企业搜索服务

用户指南

文档版本	01
发布日期	2025-08-21

华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2025。 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云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 KooSearch 文档问答服务使用流程 1

2 开通 KooSearch 文档问答服务 3

3 在控制台使用 KooSearch 实现搜索问答 9

3.1 创建和管理 KooSearch 模型服务（可选） 9

3.2 创建及修改 KooSearch 知识库 12

3.3 将本地数据上传至 KooSearch 知识库 20

3.4 体验 KooSearch 问答 25

3.5 体验 KooSearch 搜索 30

3.6 体验 AI 搜索 31

3.7 管理 KooSearch 知识库 32

3.8 管理 KooSearch 提示词 36

3.9 管理 KooSearch 对话 37

4 通过 API 使用 KooSearch 实现搜索问答 39

5 升级 KooSearch 服务 45

6 管理 KooSearch 文档问答服务 48

6.1 查看 KooSearch 文档问答服务详情 48

6.2 配置 KooSearch 文档问答服务集群路由 50

6.3 删除 KooSearch 文档问答服务 51

7 KooSearch 文档问答服务日志管理 52

1 KooSearch 文档问答服务使用流程

华为云企业搜索KooSearch是基于华为云的云搜索服务搭建的一站式智能搜索解决方案，帮助企业聚焦业务场景和应用开发，场景服务化、技术简单化、低门槛化，满足开发者基于业务场景的二次开发。在RAG（Retrieval-Augmented Generation）及搜索场景提供效果和性能突出的组件化服务，架构理想、灵活编排的机制，帮助企业客户快速构建RAG以及搜索服务。

说明

仅“西南-贵阳一”区域支持开通和使用KooSearch服务。KooSearch是受限使用阶段，如果有试用需求，请提工单申请权限。

KooSearch服务使用流程，如下图所示：

图 1-1 koosearch 使用流程

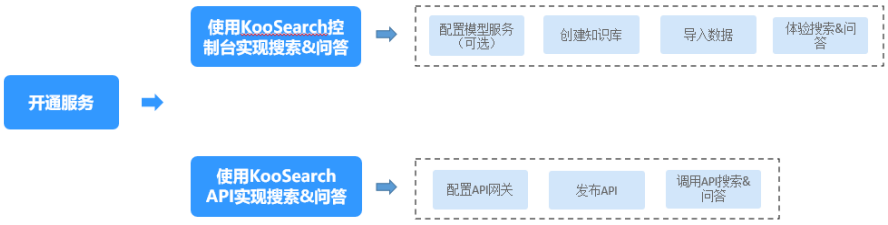


表 1-1 使用流程

步骤	操作	说明
1	开通服务	首先需要开通服务，开通服务时会选择版本规格、配置一系列参数来创建一个实例，后续可以用此实例实现搜索问答，具体请看开通KooSearch文档问答服务。

版权所有 © 华为云计算技术有限公司

步骤	操作	说明
2	在控制台使用KooSearch服务实现搜索文档	开通服务后，您可以在KooSearch控制台实现搜索问答，具体操作如下： 1. 当KooSearch服务管理员需要自定义模型服务时，可以**创建和管理KooSearch模型服务（可选）**，否则跳过该步骤。 2. **创建KooSearch知识库**。 3. **将本地文件上传至KooSearch知识库**。 4. 使用KooSearch服务进行问答和搜索。 ● **体验KooSearch问答**。 ● **体验KooSearch搜索**。 5. 管理知识库。
	使用KooSearch的API实现搜索文档	您也可以使用调用API的方式实现搜索问答，KooSearch服务提供的API支持发布到不同的环境，发布成功后支持被调用，具体操作如下： 1. **配置API网关**。 2. **发布KooSearch API**。 3. **调用已发布的KooSearch API**。 4. **编辑API**。 5. **下线API**。
3	管理KooSearch服务	在服务的基本信息页面，可以获取服务的内网访问文档解析地址、内网访问知识管理地址、计费模式等信息。除此之外，还能进行管理服务、API管理和日志管理。具体请看**管理KooSearch知识库**。
4	查看KooSearch服务日志	为了方便用户使用日志定位问题，KooSearch服务提供了日志查询功能。用户可以通过日志查询进行问题分析定位。具体请看**KooSearch文档问答服务日志管理**。

版权所有 © 华为云计算技术有限公司

2 开通 KooSearch 文档问答服务

KooSearch支持使用文档上传并进行知识问答。在使用KooSearch文档问答前需要开通服务。

说明

仅“西南-贵阳一”区域支持开通和使用KooSearch服务。KooSearch是受限使用阶段，如果有试用需求，请提工单申请权限。

进入 KooSearch 控制台

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

开通 KooSearch 服务

1. 进入KooSearch页面后，单击“服务开通”，进入服务开通页面。
2. 填写KooSearch服务的服务配置。
   - 基础配置

表 2-1 基础配置参数说明

参数	说明
计费模式	服务支持包年/包月和按需计费两种模式。 ● 按需计费：按实际使用时长计费，计费周期为一小时，不足一小时按一小时计费。 ● 包年/包月：根据服务购买时长，一次性支付集群费用。最短时长为1个月，最长时长为1年。
订购周期	选择包年/包月计费模式后，需要选择购买时长。 您可以根据需求，选择是否需要自动续费。
产品规格	支持敏捷版、基础版、专业版、企业版。

版权所有 © 华为云计算技术有限公司

参数	说明
当前区域	下拉框中选择KooSearch服务的工作区域。 不同区域的资源之间内网不互通，请选择靠近您业务的区域，可以降低网络时延，提高访问速度。
服务名称	自定义服务名称。

– 创建向量集群

负责业务数据的存储与查询。当前仅支持创建云搜索服务中Elasticsearch 7.10.2版本的集群。配置集群参数，开通文档问答服务后，会根据集群配置自动创建集群。

表 2-2 向量集群参数说明

参数	说明
节点数量	向量集群中的节点个数。可选节点数为1~32，建议节点数为3或3以上，以提升集群可用性。
CPU架构	目前支持“X86计算”和“鲲鹏计算”两种类型。具体支持的类型由实际区域环境决定。
可用区	选择集群工作区域下关联的可用区。
节点规格	选择集群的节点规格。规格的详细说明可参考弹性云服务器的实例类型与规格。
节点存储	选择节点的存储类型，支持普通I/O、高I/O、超高I/O。
节点存储容量	选择节点的存储空间大小，其取值范围与节点规格关联，不同的规格允许的取值范围不同。 节点存储容量只支持配置为20的倍数。

版权所有 © 华为云计算技术有限公司

参数	说明
磁盘加密	选择节点的数据盘是否进行KMS加密存储。 启用磁盘加密可以加强集群节点中存储的数据安全性。默认为不启用。 启用磁盘加密后，需要配置“密钥名称”，即在下拉框中选择启用状态的KMS密钥，如果没有可用的密钥，可以单击“创建密钥”跳转至数据加密控制台，创建或修改密钥。详细操作指导请参见创建密钥。 **说明** ● 仅云盘支持磁盘加密，本地盘不支持。 ● 仅支持使用“密钥算法”为AES或SM4、“密钥用途”为“ENCRYPT_DECRYPT”的自定义密钥，“密钥名称”下拉框中不可见的KMS密钥表示集群不支持。 ● 磁盘加密不会改变集群的管理和运维操作流程，但数据加解密环节会增加系统处理开销，可能会对操作性能产生一定影响。 ● 集群创建成功后不支持重新开启或关闭磁盘加密。 ● 集群创建成功后不支持修改密钥。 ● 当集群使用的KMS密钥处于非启用状态时，集群扩容、升级、节点规格变更、指定节点替换及可用区切换的操作将失败，需通过新建集群并完成数据迁移实现业务变更。
安全模式	选择是否开启集群安全模式。 ● 默认开启，则创建的是安全模式的集群。安全模式的集群会对集群进行通讯加密和安全认证。因此必须配置集群的“管理员账户名”和“管理员密码”。 – **管理员账户名**默认为admin。 – 设置并确认**管理员密码**。要记住设置的密码，后续访问集群需要输入密码。 ● 不开启，则创建的是非安全模式的集群。非安全模式的集群无需安全认证即可访问，并且采用HTTP明文传输数据。建议确认访问环境的安全性，勿将访问接口暴露到公网环境上。

– 搜索模型配置（可选）

  ▪ 开关打开：开通KooSearch文档问答服务时会自动创建文本向量模型服务和精排模型服务。

  ▪ 开关关闭：开通KooSearch文档问答服务就不会自动创建文本向量模型服务和精排模型服务，会影响KooSearch知识库的使用。如有需要，可以在模型管理页面自定义配置“搜索Embedding模型”、“搜索精排模型”和“缓存生成模型”，具体操作请参见创建和管理KooSearch模型服务（可选）。

版权所有 © 华为云计算技术有限公司

表 2-3 搜索模型参数说明

参数	说明
文本向量和精排推理实例	选择文本向量和精排推理实例个数。
模型分类	选择搜索模型的模型分类。
实例规格	选择搜索模型的实例规格，当前支持“昇腾算力”和“GP加速型”。
可用区	当“实例规格”选择“GP加速型”时，选择搜索模型工作区域下关联的可用区。
节点规格	选择搜索模型的节点规格。规格的详细说明可参考**弹性云服务器的实例类型与规格**。
节点存储	当“实例规格”选择“GP加速型”时，需要选择节点的存储类型，支持普通I/O、高I/O、超高I/O。不同可用区和实例规格所支持的存储类型不同，具体支持的存储类型由实际区域环境决定。
节点存储容量	当“实例规格”选择“GP加速型”时，需要设置节点的存储空间大小，其取值范围与节点规格关联，不同的规格允许的取值范围不同。

– 搜索规划配置（可选）

  - 开关打开：开通KooSearch文档问答服务时会自动创建搜索规划的模型服务，该模型服务提供意图识别和query改写功能。
  - 开关关闭：开通KooSearch文档问答服务就不会自动创建搜索规划的模型服务，会影响KooSearch知识库的使用。如有需要，可以在模型管理页面自定义配置“搜索规划模型”，具体操作请参见**创建和管理KooSearch模型服务（可选）**。

表 2-4 搜索规划参数说明

参数	说明
搜索规划推理实例	选择搜索规划推理实例个数。
模型分类	选择搜索规划的模型分类。
实例规格	选择搜索规划的实例规格，当前支持“昇腾算力”和“GP加速型”。
可用区	当“实例规格”选择“GP加速型”时，需要选择搜索规划推理实例工作区域下关联的可用区。
节点规格	选择搜索规划的节点规格。规格的详细说明可参考**弹性云服务器的实例类型与规格**。

版权所有 © 华为云计算技术有限公司

参数	说明
节点存储	当“实例规格”选择“GP加速型”时，需要选择节点的存储类型，支持普通I/O、高I/O、超高I/O。不同可用区和实例规格所支持的存储类型不同，具体支持的存储类型由实际区域环境决定。
节点存储容量	当“实例规格”选择“GP加速型”时，需要设置节点的存储空间大小，其取值范围与节点规格关联，不同的规格允许的取值范围不同。

– 大模型配置（可选）

- 开关打开：开通KooSearch文档问答服务时会自动创建预置的LLM大模型服务。
- 开关关闭：开通KooSearch文档问答服务就不会自动创建预置的LLM大模型服务，会影响KooSearch问答功能的使用。如有需要，可以在模型管理页面自定义配置“NLP模型-*”，具体操作请参见**创建和管理KooSearch模型服务（可选）**。

表 2-5 大模型配置参数说明

参数	说明
模型版本	当前提供koosearch-rag，基于开源大模型SFT优化后，构建的检索增强大模型，提升检索增强生成场景下的准确率。
生成模型推理实例	选择生成模型推理实例个数。
模型分类	选择生成模型推理实例的模型分类。
实例规格	选择生成模型推理实例规格，当前支持“昇腾算力”和“GP加速型”。
可用区	当“实例规格”选择“GP加速型”时，选择生成模型推理实例工作区域下关联的可用区。
节点规格	选择大模型的节点规格。规格的详细说明可参考**弹性云服务器的实例类型与规格**。
节点存储	当“实例规格”选择“GP加速型”时，需要选择节点的存储类型，支持普通I/O、高I/O、超高I/O。不同可用区和实例规格所支持的存储类型不同，具体支持的存储类型由实际区域环境决定。
节点存储容量	当“实例规格”选择“GP加速型”时，需要设置节点的存储空间大小，其取值范围与节点规格关联，不同的规格允许的取值范围不同。

– 所属企业项目

如果您开通了“企业项目”，在创建集群时，可以给服务绑定一个企业项目。您可以在右侧下拉框中选择当前用户下已创建的企业项目，也可以通过

版权所有 © 华为云计算技术有限公司

单击“查看项目管理”按钮，前往“企业项目管理”管理控制台，新建企业项目和查看已有的企业项目。

3. 单击“下一步”，设置服务的网络配置。

表 2-6 参数说明

参数	说明
虚拟私有云	VPC即虚拟私有云，是通过逻辑方式进行网络隔离，提供安全、隔离的网络环境。 选择创建集群需要的VPC，单击“查看虚拟私有云”进入VPC服务查看已创建的VPC名称和ID。如果没有VPC，需要创建一个新的VPC。 **说明** 此处您选择的VPC必须包含网段（CIDR），否则集群将无法创建成功。新建的VPC默认包含网段（CIDR）。
子网	通过子网提供与其他网络隔离的、可以独享的网络资源，以提高网络安全。 选择创建集群需要的子网，可进入VPC服务查看VPC下已创建的子网名称和ID。
安全组	指定集群使用的安全组，安全组起着虚拟防火墙的作用，为集群提供安全的网络访问控制策略。 选择合适的安全组，单击“查看安全组”可以跳转到安全组列表，了解安全组详情。 所选安全组的入方向规则中，“协议端口”必须包含30275和30277的端口范围，否则外部业务访问可能会异常。

4. 单击“下一步”，确认配置信息。
5. 单击“确认开通”。

   系统跳转至“KooSearch文档问答”页面，您开通的服务将展现在服务列表中，且服务状态为“创建中”，创建成功后服务状态会变为“可用”。

   如果服务开通失败，请根据界面提示，重新开通服务。

版权所有 © 华为云计算技术有限公司

3 在控制台使用 KooSearch 实现搜索问答

3.1 创建和管理 KooSearch 模型服务（可选）

场景描述

您可以在模型管理页面创建不同的模型服务。创建好模型服务后，可以在创建知识库时，选择您建好的模型。也可以在体验问答或搜索的时候使用您配置建好的模型，使答案更接近您想要的结果。

创建模型服务

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“配置管理 > 模型管理”，进入“模型管理”页面。
3. 单击页面的“新建模型服务”，弹出新建模型服务页面。

图 3-1 新建模型服务

4. 在新建模型服务页面。根据下表填写对应参数后，单击“确定”按钮。

表 3-1 新建模型服务

参数	说明
模型服务名称	输入模型服务名称，不能为空。

版权所有 © 华为云计算技术有限公司

参数	说明
模型类型	● NLP模型-云底座：通过华为云提供的盘古nlp大模型访问方式。 ● NLP模型-裸机：通过裸机部署提供的盘古nlp大模型访问方式。 ● 搜索Embedding模型：搜索向量化模型，支持将文本转化成向量。 ● 搜索精排模型：对搜索的召回结果进行重排序，提升向量检索的效果。 ● 搜索规划模型：提供多轮改写及意图识别功能。 ● 审核模型：提供审核服务，审核query、answer是否合规。仅可创建一个审核模型。 ● OCR模型：提供文字识别服务，将图片、扫描件或PDF、OFD文档中的文字识别成可编辑的文本。 ● NLP模型-昇腾云：通过昇腾云的MAAS服务提供的nlp大模型访问方式。如果选择此模型进行问答，建议设置模型生成最大新词数不超过512。 ● 缓存生成模型：提供query之间相似度的计算，用于知识库的缓存功能。 ● web搜索引擎服务：客户自定义的搜索引擎，提供联网搜索服务。 ● 联网增强服务：提供联网增强服务。 **说明** ● embedding模型与缓存生成模型之间存在强关联关系。在创建embedding模型时，系统会配套生成对应的缓存生成模型，若其中一个模型配置信息因意外删除，需根据相同的配置参数进行重建。例如，若embedding模型的名称为pangu_embedding，则其对应的缓存生成模型名称为pangu_embedding_faq。 ● 在创建知识库时，需要依赖embedding模型（pangu_embedding）与缓存生成模型（pangu_embedding_faq）。若缓存生成模型（pangu_embedding_faq）不存在或未授权，系统将抛出异常。此时，需由管理员检查pangu_embedding_faq模型是否存在或确认相关权限是否已授予知识库使用者。若模型缺失，需补充创建相应的pangu_embedding_faq模型；若权限不足，需为知识库使用者授予pangu_embedding_faq相关权限。
访问地址	模型的内网访问地址及端口。
是否启用	如果模型类型选择“审核模型”，会出现“是否启用”按钮。 如果启用，在“体验平台”问答的时候，模型将会审核问答中的query、answer是否合规。如果有敏感词，系统将拒答并返回默认提示。
模型描述	模型详细信息的描述。
昇腾云模型名称	如果模型类型选择“NLP模型-昇腾云”，则需要填写昇腾云模型名称。来源于昇腾云服务开通的NLP大模型的模型名称。

版权所有 © 华为云计算技术有限公司

参数	说明
上下文长度（K）	如果模型类型选择“NLP模型-云底座”、“NLP模型-裸机”，则需要填写上下文长度。 上下文长度是指：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，生成结果可以扩展上下文以生成更全面的响应。
部署ID	如果模型类型选择“NLP模型-云底座”、“NLP模型-裸机”，则需要填写部署ID。 部署ID是指：模型的部署ID信息。
认证类型	IAM认证：支持华为iam认证，系统将默认使用css资源租户进行认证。开启使用委托账号后可以通过配置委托名和委托账号的方式使用委托账号的权限进行认证。
	自定义认证：支持在调用时添加自定义请求头。
URL	设置KooSearch依赖集群：服务类型是【Embedding和rerank模型】的API管理地址。URL可以从独享集群处获取。
是否开启定时检测	定时检测模型的连通性是否正常，并将结果更新到模型列表中。

5. 单击“确定”按钮。如果选择的是NLP模型，会弹出“新建模型服务声明”，勾选同意复选框，再单击“确认”。

图 3-2 免责声明

6. 创建成功后，可以在模型管理页面中看到创建的模型服务。单击模型名称可以查看模型服务的基本信息。

编辑、删除模型服务

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“模型管理”，进入“模型管理”页面。

版权所有 © 华为云计算技术有限公司

3. 选择需要操作的模型。

   单击操作列的“编辑”，可编辑模型服务，具体参数设置可以参照表1 新建模型服务参数。

   单击操作列的“删除”，可删除模型服务。

3.2 创建及修改 KooSearch 知识库

在使用KooSearch体验平台时，首先需要新建知识库，才能进行后续的上传数据、搜索和问答体验等操作。

进入 KooSearch 控制台

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

新建知识库

1. 在KooSearch控制台，左侧导航栏选择“知识库管理”。

   进入知识库管理页面。
2. 在知识库管理页面，单击右上角“新建知识库”。

   在“新建知识库”页面设置知识库信息。
3. 在创建知识库页签填写参数，单击“下一步”。

表 3-2 新建知识库

参数	说明
知识库名称	知识库的名称。只能包含1到64位英文字母、中文、数字、中划线或者下划线，并且以字母、数字或者中文开头。
知识库语言	选择知识库使用的语言。目前支持以下几种语言： ● 中文 ● 英语 ● 泰语 ● 阿拉伯语 ● 西班牙语 ● 葡萄牙语
描述	关于此知识库的基本描述，最多可输出100个字符。

版权所有 © 华为云计算技术有限公司

参数	说明
知识库标签	添加知识库的标签，来区分知识库之间的不同，可以按照知识库标签去查找知识库，也可以按照知识库标签去授权给不同用户去使用。 ● 键名：自定义。 ● 键值：自定义。
结构化数据自定义字段	如果您需要自定义结构化数据的字段，可以在这里添加字段名和字段值，知识库创建成功后，上传结构化数据到知识库时，可以按照自定的字段去上传文件。

4. 在“解析拆分设置”页签配置解析设置和拆分设置，然后单击“下一步”。
   - **解析设置**：勾选需要解析的能力。

     表 3-3 解析设置

参数	说明
OCR增强	勾选后，即可调用OCR服务进行智能文档识别，如表格解析或扫描文件等。
解析图片	未勾选，在文档中遇到图片默认跳过，不处理图片。 勾选后，有两种解析方式可供选择： ● 提取图片文本：识别图片内文字。 ● 仅保留原图：将图片提取后上传OBS桶，便于问答图文展示。
解析页眉页脚	未勾选，解析结果中不包含页眉页脚。 勾选后，解析结果中包含页眉页脚。
解析目录页	未勾选，解析结果中不包含目录页。 勾选后，解析结果中包含目录页。

   - **拆分设置**：即分段设置，选择分段方式。

     表 3-4 拆分设置

参数	说明
自动分段	系统根据文档特点自动选择合适的分段方式。

版权所有 © 华为云计算技术有限公司

参数	说明
长度分段	默认按照段落进行拆分合并，如果段落过长则通过标识符进行分段。 ● 分段标识符：分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。自定义分段中如果未命中分段标识符，分段将会失败。 ● 分段预计长度：分段的最大长度，文档的正文如果超过设定的[最大长度]，则截取[最大长度]的片段为新文档，随后回溯[分段重叠]字符，继续向后检查，直到文档结束。
层级分段	先按照文章的标题层级分段，再按照段落进行拆分合并，如果段落过长则通过标识符进行分段。 层级解析模式：可选择自动解析和规则解析。选择规则解析需要自定义层级规则。 层级分段详情如**表3-5**所示。

表 3-5 层级分段

参数	说明
层级解析模式	自动解析：按照系统规则自动解析。
	规则解析： 由于不同文档的层次结构多样且不一致，针对不同的文档可自定义其文档层次解析规则，更好地解析切分文档从而提升基于文档知识问答的准确率。 ● 自定义默认规则 将最常见的规则，作为默认规则可选，详情请参见提供的**默认规则示例**。 ● 自定义解析规则 当前解析规则采用正则语言编写，可参见如**表3-7**示例。
标题层级深度	选择文章的标题层级深度。
标题保存方式	可选择“保存多标题组合”和“保存最后一级标题”。
分段标识符	分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。自定义分段中如果未命中分段标识符，分段将会失败。
分段预计长度	分段的最大长度，文档的正文如果超过设定的[最大长度]，则截取[最大长度]的片段为新文档，随后回溯[分段重叠]字符，继续向后检查，直到文档结束。

版权所有 © 华为云计算技术有限公司

参数	说明
跨标题合并	打开“跨标题合并”开关：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。 关闭“跨标题合并”开关：不会自动合并不同标题。 **说明** ● “层级分段”页签中有此按钮，可以自己设置开关。 ● “自动分段”页签中没有此按钮，在“自动分段”中跨标题合并功能默认打开。 ● “长度分段”不涉及此功能。

表 3-6 规则解析默认规则示例

类别	规则	描述
第一章 第一节 第一条	^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	以章的规则为例： ● 中括号内大写的阿拉伯可以匹配，例如：第一章。 ● 支持1-9的阿拉伯数字匹配，例如： 第1章。 ● 最大支持中间位数出现的位数有7位。例如：第一千一百三十七章。 节和条的规则类似。

表 3-7 自定义规则解析示例

类别	规则	描述
第一章 第一节 第一条	^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	/
1 **1.1** **1.1.1**	^(\d+\.)(?=\s) ^(\d+)(\.\d+)(?!\.)(?=\s) ^(\d+)(\.\d+)(\.\d+)(?!\.)(?=\s)	可以匹配数字开头的段落。 备注： [\u4e00-\u9fa5]+限制中文 ） 例如： 1. 简介 1.1 说明 1.1.1 详细说明

版权所有 © 华为云计算技术有限公司

类别	规则	描述
一、 （一） 1、 1）	^([零〇一二三四五六七八九十百千万]+)[、] ^（([零〇一二三四五六七八九十百千万]+)） ^[\d+]、 ^[\d+]）	/

5. 在“模型设置”页签配置好模型后，单击“下一步”。

表 3-8 模型设置

参数	说明
搜索模型设置	● Embedding模型服务：基于盘古大模型技术的文本表示模型，将文本转化为用数值表示的向量形式，用于文本检索、聚类、推荐等场景。 ● 精排模型服务：基于盘古大模型技术的文本表示模型，将文本转化为用数值表示的向量形式，用于文本检索、聚类、推荐等场景，语义搜索场景下，加入了精排模型，提升搜索的效果。 ● 搜索规划模型服务：搜索规划模型服务提供了意图分类、多轮查询改写、复杂查询分解、时间抽取等功能，在搜索增强生成任务中，通过意图分类的结果将路由到后续不同的流程；通过改写查询词及查询分解以提高搜索的准确率。 **说明** ● embedding模型与缓存生成模型之间存在强关联关系。在创建embedding模型时，系统会配套生成对应的缓存生成模型，若其中一个模型配置信息因意外删除，需根据相同的配置参数进行重建。例如，若embedding模型的名称为pangu_embedding，则其对应的缓存生成模型名称为pangu_embedding_faq。 ● 在创建知识库时，需要依赖embedding模型（pangu_embedding）与缓存生成模型（pangu_embedding_faq）。若缓存生成模型（pangu_embedding_faq）不存在或未授权，系统将抛出异常。此时，需由管理员检查pangu_embedding_faq模型是否存在或确认相关权限是否已授予知识库使用者。若模型缺失，需补充创建相应的pangu_embedding_faq模型；若权限不足，需为知识库使用者授予pangu_embedding_faq相关权限。
NLP模型设置	NLP模型服务：选择NLP模型服务。基于盘古大模型的人工智能语言模型，可进行对话互动、回答问题、协助创作。
	扩展长上下文： 如果打开了此按钮，模型在解析过程中会扩展长上下文以生成更全面结果。同时，需要设置有效输入长度，输入令牌的有效长度以保证最佳输出。
AI搜索设置	搜索服务类型：选择“web搜索引擎服务”或“联网增强服务”。 搜索服务选择：选择可用的搜索引擎服务。 深度思考模型：选择支持深度思考的模型。

版权所有 © 华为云计算技术有限公司

6. 进入“高级设置”页面，设置好后，单击“确定”。

表 3-9 高级设置

参数	说明
引用定位	打开了此按钮，可以针对回答结果定位到原文位置。
图文结合	展示原文引用关联图片。打开后，有三种解析方式可供选择： 1. 仅召回语义相关图片：引用文段中图片的上下文与生成文段语义相似即召回，否则不召回；（默认选项）。 2. 所有图片：引用文段中图片全部召回。 3. AI召回：使用大模型的能力进行图片召回。 **说明** ● 启用需勾选解析拆分设置->解析设置->解析图片，并选择仅保留原图。 ● 如果是修改知识库配置，历史版本或不兼容该功能，如果希望正常使用该功能，在确保购买了图文溯源相关服务后，修改文档解析方式-仅保留原图，按文档最新配置进行版本重构或者选择所需文档进行重试，生成可供召回的图片数据。 ● 当前除中文知识库外，其他语言的知识库仅支持AI召回。
表格问答	如果打开，可以将文档转成数据表，通过NL2SQL问答提升统计分析类的准确率。
知识库缓存	如果打开知识库缓存按钮，会将问答时的内容缓存，后期对相似问题进行检索问答时，您的搜索效率会相对高效。使用知识库缓存需要选择以下几个参数。 ● 缓存生成模型服务：选择一个模型服务。 ● 缓存阈值：达到缓存阈值，就会使用缓存，输入值必须在0.1到1之间。 ● 缓存策略：达到缓存阈值，如果有多个答案，可以设置最高分或者随机。 ● 过期策略配置：缓存过期的方式，有3种选择。 – Least Recently Used：根据当前时间与最后一次访问时间的差值超过存活时间时删除。 – First In First Out：根据当前时间与创建时间的差值超过存活时间时删除。 – Least Frequency Used：小于缓存命中阈值且当前时间与创建时间大于存活时间时清除，大于阈值时保留 ● 存活时间（秒）：可以自己设置缓存的存活时间，或者直接设置成永久。
目录管理	开启后将使用默认的目录管理功能管理知识库中的文档。 **注意** 如果已对目录管理进行二次开发开启后会导致原目录管理数据被覆盖。

知识库创建好后，可以在知识库管理页面查看到新创建的知识库基本信息，包括知识库ID、知识库名称、知识库状态等信息。

版权所有 © 华为云计算技术有限公司

修改知识库设置

针对已创建的知识库，支持修改知识库设置。

注意

修改知识库“解析拆分设置”后，仅对重试及最新上传的文档生效。

1. 在KooSearch控制台，左侧导航栏选择“知识库管理”。
   进入知识库管理页面。
2. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”。
   进入文档管理页面。
3. 单击右上角的“设置”，修改解析拆分设置和更多设置。
   - 解析拆分设置
     参考表3-3与表3-4修改设置。
   - 召回策略
     召回策略分为文本召回策略和FAQ召回策略。

表 3-10 召回策略

参数	说明
文本召回策略	是指在文档中搜索时，生成结果的策略。包含语义检索、混合检索、关键词检索。 ● 语义检索：切片使用向量检索技术，FAQ使用querytoquery相似检索技术。 ● 混合检索：切片使用向量检索和关键词检索混合检索，FAQ使用querytoquery相似检索技术。 ● 关键词检索：切片使用倒排检索技术，FAQ使用querytoquery相似检索技术。
	语义检索topk召回数量：是指语义搜索生成的片段数量。语义检索topk未配置时，将使用默认值50。 关键字topK召回数量：是指搜索生成的片段数量。 FAQ检索召回个数：通过querytoquery相似检索得到相似得分，按照配置个数进行截断检索召回默认值2。
	精排：对搜索结果进行过滤和排序后呈现给。 知识库精排开关默认为开启状态，若未配置，则为开启状态。注意：关闭精排时相关性得分范围为0-200，开启精排时相关性得分为0-1，在开启或关闭精排后需要重新设置相关性阈值和引用相关度阈值，否则会影响过滤效果！ ● 搜索页面相关性阈值：超过相关度阈值的搜索结果才能在搜索结果页展示，否则被过滤。 ● 问答相关度阈值：超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤。

版权所有 © 华为云计算技术有限公司

参数	说明
FAQ召回策略	是指在FAQ中搜索结果时，生成结果的策略。 FAQ检索召回相似阈值：通过querytoquery相似检索得到相似得分，超过阈值可以检索召回，默认值0.8。 FAQ问答直出阈值：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。默认值0.95。

– 更多设置

修改“搜索模型设置”、“NLP模型设置”、“AI搜索设置”和“高级设置”，如何修改请参考新建知识库中**步骤5**和**步骤6**。

同时支持新增“其他”设置。

表 3-11 其他设置

参数	说明
参考文档数量	设置RAG大模型参考的文档数量。 参考文档数量未配置时，将使用默认值3
Query改写	开启后，将根据用户历史多轮对话，对输入query进行问题拆分和改写，改写后的query仅用于文档检索。
意图分类	勾选意图分类。 ● 人设类：你叫什么名字? ● 天气类：今天天气怎样? ● 行业知识类： 对于行业知识类，建议使用前缀匹配，后续可能继续扩展。如：行业知识类-金融：贷款重组的定义是什么。 ● 行业知识类-制造：我国的制造业到了什么阶段? ● 行业知识类-医疗：医疗事故有哪些? ● 行业知识类-政务：《国务院关于印发新一代人工智能发展规划的通知》的指导思想是什么? ● 行业知识类-金融：今天的股市怎么样? ● 语言任务类：请创作一封约460字的邮件，主题是咨询一个新的IT项目的细节，这个邮件将被发送给公司的IT项目经理。 ● 通用知识类：豆汁和豆浆的区别。 ● 闲聊类：坐火车累死了。 **说明** 未选择分类的先使用知识库检索再进行大模型总结，选中的分类直接使用大模型回答。
拒答回复	开启后，可以自己设置拒答回复语，当搜索的问题没有答案时，则会回复设置的回复语。

版权所有 © 华为云计算技术有限公司

参数	说明
通用自定义prompt	● 使用场景：主要用于非RAG场景下的模型生成阶段。（非RAG场景：对话生成任务中，不使用检索步骤进行信息检索，直接使用生成模型生成回复。） ● 组成要素：用户问题、任务指令以及其他要求。 ● 使用方式：支持自定义prompt，如果未配置自定义prompt，则使用默认prompt。在自定义构建时，请参考默认prompt的格式。
QA问题生成自定义prompt	你是问题抽取专家，请根据下面的文档文本内容，归纳生成最多{0}个高质量问题，要求：（1）生成的问题可以根据所提供的文档文本内容进行回答（2）以知识库问答的口语化个人提问方式呈现（3）生成问题不能特指该文档文本内容（4）生成知识点丰富全面的多样性问题（5）生成的问题不能过于简单，确保生成问题的质量文档文本内容：{1} 注意：其中{0}和{1}表示占位符，且顺序固定，检索出来的文章内容将被填充至{0}所在位置，格式为【文档名称】：{title1} 【文档内容】：{content1} 【文档名称】：{title2} 【文档内容】：{content2} ..... 检索的query将被填充至{1}所在位置后进行生成。
QA答案生成自定义prompt	你是问题抽取专家，请根据下面的文档文本内容，归纳生成最多{0}个高质量问题，要求：（1）生成的问题可以根据所提供的文档文本内容进行回答（2）以知识库问答的口语化个人提问方式呈现（3）生成问题不能特指该文档文本内容（4）生成知识点丰富全面的多样性问题（5）生成的问题不能过于简单，确保生成问题的质量文档文本内容：{1} 注意：其中{0}和{1}表示占位符，且顺序固定，检索出来的文章内容将被填充至{0}所在位置，格式为 【文档名称】：{title} 【文档内容】：{content} 生成的问题将被填充至{1}所在位置后进行对应答案生成。

4. 单击“确定”，完成知识库设置的修改。
5. 修改完配置之后，已经导入的文件需要重新导入才能使知识库设置生效。

3.3 将本地数据上传至 KooSearch 知识库

创建完知识库，需要在知识库中上传知识数据。

场景描述

KooSearch知识库支持上传以下几种类型的知识。

版权所有 © 华为云计算技术有限公司

表 3-12 上传数据

上传方式	描述
上传文档	支持上传文档类型的知识，支持格式为 .doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, .md 的多个文档，单个文档不能超过128MB（超过60MB建议通过API上传）。当前版本中图片上传、文档内单张图片上传最大不超过10MB。
上传表格	如果知识库开启了“表格问答”功能。支持上传一份xls、csv、xlsx格式的表格，不能超过128M（超过60MB建议通过API上传）。 **注意** 上传的表格中列名不允许为空、表头不允许超过三行，否则解析失败。 不建议上传表头在左侧的表格。
创建FAQ	支持创建问答形式的知识。
批量导入FAQ	支持批量导入问答形式的知识，主要支持xlsx，xls，docx，doc文件类型格式。
上传结构化数据	支持上传结构化知识。支持UTF-8编码的JSONL文件格式，单个文件最大50M，自定义数据长度应在4-1024个字符之间，且文件中仅能使用一种操作。

进入 KooSearch 控制台

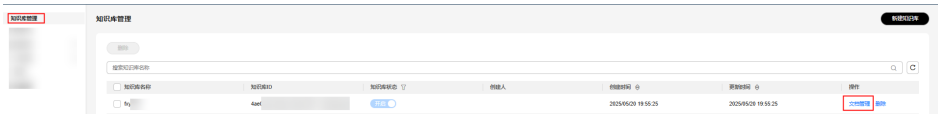
1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

上传文档

1. 提前在本地准备好待上传的文档。

   支持格式为 .doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd 的多个文档，单个文档不能超过128MB（超过60MB建议通过API上传），当前版本中图片上传、文档内单张图片上传最大不超过10MB。
2. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”，进入“文档管理”页面。

图 3-3 进入文档管理页面



版权所有 © 华为云计算技术有限公司

4. 默认进入“文档管理”页签，单击“上传”。
5. 在上传对话框中单击“选择文档”，本地选择已提前准备好的文档。重复的文档不允许上传。

图 3-4 上传文档

6. 如果需要用标签来分类或者标记文档，可以在“标签”下拉框选择标签，如果没有标签单击“添加标签”去创建一个标签。
7. 单击“确认”。文档上传后，可在文档管理页签查看已上传的文档，当“文档状态”为“正常”，说明文档已上传成功。
8. 文档上传成功后，可以单击文档名称，进入文档信息详情页查看切片效果，单击页面右边的切片内容可以跳转到原文对应的地方（此功能目前只支持.pdf文件）。

图 3-5 单击文档名称

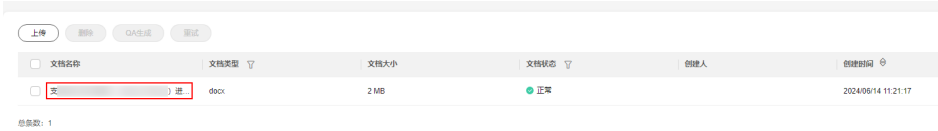
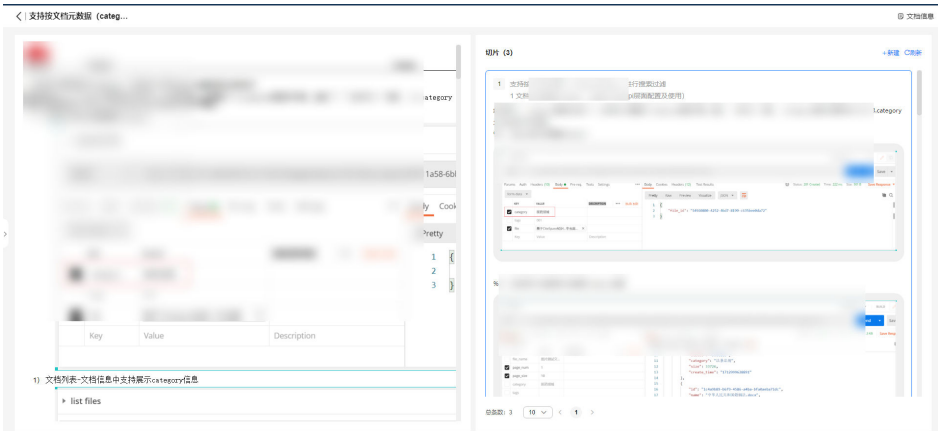


图 3-6 切片效果



9. 管理文档数据。
   - 单击操作列的“下载”，可下载文档至本地。
   - 单击操作列的“删除”，可删除已上传的文档。
   - 单击操作列的“QA生成”，可将上传的文档生成问答模式的Excel文档，生成任务在“任务管理”页签可以查看。

版权所有 © 华为云计算技术有限公司

- 单击操作列的“重试”，可对已上传的文档重新进行切片。勾选多个文档进行“批量重试”时，生成任务在“任务管理”页签可以查看。单击右侧“重试”按钮单个文档重试时不生成重试任务。
- 单击操作列的“编辑标签”，可给文档重新选择或创建标签。
- 如果知识库开启了“表格问答”开关，单击操作列的“表格生成”，可将已上传的excel文档生成表格，生成任务在“任务管理”页签可以查看，生成好后可以在“表格管理”中查看数据表详情。

10. 目录管理。

    如果知识库开启了“目录管理”开关，就可以节点后单击 ⊕ 按钮，在此创建目录，并将文档分类存放。

上传表格

如果您在创建数据库时，开启了“表格问答”按钮，那么知识库详情页面中就会出现“表格管理”页签，该页签支持上传excel文档生成表格数据，在问答过程中利用表格数据，通过NL2SQL问答可提升统计分析类问题的准确率。

支持上传一份xls、csv、xlsx格式的表格，不能超过128M（超过60MB建议通过API上传）。

注意

上传的表格中列名不允许为空、表头不允许超过三行，否则解析失败。

不建议上传表头在左侧的表格。

1. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
2. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”。

   进入“文档管理”页面。
3. 单击“表格管理”，切换至“表格管理”页签。
4. 单击“上传”，依次进行“上传数据表”、“表结构配置”、“数据预览”、“确认入库”操作。
5. 可在表格管理页签查看已上传的表格。
6. 管理表格。
   - 单击操作列的“下载”，可下载表格源文件。
   - 单击“表格名称”可预览生成的数据表内容，并且根据列名查询匹配的数据行；表格详情页面单击“导出”支持按xlsx格式导出当前表数据。
   - 单击操作列的“删除”，可删除已创建的表格。

创建 FAQ

1. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
2. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”。

   进入“文档管理”页面。

版权所有 © 华为云计算技术有限公司

3. 单击“FAQ管理”，切换至“FAQ管理”页签。
4. 单击“创建”，在“新建问答”对话框中输入“标准问题”和“答案”，单击“添加相似问题”，可输入多个相似问题。
5. 在对话框中单击“确认”。

   FAQ创建后，可在FAQ管理页签查看已创建的问答。
6. 管理FAQ。
   - 单击操作列的“编辑”，可重新编辑FAQ。
   - 单击操作列的“删除”，可删除已创建的FAQ。

批量导入 FAQ

1. 提前在本地准备好待导入的FAQ文件。

   主要支持xlsx，xls，docx，doc文件类型格式，文件内容请参考Excel样例文件或Word样例文件。Excel单个文件最大支持10000条数据，文件中不允许空行，空行后的数据将被忽略（超过60MB建议通过API上传）。Word单个文件最大为128M，Word中支持图文格式的FAQ。
2. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”。

   进入“文档管理”页面。
4. 单击“FAQ批量导入”，切换至“FAQ批量导入”页签。
5. 单击“上传”，在上传对话框中单击“选择文档”，本地选择已提前准备好的FAQ文件。

图 3-7 FAQ 批量导入

6. 在对话框中单击“确认”。

   文件上传后，可在“FAQ批量导入”页签查看已上传的文件，当“导入状态”为“正常”，说明文件已导入成功。
7. 管理FAQ文件。
   - 单击操作列的“下载”，可下载文件至本地。

版权所有 © 华为云计算技术有限公司

说明

如果导入的FAQ数据不满足格式规范将生成异常数据，可根据异常数据进行FAQ文件修改二次上传；上传的FAQ文档同样支持切片数据的增删改查，详情见文档管理下的切片数据增删改查。

– 单击操作列的“删除”，可删除已上传的文件。

上传结构化数据

1. 提前在本地准备好待上传的结构化数据文件。

   支持UTF-8无BOM编码的JSONL文件格式，单个文件最大50M，自定义数据长度应在4-1024个字符之间，且文件中仅能使用一种操作。模板如下所示：

```
{"cmd":"ADD","id":"100001","content":"content for the first data"}
{"cmd":"ADD","id":"100002","title":"title for the second data","content":"content for the second data","url":"","docTime":"2015/01/01 12:10:30","category":"category1","tags":["tag1","tag2","tag3"]}
{"cmd":"UPDATE","id":"100002","content":"The content for the second data is updated","category":"newCategory"}
{"cmd":"DELETE","id":"100002"}
```

2. 在KooSearch控制台，左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”，进入“文档管理”页面。
4. 单击“结构化数据”，切换至“结构化数据”页签。
5. 单击“上传”，在上传对话框中单击“选择文档”，本地选择已提前准备好的结构化数据文件。
6. 单击“确认”。

   文件上传后，可在结构化数据页签查看已上传的文件，当“导入状态”为“正常”，说明文件已上传成功。

3.4 体验 KooSearch 问答

当知识库有了数据以后，就可以在KooSearch体验平台进行问答体验。

前提条件

- 已开通了KooSearch服务。
- 已准备好数据库，且已上传数据。
- 待进行问答体验的知识库状态为“开启”状态。

进入 KooSearch 控制台

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

选择知识库

1. 在KooSearch控制台，左侧导航栏选择“体验平台”，进入体验平台页面。

版权所有 © 华为云计算技术有限公司

2. 单击右上角 ，在“引用来源”对话框勾选知识库，单击“确定”。可以选择单个知识库，也可以打开右边的复选框 复选 选择多个知识库作为知识来源。问答体验将在所选择的知识库中进行答案搜索。

体验问答

1. 在“体验平台”页面右上角单击“问答”，切换至问答体验页面。
2. 在输入框中输入问题。
3. 在输入框左侧单击 图标，可以选择“按标签”、“按文档”、“按表格”搜索。
   - 按标签：按文档标签筛选文档，结果只在筛选出来的文档中搜索。
   - 按文档：选择具体文档，结果只在选择出来的文档中搜索。
   - 按表格：当知识库开启了表格问答功能时，可以选择表格问答，当问题命中含表格数据的excel文档时会触发nl2sql，结果只在选择出来的表格中搜索。

     说明

     按表格问答时，建议在问题中明确表格名称、列名，以提高问答效果。
4. 单击 ，查看返回的答案。

图 3-8 体验问答

表 3-13 图标说明

图标	说明
	认同内容，直接单击即可。
	反馈建议。不认同内容，在针对问题、针对搜索、针对回答中选出您认为的不合理的意见，也可以在对话框中输入您认为更理想的回答，单击“提交”。
	复制内容。
	刷新内容。

版权所有 © 华为云计算技术有限公司

图标	说明
	查看答案参考源。在参考列表中，单击“阅读全文”，可查看文档原文。 **说明** 当前针对上传的多栏排版docx文档，查看文档原文时存在内容显示错位及显示不全的问题。

5. 问答体验页面上还有“对话配置”和“对话清空”按钮，如下图所示。

图 3-9 按钮说明

- **“对话配置”**：如果您在对话过程中想修改配置，可以单击“对话配置”按钮，具体的配置参数请参考**配置问答**小节。
- **“对话清空”**：单击“对话清空”按钮可以清空当前对话页面，清空之后再进行问答，会默认进行下一轮问答。

配置问答

1. 在“体验平台”页面，单击右上角⚙，在配置页面进行问答配置。

表 3-14 召回策略

参数	说明
文本召回策略	是指在文档中搜索时，生成结果的策略。包含语义检索、混合检索、关键词检索。 ● 语义检索：切片使用向量检索技术，FAQ使用querytoquery相似检索技术。 ● 混合检索：切片使用向量检索和关键词检索混合检索，FAQ使用querytoquery相似检索技术。 ● 关键词检索：切片使用倒排检索技术，FAQ使用querytoquery相似检索技术。
	语义检索topk召回数量：是指语义搜索生成的片段数量。语义检索topk未配置时，将使用默认值50。 关键字topK召回数量：是指搜索生成的片段数量。 FAQ检索召回个数：通过querytoquery相似检索得到相似得分，按照配置个数进行截断检索召回默认值2。

版权所有 © 华为云计算技术有限公司

参数	说明
	精排：对搜索结果进行过滤和排序后呈现给。 知识库精排开关默认为开启状态，若未配置，则为开启状态。注意：关闭精排时相关性得分范围为0-200，开启精排时相关性得分为0-1，在开启或关闭精排后需要重新设置相关性阈值和引用相关度阈值，否则会影响过滤效果！ ● 搜索页面相关性阈值：超过相关度阈值的搜索结果才能在搜索结果页展示，否则被过滤。 ● 问答相关度阈值：超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤。
FAQ召回策略	是指在FAQ中搜索结果时，生成结果的策略。 FAQ检索召回相似阈值：通过querytoquery相似检索得到相似得分，超过阈值可以检索召回，默认值0.8。 FAQ问答直出阈值：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。默认值0.95。

表 3-15 问答配置

参数	说明
NLP模型服务	选择NLP模型服务。
Query改写	开启后，将根据用户历史多轮对话，对query进行多轮改写和分解，改写后的query仅用于文档检索。
意图分类	勾选意图分类。 ● 人设类：你叫什么名字? ● 天气类：今天天气怎样? ● 行业知识类： 对于行业知识类，建议使用前缀匹配，后续可能继续扩展。如：行业知识类-金融：贷款重组的定义是什么。 ● 行业知识类-制造：我国的制造业到了什么阶段? ● 行业知识类-医疗：医疗事故有哪些? ● 行业知识类-政务：《国务院关于印发新一代人工智能发展规划的通知》的指导思想是什么? ● 行业知识类-金融：今天的股市怎么样? ● 语言任务类：请创作一封约460字的邮件，主题是咨询一个新的IT项目的细节，这个邮件将被发送给公司的IT项目经理。 ● 通用知识类：豆汁和豆浆的区别。 ● 闲聊类：坐火车累死了。 **说明** 未选择分类的先使用知识库检索再进行大模型总结，选中的分类直接使用大模型回答。

版权所有 © 华为云计算技术有限公司

参数	说明
拒答回复	开启后，可以自己设置拒答回复语，当搜索的问题没有答案时，则会回复设置的回复语。
通用自定义prompt	● 使用场景：主要用于非RAG场景下的模型生成阶段。（非RAG场景：对话生成任务中，不使用检索步骤进行信息检索，直接使用生成模型生成回复。） ● 组成要素：用户问题、任务指令以及其他要求。 ● 使用方式：支持自定义prompt，如果未配置自定义prompt，则使用默认prompt。在自定义构建时，请参考默认prompt的格式。
QA问题生成自定义prompt	你是问题抽取专家，请根据下面的文档文本内容，归纳生成最多{0}个高质量问题，要求：（1）生成的问题可以根据所提供的文档文本内容进行回答（2）以知识库问答的口语化个人提问方式呈现（3）生成问题不能特指该文档文本内容（4）生成知识点丰富全面的多样性问题（5）生成的问题不能过于简单，确保生成问题的质量文档文本内容：{1} 注意：其中{0}和{1}表示占位符，且顺序固定，检索出来的文章内容将被填充至{0}所在位置，格式为【文档名称】：{title1} 【文档内容】：{content1} 【文档名称】：{title2} 【文档内容】：{content2} ..... 检索的query将被填充至{1}所在位置后进行生成。
QA答案生成自定义prompt	你是问题抽取专家，请根据下面的文档文本内容，归纳生成最多{0}个高质量问题，要求：（1）生成的问题可以根据所提供的文档文本内容进行回答（2）以知识库问答的口语化个人提问方式呈现（3）生成问题不能特指该文档文本内容（4）生成知识点丰富全面的多样性问题（5）生成的问题不能过于简单，确保生成问题的质量文档文本内容：{1} 注意：其中{0}和{1}表示占位符，且顺序固定，检索出来的文章内容将被填充至{0}所在位置，格式为 【文档名称】：{title} 【文档内容】：{content} 生成的问题将被填充至{1}所在位置后进行对应答案生成。

表 3-16 模型配置

参数	说明
文本多样性(top_p)	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。默认值为0.1。
模型生成最大新词数(max_tokens)	控制文本的最大生成长度，值越大有助于生成较长或完整的回复；值较小，生成的内容越简洁。默认值为2048。 **说明** 如果选择NLP模型-昇腾云类型的模型服务进行问答，建议设置模型生成最大新词数不超过512。

版权所有 © 华为云计算技术有限公司

参数	说明
非搜索增强模型生成多样性(temperature)	控制非搜索增强模型文本的随机性，值越高，文本随机性越、多样性和创造性越高。默认值为0.6。
搜索增强模型生成多样性(temperature)	控制搜索增强模型文本的随机性，值越高，文本随机性越、多样性和创造性越高。默认值为0.6。
文本重复度(presence_penalty)	用于控制生成文本中特定单词或短语出现的频率。值越高生成的文本会使用更多样的单词和短语，减少重复性。默认值为0。

2. 单击“确定”，完成配置。

3.5 体验 KooSearch 搜索

当知识库有了数据以后，就可以在KooSearch体验平台支持进行搜索体验。

前提条件

- 已开通了KooSearch服务。
- 已准备好数据库，且已上传数据。
- 待进行问答体验的知识库状态为“开启”状态。

进入 KooSearch 控制台

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

选择知识库

1. 在KooSearch控制台，左侧导航栏选择“体验平台”，进入体验平台页面。
2. 单击右上角，在“引用来源”对话框勾选知识库，单击“确定”。可以选择单个知识库，也可以打开右边的复选框 复选 选择多个知识库作为知识来源。问答体验将在所选择的知识库中进行答案搜索。

配置搜索

1. 在“体验平台”页面单击右上角，在配置页面设置搜索配置。具体配置请参考配置搜索。

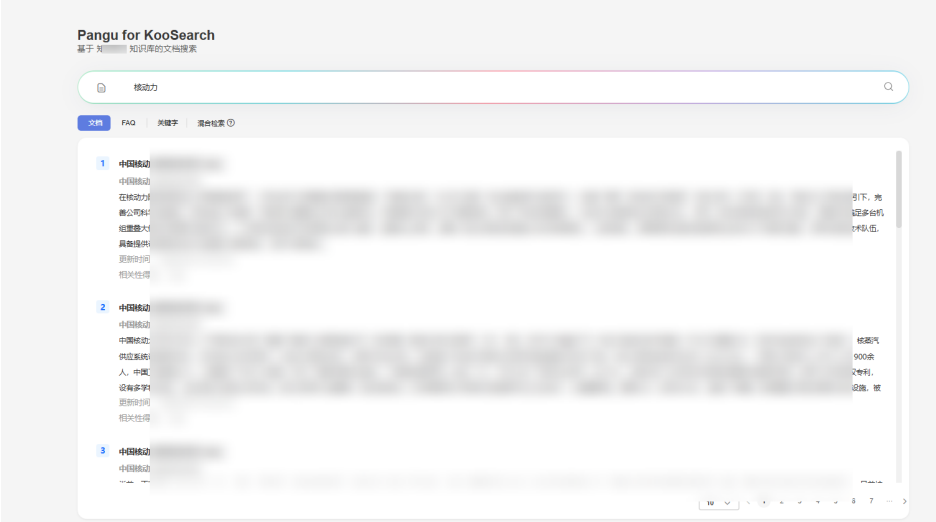
版权所有 © 华为云计算技术有限公司

2. 单击“确定”。

体验搜索

1. 在“体验平台”页面右上角单击“搜索”，切换至搜索体验。
2. 在输入框中输入问题，单击 ，查看搜索结果。

图 3-10 体验搜索



单击搜索结果，可查看更详细的内容。单击“阅读全文”，可查看文档原文。

说明

当前针对上传的多栏排版docx文档，查看文档原文时存在内容显示错位及显示不全的问题。

3.6 体验 AI 搜索

AI搜索是除了将知识库作为知识来源，还可以配置“深度思考”和“联网搜索”模型，突破预训练数据的时间边界，提供深度思考、时效精准的智能搜索服务，让您的搜索体验更佳。

- 深度思考：深度思考是通过配置“深度思考”模型，来模拟人类的深度思考过程，以便更有效地解决问题或做出决策。
- 联网搜索：联网搜索是通过配置“联网搜索”模型，通过互联网上的搜索引擎来查找信息的过程，可以帮助您在庞大的互联网信息资源中快速找到所需的内容。

前提条件

- 已开通了KooSearch服务。
- 已经在模型管理中配置了“联网增强服务”或“web搜索引擎服务”和支持深度思考的NLP模型。具体配置详情，请参考**表3-1**。
- 已经在需要使用的知识库中配置了“联网增强服务”或“web搜索引擎服务”和“深度思考模型”。具体配置详情，请参考**表3-8**。
- 已经在需要使用的知识库中配置了适用于AI搜索的“通用自定义prompt”。

版权所有 © 华为云计算技术有限公司

- 待进行问答体验的知识库状态为“开启”状态。

进入 KooSearch 控制台

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch文档问答页面。
3. 选择已创建好的文档问答服务，单击操作列的“问答”，前往KooSearch控制台。

选择知识库

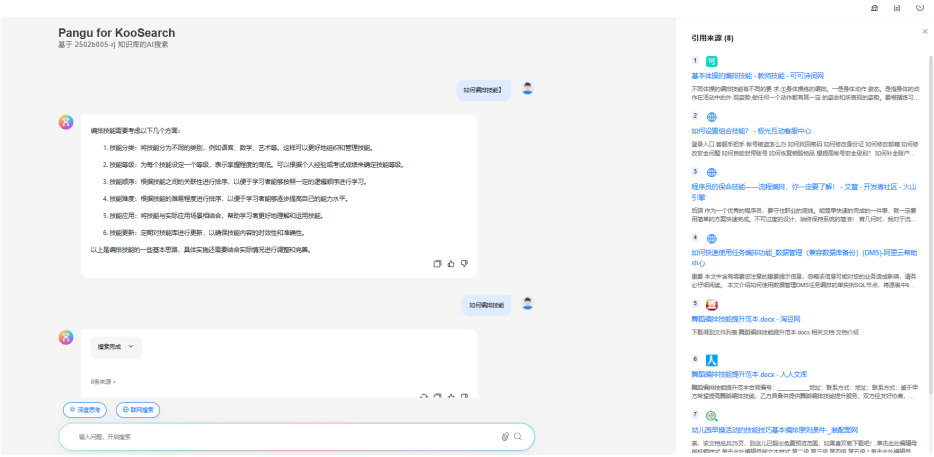
1. 在KooSearch控制台，左侧导航栏选择“AI搜索”，进入“AI搜索”页面。
2. 单击右上角，在“引用来源”对话框勾选知识库，单击“确定”。可以选择单个知识库，也可以打开右边的复选框按钮“复选”选择多个知识库作为知识来源。将在所选择的知识库中进行答案搜索。

配置搜索

1. 在“AI搜索”页面单击右上角，在配置页面设置搜索配置。具体配置请参考配置搜索。
2. 单击“确定”。

体验 AI 搜索

1. 在输入框中输入问题，单击搜索框上方的“深度思考”、“联网搜索”搜索更全面的回答。
2. 单击，查看搜索结果。



3.7 管理 KooSearch 知识库

创建好的知识库可以进行查看、修改、关闭、开启、引用、删除等操作。

版权所有 © 华为云计算技术有限公司

查看知识库详情

1. 进入KooSearch控制台。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 单击需要查看的知识库的操作列的“文档管理”，进入知识库详情页面。

图 3-11 查看知识库详情

4. 进入详情页面后，除了查看知识库详细信息，也可以进行开启关闭知识库、引用知识库、设置知识库、上传文件、任务管理、版本管理的操作。

修改知识库名称

1. 进入KooSearch控制台。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 单击“知识库名称”后面的 ，修改知识库名称后，单击“确认”。

引用知识库

如果您有知识共享诉求，比如同级部门间的共享A引用B的知识库，或者各部门分权独立维护知识库、但作为整体对外。您可以引用知识库，具体操作如下：

1. 进入KooSearch控制台。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 单击“引用知识库”后面的 ，选择需要引用的知识库后，单击“确认”。

图 3-12 引用知识库

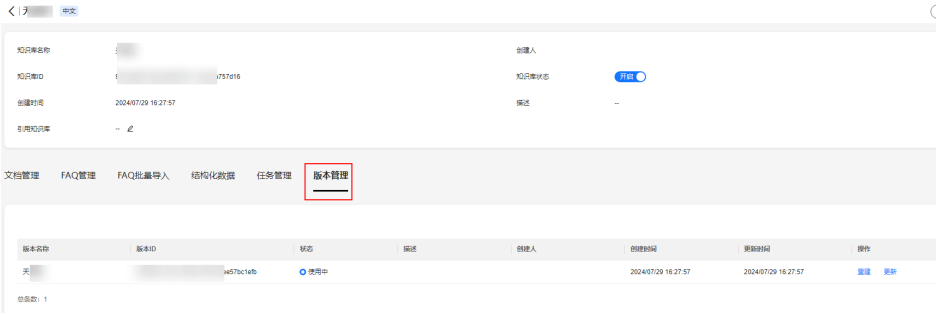
版权所有 © 华为云计算技术有限公司

版本管理

您创建成功知识库的时候，系统会自动创建一个知识库初始版本，对版本的一些管理操作如下所示：

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 单击“版本管理”页签。

   您创建好知识库后，系统会默认创建一个初始版本。



4. 如果您需要再创建一个版本，单击操作列的“重建”按钮，依次选好参数，即可创建成功。

图 3-13 重建版本

- 版本名称：版本的名称。
- 重建来源：选择“索引”或者“文档”。
  - 索引：按照索引重建版本，会直接复用已经完成的向量数据库索引。
  - 文档：按照文档重建版本。选择文档的话，要设置好解析规则是继承原有的规则还是使用最新的规则。如果需要进行全部文档重试，可以在知识库配置好“解析拆分设置”后按文档-最新重建版本。
- 是否立即激活：选择是否立即激活。
- 描述：对于重建版本的描述。

5. 重建好版本后，可以对版本进行以下操作。

图 3-14 版本操作

版权所有 © 华为云计算技术有限公司

状态为“使用中”的版本可进行如下操作：

- “重建”：按照**步骤4**重新建一个版本。
- “更新”：可以单击此按钮更新版本描述。

状态为“可用”的版本除了“重建”、“更新”还可以进行如下操作：

- “关闭”：当版本不用时，可以关闭版本释放索引资源。
- “删除”：当不再需要此版本时，可以删除版本。
- “激活”：可用状态的版本可以激活，激活后此版本的状态变为“使用中”，之前“使用中”的版本状态变为“可用”。

被关闭的版本可进行如下操作：

- “启用”：被关闭版本如果想再次使用，可以单击此按钮启用，启用后版本状态会变成“可用”。
- “删除”：当不再需要此版本时，可以删除版本。
- “更新”：可以单击此按钮更新版本描述。

任务管理

在“文档管理”页签单击“QA生成”、“重试”的任务，都可以在任务管理中查看。“QA生成”的文件可以下载、删除。“重试”生成的任务只支持删除。

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 单击“任务管理”页签，勾选需要操作的任务。可进行下载、删除操作。

图 3-15 任务管理

4. 下载的文档可以在“FAQ批量导入”页签中上传使用。

关闭知识库

创建的知识库默认是开启知识库的，如果问答和搜索时不需要使用该知识库时，可关闭知识库。

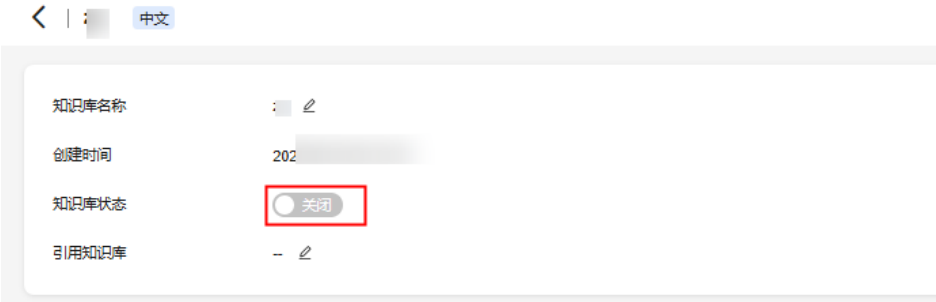
1. **进入KooSearch控制台**。
2. 左侧导航栏选择“知识库管理”，进入“知识库管理”页面。
3. 选择知识库，在“知识库状态”列，关闭开关，使知识库状态处于“关闭”状态。

图 3-16 关闭知识库

版权所有 © 华为云计算技术有限公司

或者单击操作列的“文档管理”，进入“文档管理”页面，关闭知识库状态的开关，使知识库状态处于“关闭”状态。

图 3-17 关闭知识库



删除知识库

如果无需使用已创建的知识库，可删除知识库释放资源。

说明

删除知识库同时会删除知识库的业务数据，请谨慎操作。

1. 在KooSearch控制台，左侧导航栏选择“知识库管理”。

   进入知识库管理页面。
2. 在知识库管理页面，选择已创建的知识库，单击操作列的“删除”，确认需要删除的知识库信息，在输入框输入要删除的知识库名称后，单击“确认”，删除知识库。
3. 在知识库管理页面，选择已创建的知识库，单击操作列的“文档管理”。

   进入文档管理页面。
4. 单击页面右上角的“删除”，确认需要删除的知识库信息，在输入框输入要删除的知识库名称后，单击“确认”，删除知识库。

3.8 管理 KooSearch 提示词

提示词是指导大模型生成答案的关键词。拥有好的数据并在好的提示词指导下，模型就能给出更准确的回应，从而产生更好的结果。精心设计的提示词可以让人们快速获得所需信息，更好地理解所寻找的内容，并减少结果中的错误。

KooSearch服务提供管理提示词的功能，您可以将常用的提示词放在此管理。

新建提示词

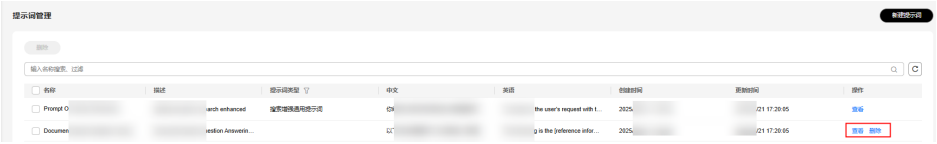
1. **进入KooSearch控制台**。
2. 左侧导航栏选择“配置管理 >提示词管理”，进入“提示词管理”页面。
3. 单击“新建提示词”输入参数，单击“创建”即可创建成功。

参数	说明

版权所有 © 华为云计算技术有限公司

提示词名称	提示词名称。
提示词类型	支持的类型有：搜索增强通用提示词、QA问题生成提示词、QA答案生成提示词。
描述	关于提示词的描述。
默认提示词	服务提供默认的提示词，如果默认提示词可以满足您的诉求，单击“一键导入”，导入默认提示词。您也可以根据您的需求自定义提示词。 目前支持：中文、英语、阿拉伯语、泰语、西班牙语、葡萄牙语提示词。

4. 新建好的提示词也可以查看、删除。单击操作列的“查看”、“删除”即可。已经被引用的提示词不能删除。

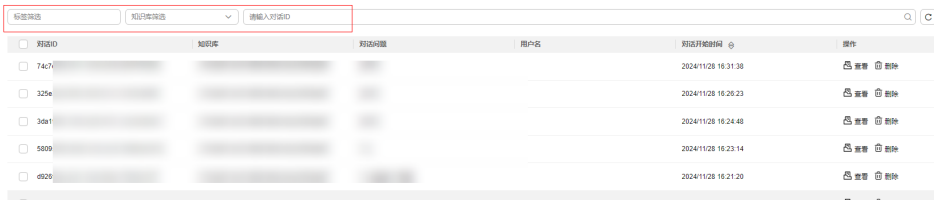


3.9 管理 KooSearch 对话

KooSearch服务提供管理对话和管理反馈结果的功能，您通过体验平台生成的对话以及对结果的反馈，可以在此管理。

查看对话历史

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“对话管理>对话历史”，进入“对话历史”页面。可以在此页面查看历史对话。
3. 单击操作列的“查看”可以查看具体的对话详情。
4. 也可以根据标签、知识库、对话ID来筛选你想查找的对话。



5. 单击操作列的“删除”按钮，或者选中想删除的对话，单击左上方的“删除”按钮，可以删除对话。

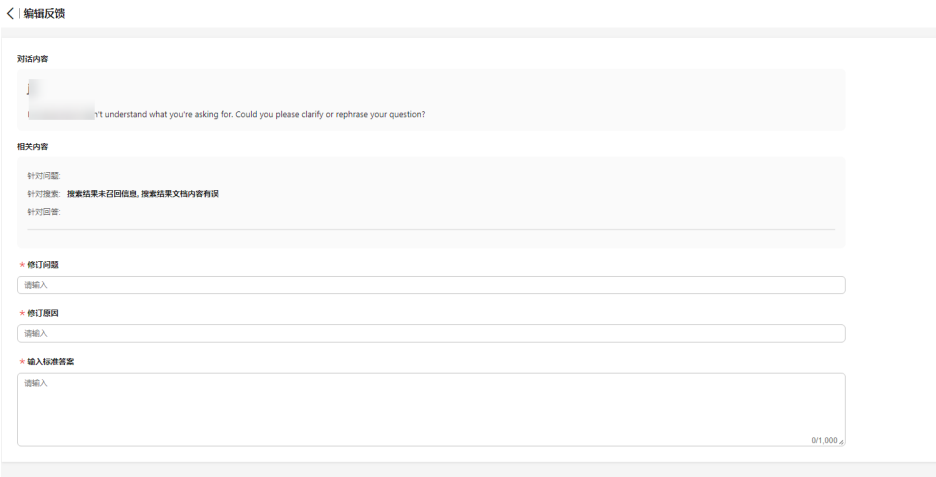
版权所有 © 华为云计算技术有限公司

管理用户反馈

用户在体验平台对问答的反馈可以在此管理。

1. **进入KooSearch控制台**。
2. 左侧导航栏选择“对话管理>反馈管理”，进入“反馈管理”页面。
3. 可以根据标签、相关问题筛选反馈。也可以按筛选条件导出反馈。

4. 对于反馈类型为“点踩”的对话，可以单击操作列的“编辑”按钮，去修正对话。



- “修订问题”：管理员经过分析，出现“点踩”的问题是什么。
- “修订原因”：管理员经过分析，出现“点踩”的原因是什么。
- “输入标准答案”：修正后的答案。

记录修正后，单击“确定”，页面自动返回到反馈列表页面，“反馈状态”变成了“已修正”。

版权所有 © 华为云计算技术有限公司

4 通过 API 使用 KooSearch 实现搜索问答

KooSearch服务提供的API支持发布到不同的环境，发布成功后支持被调用。

场景描述

KooSearch服务开通成功后，会自动创建KooSearch API。在KooSearch服务详情页的API管理页签，可以看到知识管理和文档解析两类API。

- 知识管理：该类API主要用于知识库管理，例如上传文件、查询文件等。
- 文档解析：该类API主要用于对文档数据进行处理，例如解析文档内容。

将KooSearch API发布到不同环境后，支持在环境中调用API使用KooSearch服务。操作流程如下：

1. 在APIG服务配置API网关：**配置API网关**。
2. 在CSS服务发布KooSearch API：**发布KooSearch API**。
3. 在业务环境中调用已发布的KooSearch API：**调用已发布的KooSearch API**。

当已发布的KooSearch API需要修改安全认证方式时，可以**编辑API**。

当已发布的KooSearch API不希望被调用时，可以**下线API**。

配置 API 网关

1. **创建实例**：使用API网关，需要先购买实例。

   说明

   实例需要跟KooSearch服务在同一个VPC和子网。

2. **创建API分组**：API分组相当于API的集合，您在创建API前，需要先创建API分组。
3. **创建环境**：API可以同时提供给不同的场景调用，如生产环境（RELEASE）及其他自定义环境。RELEASE是默认存在的环境，无需创建。

发布 KooSearch API

将KooSearch API发布到环境。

1. 进入KooSearch服务详情页面。
   a. 登录**云搜索服务管理控制台**。

版权所有 © 华为云计算技术有限公司

b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
c. 选择目标服务，单击服务名称，进入服务详情页。

2. 单击“API管理”，进入API管理页签。
3. 选择待发布的API，单击操作列的“发布”。
4. 在“发布”页面配置API网关信息。

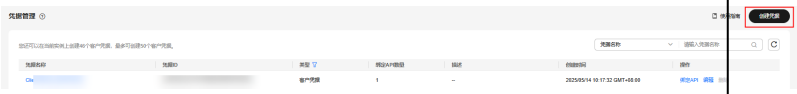
表 4-1 发布服务

参数	说明
实例	使用API网关，需要先购买实例。可以单击右边的“实例管理”去创建实例。具体创建步骤请看**创建实例**。 **说明** 实例需要跟KooSearch服务在同一个VPC和子网。
发布环境	API可以同时提供给不同的场景调用，如生产环境（RELEASE）及其他自定义环境。 建议选择默认存在的环境“RELEASE”，RELEASE是默认的线上环境，是正式发布API的环境，只有发布在RELEASE上的API，才能上架售卖。 如果您需要创建自定义环境，具体创建步骤请看**创建环境**。
分组	API分组相当于API的集合，API提供者以API分组为单位，管理分组内的所有API。 建议选择默认分组“DEFAULT”，该分组为系统自动生成，组内所有API均支持通过弹性公网IP（EIP）或私有IP地址两种方式访问。 如果您需要创建自定义分组，具体创建步骤请看**创建API分组**。

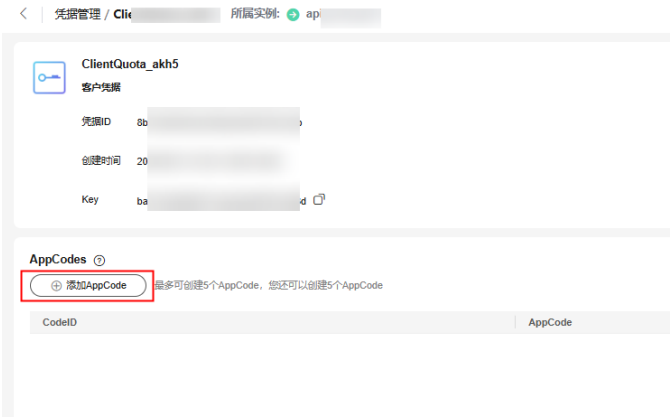
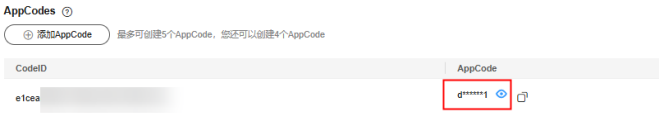
版权所有 © 华为云计算技术有限公司

参数	说明
配置委托	选择IAM委托，授权当前账号访问和使用APIG的权限。 ● 当首次配置委托时，可以单击“自动创建委托”新建委托“css_apig_agency”直接使用。 ● 当已有自动创建的委托时，可以单击“委托一键授权”，自动删除委托中APIG Administrator系统角色或APIG FullAccess系统策略的权限，并自动新增如下自定义策略授权委托到最小化权限。 "apig:vpcChannels:*", "apig:apis:*", "apig:instances:*", "apig:envs:*", "apig:groups:*", "apig:apps:*" ● 执行“自动创建委托”和“委托一键授权”的用户需要如下最小权限。 "iam:agencies:listAgencies", "iam:roles:listRoles", "iam:agencies:getAgency", "iam:agencies:createAgency", "iam:permissions:listRolesForAgency", "iam:permissions:grantRoleToAgency", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:revokeRoleFromAgency", "iam:roles:createRole" ● 使用委托的用户需要如下最小权限。 "iam:agencies:listAgencies", "iam:agencies:getAgency", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:listRolesForAgency"

版权所有 © 华为云计算技术有限公司

参数	说明
安全认证	有APP认证和IAM认证两种方式，推荐使用APP认证方式。 ● “APP认证”：表示由API网关服务负责接口请求的安全认证。APP认证方式具体有多种认证路径。推荐使用AppCode简易认证。 – 使用AppCode简易认证（推荐）：简易认证指在调用API时，HTTP请求头部消息增加一个参数X-Apig-AppCode（参数值填凭据详情中“AppCode”的值），而不需要对请求内容签名，API网关也仅校验AppCode，不校验请求签名，从而实现快速响应。以下为操作步骤： 1. 单击“凭据管理”，进入凭据管理页面。 **图 4-1** 凭据管理 2. 在“凭据管理”页面，单击“创建凭据”按钮，进入“创建凭据页面”，根据下面参数说明，填写凭据信息。 凭据名称：凭据的名称。支持英文、中文、数字、下划线，且只能以英文或中文开头，长度为3~64个字符。 描述：对凭据的介绍。长度为1~255个字符。  3. 单击“确定”，创建凭据。 4. 创建完成后单击创建好的凭据名称进入凭据详情页面。 5. 单击“添加AppCode”按钮。

版权所有 © 华为云计算技术有限公司

参数	说明
	 6. 生成方式选择“自动生成”，单击“确定”。 7. 生成好的AppCode出现在列表中。  AppCode添加完成后，您可以使用AppCode进行API请求的简易认证。 发送请求时，增加请求头部参数“X-Apig-AppCode”，省略请求签名相关信息。 以Curl方式为例，增加头部参数名称：X-Apig-AppCode，参数值填已生成的AppCode。 curl -X GET "https://api.exampledemo.com/testapi" -H "content-type: application/json" -H "host: api.exampledemo.com" -H "X-Apig-AppCode: xhrJVJKABSOxc7d**********FZL4gSHEXkCMQC" 8. 创建好凭据后，返回到发布页面，选择创建好的凭据，进行下一步操作。 – 使用密钥对（Key、Secret）认证。需要在API网关中创建一个凭据，以生成凭据ID和密钥对，将创建的凭据绑定API后，才可以使用APP认证调用API。客户端（API调用者）在调用API过程中，把密钥对替换SDK中的密钥对，API网关服务根据密钥对进行身份核对，完成鉴权。具体创建凭据的步骤，请参考**配置APIG的API认证凭据**。 ● “华为IAM认证”：表示借助IAM服务进行安全认证。 **说明** 选择“华为IAM认证”时，任何API网关租户均可以访问此API，可能存在恶意刷流量，导致过量计费的风险。

版权所有 © 华为云计算技术有限公司

5. 单击“确定”。

   当API的“状态”变成“已发布”时，表示该API已发布到环境，支持被调用。

调用已发布的 KooSearch API

在业务环境中调用已发布的KooSearch API。

具体操作请参见《KooSearch API参考》。

编辑 API

已发布的API，支持修改安全认证方式。

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 单击“API管理”，进入API管理页签。
3. 选择已发布的API，单击操作列的“编辑”。
4. 在“编辑”页面修改API的安全认证方式。
   – “华为IAM认证”：表示借助IAM服务进行安全认证。

     说明

     选择“华为IAM认证”时，任何API网关租户均可以访问此API，可能存在恶意刷流量，导致过量计费的风险。
5. 单击“确定”完成修改。

下线 API

已发布的API因为其他原因需要暂停对外提供服务，可以暂时将API从相关环境中下线。

说明

该操作将导致此API在指定的环境无法被访问，请确保已经告知使用此API的用户。

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 单击“API管理”，进入API管理页签。
3. 选择待下线的API，单击操作列的“下线”。
4. 在滑出的页面中选择下线API的“实例”、“下线环境”、“配置委托”后单击“确定”。
5. 待API的状态变为“未发布”，下线成功。

版权所有 © 华为云计算技术有限公司

5 升级 KooSearch 服务

KooSearch服务的升级功能升级的是集群的内核补丁。

场景描述

升级原理

升级过程采用的是one-by-one的方式。升级时，先下线一个节点，然后对该节点执行切换OS镜像的动作，再将已下线节点的网卡port挂载回来，以此保留节点IP地址，再进行初始化节点启动进程，待节点信息更新后，再依次将其余节点镜像进行替换。升级过程中存在节点下线再上线的动作，可能会中断服务，请在业务低峰期执行。

升级流程

步骤1 进行升级前检查：**升级前检查**。

升级前检查大部分支持系统检查，少部分需要人工检查。

步骤2 创建升级任务，启动升级：**创建升级任务**。

----结束

约束限制

- 最多同时支持20个集群升级，建议在业务低峰期进行升级操作。
- 待升级的集群不能存在正在进行中的任务。
- 升级任务一旦启动就无法中止，直到升级任务的“任务状态”显示“失败”或“成功”才结束。

升级前检查

为了保证升级成功，需要做升级前检查，升级前检查主要包括如下事项：

版权所有 © 华为云计算技术有限公司

表 5-1 升级前检查项

检查项	检查方式	描述	正常状态
集群状态	系统检查	升级任务启动后，系统会自动检查集群状态。“集群状态”为“可用”，表示集群可以正常提供服务。	“集群状态”为“可用”。
资源充足	系统检查	升级任务启动后，系统会自动检查资源。升级过程中会切换OS镜像，需要保证有资源可用。	资源可用且配额充足。
非标操作	人工检查	确认是否存在非标操作。非标操作指的是没有被记录下来的手动操作，这些操作在升级过程中无法自动传递，比如修改系统配置、回程路由等。	未记录到系统中的非标改动，在升级过程中将不会继承下来，升级后可能会影响您的业务，需要提前备份。

创建升级任务

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 选择“升级”页签。
3. 在升级页面，配置升级参数。

表 5-2 升级参数说明

参数	描述
目标镜像	选择目标版本的镜像。选中镜像后，下方会显示镜像名称和目标版本的详细说明。 实际支持的目标版本请以升级页面中“目标镜像”的可选值为准。如果无法选择目标镜像，有如下几个原因： ● 当前集群已是最新版本集群。 ● 当前局点暂未录入新版本镜像。

版权所有 © 华为云计算技术有限公司

参数	描述
配置委托	删除节点会释放网卡，需要VPC的操作权限。选择IAM委托，授权当前账号访问和使用VPC的权限。 ● 当首次配置委托时，可以单击“自动创建委托”新建委托“css_upgrade_agency”直接使用。 ● 当已有自动创建的委托时，可以单击“委托一键授权”，自动删除委托中VPC Administrator系统角色和VPC FullAccess系统策略的权限，并自动新增如下自定义策略授权委托到最小化权限。 "vpc:subnets:get", "vpc:ports:*" ● 执行“自动创建委托”和“委托一键授权”的用户需要如下最小权限。 "iam:agencies:listAgencies", "iam:roles:listRoles", "iam:agencies:getAgency", "iam:agencies:createAgency", "iam:permissions:listRolesForAgency", "iam:permissions:grantRoleToAgency", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:revokeRoleFromAgency", "iam:roles:createRole" ● 使用委托的用户需要如下最小权限。 "iam:agencies:listAgencies", "iam:agencies:getAgency", "iam:permissions:listRolesForAgencyOnProject", "iam:permissions:listRolesForAgency"

4. 配置完成后，单击“确认提交”。
5. 在“任务记录”列表，显示当前升级任务。当“任务状态”为“运行中”时，可以展开任务列表，单击“查看进度”查看详细的升级进度。

   当“任务状态”为“失败”时，可以重试任务或者直接终止任务。

   – 重试升级：在任务列表的操作列，单击“重试”，重新升级。
   – 终止升级：在任务列表的操作列，单击“终止”，结束升级。

     当升级任务终止后，请联系技术支持处理升级失败的任务。

版权所有 © 华为云计算技术有限公司

6 管理 KooSearch 文档问答服务

6.1 查看 KooSearch 文档问答服务详情

在服务的基本信息页面，可以获取服务的内网访问文档解析地址、内网访问知识管理地址、计费模式等信息。除此之外，还能进行管理服务、API管理和日志管理。

- **管理服务**：KooSearch文档问答服务针对已创建的服务所配置的集群，可前往云搜索服务控制台对集群进行管理。
- **API管理**：KooSearch服务开通成功后，会自动创建KooSearch API，将API发布到不同环境后，支持在环境中调用API使用KooSearch服务。
- **日志管理**：为了方便用户使用日志定位问题，KooSearch服务提供了日志查询功能。用户可以通过日志查询进行问题分析定位。

查看 KooSearch 服务信息

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 查看KooSearch服务基本信息和配置信息。

表 6-1 基本信息

参数	描述
名称	服务名称。
ID	服务的唯一标识，是系统自动生成的。
集群状态	服务当前的状态。
内网访问文档解析地址	服务的内网访问文档解析地址。
产品规格	服务的产品规格。

版权所有 © 华为云计算技术有限公司

参数	描述
计费模式	服务的计费模式。
任务状态	服务当前的任务状态，如果没有进行中的任务则显示“--”。
区域	服务所在区域。
创建时间	服务创建的时间。
内网访问知识管理地址	服务的内网访问知识管理地址。

表 6-2 配置信息

参数	描述
虚拟私有云	服务所属的虚拟私有云。
企业项目	服务所属的企业项目。 单击项目名称可以跳转到项目管理页面查看企业项目的基本信息。
子网	服务所属的子网。
集群路由	KooSearch路由信息，可以查看、添加或修改集群路由，配置指导请参见**配置KooSearch文档问答服务集群路由**。
安全组	服务所属的安全组。 单击右侧的“更改安全组”可以修改服务的安全组信息。

管理依赖服务

依赖服务属于KooSearch服务的子服务，KooSearch作为其父服务存在。例如，KooSearch服务依赖Elasticsearch向量数据库，此时Elasticsearch集群即为KooSearch的依赖服务。

对KooSearch服务进行以下操作时，系统将同步对依赖服务执行对应操作：删除集群、退订、续费、变更计费模式。

KooSearch服务支持前往云搜索服务管理控制台对依赖服务进行管理。

1. 进入KooSearch服务详情页面。
   a. 登录**云搜索服务管理控制台**。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 选择“依赖服务管理”页签，可查看KooSearch的依赖服务。

版权所有 © 华为云计算技术有限公司

3. 单击操作列的“跳转”，可跳转至云搜索服务控制台对集群进行管理操作，详情请见《云搜索服务用户指南》。

注意

- KooSearch的依赖服务，禁止单独执行以下操作：更改集群安全模式、删除集群、退订、续费、变更计费模式。
- 当KooSearch服务所依赖的向量数据库在执行扩容或缩容操作后，在KooSearch服务的模型管理页面中，部分模型服务的“连通性”可能会显示为“异常”。此时执行以下步骤可恢复正常：
  1. 在异常模型服务右侧单击“编辑”。
  2. 不修改任何参数，直接单击“确定”。

  系统将自动重新建立连接，状态将恢复为"正常"。

6.2 配置 KooSearch 文档问答服务集群路由

当KooSearch服务需要主动访问公网，或者是需要跨网络访问KooSearch API，则需要配置KooSearch服务的集群路由，连通网络。

操作步骤

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 单击“集群路由”后面的“添加路由”。
3. 在“添加路由”弹窗中，配置路由信息。

表 6-3 配置集群路由

参数	说明
ip地址	填写远程服务器的IP地址，取前16位或者24位，例如源IP为“192.168.1.1”，可以填“192.168.0.0”。
子网掩码	填写IP地址的子网掩码。 ● 当IP地址取的是16位，则子网掩码填“255.255.0.0”。 ● 当IP地址取的是24位，则子网掩码填“255.255.255.0”。 **说明** 子网掩码必须要覆盖IP网段，即子网掩码和IP地址转换为二进制后，IP地址最后的0个数一定要比子网掩码的最后为0的个数多。

版权所有 © 华为云计算技术有限公司

4. 单击“确定”完成集群路由配置。
5. 单击“集群路由”后面的“查看路由”，在“查看路由”弹窗中可以了解集群路由更新信息。

6.3 删除 KooSearch 文档问答服务

当无需使用KooSearch文档问答服务时，可删除服务释放资源。

约束限制

删除服务时，会清理业务数据也会删除依赖集群，请谨慎操作。

删除按需计费的服务

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
3. 在对应服务的“操作”列中单击“删除”。
4. 在弹出的确认提示框中，输入**DELETE**，单击“确定”完成服务删除。

删除包年/包月的服务

包年包月计费方式的服务支持退订/释放，退订/释放服务后，服务将会释放资源并清空数据，且无法恢复，即删除了该服务。

1. 登录云搜索服务管理控制台。
2. 选择需要退订的服务，在操作列单击“更多”>“退订/释放”。
3. 在弹窗中输入**RETREAT**，单击“确定”。

   进入退订资源页面，可以在该页面核对资源信息以及退费金额。
4. 填写退订原因，勾选相关协议后，单击“退订”。

   在弹出确认退订提示信息后，再次单击“退订”。

说明

当服务处于生效中状态时，则走退订流程，此时会产生一个订单进行退费，然后删除服务。当服务处于已过期或者已冻结状态时，则走释放流程，直接删除服务。服务退订的使用说明请参考费用中心相关描述。

版权所有 © 华为云计算技术有限公司

7 KooSearch 文档问答服务日志管理

为了方便用户使用日志定位问题，KooSearch服务提供了日志查询功能。用户可以通过日志查询进行问题分析定位。

日志查询

1. 进入KooSearch服务详情页面。
   a. 登录云搜索服务管理控制台。
   b. 在左侧导航栏选择“KooSearch>KooSearch文档问答”，进入KooSearch服务列表。
   c. 选择目标服务，单击服务名称，进入服务详情页。
2. 单击“日志管理”，进入日志管理页签。
3. 在日志管理页面进行日志查询。

   选择需要查询的节点后，单击 ，显示查询结果。
   - 查询日志时，是从最近时刻的1万条日志中进行匹配，查询结果最多显示100条。
   - 可在搜索框中通过搜索关键词定位日志。

版权所有 © 华为云计算技术有限公司