
  } ],
  "total": 2,
  "size": 10,
  "current": 1,
  "optimize_count_sql": true,
  "search_count": true,
  "pages": 1
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

7.5 删除 FAQ

功能介绍

根据知识库ID和FAQ的ID, 删除FAQ。

URI

```
DELETE /v1/koosearch/repos/{repo_id}/faqs/{faq_id}
```

表 7-17 路径参数

参数	是否必选	参数类型	描述
faq_id	是	String	FAQ编号 最小长度：1 最大长度：64
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 7-18 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

无

请求示例

无

响应示例

无

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

7.6 批量删除 FAQ

功能介绍

选择多个需要删除的FAQ，根据ID批量删除。

URI

DELETE /v1/koosearch/repos/{repo_id}/faqs

表 7-19 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度： 1 最大长度： 64

请求参数

表 7-20 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 7-21 请求 Body 参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64
faq_ids	是	Array of strings	待删除FAQ编号数组 最小长度：1 最大长度：64 数组长度：1 - 1000

响应参数

状态码：200

表 7-22 响应 Body 参数

参数	参数类型	描述
total_count	Integer	总数量 最小值：0 最大值：65535
deleted_count	Integer	删除成功数量 最小值：0 最大值：65535

请求示例

无

响应示例

状态码: 200

成功

```
{  
  "totalCount" : 2,  
  "deletedCount" : 2  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

8 FAQ 批量管理

- [批量FAQ上传](#)
- [列举批量上传FAQ文件](#)
- [下载上传失败的FAQ](#)
- [下载批量导入FAQ原文件](#)
- [删除批量导入FAQ文件](#)

8.1 批量 FAQ 上传

功能介绍

上传xlsx/xls文件，用于FAQ的批量导入。

URI

POST /v1/koosearch/repos/{repo_id}/faqs/files

表 8-1 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

请求参数

表 8-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 8-3 FormData 参数

参数	是否必选	参数类型	描述
file	是	File	仅支持JSON文件上传
rule	否	String	问题匹配规则 最小长度: 0 最大长度: 64

响应参数

状态码: 200

表 8-4 响应 Body 参数

参数	参数类型	描述
file_id	String	文件ID 最小长度: 1 最大长度: 64

状态码: 400

表 8-5 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 8-6 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
https://console.ulanhqab.huawei.com/elasticsearch/unisearch-dev/v1/729cbd739854470da5426ed26bd900ca/
unisearch-dev/v1/1ed40ceefc8d40f8b884edb6a84e7768/applications/fb9731ab-7085-474f-
b6c7-64473586f0f3/uni-search/e8196714-3e6b-454e-8c23-61baff601ec9/faq/batch/upload
file
```

响应示例

状态码：200

上传文档接口响应体

```
{
  "file_id": "60fa0fbf-a809-4e4e-8d26-ec389a497bf4"
}
```

状态码

状态码	描述
200	上传文档接口响应体
400	参数错误
500	内部异常

错误码

请参见[错误码](#)。

8.2 列举批量上传 FAQ 文件

功能介绍

列举批量上传FAQ文件信息。

URI

GET /v1/koosearch/repos/{repo_id}/faqs/files

表 8-7 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

表 8-8 Query 参数

参数	是否必选	参数类型	描述
file_name	否	String	文件名 最小长度：0 最大长度：255

参数	是否必选	参数类型	描述
file_status	否	String	文件状态 SUCCESS 上传成功 ERROR 上传失败 PENDING 上传等待 RUNNING 上传进行中 IMPORT_EXCEPTION 导入失败 FILE_ENCODING_ERROR 文件编码错误 最小长度: 0 最大长度: 128 枚举值: <ul style="list-style-type: none">• SUCCESS• ERROR• PENDING• RUNNING• IMPORT_EXCEPTION• FILE_ENCODING_ERROR
page_num	否	Integer	请求页码 最小值: 1 最大值: 65535 缺省值: 1
page_size	否	Integer	请求限定响应结果的分页大小, 例如5条/页, 10条/页 最小值: 1 最大值: 65535 缺省值: 10
ids	否	Array	精准查询id列表 最小长度: 1 最大长度: 64 数组长度: 1 - 65535

请求参数

表 8-9 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码： 200

表 8-10 响应 Body 参数

参数	参数类型	描述
total	Integer	返回的文档总数 最小值： 1 最大值： 65535
page_num	Integer	页码 最小值： 1 最大值： 65535
page_size	Integer	每页数量 最小值： 1 最大值： 65535
files	Array of FileInfo objects	文件列表 数组长度： 0 - 65535

表 8-11 FileInfo

参数	参数类型	描述
id	String	文件ID 最小长度： 1 最大长度： 64
task_id	String	任务id 最小长度： 0 最大长度： 64
name	String	文件名 最小长度： 1 最大长度： 255
repo_id	String	知识库id 最小长度： 0 最大长度： 64

参数	参数类型	描述
project_id	String	项目id 最小长度: 1 最大长度: 64
application_id	String	应用id 最小长度: 1 最大长度: 64
status	String	文件状态 <ul style="list-style-type: none"> • SUCCESS 执行成功 • ERROR 执行失败 • PENDING 初始状态,未处理 • RUNNING 执行中 • IMPORT_EXCEPTION 导入异常 • FILE_ENCODING_ERROR 编码错误 最小长度: 0 最大长度: 64 枚举值: <ul style="list-style-type: none"> • SUCCESS • ERROR • PENDING • RUNNING • IMPORT_EXCEPTION • FILE_ENCODING_ERROR
type	String	文件类型 最小长度: 1 最大长度: 64
size	Long	文件大小 (bytes) 最小值: 1 最大值: 99999999
category	String	文档目录, 单值, 对应目录树中的叶子节点。建议格式为 叶子节点目录名称(目录 ID), 如: 专利 (3166-1)。 最小长度: 0 最大长度: 255
create_user	String	创建者, 指上传文件的用户, 可能不存在该键值 最小长度: 1 最大长度: 64

参数	参数类型	描述
create_time	String	创建时间，指上传文件的时间，例如： 1692848139119 最小长度：1 最大长度：13
update_time	String	更新时间，可能不存在该键值 最小长度：1 最大长度：13
file_path	String	文件地址，可能不存在该键值 最小长度：0 最大长度：65535
upload_desc	String	上传说明，可能不存在该键值 最小长度：0 最大长度：65535
file_extract_conf	FileExtractConf object	文件抽取配置项
tags	Array of strings	文档标签，可以使用标签对文档进行自动分组，用于筛选过滤。 注意事项：1.建议不区分大小写，例如，Approved and approved 是同一个标签 2.文档可以有一个或多个标签 3.建议格式为：标签名称:标签值，如果能够保证标签值不冲突，也可以直接使用 标签值，例如 《冰箱使用说明书》对应的产品型号： ProductModel:BCD-551WLCTDAFA5U1 《一句顶一万句》的作者是：刘震云 《外观设计专利交底书模板》的专业领域是：法务领域 最小长度：1 最大长度：255 数组长度：0 - 250
fail_count	Integer	上传失败的数据条数 最小值：0 最大值：65535
fail_records_expire_time	String	上传失败的时间戳 最小长度：1 最大长度：13

表 8-12 FileExtractConf

参数	参数类型	描述
parse_conf	ParseConf object	文档解析配置，包含是否使用OCR增强、是否解析图片、解析图片是否需要提取文字、是否解析页眉页脚、是否解析目录页
split_conf	SplitConf object	拆分配置，包括分段方式设置、层级解析模式设置、标题层级深度设置、标题保存方式设置、分段长度配置、标题匹配pattern配置等
id	String	文档解析ID 最小长度： 0 最大长度： 64

表 8-13 ParseConf

参数	参数类型	描述
ocr_enabled	Boolean	ocr增强 缺省值： false
image_enabled	Boolean	图片解析 缺省值： false
header_footer_enabled	Boolean	解析页眉页脚 缺省值： false
catalog_enabled	Boolean	解析目录页 缺省值： false
image_conf	String	图片解析开启后（TEXT 提取图片文本、IMAGE 保留原图） 缺省值： TEXT 枚举值： <ul style="list-style-type: none">● TEXT● IMAGE

表 8-14 SplitConf

参数	参数类型	描述
split_mode	String	分段设置/层级解析模型 <ul style="list-style-type: none">● LENGTH-长度拆分，即为字数拆分● CATALOG-层级分段下的自动解析● RULE- 层级分段下的规则解析● AUTO- 自动拆分，自动识别文档格式匹配适合的拆分解析方式 缺省值： AUTO 枚举值： <ul style="list-style-type: none">● LENGTH● CATALOG● RULE● AUTO

参数	参数类型	描述
separator_ids	Array of strings	<p>分段标识符ID列表</p> <p>文件解析分类</p> <p>default.separator.list[0].id=space_en default.separator.list[0].separator=\u0020 default.separator.list[0].locale=en</p> <p>default.separator.list[1].id=period_zh default.separator.list[1].separator=。 default.separator.list[1].locale=zh</p> <p>default.separator.list[2].id=period_en default.separator.list[2].separator=. default.separator.list[2].locale=en</p> <p>default.separator.list[3].id=exclamation_mark_zh default.separator.list[3].separator=! default.separator.list[3].locale=zh</p> <p>default.separator.list[4].id=exclamation_mark_en default.separator.list[4].separator=! default.separator.list[4].locale=en</p> <p>default.separator.list[5].id=question_mark_zh default.separator.list[5].separator=? default.separator.list[5].locale=zh</p> <p>default.separator.list[6].id=question_mark_en default.separator.list[6].separator=? default.separator.list[6].locale=en</p> <p>default.separator.list[7].id=comma_zh default.separator.list[7].separator=, default.separator.list[7].locale=zh</p> <p>default.separator.list[8].id=comma_en default.separator.list[8].separator=, default.separator.list[8].locale=en</p> <p>最小长度: 1 最大长度: 128 数组长度: 1 - 100</p>
rule_regex_id	String	<p>用户定义解析规则ID</p> <p>最小长度: 1 最大长度: 256</p>
chunk_size	Integer	<p>分段预计长度</p> <p>最小值: 0 最大值: 6000 缺省值: 500</p>

参数	参数类型	描述
title_level	Integer	标题层级深度 最小值：1 最大值：10 缺省值：3
combine_title	Boolean	false 保留最后一级 true 保存多标题组合 缺省值：false
merge_titles	Boolean	是否跨标题合并，不同标题段落文字较少时，会自动合并到指定的分段长度，有助于生成更全面的 结果

状态码：400

表 8-15 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 8-16 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

<https://console.ulanhqab.huawei.com/elasticsearch/unisearch-dev/v1/729cbd739854470da5426ed26bd900ca/unisearch-dev/v1/1ed40ceefc8d40f8b884edb6a84e7768/applications/fb9731ab-7085-474f->

```
b6c7-64473586f0f3/uni-search/e8196714-3e6b-454e-8c23-61baff601ec9/faq/batch/search?  
page_num=1&page_size=10
```

响应示例

状态码: 200

文档列表响应体

```
{  
  "total": 1,  
  "page_num": 1,  
  "page_size": 1,  
  "files": [{  
    "id": "6e9abdb7-4626-492d-9f18-abad3b95572d",  
    "name": "successData.xlsx",  
    "type": "faq",  
    "size": 9834,  
    "create_time": "1700481265798"  
  }]  
}
```

状态码

状态码	描述
200	文档列表响应体
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

8.3 下载上传失败的 FAQ

功能介绍

下载给定上传的批量FAQ文件中，上传失败的条目。

URI

GET /v1/koosearch/repos/{repo_id}/faqs/files/{file_id}/fails

表 8-17 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 获取方式: 账号登录KooSearch体验平台, 在左侧导航栏选择: "知识库管理", 在此界面获取"知识库ID"字段的值, 表示该知识库在向量数据库中存储的ID, 具有唯一性 最小长度: 1 最大长度: 64
file_id	是	String	文件ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 最小长度: 1 最大长度: 64

请求参数

表 8-18 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 8-19 响应 Body 参数

参数	参数类型	描述
-	File	下载的记录有失败条目的文件

请求示例

```
https://console.ulanhqab.huawei.com/elasticsearch/unisearch-dev/v1/729cbd739854470da5426ed26bd900ca/  
unisearch-dev/v1/1ed40ceefc8d40f8b884edb6a84e7768/applications/fb9731ab-7085-474f-  
b6c7-64473586f0f3/uni-search/e8196714-3e6b-454e-8c23-61baff601ec9/faq/batch/fail-records/  
f01d3089-7c36-4f03-94ac-38b8df2ff016
```

响应示例

无

状态码

状态码	描述
200	下载的记录有失败条目的文件

错误码

请参见[错误码](#)。

8.4 下载批量导入 FAQ 原文件

功能介绍

下载批量导入FAQ的原始文件。

URI

GET /v1/koosearch/repos/{repo_id}/faqs/files/{file_id}

表 8-20 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64

参数	是否必选	参数类型	描述
file_id	是	String	文件ID, 1~64个字符, 只能包含数字、字母、中划线和下划线。 最小长度: 1 最大长度: 64

请求参数

表 8-21 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 8-22 响应 Body 参数

参数	参数类型	描述
-	File	下载的FAQ原始文件

请求示例

```
https://console.ulanqab.huawei.com/elasticsearch/unisearch-dev/v1/729cbd739854470da5426ed26bd900ca/unisearch-dev/v1/1ed40ceefc8d40f8b884edb6a84e7768/applications/fb9731ab-7085-474f-b6c7-64473586f0f3/uni-search/faq/batch/6e9abdb7-4626-492d-9f18-abad3b95572d
```

响应示例

无

状态码

状态码	描述
200	下载的FAQ原始文件

错误码

请参见[错误码](#)。

8.5 删除批量导入 FAQ 文件

功能介绍

删除指定的FAQ文件。

URI

DELETE /v1/koosearch/repos/{repo_id}/faqs/files/{file_id}

表 8-23 路径参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID，1~64个字符，只能包含数字、字母、中划线和下划线。 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64
file_id	是	String	文件ID，1~64个字符，只能包含数字、字母、中划线和下划线。 最小长度：1 最大长度：64

请求参数

表 8-24 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码： 400

表 8-25 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

状态码： 500

表 8-26 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度： 8 最大长度： 36
error_msg	String	错误描述 最小长度： 2 最大长度： 512

请求示例

```
https://console.ulanqab.huawei.com/elasticsearch/unisearch-dev/v1/729cbd739854470da5426ed26bd900ca/unisearch-dev/v1/1ed40ceefc8d40f8b884edb6a84e7768/applications/fb9731ab-7085-474f-b6c7-64473586f0f3/uni-search/e8196714-3e6b-454e-8c23-61baff601ec9/faq/batch/6e9abdb7-4626-492d-9f18-abad3b95572d
```

响应示例

无

状态码

状态码	描述
200	删除成功
400	请求体参数错误

状态码	描述
500	内部异常

错误码

请参见[错误码](#)。

9 搜索与问答

搜索知识库

生成答案

9.1 搜索知识库

功能介绍

根据query，在指定知识库中搜索得到topN相关的分片文档。

URI

POST /v1/koosearch/experience/search

请求参数

表 9-1 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 9-2 请求 Body 参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID 获取方式： 账号登录KooSearch体验平台， 在左侧导航栏选择：“知识库管理”， 在此界面获取“知识库ID”字段的值， 表示该知识库在向量数据库中存储的ID， 具有唯一性 最小长度：1 最大长度：64
content	是	String	搜索内容 最小长度：1 最大长度：10000
page_num	是	Integer	页码，范围1~65535 最小值：1 最大值：65535
page_size	是	Integer	页面大小，范围1~65535 最小值：1 最大值：65535
filter_string	否	String	带语法的查询字符串，使用具有严格语法的解析器， 根据提供的查询字符串返回文档。 例如： <i>tags:(魔都 OR 首都) AND category:"城市(city)"</i> 例如：需要依据文件ID过滤分片， 则使用： <i>"fileId": "f6a050ffef5512e80a6dea57e386f194"</i> 注意事项：1.支持布尔运算符 AND、OR 和 NOT（也写为 &&、 和 !）， 但请注意它们不遵守通常的优先级规则， 因此只要一起使用多个运算符，就应使用括号。 2.保留字符查询时需要使用 转义： <i>+ - = && // > < ! () { } [] ^ " ~ * ? : /</i> ， 如 3.详细语法参考 query_string 语法 最小长度：4 最大长度：1024

参数	是否必选	参数类型	描述
scope	否	String	搜索范围，定义搜索索引中的信息子集，默认文档库 doc: 文档知识，使用query2doc模型 faq: FAQ，常见问答集，使用query2query模型 web: 网络来源，来自于web搜索引擎 缺省值: doc 最小长度: 1 最大长度: 16 枚举值: <ul style="list-style-type: none">• doc• faq• web

响应参数

状态码: 200

表 9-3 响应 Body 参数

参数	参数类型	描述
doc_list	Array of ChatReferenceInfo objects	文档列表 数组长度: 1 - 65535
total	Integer	总数 最小值: 1 最大值: 65535

表 9-4 ChatReferenceInfo

参数	参数类型	描述
file_id	String	文件id (或FAQ id) 最小长度: 1 最大长度: 64
title	String	标题 最小长度: 1 最大长度: 65535

参数	参数类型	描述
content	String	文本内容 最小长度: 1 最大长度: 65535
big_content	String	内容上下文, 比如未分片的原始表格。 注意: (1) 搜索接口可能有值 (2) 问答接口 会将big_content替换到content, 此字段清空 最小长度: 0 最大长度: 100000
update_date_time	String	更新时间 最小长度: 1 最大长度: 64
doc_type	String	文档类型 (doc-文档, faq-问答) 最小长度: 1 最大长度: 64
file_path	String	文档归档路径 最小长度: 1 最大长度: 64
category	String	文档目录, 单值, 对应目录树中的叶子节点 最小长度: 4 最大长度: 255
tags	Array of strings	文档标签列表 最小长度: 4 最大长度: 255 数组长度: 0 - 250
score	Float	ES打分 最小值: 0 最大值: 10000
subtitle	String	文档子标题 最小长度: 1 最大长度: 65535
repo_id	String	知识库标识 最小长度: 1 最大长度: 64

请求示例

```
/v1/123456/applications/123456/uni-search/experience/searchtext
```

```
{
  "repo_id": "1234abdc",
  "content": "写一个穿越到宋朝的故事。",
  "page_num": 1,
  "page_size": 10
}
```

响应示例

状态码： 200

搜索结果返回体

```
{
  "doc_list": [ {
    "file_id": "123abc",
    "title": "人工智能.doc",
    "content": "人工智能",
    "update_date_time": "134554",
    "doc_type": "doc",
    "file_path": "uni-search/v1/人工智能.doc",
    "score": 0.9999
  } ],
  "total": 12
}
```

状态码

状态码	描述
200	搜索结果返回体

错误码

请参见[错误码](#)。

9.2 生成答案

功能介绍

将content输入给模型，使用模型能力生成答案。

注意：

1. 当"chat_id"为空时，"chat_create_flag"应为"1"，表示需要开启新的对话；当"chat_id"有值时，"chat_create_flag"可以为"0"，表示不需要开启新的对话。
2. "messages"中的元素必须为奇数个，单轮对话时为1，多轮对话时为3、5、7、...，如 [question1, answer1, question2]
3. 多轮对话中：对话排列顺序从旧到新

URI

POST /v1/koosearch/experience/chat

请求参数

表 9-5 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 9-6 请求 Body 参数

参数	是否必选	参数类型	描述
repo_id	是	String	知识库ID 获取方式： 账号登录KooSearch体验平台，在左侧导航栏选择：“知识库管理”，在此界面获取“知识库ID”字段的值，表示该知识库在向量数据库中存储的ID，具有唯一性 最小长度：1 最大长度：64
chat_id	否	String	对话ID 最小长度：1 最大长度：64
messages	是	Array of ChatMessage objects	对话内容 数组长度：1 - 19
chat_create_flag	是	String	对话开始标志（0-否，1-是） 最小长度：1 最大长度：1
refresh_flag	是	String	问答刷新标志（0-否，1-是） 最小长度：1 最大长度：1

参数	是否必选	参数类型	描述
max_tokens	否	Integer	生成文本的最大token数量。输入的文本加上生成的文本总量不能超过模型所能处理的最大长度。最小值: 1 最小值: 1 最大值: 131072 缺省值: 2048
stream	否	Boolean	流式输出标志 缺省值: false
filter_string	否	String	RAG搜索环节过滤。 带语法的查询字符串, 使用具有严格语法的解析器, 根据提供的查询字符串返回文档。 例如: <i>tags:(魔都 OR 首都) AND category:"城市(city)"</i> 注意事项: 1.支持布尔运算符 AND、OR 和 NOT (也写为 &&、 和 !), 但请注意它们不遵守通常的优先级规则, 因此只要一起使用多个运算符, 就应使用括号。2. 详细语法参考 query_string语法 最小长度: 4 最大长度: 1024
top_p	否	Float	一种替代温度采样的方法, 称为 nucleus sampling, 其中模型考虑具有top_p 概率质量的标记的结果。通常建议更改此值或温度, 但不要同时更改两者。通常建议更改top_p或temperature 来调整生成文本的倾向性, 但不要同时更改这两个参数。取值范围: (0, 1] 最小值: 0.1 最大值: 1 缺省值: 0.1

参数	是否必选	参数类型	描述
presence_penalty	否	Float	用于调整模型对新Token的处理方式。即如果一个Token已经在之前的文本中出现过，那么模型在生成这个Token时会受到一定的惩罚。当presence_penalty的值为正数时，模型会更倾向于生成新的、未出现过的Token，即模型会更倾向于谈论新的话题。 最小值: -2 最大值: 2 最小值: -2 最大值: 2 缺省值: 0
search_temperature	否	Float	搜索增强场景用于控制生成文本的多样性和创造力。取值接近0表示最低的随机性，1表示最高的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，例如0.9，适合完成创造性的任务。 最小值: 0 最大值: 1 缺省值: 0.3
chat_temperature	否	Float	非搜索增强场景用于控制生成文本的多样性和创造力。取值接近0表示最低的随机性，1表示最高的随机性。一般来说，temperature越低，适合完成确定性的任务。temperature越高，例如0.9，适合完成创造性的任务。 最小值: 0 最大值: 1 缺省值: 0.8
extra_repo_ids	否	Array of strings	多知识库问答 最小长度: 1 最大长度: 64 数组长度: 1 - 10

表 9-7 ChatMessage

参数	是否必选	参数类型	描述
role	否	String	对话的角色，取值为system、user 最小长度：1 最大长度：100
content	是	String	对话内容 最小长度：1 最大长度：4096

响应参数

状态码：200

表 9-8 响应 Body 参数

参数	参数类型	描述
chat_id	String	对话ID 最小长度：1 最大长度：64
chat_result	ChatResult object	对话结果
references	Array of ChatReferencelInfo objects	参考结果 数组长度：1 - 100
reference_total	Integer	参考来源总数 最小值：1 最大值：65535
sub_queries	Array of strings	搜索使用的query，包括：原始query和改写后的query 最小长度：1 最大长度：65535 数组长度：1 - 65535
category	String	搜索规划返回的query类别 最小长度：1 最大长度：65535

表 9-9 ChatResult

参数	参数类型	描述
index	Integer	序号 最小值：1 最大值：65535
message	String	文本内容 最小长度：1 最大长度：65535
question_id	String	问答ID 最小长度：1 最大长度：64

表 9-10 ChatReferenceInfo

参数	参数类型	描述
file_id	String	文件id (或FAQ id) 最小长度：1 最大长度：64
title	String	标题 最小长度：1 最大长度：65535
content	String	文本内容 最小长度：1 最大长度：65535
big_content	String	内容上下文，比如未分片的原始表格。 注意：（1）搜索接口可能有值（2）问答接口会将big_content替换到content，此字段清空 最小长度：0 最大长度：100000
update_date_time	String	更新时间 最小长度：1 最大长度：64
doc_type	String	文档类型（doc-文档，faq-问答） 最小长度：1 最大长度：64

参数	参数类型	描述
file_path	String	文档归档路径 最小长度: 1 最大长度: 64
category	String	文档目录, 单值, 对应目录树中的叶子节点 最小长度: 4 最大长度: 255
tags	Array of strings	文档标签列表 最小长度: 4 最大长度: 255 数组长度: 0 - 250
score	Float	ES打分 最小值: 0 最大值: 10000
subtitle	String	文档子标题 最小长度: 1 最大长度: 65535
repo_id	String	知识库标识 最小长度: 1 最大长度: 64

请求示例

上述是一个多轮对话的请求体示例。注意: 1)、当"chat_id"为空时,"chat_create_flag"应为"1",表示需要开启新的对话;当"chat_id"有值时,"chat_create_flag"可以为"0",表示不需要开启新的对话。2)、"messages"中的元素始终为奇数,单轮对话时为1,多轮对话时为3、5、7、...如上述示例,message[0]["content"]是第一轮的query,message[1]["content"]是第一轮的answer,message[2]["content"]是第二轮的query

```
/v1/123/applications/1233453/uni-search/experience/chat
{
  "repo_id": "9d261b1c-e9b6-464f-881a-687251504edd",
  "chat_create_flag": "0",
  "chat_id": "cc84be57-3fc1-4365-a54d-417f098e8a19",
  "messages": [ {
    "role": "user",
    "content": "2023款本田雅阁的维修保养周期建议?"
  }, {
    "role": "user",
    "content": "对于2023款本田雅阁,我建议你可以注意以下几个方面的定期维护:定期更换机油和机油滤芯。根据厂家建议,大约每5,000-10,000公里更换一次。"
  }, {
    "role": "user",
    "content": "还有什么需要注意的?"
  } ],
}
```

```
"refresh_flag": "0"  
}
```

响应示例

状态码： 200

模型生成结果响应体

```
{  
  "chat_id": "cc84be57-3fc1-4365-a54d-417f098e8a19",  
  "chat_result": {  
    "index": 0,  
    "message": "在进行维修保养时，需要注意：检查轮胎状况,包括胎压、胎纹深度等。及时更换磨损严重的轮胎。检查制动系统,包括刹车片、制动液等。如有磨损需要及时维修更换。定期检查发动机冷却系统,更换冷却液和防冻液。检查蓄电池状态,确保电瓶能够正常工作。保持定期保养,比如四轮定位、空调保养等,延长车辆使用寿命。",  
    "question_id": "12453567446"  
  },  
  "references": [ {  
    "file_id": "123",  
    "title": "汽车包养注意事项.pdf",  
    "content": "检查轮胎状况,包括胎压、胎纹深度等。及时更换磨损严重的轮胎。检查制动系统,包括刹车片、制动液等。",  
    "update_date_time": "1234554456",  
    "doc_type": "doc",  
    "file_path": "uni-search/v2/文档名称.doc"  
  }, {  
    "file_id": "123",  
    "title": "汽车包养注意事项.pdf",  
    "content": "如有磨损需要及时维修更换。定期检查发动机冷却系统,更换冷却液和防冻液。",  
    "update_date_time": "1234554456",  
    "doc_type": "doc",  
    "file_path": "uni-search/v2/文档名称.doc"  
  }, {  
    "file_id": "123",  
    "title": "汽车包养注意事项.pdf",  
    "content": "检查蓄电池状态,确保电瓶能够正常工作。保持定期保养,比如四轮定位、空调保养等,延长车辆使用寿命。",  
    "update_date_time": "1234554456",  
    "doc_type": "doc",  
    "file_path": "uni-search/v2/文档名称.doc"  
  } ],  
  "reference_total": "3"  
}
```

状态码

状态码	描述
200	模型生成结果响应体

错误码

请参见[错误码](#)。

10 对话历史

[查询对话历史列表](#)

[获取对话历史](#)

[删除对话历史](#)

[批量删除对话历史](#)

10.1 查询对话历史列表

功能介绍

查询历史问题列表。包含对话ID、对话问题、用户名、对话开始时间。

URI

GET /v1/koosearch/chat-history

表 10-1 Query 参数

参数	是否必选	参数类型	描述
user_name	否	String	用户名 最小长度：1 最大长度：64
repo_id	否	String	知识库ID 获取方式： 账号登录KooSearch体验平台， 在左侧导航栏选择：“知识库管理”， 在此界面获取“知识库ID”字段的值， 表示该知识库在向量数据库中存储的ID， 具有唯一性 最小长度：1 最大长度：64

参数	是否必选	参数类型	描述
page_num	否	Integer	请求页码 最小值：1 最大值：65535 缺省值：1
page_size	否	Integer	请求限定响应结果的分页大小， 例如5条/页，10条/页 最小值：1 最大值：65535 缺省值：10

请求参数

表 10-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码：200

表 10-3 响应 Body 参数

参数	参数类型	描述
total	Integer	会话历史总数 最小值：1 最大值：65535
size	Integer	当前页面返回条数 最小值：1 最大值：65535
pages	Integer	总页数 最小值：1 最大值：65535

参数	参数类型	描述
current	Integer	当前页码 最小值：1 最大值：65535
optimize_count_sql	Boolean	当前无实际意义
search_count	Boolean	当前无实际意义
records	Array of ChatHistoryRecordInfo objects	会话历史详情 数组长度：0 - 65535

表 10-4 ChatHistoryRecordInfo

参数	参数类型	描述
chat_id	String	会话ID，会话的唯一标识 最小长度：1 最大长度：64
knowledge_repo_id	String	知识库ID，标识当前会话参考的知识库 最小长度：1 最大长度：64
chat_title	String	当前会话中的query 最小长度：1 最大长度：65535
create_datetime	String	会话发生的时间戳 最小长度：1 最大长度：13
project_id	String	项目ID，1~64个字符，只能包含数字、字母、中划线和下划线，且必须以字母开头。预留字段，可联系负责开通服务的工作人员获取该字段的值 最小长度：1 最大长度：64
is_del	Integer	当前对话是否删除标记，0：未删除，1：已删除 枚举值： <ul style="list-style-type: none">01

请求示例

无

响应示例

状态码： 200

成功

```
{
  "records": [ {
    "chat_id": "e447ed24-b3e0-4184-96cc-e06a9dc6cf89",
    "knowledge_repo_id": "176e63a2-73bd-415c-aab2-f688030855e3",
    "chat_title": "今天天气",
    "create_date_time": "1695196394925",
    "is_del": 0
  }, {
    "chat_id": "2c57dd12-0384-46c8-8d13-ae78dc857d94",
    "knowledge_repo_id": "176e63a2-73bd-415c-aab2-f688030855e3",
    "chat_title": "人工智能会取代人类吗",
    "create_date_time": "1695195832442",
    "is_del": 0
  }, {
    "chat_id": "ff193ad6-c812-402c-8233-f7ab5f4b0b11",
    "knowledge_repo_id": "176e63a2-73bd-415c-aab2-f688030855e3",
    "chat_title": "人工智能会取代人类吗",
    "create_date_time": "1695195733199",
    "is_del": 0
  }, {
    "chat_id": "f557e561-cc8f-47b1-bbd7-9d6064293588",
    "knowledge_repo_id": "176e63a2-73bd-415c-aab2-f688030855e3",
    "chat_title": "今天天气好吗",
    "create_date_time": "1695195285924",
    "is_del": 0
  }, {
    "chat_id": "a0330214-ea10-4aba-874a-06cefa4bc77a",
    "knowledge_repo_id": "852947fe-6c60-4d91-af65-839a0292127e",
    "chat_title": "神经网络是什么",
    "create_date_time": "1695193989920",
    "is_del": 0
  }, {
    "chat_id": "fbc01f0d-a9c4-4d80-a8e4-7c2e191472e9",
    "knowledge_repo_id": "fba6ea89-5d9a-4835-9a8d-3ecebfc023e9",
    "chat_title": "你好吗",
    "create_date_time": "1695193289515",
    "is_del": 0
  }, {
    "chat_id": "475ad732-d682-4899-8263-6643e1f1d045",
    "knowledge_repo_id": "d3e0e470-0032-4907-ba64-9aba4521ed0c",
    "chat_title": "你好吗",
    "create_date_time": "1695193277246",
    "is_del": 0
  }, {
    "chat_id": "73e00b01-df45-4222-81f1-018726e66315",
    "knowledge_repo_id": "fba6ea89-5d9a-4835-9a8d-3ecebfc023e9",
    "chat_title": "杭州在哪里",
    "create_date_time": "1695192613055",
    "is_del": 0
  }, {
    "chat_id": "f6bee59c-a8b9-41d7-bca3-06b4e28fd70b",
    "knowledge_repo_id": "d3e0e470-0032-4907-ba64-9aba4521ed0c",
    "chat_title": "杭州在哪里",
    "create_date_time": "1695192600724",
    "is_del": 0
  }, {
    "chat_id": "8b8d345e-833c-4b71-9bdd-e4ba75a6457e",
    "knowledge_repo_id": "dbac6a40-6142-4447-b7ad-6718c089faf4",
```

```
"chat_title" : "something",  
"create_date_time" : "1695115924537",  
"is_del" : 0  
}],  
"total" : 222,  
"size" : 10,  
"current" : 1,  
"optimize_count_sql" : true,  
"search_count" : true,  
"pages" : 23  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

10.2 获取对话历史

功能介绍

根据chat_id获取对话历史，返回历史的question和answer数据。

URI

GET /v1/koosearch/chat-history/{chat_history_id}

表 10-5 路径参数

参数	是否必选	参数类型	描述
chat_history_id	是	String	对话历史ID。获取方式：【对话管理】-【对话历史】中的【对话ID】，即为【对话问题】所对应的【对话ID】 最小长度：1 最大长度：64

请求参数

表 10-6 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 10-7 响应 Body 参数

参数	参数类型	描述
chat_id	String	会话ID, 会话的唯一标识 最小长度: 1 最大长度: 64
knowledge_repo_id	String	知识库ID, 当前会话参考的知识库ID 最小长度: 1 最大长度: 64
create_datetime	String	会话发生的时间戳 最小长度: 1 最大长度: 13
is_del	Integer	当前对话是否删除标记, 0: 未删除, 1: 已删除 枚举值: <ul style="list-style-type: none">• 0• 1
question_models	Array of QuestionModel objects	当前会话详情, 包括问题、答案、参考分片信息等 数组长度: 0 - 65535

表 10-8 QuestionModel

参数	参数类型	描述
question_id	String	问题ID 最小长度: 1 最大长度: 64
chat_id	String	会话ID 最小长度: 1 最大长度: 64
question	String	问题 最小长度: 1 最大长度: 65535
answer	String	答案 最小长度: 1 最大长度: 65535
create_date_time	String	对话发生时间 最小长度: 1 最大长度: 13
feedback_status	Integer	是否有问题反馈 枚举值: <ul style="list-style-type: none">• 0• 1
question_reference_models	Array of QuestionReferenceModel objects	当前答案的参考文件和分片详情 数组长度: 0 - 65535

表 10-9 QuestionReferenceModel

参数	参数类型	描述
id	Integer	参考分片ID 最小值: 0 最大值: 2147483647
reference_id	String	参考文件ID 最小长度: 1 最大长度: 64

参数	参数类型	描述
doc_title	String	参考文件名称 最小长度: 1 最大长度: 65535
subtitle	String	参考分片子标题 最小长度: 1 最大长度: 65535
doc_content	String	参考分片内容 最小长度: 1 最大长度: 65535
score	Float	参考分片与问题的相关性得分 最小值: 0.0 最大值: 200.0

请求示例

无

响应示例

状态码: 200

成功

```
{
  "chat_id": "9eafef43-0fc5-4f62-891a-34810ada9f20",
  "knowledge_repo_id": "2e1524cc-cf40-4607-9134-1204f320b868",
  "create_date_time": "1717641946646",
  "is_del": 0,
  "question_models": [ {
    "question_id": "7ae2d834-591a-4a2b-bba5-53802c93998f",
    "chat_id": "9eafef43-0fc5-4f62-891a-34810ada9f20",
    "question": "人身保险业务包括哪些子业务?",
    "answer": "根据提供的文档内容, 人身保险业务包括以下子业务:\n\n-人寿保险: 这是指在被保险人生存期间, 保险人按照合同约定的时间, 向被保险人或者其指定的受益人支付保险金的保险。
\n-健康保险: 这类保险主要是为了减轻因疾病、意外伤害或其他健康问题导致的个人或家庭的经济负担。
\n-意外伤害保险: 这种保险是为了保障被保险人因意外伤害导致身体损害或死亡时, 保险人给予的经济补偿。
\n\n文档中还提到, 经营财产保险业务的保险公司经国务院保险监督管理机构批准, 可以经营短期健康保险业务和意外伤害保险业务。这意味着在特定条件下, 健康保险和意外伤害保险可以与财产保险业务相结合。",
    "create_date_time": "1717641952936",
    "feedback_status": 0,
    "question_reference_models": [ {
      "id": 2141249532,
      "reference_id": "aa1298dfe8a2cfd3f4392019231137d2",
      "doc_title": "中华人民共和国保险法.pdf",
      "subtitle": "中华人民共和国保险法\n中华人民共和国保险法 第四章 保险经营规则(1)",
      "doc_content": "第九十五条 保险公司的业务范围:\n(一) 人身保险业务, 包括人寿保险、健康保险、意外伤害保险等保险业务; \n(二) 财产保险业务, 包括财产损失保险、责任保险、信用保险、保证保险等保险业务; \n(三) 国务院保险监督管理机构批准的与保险有关的其他业务。
\n保险人不得兼营人身保险业务和财产保险业务。但是, 经营财产保险业务的保险公司经国务院保险监督管理机构批准, 可以经营短期健康保险业务和意外伤害保险业务。
\n\n保险公司应当在国务院保险监督管理机构依法批准的业务范围内从事保险经营活动。
\n第九十六条 经国务院保险监督管理机构批准, 保险公司可以经营本法第九十五条规定的保险业务的下列再保险业务:\n(一) 分出保险; \n(二) 分入保险。
\n第九十七条 保险公司应当按照其注册资本总额的百分之二十提取保证
```

```
金, 存入国务院保险监督管理机构指定的银行, 除公司清算时用于清偿债务外, 不得动用。\\n第九十八条 保险公  
司应当根据保障被保险人利益、保证偿付能力的原则, 提取各项责任准备金。\\n保险公司提取和结转责任准备金  
的具体办法,由国务院保险监督管理机构制定。\\n第九十九条 保险公司应当依法提取公积金。\\n第一百条 保险公  
司应当缴纳保险保障基金。",  
  "score": 0.478392  
}, {  
  "id": 2141249533,  
  "reference_id": "aa1298dfe8a2cfd3f4392019231137d2",  
  "doc_title": "中华人民共和国保险法.pdf",  
  "subtitle": "中华人民共和国保险法\\n中华人民共和国保险法 第一节 一般规定(3)",  
  "doc_content": "保险人在合同订立时已经知道投保人未如实告知的情况的, 保险人不得解除合同; 发生保险  
事故的, 保险人应当承担赔偿或者给付保险金的责任。\\n保险事故是指保险合同约定的保险责任范围内的事故。  
\\n第十七条 订立保险合同, 采用保险人提供的格式条款的, 保险人向投保人提供的投保单应当附格式条款, 保险人  
应当向投保人说明合同的内容。\\n对保险合同中免除保险人责任的条款, 保险人在订立合同时应当在投保单、保险  
单或者其他保险凭证上作出足以引起投保人注意的提示, 并对该条款的内容以书面或者口头形式向投保人作出明确  
说明; 未作提示或者明确说明的, 该条款不产生效力。\\n第十八条 保险合同应当包括下列事项:\\n(一) 保险人的  
名称和住所; \\n(二) 投保人、被保险人的姓名或者名称、住所, 以及人身保险的受益人的姓名或者名称、住所;  
\\n(三) 保险标的; \\n(四) 保险责任和责任免除; \\n(五) 保险期间和保险责任开始时间; \\n(六) 保险金  
额; \\n(七) 保险费以及支付办法; \\n(八) 保险金赔偿或者给付办法; \\n(九) 违约责任和争议处理; \\n  
(十) 订立合同的年、月、日。\\n投保人和保险人可以约定与保险有关的其他事项。\\n受益人是指人身保险合同  
中由被保险人或者投保人指定的",  
  "score": 0.0246569  
}, {  
  "id": 2141249534,  
  "reference_id": "aa1298dfe8a2cfd3f4392019231137d2",  
  "doc_title": "中华人民共和国保险法.pdf",  
  "subtitle": "中华人民共和国保险法\\n中华人民共和国保险法 第一章 总则",  
  "doc_content": "第一条 为了规范保险活动, 保护保险活动当事人的合法权益, 加强对保险业的监督管理, 维  
护社会经济秩序和社会公共利益, 促进保险事业的健康发展, 制定本法。\\n第二条 本法所称保险, 是指投保人根  
据合同约定, 向保险人支付保险费, 保险人对于合同约定的可能发生的事故因其发生所造成的财产损失承担赔偿责任  
保险金责任, 或者当被保险人死亡、伤残、疾病或者达到合同约定的年龄、期限等条件时承担给付保险金责任的商  
业保险行为。\\n第三条 在中华人民共和国境内从事保险活动, 适用本法。\\n第四条 从事保险活动必须遵守法律、  
行政法规, 尊重社会公德, 不得损害社会公共利益。\\n第五条 保险活动当事人行使权利、履行义务应当遵循诚实  
信用原则。\\n第六条 保险业务由依照本法设立的保险公司以及法律、行政法规规定的其他保险组织经营, 其他单位  
和个人不得经营保险业务。\\n第七条 在中华人民共和国境内的法人和其他组织需要办理境内保险的, 应当向中华  
人民共和国境内的保险公司投保。\\n第八条 保险业和银行业、证券业、信托业实行分业经营、分业管理, 保险公  
司与银行、证券、信托业务机构分别设立。国家另有规定的除外。\\n第九条 国务院保险监督管理机构依法对保险  
业实施监督管理。\\n国务院保险监督管理机构根据履行职责的需要设立派出机构。派出机构按照国务院保险监督  
管理机构的授权履行监督管理职责。",  
  "score": 0.014672  
  }  
  }  
  }  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

10.3 删除对话历史

功能介绍

根据指定ID, 删除对应的对话历史。

URI

DELETE /v1/koosearch/chat-history/{chat_history_id}

表 10-10 路径参数

参数	是否必选	参数类型	描述
chat_history_id	是	String	对话历史ID 最小长度：1 最大长度：64

请求参数

表 10-11 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

无

请求示例

无

响应示例

无

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

10.4 批量删除对话历史

功能介绍

根据对话ID，批量删除对话历史信息。

URI

DELETE /v1/koosearch/chat-history/batch

请求参数

表 10-12 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 10-13 请求 Body 参数

参数	是否必选	参数类型	描述
chat_ids	是	Array of strings	待删除的会话ID列表 最小长度：1 最大长度：64 数组长度：1 - 65535

响应参数

状态码：200

表 10-14 响应 Body 参数

参数	参数类型	描述
total_count	Integer	需要删除对话历史的总数 最小值：0 最大值：65535

参数	参数类型	描述
deleted_count	Integer	成功删除的对话历史数量 最小值： 0 最大值： 65535

请求示例

```
/v1/koosearch/chat-history/batch  
  
{  
  "chat_ids": [ "dae31d83-36f6-4972-ac50-6034e04c87a2", "be36c420-ee12-4e35-85d4-00aa9039091e" ]  
}
```

响应示例

状态码： 200

成功

```
{  
  "totalCount": 2,  
  "deletedCount": 2  
}
```

状态码

状态码	描述
200	成功

错误码

请参见[错误码](#)。

11 图片管理

[上传图片](#)
[获取图片内容](#)

11.1 上传图片

功能介绍

上传图片入库。

URI

POST /v1/koosearch/image/{img_id}

表 11-1 路径参数

参数	是否必选	参数类型	描述
img_id	是	String	图片ID，格式为img-[0-9a-z]{32} 最小长度：1 最大长度：64

请求参数

表 11-2 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

表 11-3 FormData 参数

参数	是否必选	参数类型	描述
image	是	File	需要上传的图片

响应参数

状态码: 400

表 11-4 响应 Body 参数

参数	参数类型	描述
[数组元素]	Array of ErrorRsp objects	请求错误

表 11-5 ErrorRsp

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

请求示例

无

响应示例

无

状态码

状态码	描述
200	上传成功
400	请求错误
500	内部异常

错误码

请参见[错误码](#)。

11.2 获取图片内容

功能介绍

根据ID获取图片内容。

URI

GET /v1/koosearch/image/{img_id}

表 11-6 路径参数

参数	是否必选	参数类型	描述
img_id	是	String	图片ID，格式为img-[0-9a-z]{32} 最小长度：1 最大长度：64 枚举值： <ul style="list-style-type: none">img-[0-9a-z]{32}

请求参数

表 11-7 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 400

表 11-8 响应 Body 参数

参数	参数类型	描述
[数组元素]	Array of ErrorRsp objects	请求错误

表 11-9 ErrorRsp

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

状态码: 404

表 11-10 响应 Body 参数

参数	参数类型	描述
[数组元素]	Array of ErrorRsp objects	图片不存在

表 11-11 ErrorRsp

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 11-12 响应 Body 参数

参数	参数类型	描述
[数组元素]	Array of ErrorRsp objects	内部异常

表 11-13 ErrorRsp

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

无

响应示例

无

状态码

状态码	描述
200	正常响应
400	请求错误
404	图片不存在
500	内部异常

错误码

请参见[错误码](#)。

12 模型管理

- [新增模型](#)
- [修改模型配置](#)
- [列举模型](#)
- [删除模型](#)

12.1 新增模型

功能介绍

新增配置模型，包括模型名称、模型描述、模型endpoint、模型配置。

URI

POST /v1/koosearch/models

请求参数

表 12-1 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 12-2 请求 Body 参数

参数	是否必选	参数类型	描述
name	是	String	模型名称 最小长度：1 最大长度：32
detail	是	String	模型描述 最小长度：1 最大长度：512
endpoint	是	String	模型endpoint 最小长度：1 最大长度：512
type	是	String	模型类型 pangu_chat 盘古对话模型 pangu_llm 盘古裸机模型 embedding 向量模型 query2query query2query模型 rerank 精排模型 search_plan 搜索规划模型 llm 原生模型 最小长度：1 最大长度：64 枚举值： <ul style="list-style-type: none">• pangu_chat• pangu_llm• embedding• query2query• rerank• search_plan• llm
extend_config	是	ModelExtend Config object	模型的自定义配置，json格式
alias	否	String	昇腾模型场景请求模型需要的 name 最小长度：1 最大长度：32

表 12-3 ModelExtendConfig

参数	是否必选	参数类型	描述
iam_info	是	IamInfo object	iam鉴权委托信息

参数	是否必选	参数类型	描述
custom_header	否	map<string, object>	扩展请求头
deployment_id	否	String	模型的deployment_id信息 最小长度：1 最大长度：128
vector_config	否	VectorConfig object	向量模型的相关配置
nlp_config	否	NlpConfig object	NLP大模型相关配置

表 12-4 iamInfo

参数	是否必选	参数类型	描述
iam_enabled	否	Boolean	是否启用iam鉴权
agency_enabled	否	Boolean	是否使用委托账号进行鉴权
project_id	否	String	iam用户的project_id信息, iam_enable=true是为必填字段 最小长度：1 最大长度：256
domain_name	否	String	委托账号的账号名 最小长度：1 最大长度：256
agency_name	否	String	委托名 最小长度：1 最大长度：256

表 12-5 VectorConfig

参数	是否必选	参数类型	描述
embedding_url	是	String	embedding或query2query模型的调用url 最小长度：1 最大长度：512

参数	是否必选	参数类型	描述
rerank_url	是	String	rerank模型的调用url 最小长度：1 最大长度：512

表 12-6 NlpConfig

参数	是否必选	参数类型	描述
context_length	是	Integer	上下文长度，即NLP大模型在进行一次特定的推理时可以考虑的最大令牌数 最小值：4 最大值：128 缺省值：4

响应参数

状态码：400

表 12-7 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 12-8 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/111/applications/222/uni-search/nlp-models  
  
{  
  "name": "pangu_38b",  
  "detail": "pangu_38b",  
  "type": "nlp",  
  "endpoint": "http://10.0.0.169:31010",  
  "extend_config": {  
    "deployment_id": "123",  
    "iam_info": {  
      "iam_enabled": false,  
      "agency_enabled": false  
    }  
  }  
}
```

响应示例

无

状态码

状态码	描述
200	新增成功
400	请求体错误
500	内部异常

错误码

请参见[错误码](#)。

12.2 修改模型配置

功能介绍

修改模型配置

URI

PUT /v1/koosearch/models/{model_name}

表 12-9 路径参数

参数	是否必选	参数类型	描述
model_name	是	String	模型名称 最小长度：1 最大长度：64

请求参数

表 12-10 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 12-11 请求 Body 参数

参数	是否必选	参数类型	描述
detail	否	String	模型描述 最小长度：1 最大长度：512
endpoint	否	String	模型endpoint 最小长度：1 最大长度：512
extend_config	否	ModelExtendConfig object	模型的自定义配置，json格式
alias	否	String	昇腾模型场景请求模型需要的name 最小长度：1 最大长度：32

表 12-12 ModelExtendConfig

参数	是否必选	参数类型	描述
iam_info	是	IamInfo object	iam鉴权委托信息

参数	是否必选	参数类型	描述
custom_header	否	map<string, object>	扩展请求头
deployment_id	否	String	模型的deployment_id信息 最小长度：1 最大长度：128
vector_config	否	VectorConfig object	向量模型的相关配置
nlp_config	否	NlpConfig object	NLP大模型相关配置

表 12-13 iamInfo

参数	是否必选	参数类型	描述
iam_enabled	否	Boolean	是否启用iam鉴权
agency_enabled	否	Boolean	是否使用委托账号进行鉴权
project_id	否	String	iam用户的project_id信息, iam_enable=true是为必填字段 最小长度：1 最大长度：256
domain_name	否	String	委托账号的账号名 最小长度：1 最大长度：256
agency_name	否	String	委托名 最小长度：1 最大长度：256

表 12-14 VectorConfig

参数	是否必选	参数类型	描述
embedding_url	是	String	embedding或query2query模型的调用url 最小长度：1 最大长度：512

参数	是否必选	参数类型	描述
rerank_url	是	String	rerank模型的调用url 最小长度：1 最大长度：512

表 12-15 NlpConfig

参数	是否必选	参数类型	描述
context_length	是	Integer	上下文长度，即NLP大模型在进行一次特定的推理时可以考虑的最大令牌数 最小值：4 最大值：128 缺省值：4

响应参数

状态码：400

表 12-16 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 12-17 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/111/applications/222/uni-search/models  
  
{  
  "detail": "pangu_38b",  
  "endpoint": "http://10.0.0.169:31010",  
  "extend_config": {  
    "deployment_id": "123",  
    "iam_info": {  
      "iam_enabled": false,  
      "agency_enabled": false  
    }  
  }  
}
```

响应示例

无

状态码

状态码	描述
200	新增成功
400	请求体错误
500	内部异常

错误码

请参见[错误码](#)。

12.3 列举模型

功能介绍

列举模型

URI

GET /v1/koosearch/models/search

表 12-18 Query 参数

参数	是否必选	参数类型	描述
model_name	否	String	模型名称 最小长度：1 最大长度：64
model_type	否	String	模型类型，枚举值： embedding, rerank, nlp, search-plan, query2query 最小长度：1 最大长度：64
model_status	否	String	模型状态 最小长度：1 最大长度：64
page_num	否	Integer	请求页码 最小值：1 最大值：65535 缺省值：1
page_size	否	Integer	请求限定响应结果的分页大小， 例如5条/页，10条/页 最小值：1 最大值：65535 缺省值：10

请求参数

表 12-19 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

响应参数

状态码：200

表 12-20 响应 Body 参数

参数	参数类型	描述
models	Array of ModelInfo objects	模型列表 数组长度: 1 - 65535
total	Integer	返回总条数 最小值: 1 最大值: 65535
page_num	Integer	页码 最小值: 1 最大值: 65535
page_size	Integer	每页返回数量 最小值: 1 最大值: 65535

表 12-21 ModelInfo

参数	参数类型	描述
name	String	模型名称 最小长度: 1 最大长度: 64
detail	String	模型详情 最小长度: 1 最大长度: 512
type	String	模型类型 最小长度: 1 最大长度: 64
endpoint	String	模型endpoint 最小长度: 1 最大长度: 512
status	String	模型状态 最小长度: 1 最大长度: 32
create_time	String	模型创建时间 最小长度: 1 最大长度: 13

参数	参数类型	描述
update_time	String	模型更新时间 最小长度：1 最大长度：13
extend_config	ModelExtendConfig object	模型扩展配置
alias	String	昇腾模型请求时使用的名称 最小长度：1 最大长度：64

表 12-22 ModelExtendConfig

参数	参数类型	描述
iam_info	IamInfo object	iam鉴权委托信息
custom_header	map<string, object>	扩展请求头
deployment_id	String	模型的deployment_id信息 最小长度：1 最大长度：128
vector_config	VectorConfig object	向量模型的相关配置
nlp_config	NlpConfig object	NLP大模型相关配置

表 12-23 IamInfo

参数	参数类型	描述
iam_enabled	Boolean	是否启用iam鉴权
agency_enabled	Boolean	是否使用委托账号进行鉴权
project_id	String	iam用户的project_id信息，iam_enable=true是为必填字段 最小长度：1 最大长度：256

参数	参数类型	描述
domain_name	String	委托账号的账号名 最小长度：1 最大长度：256
agency_name	String	委托名 最小长度：1 最大长度：256

表 12-24 VectorConfig

参数	参数类型	描述
embedding_url	String	embedding或query2query模型的调用url 最小长度：1 最大长度：512
rerank_url	String	rerank模型的调用url 最小长度：1 最大长度：512

表 12-25 NlpConfig

参数	参数类型	描述
context_length	Integer	上下文长度，即NLP大模型在进行一次特定的推理时可以考虑的最大令牌数 最小值：4 最大值：128 缺省值：4

请求示例

无

响应示例

状态码：200

模型列表

```
{
  "page_num": 1,
  "page_size": 1,
  "total": 1,
  "models": [ {
```

```
"name": "pangu_38b",
"detail": "pangu_38b",
"type": "nlp",
"status": "ready",
"create_time": "1716946517167",
"update_time": "1716946517167",
"endpoint": "http://10.0.0.169:31010",
"extend_config": {
  "deployment_id": "123",
  "iam_info": {
    "iam_enabled": false,
    "agency_enabled": false,
    "project_id": "456"
  }
}
}]
}
```

状态码

状态码	描述
200	模型列表
400	请求体错误
500	内部异常

错误码

请参见[错误码](#)。

12.4 删除模型

功能介绍

删除模型

URI

DELETE /v1/koosearch/models/{model_name}

表 12-26 路径参数

参数	是否必选	参数类型	描述
model_name	是	String	模型名称 最小长度：1 最大长度：64

请求参数

表 12-27 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

无

请求示例

无

响应示例

无

状态码

状态码	描述
200	删除成功
400	请求体错误
500	内部异常

错误码

请参见[错误码](#)。

13 用户的文档解析规则定义

[创建用户解析规则](#)

[修改用户解析规则](#)

[列举用户规则列表](#)

[删除用户定义规则解析接口](#)

13.1 创建用户解析规则

功能介绍

用于创建用户自定义的解析规则，解析规则用于解析需要作为subtitle的文本，规则由上而下依次校验。

URI

POST /v1/koosearch/rule-regex

请求参数

表 13-1 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token，Token获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 13-2 请求 Body 参数

参数	是否必选	参数类型	描述
regexs	是	Array of strings	规则list 最小长度：1 最大长度：200 数组长度：1 - 10

响应参数

状态码：200

表 13-3 响应 Body 参数

参数	参数类型	描述
id	String	规则ID 最小长度：1 最大长度：64

状态码：400

表 13-4 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 13-5 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36

参数	参数类型	描述
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

```
/v1/{project_id}/applications/{application_id}/uni-search/rule-regex  
{  
  "regexs": [ "string", "string" ]  
}
```

响应示例

状态码：200

创建成功

```
{  
  "id": "1777619994213965826"  
}
```

状态码

状态码	描述
200	创建成功
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

13.2 修改用户解析规则

功能介绍

用于修改已配置的解析规则。

URI

PUT /v1/koosearch/rule-regex/{id}

表 13-6 路径参数

参数	是否必选	参数类型	描述
id	是	String	用户解析规则ID 最小长度：1 最大长度：64

请求参数

表 13-7 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考： 获取IAM用户Token 最小长度：1 最大长度：1000000

表 13-8 请求 Body 参数

参数	是否必选	参数类型	描述
regexs	是	Array of strings	规则list 最小长度：1 最大长度：200 数组长度：1 - 10

响应参数

状态码：200

表 13-9 响应 Body 参数

参数	参数类型	描述
id	String	修改的规则id 最小长度：1 最大长度：64

状态码：400

表 13-10 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码： 500

表 13-11 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

样例

```
/v1/{project_id}/applications/{application_id}/uni-search/rule-regex/{id}
{
  "regexs" : [ "string", "string" ]
}
```

响应示例

状态码： 200

修改请求

```
{
  "id" : "1777619994213965826"
}
```

状态码

状态码	描述
200	修改请求
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

13.3 列举用户规则列表

功能介绍

列举用户配置的所有解析规则列表。

URI

GET /v1/koosearch/rule-regex

请求参数

表 13-12 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 13-13 响应 Body 参数

参数	参数类型	描述
rules	Array of RuleRegexInfo objects	已配置的规则解析列表 数组长度: 0 - 65535

表 13-14 RuleRegexInfo

参数	参数类型	描述
id	String	规则ID 最小长度：1 最大长度：65535
rule_regexs	Array of strings	用户自定义规则内容 最小长度：1 最大长度：65535 数组长度：1 - 65535

状态码：400

表 13-15 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

状态码：500

表 13-16 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度：8 最大长度：36
error_msg	String	错误描述 最小长度：2 最大长度：512

请求示例

无

响应示例

状态码： 200

用户定义规则列表

```
{
  "rules": [{
    "id": "string",
    "rule_regexs": [ "string" ]
  }]
}
```

状态码

状态码	描述
200	用户定义规则列表
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

13.4 删除用户定义规则解析接口

功能介绍

用于删除用户自定义解析规则。

URI

DELETE /v1/koosearch/rule-regex/{id}

表 13-17 路径参数

参数	是否必选	参数类型	描述
id	是	String	用户规则ID 最小长度：1 最大长度：64

请求参数

表 13-18 请求 Header 参数

参数	是否必选	参数类型	描述
X-Auth-Token	是	String	接口鉴权使用的Token, Token 获取方式参考: 获取IAM用户Token 最小长度: 1 最大长度: 1000000

响应参数

状态码: 200

表 13-19 响应 Body 参数

参数	参数类型	描述
id	String	需要删除的规则ID 最小长度: 1 最大长度: 64

状态码: 400

表 13-20 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

状态码: 500

表 13-21 响应 Body 参数

参数	参数类型	描述
error_code	String	错误码 最小长度: 8 最大长度: 36
error_msg	String	错误描述 最小长度: 2 最大长度: 512

请求示例

无

响应示例

状态码: 200

删除成功

```
{  
  "id": "1777619994213965826"  
}
```

状态码

状态码	描述
200	删除成功
400	请求体参数错误
500	内部异常

错误码

请参见[错误码](#)。

14 公共参数

状态码
错误码

14.1 状态码

状态码如表14-1所示。

表 14-1 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。

状态码	编码	状态码说明
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。
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。

状态码	编码	状态码说明
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。
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	Unprocessable Entity	请求格式正确，但是由于含有语义错误，无法响应。
429	Too Many Requests	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	Service Unavailable	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	Server Timeout	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。

14.2 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	KOS.00010001	JSON格式异常	JSON格式异常。	根据具体提示信息修改参数。
400	KOS.00010002	请求参数不合法	请求参数不合法。	根据具体提示信息修改参数。
400	KOS.00010003	请求参数为空	请求参数为空。	根据具体提示信息修改参数。
400	KOS.00010004	url参数解析异常	url参数解析异常。	根据具体提示信息修改参数。
400	KOS.00010005	url路径不合法	url路径不合法。	参考文档进行url路径修改。
400	KOS.00010006	参数类型异常	参数类型异常。	根据具体提示信息修改参数。
400	KOS.00020001	es查询异常	es查询异常。	建议检查es节点状态。
400	KOS.00020002	es索引创建异常	es索引创建异常。	建议检查es节点状态。
400	KOS.00020003	es更新异常	es更新异常。	建议检查es节点状态。
400	KOS.00020004	es删除索引异常	es删除索引异常。	建议检查es节点状态。
400	KOS.00020005	es批量删除异常	es批量删除异常。	建议检查es节点索引状态。
400	KOS.00020006	es删除索引异常	es删除索引异常。	建议检查es节点索引状态。
400	KOS.00020007	es批量创建索引异常	es批量创建索引异常。	建议检查es节点状态。
400	KOS.00020008	es删除文档异常	es删除文档异常。	建议检查es节点状态。
400	KOS.00020009	es更新文档异常	es更新文档异常。	建议检查es节点状态。
400	KOS.00020010	es批量操作异常	es批量操作异常。	建议检查es节点状态。

状态码	错误码	错误信息	描述	处理措施
400	KOS.00020011	连接es失败	连接es失败。	建议检查es节点状态。
400	kos.00020013	更新es索引配置异常	更新es索引配置异常。	建议检查es节点状态。
400	KOS.00040001	调用盘古nlp模型异常	调用盘古nlp模型异常。	模型状态异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00040002	调用多轮改写服务异常	调用多轮改写服务异常。	模型状态异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00040003	调用文档解析服务异常	调用文档解析服务异常。	文档解析服务异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	kos.00050003	文件信息写入数据库异常	文件信息写入数据库异常。	服务数据库异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。

状态码	错误码	错误信息	描述	处理措施
400	kos.00050004	上传obs异常	上传obs异常。	obs服务异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	kos.00050005	json处理异常	json处理异常。	请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00050006	处理中的文件不支持删除	处理中的文件不支持删除。	文件正在处理中，请稍后重试。
400	kos.00060001	数据库查询结果为空	数据库查询结果为空。	数据库查询结果为空。
400	kos.00060002	数据库操作失败	数据库操作失败。	服务数据库异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	kos.00060003	数据库更新异常	数据库更新异常。	服务数据库异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	kos.00060004	操作资源不存在	操作资源不存在。	资源不存在。

状态码	错误码	错误信息	描述	处理措施
400	kos.00060005	模型不可用	模型不可用。	模型可能存在异常，请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00070004	下载json文件异常	下载json文件异常。	请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00070005	写入json文件异常	写入json文件异常。	请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.00080001	同名知识库已存在	同名知识库已存在。	知识库已经存在。
400	KOS.00080002	知识库为空时不支持开启	知识库为空时不支持开启。	知识库为空。
400	KOS.00080003	知识库关闭时不允许当前操作	知识库关闭时不允许当前操作。	知识库为关闭状态。
400	KOS.00080004	对话时间超期	对话时间超期。	建议重新开启新对话。
400	KOS.00080005	知识库更新失败	知识库更新失败。	请查看服务日志，根据具体问题分析服务异常。

状态码	错误码	错误信息	描述	处理措施
400	KOS.0009000 1	内部异常	内部异常。	请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。
400	KOS.0010000 2	权限不足	权限不足。	请联系技术支持。您可以登录华为云官网，在华为云首页，单击右上角的“联系我们”，反馈您所遇到的问题，技术支持会快速为您定位问题，解决问题。