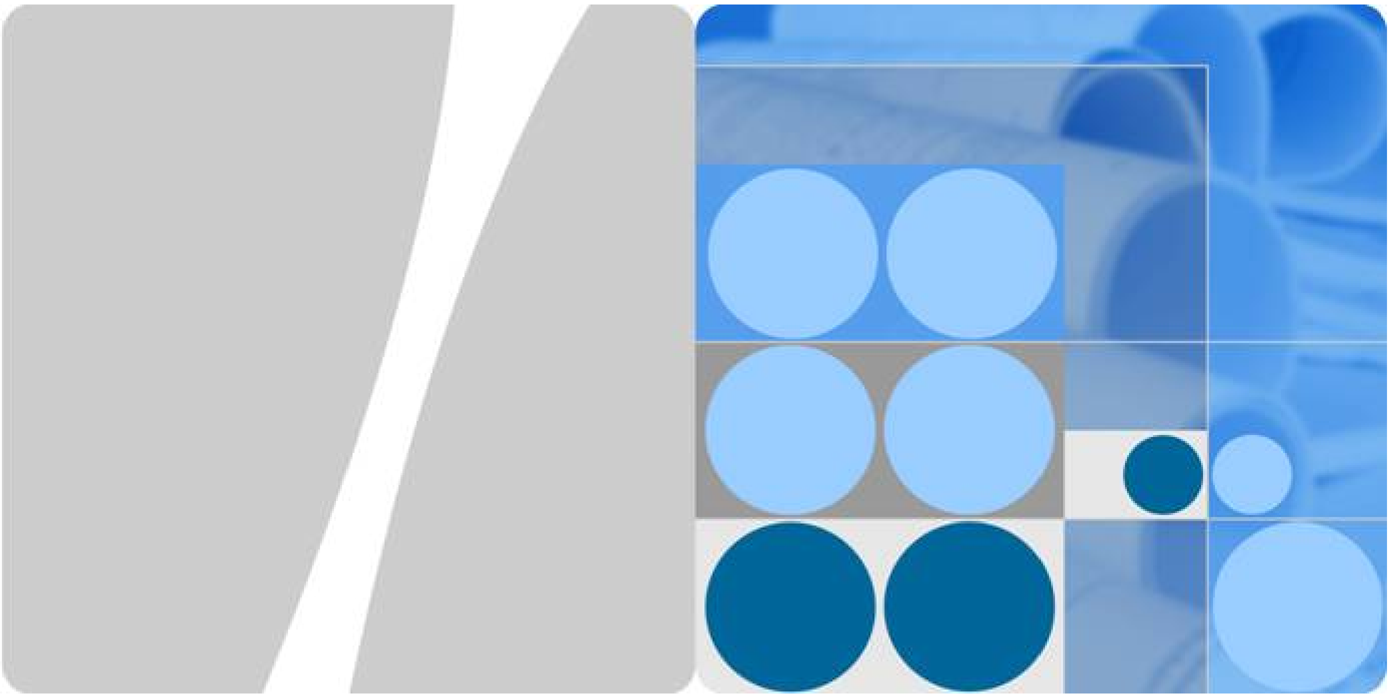


KooSearch

User Guide

Date **2025-08-22**

Contents

1 Procedure for Using KooSearch Document Q&A Service

Huawei Cloud KooSearch is a one-stop intelligent enterprise search solution built on Huawei Cloud's Cloud Search Service (CSS). This solution helps enterprises set up their own enterprise search service effortlessly, by eliminating all the technical complexity and allowing them to focus on the core service applications and use cases. It also allows scenario-oriented secondary development by developers. KooSearch offers modular services that deliver excellent performance in Retrieval-Augmented Generation (RAG) and search scenarios. The modular design and flexible orchestration enable enterprises to quickly implement tailored RAG and search services.

NOTE

KooSearch is available only in CN-Hong Kong and AP-Singapore. KooSearch is in the open beta test (OBT) phase. To trial-use it, submit a **service ticket**.

The following figure shows the procedure for using KooSearch.

Table 1-1 Procedure for using KooSearch

Step	Operation	Description
1	Enabling the service	To use the service, you must enable it first. When enabling the service, you need to select a version and specifications and configure the necessary parameters to create an instance. Then, you use this instance to use the search and document-based Q&A services. For details, see **Enabling KooSearch Document Q&A Service**.

Step	Operation	Description
2	Using search and Q&A on the KooSearch console	After the KooSearch service is enabled, you can use the search and Q&A services on the KooSearch console. The procedure is as follows: 1. (Optional) To use custom models, perform **Creating and Managing Models on KooSearch**. Otherwise, skip this step. 2. **Create a KooSearch knowledge base**. 3. **Upload local documents to the KooSearch knowledge base**. 4. Use KooSearch for Q&A and search. ● **Experiencing KooSearch Document Q&A** ● **Experiencing KooSearch Document Search** 5. Manage knowledge bases.
	Using search and Q&A via KooSearch APIs	You can use the KooSearch document search and Q&A services by calling their APIs. KooSearch APIs can be published to different environments, where they can be called to access the corresponding KooSearch services. The procedure is as follows: 1. **Configuring an API Gateway** 2. **Publishing a KooSearch API** 3. **Calling a Published KooSearch API** 4. **Editing an API** 5. **Taking an API Offline**
3	Managing a KooSearch service	On the service's basic information page, you can obtain the internal IP addresses for accessing the document parsing and knowledge management services, as well as the billing mode. Additionally, you can manage services, APIs, and logs. For more information, see **Managing KooSearch Knowledge Bases**.
4	Viewing KooSearch service logs	You can query KooSearch service logs to locate and diagnose issues. For more information, see **Managing the Logs of KooSearch Document Q&A Service**.

2 Enabling KooSearch Document Q&A Service

On KooSearch, you can create a knowledge base, upload documents to it, and then use this knowledge base for document-based Q&A. However, before you can use KooSearch Document Q&A, you need to enable the service.

📖 **NOTE**

KooSearch is available only in CN-Hong Kong and AP-Singapore. KooSearch is in the open beta test (OBT) phase. To trial-use it, submit a **service ticket**.

Accessing the KooSearch Console

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Enabling KooSearch Document Q&A Service

1. On the KooSearch page, click **Enable**.
2. On the displayed page, configure the service.
   - Basic Configuration

Table 2-1 Basic settings

Parameter	Description
Billing Mode	Select **Yearly/Monthly** or **Pay-per-use**. ● **Pay-per-use**: You are billed by actual duration of use, with a billing cycle of one hour. For example, 58 minutes of usage will be rounded up to an hour and billed. ● **Yearly/Monthly**: You pay for the service by year or month, in advance. The service duration ranges from one month to one year.
Required Duration	Select a duration when **Billing Mode** is set to **Yearly/ Monthly**. If necessary, select **Auto-renew** to enable automatic renewal of the service.
Specifications	Select Agile, Basic, Professional, or Enterprise.
Region	Select the region of the KooSearch service from the drop-down list. Regions are geographic areas isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections. Select a region near you to ensure the lowest latency possible.
Service Name	Set a custom service name.

– Vector Cluster Configuration

A vector database stores and retrieves vectorized data. Currently, only Elasticsearch 7.10.2 clusters are supported. Configure cluster parameters. After the document Q&A service is enabled, a cluster is automatically created based on the settings you configure here.

Table 2-2 Vector cluster parameters

Parameter	Description
Nodes	Number of nodes in the vector cluster. Select a number from 1 to 32. You are advised to configure three or more nodes to ensure high availability of the cluster.
CPU Architecture	Select **x86** or **Kunpeng**. The architectures actually supported depend on your region.
AZ	Select an AZ associated with the cluster's region.
Flavor	Select a node flavor for the cluster. For details, see **ECS Types**.

Parameter	Description
Node Storage Type	Select a node storage type, which can be Common I/O, High I/O, or Ultra-high I/O.
Node Storage Capacity	Select a node storage capacity. Its value range varies with different node specifications. The node storage capacity must be a multiple of 20.
Disk Encryption	Whether to encrypt the data disks of nodes using Key Management Service (KMS). Enabling disk encryption enhances the security of the data stored on cluster nodes. By default, disk encryption is disabled. After disk encryption is enabled, you need to configure **Key Name** by selecting an enabled KMS key from the drop-down list. If no key is available, click **Create key** to go to the Data Encryption Workshop (DEW) console and create a new key or modify an existing key. For details, see **Creating a Key**. **NOTE** ● Only cloud disks support disk encryption. Local disks do not support disk encryption. ● Only custom keys whose **Key Algorithm** is AES or SM4 and **Usage** is **ENCRYPT_DECRYPT** are supported. KMS keys that are unavailable in the **Key Name** drop-down list are not supported by the cluster. ● Disk encryption and decryption do not alter cluster management or O&M processes. However, they do increase the system's processing load, potentially affecting the system's operational performance. ● Once a cluster is already created, disk encryption cannot be enabled or disabled. ● After cluster creation, the KMS key cannot be changed. ● If the KMS key used by the cluster is disabled, the cluster cannot be scaled or upgraded, its node specifications or AZs cannot be changed, and its nodes cannot be replaced (by specifying the nodes that need replacement). To solve this problem, you will have to create a new cluster and migrate your data to that new cluster.

Parameter	Description
Security Mode	Whether to enable security mode for the cluster. ● The security mode is enabled by default. In security mode, a cluster's communication is encrypted and access to the cluster requires user authentication. This is why the **Administrator Username** and **Administrator Password** of the cluster must be configured. – The default administrator username is **admin**. – Set and confirm the **Administrator Password**. This password will be required when you access this cluster. ● If **Security Mode** is disabled, a cluster in non-security mode will be created. With such a cluster, access to the cluster will not require user authentication, and data will be transmitted in plaintext using HTTP. Make sure the customer is in a secure environment, and do not expose the cluster access interface to the public network.

– (Optional) Search Model Configuration

- **Enabled**: When the KooSearch Document Q&A service is enabled, a text vectorization model and a search reranking model will be created automatically.
- **Disabled**: When the KooSearch Document Q&A service is enabled, these two models will not be created, which will affect the use of knowledge bases. If necessary, configure a search embedding model, search reranking model, and cache generation model on the **Model Management** page. For details, see **(Optional) Creating and Managing Models on KooSearch**.

Table 2-3 Parameters for search model configuration

Parameter	Description
Text Vector and Reranking Inference Instance	Select the number of inference instances for the text vectorization and reranking models.
Model Type	Select the model language.
Instance Specifications	Select **Ascend** or **GP-accelerated**.
AZ	When **Instance Specifications** is set to **GP-accelerated**, select an AZ under the current region.

Parameter	Description
Flavor	Select a node flavor for the search models. For details, see **ECS Types**.
Node Storage Type	When **Instance Specifications** is set to **GP-accelerated**, select a node storage type, which can be Common I/O, High I/O, or Ultra-high I/O. The node storage types available vary depending on the selected AZ and instance specifications, as well as the regional environment.
Node Storage Capacity	When **Instance Specifications** is set to **GP-accelerated**, set a node storage capacity. The value range varies depending on the node flavor selected.

– (Optional) Search Planning Configuration

- **Enabled**: When the KooSearch Document Q&A service is enabled, a search planning model will be created automatically. This model provides intent recognition and query rewriting.
- **Disabled**: When the KooSearch Document Q&A service is enabled, this model will not be created, which will affect the use of knowledge bases. If necessary, configure a search planning model on the **Model Management** page. For details, see **(Optional) Creating and Managing Models on KooSearch**.

Table 2-4 Parameters for search planning configuration

Parameter	Description
Search Planning Inference Instance	Select the number of inference instances for the search planning model.
Model Type	Select the model language.
Instance Specifications	Select **Ascend** or **GP-accelerated**.
AZ	When **Instance Specifications** is set to **GP-accelerated**, select an AZ under the current region.
Flavor	Select a node flavor for the search planning model. For details, see **ECS Types**.
Node Storage Type	When **Instance Specifications** is set to **GP-accelerated**, select a node storage type, which can be Common I/O, High I/O, or Ultra-high I/O. The node storage types available vary depending on the selected AZ and instance specifications, as well as the regional environment.

Parameter	Description
Node Storage Capacity	When **Instance Specifications** is set to **GP-accelerated**, set a node storage capacity. The value range varies depending on the node flavor selected.

– (Optional) Large Model Configuration

- **Enabled**: When the KooSearch Document Q&A service is enabled, a preset LLM service will be created automatically.
- **Disabled**: When the KooSearch Document Q&A service is enabled, no such service will be created, which will affect the use of the KooSearch Document Q&A service. If necessary, configure the needed NLP model on the **Model Management** page. For details, see **(Optional) Creating and Managing Models on KooSearch**.

Table 2-5 LLM configuration parameters

Parameter	Description
Model Version	Currently, koosearch-rag is provided. Developed by fine-tuning an open-source model, this model significantly improves search accuracy through retrieval-augmented generation.
Generative Model Inference Instance	Select the number of inference instances for the generative model.
Model Type	Select the model language.
Instance Specifications	Select **Ascend** or **GP-accelerated**.
AZ	When **Instance Specifications** is set to **GP-accelerated**, select an AZ under the current region.
Flavor	Select a node flavor for the LLM. For details, see **ECS Types**.
Node Storage Type	When **Instance Specifications** is set to **GP-accelerated**, select a node storage type, which can be Common I/O, High I/O, or Ultra-high I/O. The node storage types available vary depending on the selected AZ and instance specifications, as well as the regional environment.
Node Storage Capacity	When **Instance Specifications** is set to **GP-accelerated**, set a node storage capacity. The value range varies depending on the node flavor selected.

– Enterprise Project

  When creating a cluster, you can bind an enterprise project to the cluster if you have enabled the enterprise project function. You can select an enterprise project created by the current user from the drop-down list or click **View Enterprise Projects** to go to the **Enterprise Project Management Service** console and create a new project or view existing projects.

3. Click **Next** and configure the network settings of the service.

Table 2-6 Network settings

Parameter	Description
VPC	A VPC is a secure, isolated logical network environment. Select the target VPC. Click **View VPC** to go to the VPC management console and check the name and ID of existing VPCs. If no VPCs are available, create one. **NOTE** The VPC must contain CIDRs. Otherwise, cluster creation will fail. By default, a VPC contains CIDRs.
Subnet	A subnet provides dedicated network resources that are isolated from other networks, improving network security. Select the destination subnet. You can access the VPC management console to check the names and IDs of existing subnets in the VPC.
Security Group	Select a security group for the cluster. A security group serves as a virtual firewall that provides access control policies for clusters. To select a security group that meets your requirements, click **View Security Group** to go to the security group list, where you can check the details of each security group. The selected security group must allow ports 30275 and 30277 in the inbound direction. Otherwise, the cluster may be inaccessible to external services.

4. Click **Next** and confirm the configuration.

5. Click **Confirm**.

   The KooSearch Document Q&A page is displayed. The service you enabled is displayed in the service list and its status is **Creating**. When its status changes to **Available**, the service is created successfully.

   If the service fails to be enabled, try correcting errors by following the instructions on the web console.

3 Using Search and Q&A on the KooSearch Console

3.1 (Optional) Creating and Managing Models on KooSearch

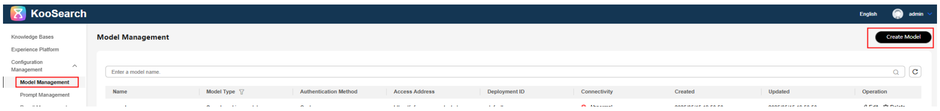
Scenarios

You can configure the models you need on the model management page. When you create a knowledge base, you can select the models you want to use with it. Or you can choose models when you trial-use KooSearch's intelligent Q&A and search services to get better results.

Creating a Model

1. Assess the KooSearch console.
2. In the navigation pane on the left, choose **Configuration Management > Model Management**. The **Model Management** page is displayed.
3. Click **Create Model**.

Figure 3-1 Creating a model



4. On the displayed page, set the parameters described in the following table. Then click **OK**.

Table 3-1 Creating a model

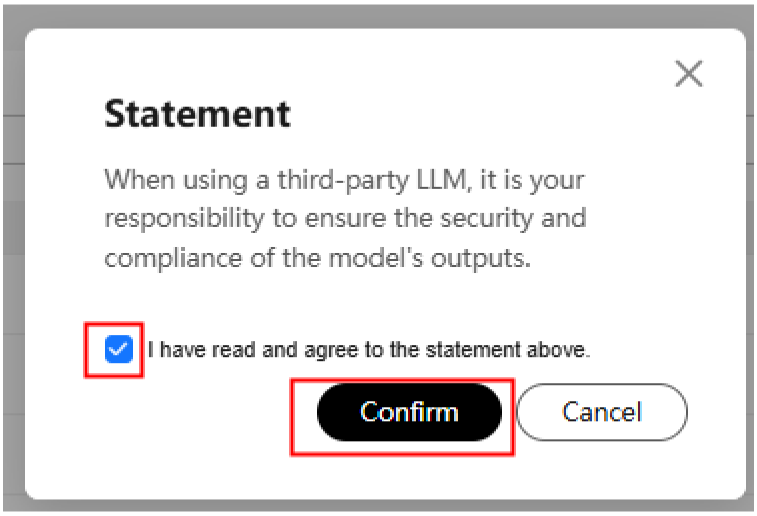
Parameter	Description
Model Name	Enter a model name. The value cannot be empty.

Parameter	Description
Model Type	● NLP models - Cloud base: Pangu NLP models provided by Huawei Cloud. ● NLP models - Bare metal: Pangu NLP models deployed on bare metal servers. ● Search embedding model: a vector search model that converts text into numerical representations called embeddings, which are essentially vectors. ● Search reranking model: Reranks the search results to provide more relevant results. ● Search planning model: Provides multi-turn rewriting and intent recognition. ● Moderation model: Provides a content moderation service that checks the compliance of questions and answers. Only one moderation model can be created. ● OCR model: Extracts text from images, scanned documents, PDFs, and OFD files and outputs the text in an editable format. ● NLP Model - Ascend Cloud: The NLP model is available as a MaaS service on the Ascend cloud. If you select this model to provide Q&A, you are advised to set the maximum number of new tokens generated by the model in response to a given input to 512. ● Cache generation model: Calculates the similarity between queries and provides a cache for the knowledge base. ● Web search engine service: Allows users to add a custom search engine service. ● Enhanced web search service: Provides an enhanced web search service. **NOTE** ● There is a strong connection between the embedding model and the cache generation model. When an embedding model is created, the system automatically generates a cache generation model. If any configuration information is deleted by mistake, the model must be rebuilt using the same configuration parameters. For example, if the name of the embedding model is pangu_embedding, the name of the matching cache generation model is pangu_embedding_faq. ● When creating a knowledge base, both the embedding model (pangu_embedding) and cache generation model (pangu_embedding_faq) are required. If the cache generation model (pangu_embedding_faq) does not exist or is not accessible, an error is returned. In this case, the administrator needs to check whether the pangu_embedding_faq model exists or whether the knowledge base user has access to it. If the model is missing, create it. If the knowledge base user does not have access to it, grant them the access permission.

Parameter	Description
Access Address	Internal network address and port number of the model.
Enable	If you select Review model, the **Enable** button is displayed. When Enable is selected, the Moderation Model will check the compliance of questions asked by users on the Experience Platform as well as the answers generated by AI. KooSearch will refuse to answer questions that contain sensitive words by returning a preset response.
Description	A detailed description of the model.
Ascend Cloud Model Name	If Model Type is set to NLP Model-Ascend Cloud, you need to specify Ascend Cloud Model Name. Select an NLP model provided via the Ascend AI Cloud Service.
Context Length (K)	If Model Type is set to NLP model-Cloud base or NLP model-Bare Metal, you need to set Context Length. Context length refers to the maximum number of tokens the NLP model can process in a single input sequence. A larger context window enables an AI model to process longer inputs and generate more comprehensive outputs.
Deployment ID	If Model Type is set to NLP model-Cloud base or NLP model-Bare Metal, you need to set Deployment ID. Deployment ID indicates the model deployment ID.
Authentication Type	IAM authentication: Huawei IAM authentication is supported. By default, the CSS resource tenant is authenticated. When entrusted account authentication is enabled, you may use an entrusted account to perform the authentication.
	Custom authentication: Custom request headers can be added during API calling.
URL	API management address of the embedding or reranking model when you configure the KooSearch dependent cluster. You can obtain the URL from the dedicated cluster.
Enable Periodic Detection	Periodically checks connectivity to all models and updates the status to the model list.

5. Click **OK**. If you have selected an NLP model, the Statement dialog box shown below is displayed. Select the check box below to agree to the statement, and click **Confirm**.

Figure 3-2 Disclaimer



Statement

When using a third-party LLM, it is your responsibility to ensure the security and compliance of the model's outputs.

☑ I have read and agree to the statement above.

6. After the model is created, find it on the model management page. You can click the model name to check its information.

Editing and Deleting a Model

1. Assess the KooSearch console.
2. In the navigation pane on the left, choose **Model Management**. The **Model Management** page is displayed.
3. Select the target model.

   Click **Edit** in the **Operation** column to edit the model. For details about its parameters, see **Table 1**.

   Click **Delete** in the **Operation** column to delete a model.

3.2 Creating and Modifying a KooSearch Knowledge Base

To use the KooSearch experience platform, start by creating a knowledge base. Once set up, you can upload your data to it, search the data, and ask questions.

Accessing the KooSearch Console

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Creating a Knowledge Base

1. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.

The **Knowledge Bases** page is displayed.

2. Click **Create Knowledge Base** in the upper-right corner.

   On the displayed page, set the knowledge base information.

3. On the **Create** tab, set the parameters and click **Next**.

Table 3-2 Creating a knowledge base

Parameter	Description
Knowledge Base Name	Name of the knowledge base. The value can contain 1 to 64 characters, including letters, digits, hyphens (-), and underscores (_), and must start with a letter or digit.
Knowledge Base Language	Language of the knowledge base. The following languages are supported: ● Chinese ● English ● Thai ● Arabic ● Spanish ● Portuguese
Description	A brief description of the knowledge base. A maximum of 100 characters are allowed.
Knowledge Base Tags	Tags that identify the knowledge base. You can search for knowledge bases by tags or grant access to knowledge bases to different users by their tags. ● Key: custom ● Value: custom
Custom Fields of Structured Data	Adds custom fields of structure data. Click **Add Custom Field**, and set **Field** and **Value**. After the knowledge base is created, you can upload structured data to the knowledge base based on these custom fields.

4. On the **Parse and Split Settings** tab, configure **Parsing Settings** and **Splitting Settings**, and click **Next**.

   – **Parsing Settings**: Select the needed capabilities.

Table 3-3 Parsing settings

Parameter	Description
OCR Enhancement	Calls the OCR service for intelligent document recognition, such as table parsing and file scanning.

Parameter	Description
Image Parsing	If unselected, images in documents will be skipped by default. If selected, two parsing methods are available: ● **Extract Image Text**: Recognize and extract the text in images. ● **Retain Original Images**: Recognize image content and then upload the original images to OBS. The original images will be used in answers.
Header and Footer Parsing	If unselected, the parsing result does not contain document headers or footers. If selected, the parsing result contains document headers and footers.
Contents Page Parsing	If unselected, the parsing result does not contain the contents page. If selected, the parsing result contains the contents page.

– **Splitting Settings**: Select a segmentation method.

Table 3-4 Splitting settings

Parameter	Description
Auto Segmentation	The system automatically selects a proper segmentation method based on the characteristics of the document.
By Length	By default, a document is segmented and merged by paragraph. If a paragraph is too long, it is segmented and merged by identifiers. You need to further set the following parameters: ● **Segment Identifier**: A paragraph is segmented when a selected identifier is encountered. There are no priorities between the selected identifiers. Short segments will be combined to a specified maximum length. If there is no hit on any of the user-specified segment identifiers, the segmentation fails. ● **Estimated Segment Length**: Specifies the maximum segment length. A document will be split to segments of this length. Two adjacent segments will have certain overlapped characters.

Parameter	Description
By Hierarchy	Split the document by title hierarchy, and then split and merge the document by paragraph. Over-long paragraphs will be split by identifiers. Further set the following parameters for detailed splitting methods: Hierarchical Parsing Mode: Select Automatic Parsing or Rule Parsing. If you select **Rule Parsing**, you need to define the rules. For more information about the Rule Parsing mode, see **Table 3-5**.

Table 3-5 By Hierarchy

Parameter	Description
Hierarchical Parsing Mode	Automatic Parsing: Automatically parses documents by system-defined rules.
	Rule explanation: Different types of documents have different hierarchical structures. You can customize parsing rules for different types of documents to enable better parsing and splitting of the documents, thus improving the accuracy of document-based Q&A. ● Default rules Define the most typical rules as default rules. For details, see **Examples of Default Rules**. ● Custom rules Define custom rules using regular expressions. For details, see **Table 3-7**.
Title Level	Select the title level depth of documents.
Title Saving Mode	Select **Save Multi-Title** or **Save Last-Level Title**.
Segment Identifier	A paragraph is segmented when a selected identifier is encountered. There are no priorities between the selected identifiers. Short segments will be combined to a specified maximum length. If there is no hit on any of the user-specified segment identifiers, the segmentation fails.
Estimated Segment Length	Specifies the maximum segment length. A document will be split to segments of this length. Two adjacent segments will have certain overlapped characters.

Parameter	Description
Cross-Title Merge	When two short adjacent paragraphs appear under different section titles, they will be automatically merged into a single segment of predefined length. This helps AI generate better, more comprehensive answers. When **Cross-Title Merge** is disabled, paragraphs under different titles will not be merged automatically. **NOTE** ● This setting is available under (Segmentation) **By Hierarchy**, where you can enable or disable it. ● It is unavailable under **Auto Segmentation**. When **Auto Segmentation** is used, it is enabled by default. ● It does not apply when (Segmentation) **By Length** is used.

Table 3-6 Examples of default parsing rules

Type	Rule	Description
Chapter 1 Section 1 Article 1	^Chapter([01234567891-9]{1,7}) ^Section([01234567891-9]{1,7}) ^Article([01234567891-9]{1,7})	Take the rules for chapters as an example: ● Characters indicating numbers in square brackets can be identified as chapter numbers. ● Arabic numerals from 1 to 9 can be identified as chapter numbers. ● The maximum number of characters indicating a chapter number is 7. The rules for sections and articles are similar.

Table 3-7 Example of custom parsing rules

Type	Rule	Description
Chapter 1 Section 1 Article 1	^Chapter([01234567891-9]{1,7}) ^Section([01234567891-9]{1,7}) ^Article([01234567891-9]{1,7})	/

Type	Rule	Description
1 **1.1** **1.1.1**	^(\d+\.)(?=\s) ^(\d+)(\.\d+)(?!\.)(?=\s) ^(\d+)(\.\d+)(\.\d+)(?!\.)(?=\s)	Matches paragraphs that start with a digit. Note: [\u4e00-\u9fa5]+Chinese characters) Example: 1. Overview 1.1 Description 1.1.1 Detailed Explanation

5. On the **Model Settings** tab, configure the models to use. Then click **Next**.

Table 3-8 Model Settings

Parameter	Description
Search Model Settings	● Embedding model: A Pangu-based text representation model. It converts text into vectors represented by numerics and uses them for purposes such as text retrieval, clustering, and recommendations. ● Reranking model: A Pangu-based text representation model that converts text into vectors represented by numerics and uses them for purposes such as text retrieval, clustering, and recommendations. In the case of semantic search, a reranking model improves search results. ● Search planning model: The model provides capabilities such as intent classification, multi-turn query rewriting, complex query decomposition, and time extraction. In a retrieval augmented generation (RAG) task, intent classification enables queries to be routed to the correct logic and processes; query rewriting and decomposition help improve search accuracy. **NOTE** ● There is a strong connection between the embedding model and the cache generation model. When an embedding model is created, the system automatically generates a cache generation model. If any configuration information is deleted by mistake, the model must be rebuilt using the same configuration parameters. For example, if the name of the embedding model is pangu_embedding, the name of the matching cache generation model is pangu_embedding_faq. ● When creating a knowledge base, both the embedding model (pangu_embedding) and cache generation model (pangu_embedding_faq) are required. If the cache generation model (pangu_embedding_faq) does not exist or is not accessible, an error is returned. In this case, the administrator needs to check whether the pangu_embedding_faq model exists or whether the knowledge base user has access to it. If the model is missing, create it. If the knowledge base user does not have access to it, grant them the access permission.
NLP Model Settings	**NLP model**: Select an NLP model. The Pangu NLP model can be used for interactive dialogues, question answering, and content creation.
	Extend Long Context: When enabled, the context length may be extended during document parsing to generate more comprehensive results. Additionally, **Effective Context Length** needs to be set to ensure optimal outputs.
AI Search Settings	Search service type: Select Web search engine service or Enhanced web search service. Select search service: Select an available search engine service. Deep Thinking Model: Select a deep thinking model.

6. Go to the **Advanced Settings** tab, set the parameters, and then click **OK**.

Table 3-9 Advanced Settings

Parameter	Description
Reference Location	When enabled, generated answers will contain hyperlinks that point to the source text.
Image + Text	When enabled, the answer will include both text and images from the original documents. There are three image recalling methods: 1. Recall only semantically related images (default): If an image in the referenced paragraph has semantically related context as the generated text, the image will be recalled. Otherwise, it will not be recalled. 2. All images: Recall all images in the referenced text. 3. AI recall: Recall images using AI. **NOTE** ● To enable this function, select Parse and Split Settings > Parse Settings > Image Parsing and select Retain only Original Images. ● If you are modifying an existing knowledge base, this setting is likely to be unavailable because an old version is used. To use this setting, ensure that you have purchased the image + text service, modify the document parsing mode (Retain only Original Images), reconstruct the knowledge base version based on the latest document configuration, or select the required documents to try again. ● Currently, only AI recall is supported for knowledge bases whose language is not Chinese.
Tabular Q&A	When enabled, documents can be converted into tables, and NL2SQL can enable more accurate statistical analysis.

Parameter	Description
Knowledge Base Cache	When enabled, Q&A history will be cached. This enables the knowledge base to answer similar questions faster in the future. To use a knowledge base cache, you need to further set the following parameters: ● **Cache Generation Model**: Choose a model. ● **Cache Threshold**: Triggers the cache policy when this threshold is reached. Select a value from 0.1 to 1. ● **Cache Policy**: Select **Highest Score** or **Random**. This is the policy for choosing from multiple answers. ● **Expiration Policy**: Specifies how the cache is cleared. There are three options: – **Least Recently Used**: Delete items that are least recently used. – **First In First Out**: Delete the oldest data. – **Least Frequently Used**: Delete the least frequently accessed cache items (with the least hits) when the cache capacity is about to run out. ● **Keepalive Time (s)**: TTL of the cache. You can set it to **Permanent**.
Directory Management	When enabled, the default directory management function will be used to manage documents. CAUTION A secondary development of the directory management settings will overwrite existing ones.

You can check the basic information about the newly created knowledge base on the knowledge base management page, including the knowledge base ID, name, and status.

Modifying Knowledge Base Settings

You can modify the settings of an existing knowledge base.

⚠ CAUTION

New document parsing and splitting rules will be applied only to newly uploaded documents or retried documents.

1. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.

   The **Knowledge Bases** page is displayed.

2. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

The **Document Management** page is displayed.

3. Click **Configure** in the upper-right corner to modify parsing and splitting settings, and more.
   - Parsing and splitting settings

     See **Table 3-3** and **Table 3-4**.
   - Recall policies

     Recall policies are classified into text recall policies and FAQ recall policies.

Table 3-10 Recall policies

Parameter	Description
Text Recall Policy	Recall policy used for document searches. Options include semantic search, hybrid search, and keyword search. ● Semantic search: Queries document chunks using vector search, and FAQs using query-to-query similarity-based search. ● Hybrid search: Queries document chunks using the hybrid of vector search and keyword search, and FAQs using query-to-query similarity-based search. ● Keyword search: Queries document chunks using inverted index search, and FAQs using query-to-query similarity-based search.
	Top-k recalls for semantic search: the number of recalls for each semantic search. If not specified, the default value 50 is used. Top-k recalls by keyword: the number of recalls for each keyword-based search. FAQ recalls: Obtains the similarity score through query-to-query similarity-based search and recalls the specified number of results. The default value is 2.

Parameter	Description
	Refined Ranking: filters and ranks search results before displaying them. Reranking is enabled by default. Note that when reranking is disabled, the relevance score ranges from 0 to 200. When it is enabled, this score ranges from 0 to 1. After enabling or disabling reranking, you must reconfigure the relevance threshold and reference relevance threshold. Otherwise, relevance-based result filtering will be affected. ● Search Page Correlation Threshold: Only search results with a relevance score higher than the correlation (relevance) threshold will be displayed on the search results page. ● Q&A Reference Correlation Threshold: The search results with a relevance score higher than the correlation (relevance) threshold will be submitted to the LLM for summarization.
FAQ Recall Policy	Recall policy used for FAQ searches. FAQ Recall Similarity Threshold: Obtains the similarity score through query-to-query similarity-based search and recalls results based on a similarity threshold. The default value is 0.8. FAQs with a relevance score exceeding this threshold will be provided as answers directly. There is no need for the LLM to summarize the answer. Default value: **0.95**.

– More settings

Modify **Search Model Settings**, **NLP Model Settings**, **AI Search Settings**, and **Advanced Settings**. For details, see **step 5** and **step 6** in section "Creating a Knowledge Base."

You can also configure the settings under **Others**.

Table 3-11 Other settings

Parameter	Description
Reference Documents	Sets the number of reference documents for the RAG model. If not configured, the default value 3 is used.
Query Rewriting	User queries are split and rewritten based on the multi-turn dialog. The rewritten queries are used for document retrieval only.

Parameter	Description
Intent Classification	Select an intent category. ● Human interaction: What's your name? ● Weather: What is the weather today? ● Industry knowledge: Prefix matching is recommended, allowing for future extensions. For example, "Industry knowledge-Finance: What is the definition of loan restructuring?" ● Industry knowledge-Manufacturing: What is the current stage of China's manufacturing? ● Industry knowledge-Healthcare: What types of medical errors are there? ● Industry knowledge-Government: What are the main guidelines in the New-Generation Artificial Intelligence Development Plan issued by the State Council of China? ● Industry Knowledge-Finance: How is the stock market doing today? ● NLP task: Please write an email of about 460 words asking for details about a new IT project. This email will be sent to the company's IT project manager. ● General knowledge: What is the difference between soybean juice and soy milk? ● Chit-chat: It's so exhausting taking a long-distance train. **NOTE** Questions with identified intents are answered by an LLM directly. For questions with unidentified intents, they will first be searched in the knowledge base. Then, the LLM summarizes the results to generate answers.
Refuse Certain Questions	When enabled, you can set **Response When Refusing a Question**. If no answer is found for a question, this preset response is returned.
General Prompt	● Use scenarios: non-RAG. In non-RAG scenarios, there are no search processes. The generative AI model generates answers directly. ● Elements: The prompt must contain the question, task instructions, and other requirements. ● Usage: The prompt can be customized. If not specified, the default prompt is used. Refer to the format of the default prompt when you write a custom prompt.

Parameter	Description
Custom Prompt for Question Generation	You are an expert in question extraction. Please summarize and generate up to {0} high-quality questions based on the content of the document text provided below. The requirements are as follows: (1) The generated questions should be answerable based on the provided document text. (2) Present the questions in a conversational and personalized manner, suitable for a knowledge base Q&A format. (3) Avoid revealing that your answer is based on some reference material. (4) Make sure the questions are diverse in terms of the knowledge points they cover. (5) Avoid overly simple questions; maintain high quality in the generated questions. Document text: {1} Note: {0} and {1} are placeholders in a fixed sequence. The retrieved document content will be filled to the location indicated by {1}. The format is as follows: [Document name]: {title1} [Document content]: {content1} [Document name]: {title2} [Document content]: {content2} ...... The number of questions generated will be filled to the location indicated by {0}.
Custom Prompt for Answer Generation	You are an expert in question extraction. Please summarize and generate up to {0} high-quality questions based on the content of the document text provided below. The requirements are as follows: (1) The generated questions should be answerable based on the provided document text. (2) Present the questions in a conversational and personalized manner, suitable for a knowledge base Q&A format. (3) Avoid revealing that your answer is based on some reference material. (4) Make sure the questions are diverse in terms of the knowledge points they cover. (5) Avoid overly simple questions; maintain high quality in the generated questions. Document text: {1} Note: {0} and {1} are placeholders in a fixed sequence. The retrieved document content will be filled to the location indicated by {1}. The format is as follows: [Document name]: {title} [Document content]: {content} The number of questions generated will be filled to the location indicated by {0} before answers are generated.

4. Click **OK** to confirm the modification.
5. After the modification, you need to re-import the required documents and files for the new knowledge base settings to take effect.

3.3 Uploading Local Data to a KooSearch Knowledge Base

After creating a knowledge base, the next step is to add knowledge to it by uploading documents.

Scenarios

The following types of knowledge can be uploaded to a KooSearch knowledge base.

Table 3-12 Uploading data

Method	Description
Uploading Documents	Supported document formats are as follows: .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. Multiple documents can be uploaded at once. The size of each single document cannot exceed 128 MB. (Upload a document through an API if its size exceeds 60 MB.) In the current version, the maximum size of each image in an uploaded document is 10 MB.
Uploading Tables	If tabular Q&A is enabled for the knowledge base, upload tables in XLS, CSV, or XLSX format. The size cannot exceed 128 MB. (If the size exceeds 60 MB, we recommend uploading the table via an API.) **CAUTION** The uploaded table cannot contain empty column names, and the table header cannot contain more than three rows. Otherwise, it cannot be parsed. Do not upload tables whose header is on the left.
Creating an FAQ	Create knowledge in the form of question-answer (Q&A) pairs.
Importing FAQs In Batches	Import Q&A pairs in batches through .xlsx, .xls, .docx, or .doc documents.
Uploading Structured Data	Upload knowledge in the form of structured data. JSONL files encoded using UTF-8 are supported. A single file cannot exceed 50 MB. The length of a user-defined data item is from 4 to 1024 characters. Furthermore, the file can only be used for one type of operation.

Accessing the KooSearch Console

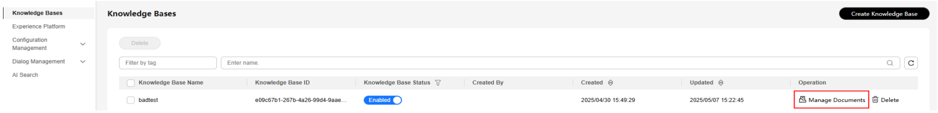
1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Uploading Documents

1. Prepare documents you want to upload to KooSearch on your local PC.

   Supported document formats are as follows: .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. Multiple documents can be uploaded at once. The size of each single document cannot exceed 128 MB. (Upload a document through an API if its size exceeds 60 MB.) In the current version, the maximum size of each image allowed in an uploaded document is 10 MB.
2. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.
3. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

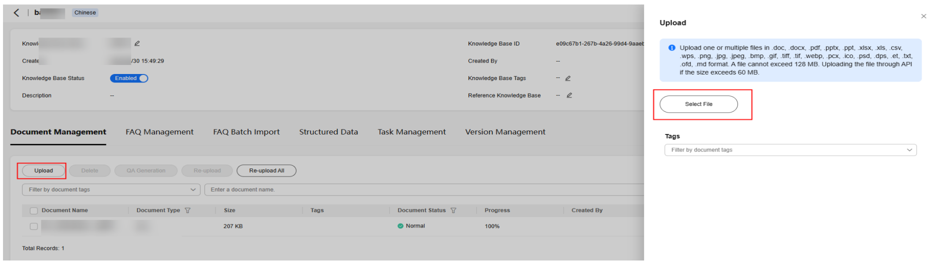
Figure 3-3 Manage Documents



Knowledge Base Name	Knowledge Base ID	Knowledge Base Status	Created By	Created	Updated	Operation
badtest	e09c67b1-267b-4a26-99d4-9aae...	Enabled		2025/04/30 15:49:29	2025/05/07 15:22:45	Manage Documents Delete

4. On the **Document Management** tab displayed by default, click **Upload**.
5. In the **Upload** dialog box, click Select File, and select the prepared documents on the local PC. Duplicate documents cannot be uploaded.

Figure 3-4 Uploading documents



Knowl [redacted]
Create [redacted] /30 15:49:29
Knowledge Base Status: Enabled
Description: --
Knowledge Base ID: e09c67b1-267b-4a26-99d4-9aaeb
Created By: --
Knowledge Base Tags: --
Reference Knowledge Base: --

Document Name	Document Type	Size	Tags	Document Status	Progress	Created By
[redacted]	[redacted]	207 KB		Normal	100%	

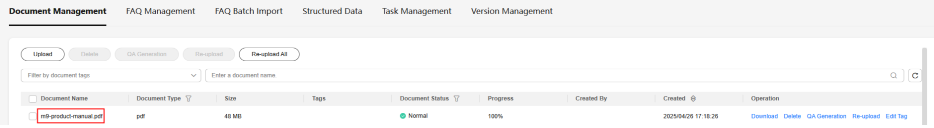
Total Records: 1

Upload one or multiple files in .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 format. A file cannot exceed 128 MB. Uploading the file through API if the size exceeds 60 MB.

6. To classify or mark documents using tags, select a tag from the Tags drop-down list. If no tag is available, click Add Tag to create new tags.
7. Click **Confirm**. Check the uploaded documents on the **Document Management** tab. If **Document Status** is **Normal**, the document is uploaded successfully.

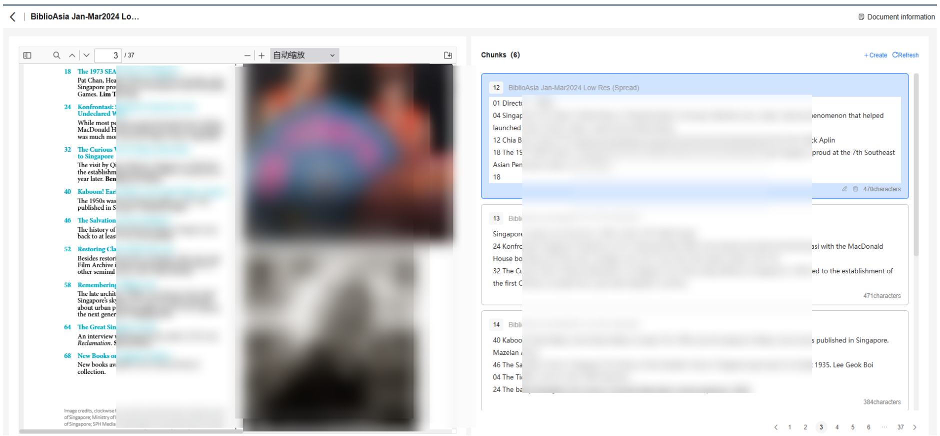
8. Click the name of an uploaded document to check the document segments. For a .pdf document, you can click a segment on the right to get redirected to the corresponding paragraph in the original document.

Figure 3-5 Clicking a document name



□ Document Name	Document Type	Size	Tags	Document Status	Progress	Created By	Created	Operation
□ m9-product-manual.pdf	pdf	48 MB		● Normal	100%		2025/04/26 17:18:26	Download Delete QA Generation Re-upload Edit Tag

Figure 3-6 Checking document segments



9. Manage documents.
   - Click **Download** in the **Operation** column to download a document to the local PC.
   - Click **Delete** in the **Operation** column to delete an uploaded document.
   - Click **QA Generation** in the **Operation** column to generate an Excel document containing Q&A pairs based on the uploaded document. You can check the QA generation task on the Task Management tab.
   - Click **Re-upload** in the **Operation** column to re-segment an uploaded document. You can select multiple documents and click Re-upload All. You can check the generated task on the **Task Management** tab. When you click **Re-upload** in the **Operation** column to re-upload a single document, no task will be generated.
   - Click **Edit Tag** in the **Operation** column to reselect or create tags for the document.
   - If tabular Q&A is enabled for the knowledge base, you can click Generate Table in the **Operation** column to generate a table from the uploaded Excel file. You can check the generation task on the Task Management tab page. After the table is generated, you can check the table details on the Table Management tab.
10. Manage directories.

    If Directory Management is enabled for the knowledge base, you can click the ⊕ button to create directories. This enables better document management by category.

Uploading Tables

If tabular Q&A is enabled when a knowledge base is created, the Table Management tab is available on the knowledge base details page. On this tab, you can upload Excel files to generate tables. These tables will be used during Q&A, and NL2SQL can enable more accurate statistical analysis.

Upload tables in XLS, CSV, or XLSX format. The size cannot exceed 128 MB. (If the size exceeds 60 MB, we recommend uploading the table via an API.)

⚠ CAUTION

The uploaded table cannot contain empty column names, and the table header cannot contain more than three rows. Otherwise, it cannot be parsed.

Do not upload tables whose header is on the left.

1. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.
2. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

   The **Document Management** page is displayed.
3. Click the **Table Management** tab.
4. Click **Upload** and perform the following operations in sequence: upload tables, configure table structure, preview data, and confirm.
5. You can check the uploaded tables on the Table Management tab.
6. Manage tables.
   - Click **Download** in the **Operation** column to download the source file of a table.
   - Click a table name to preview the table content, and query the matched rows by column name. On the table details page, you can click Export to export the table in XLSX format.
   - Click **Delete** in the **Operation** column to delete existing tables.

Creating an FAQ

1. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.
2. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

   The **Document Management** page is displayed.
3. Click the **FAQ Management** tab.
4. Click **Create**. In the **Create Q&A** dialog box, enter the standard question and the answer. Click **Add Extended Question** to create similar related questions.
5. Click **Confirm** in the dialog box.

   Check the newly created FAQ on the **FAQ Management** tab.

6. Manage FAQs.
   - Click **Edit** in the **Operation** column to edit an existing FAQ.
   - Click **Delete** in the **Operation** column to delete an FAQ.

Importing FAQs In Batches

1. Prepare an FAQ file you want to upload to KooSearch on your local PC.

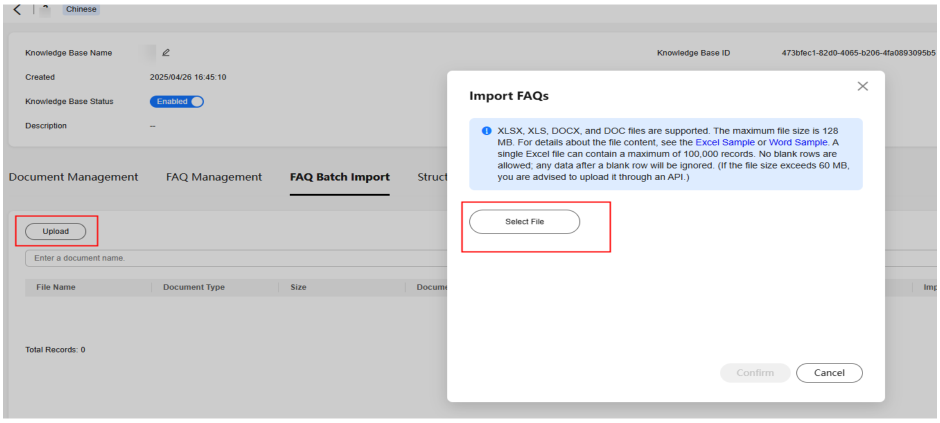
   The supported file formats include .xlsx, .xls, .docx, and .doc. For the file content, see the Excel or Word sample document. An Excel document can contain a maximum of 10,000 records. Blank lines are not allowed because any data following a blank line will be ignored. If the document size exceeds 60 MB, you are advised to upload it via an API. The maximum size of a single Word document is 128 MB. A Word document can contain FAQs consisting of both image and text.

2. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.
3. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

   The **Document Management** page is displayed.

4. Click the **FAQ Batch Import** tab.
5. Click **Upload**. In the **Import FAQs** dialog box, click **Select File**, and select the prepared FAQ file on the local PC.

   Figure 3-7 FAQ Batch Import

   Knowledge Base Name
   Knowledge Base ID 473bfec1-82d0-4065-b206-4fa0893095b5
   Created 2025/04/26 16:45:10
   Knowledge Base Status Enabled
   Description --

   Import FAQs

   XLSX, XLS, DOCX, and DOC files are supported. The maximum file size is 128 MB. For details about the file content, see the Excel Sample or Word Sample. A single Excel file can contain a maximum of 100,000 records. No blank rows are allowed; any data after a blank row will be ignored. (If the file size exceeds 60 MB, you are advised to upload it through an API.)

   Total Records: 0

6. Click **Confirm** in the dialog box.

   Check the uploaded file on the **FAQ Batch Import** tab. If **Import Status** is **Normal**, the file is uploaded successfully.

7. Manage FAQ files.
   - Click **Download** in the **Operation** column to download a file to the local PC.

     NOTE

     If the imported FAQ data does not meet the format requirements, abnormal data will be generated. You can modify the FAQ file accordingly and upload it again. For the uploaded FAQ file, you can add, delete, modify, and query data. For details, see the descriptions under Manage Documents.

– Click **Delete** in the **Operation** column to delete an FAQ file.

Uploading Structured Data

1. Prepare a structured data file you want to upload to KooSearch on your local PC.

   JSONL files encoded using UTF-8 without BOM are supported. A single file cannot exceed 50 MB. The length of a user-defined data item is from 4 to 1024 characters. Furthermore, the file can only be used for one type of operation. The following is a template:

   ```
   {"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. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.
3. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.
4. Click the **Structured Data** tab.
5. Click **Upload**. In the **Upload** dialog box, click **Select File**, and select the prepared structured data file on the local PC.
6. Click **Confirm**.

   Check the uploaded file on the **Structured Data** tab. If **Import Status** is **Normal**, the file is uploaded successfully.

3.4 Experiencing KooSearch Document Q&A

With a knowledge base that contains the necessary knowledge in it, you can try the Document Q&A service on the KooSearch Experience Platform.

Prerequisites

- KooSearch has been enabled.
- A knowledge base has been created, and knowledge has been uploaded to it.
- The knowledge base that you plan to use to experience the Document Q&A service is **Enabled**.

Accessing the KooSearch Console

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Selecting a Knowledge Base

1. On the KooSearch console, choose **Experience Platform** from the left navigation pane.

2. Click ⫩ in the upper right corner. In the displayed **Sources** dialog box, select a knowledge base and click **OK**. You can select a single knowledge base or select multiple ones after toggling on ● Multi-choice .

   KooSearch will search for answers in response to your questions in the selected knowledge base.

Experiencing Document Q&A

1. In the upper right corner of the Experience Platform, click **Q&A**.
2. Enter your question in the search box.
3. On the left end of the question box, click 🗎, and click to search by tag, document, or table.
   - By Tag: Filter documents by document tags. The answer will only come from documents that contain the specified tags.
   - By Document: Select specific documents. The answer will come from the selected documents only.
   - By Table: When the question matches an existing table (Excel file), NL2SQL is triggered. The answer will come from the selected table only.

   NOTE

   For tabular Q&A, you are advised to specify the table name and column name in the question to improve accuracy.
4. Click 🔍, and check the generated answer.

Figure 3-8 Experiencing Document Q&A

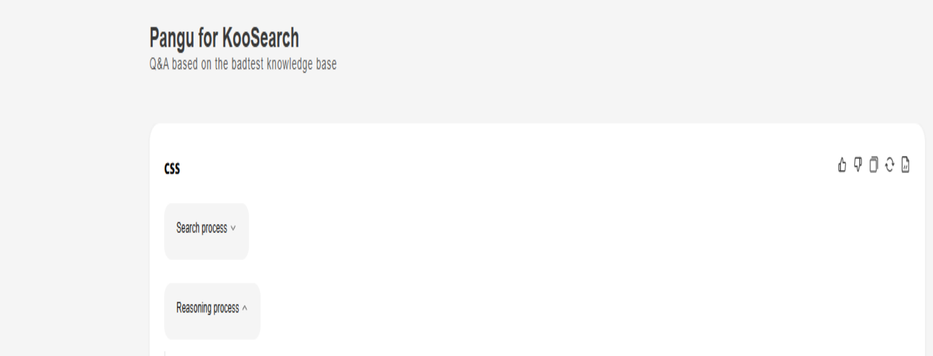


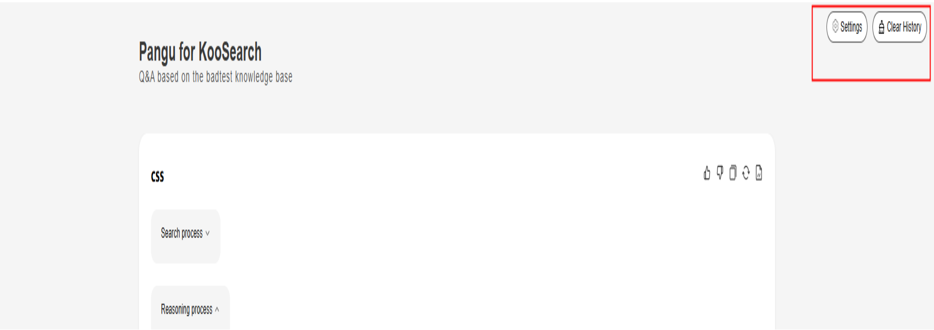
Table 3-13 Icons

Icon	Description
👍	Click to agree with the answer.

Icon	Description
👎	Provide feedback or suggestions. Click to disagree with the answer. In the displayed dialog box, give your feedback on three perspectives: On question, On search, or On answer. Or provide your own answer in the text box below. Then click **Submit**.
⧉	Copy the answer.
⟳	Refresh content.
🗎	Check the reference sources of the answer. In the reference list, click **Full text** to check the original document. **NOTE** The system currently cannot properly display content from a DOCX document that has multiple columns.

5. In the upper right corner of the Q&A page, the **Settings** and **Clear History** buttons are available.

Figure 3-9 Buttons



- **Settings**: Click **Settings** to modify related settings in the middle of a dialog with KooSearch. For more information about the parameters, see **Configuring Q&A Settings**.
- **Clear History**: Click **Clear History** to clear the current dialog page. After clearing, a new dialog begins by default.

Configuring Q&A Settings

1. On the **Experience Platform** page, click ⚙ in the upper right corner. On the **Configure** page, configure the Q&A settings.

Table 3-14 Recall policies

Parameter	Description
Text Recall Policy	Recall policy used for document searches. Options include semantic search, hybrid search, and keyword search. ● Semantic search: Queries document chunks using vector search, and FAQs using query-to-query similarity-based search. ● Hybrid search: Queries document chunks using the hybrid of vector search and keyword search, and FAQs using query-to-query similarity-based search. ● Keyword search: Queries document chunks using inverted index search, and FAQs using query-to-query similarity-based search.
	Top-k recalls for semantic search: the number of recalls for each semantic search. If not specified, the default value 50 is used. Top-k recalls by keyword: the number of recalls for each keyword-based search. FAQ recalls: Obtains the similarity score through query-to-query similarity-based search and recalls the specified number of results. The default value is 2.
	Refined Ranking: filters and ranks search results before displaying them. Reranking is enabled by default. Note that when reranking is disabled, the relevance score ranges from 0 to 200. When it is enabled, this score ranges from 0 to 1. After enabling or disabling reranking, you must reconfigure the relevance threshold and reference relevance threshold. Otherwise, relevance-based result filtering will be affected. ● Search Page Correlation Threshold: Only search results with a relevance score higher than the correlation (relevance) threshold will be displayed on the search results page. ● Q&A Reference Correlation Threshold: The search results with a relevance score higher than the correlation (relevance) threshold will be submitted to the LLM for summarization.
FAQ Recall Policy	Recall policy used for FAQ searches. FAQ Recall Similarity Threshold: Obtains the similarity score through query-to-query similarity-based search and recalls results based on a similarity threshold. The default value is 0.8. FAQ Relevance Threshold: FAQs with a relevance score exceeding this threshold will be provided as answers directly. There is no need for the LLM to summarize the answer. Default value: **0.95**.

Table 3-15 Q&A settings

Parameter	Description
NLP model	Select an NLP model.
Query Rewriting	User queries are split and rewritten based on the multi-turn dialog. The rewritten queries are used for document retrieval only.
Intent Classification	Select an intent category. ● Human interaction: What's your name? ● Weather: What is the weather today? ● Industry knowledge: Prefix matching is recommended, allowing for future extensions. For example, "Industry knowledge-Finance: What is the definition of loan restructuring?" ● Industry knowledge-Manufacturing: What is the current stage of China's manufacturing? ● Industry knowledge-Healthcare: What types of medical errors are there? ● Industry knowledge-Government: What are the main guidelines in the New-Generation Artificial Intelligence Development Plan issued by the State Council of China? ● Industry Knowledge-Finance: How is the stock market doing today? ● NLP task: Please write an email of about 460 words asking for details about a new IT project. This email will be sent to the company's IT project manager. ● General knowledge: What is the difference between soybean juice and soy milk? ● Chit-chat: It's so exhausting taking a long-distance train. **NOTE** Questions with identified intents are answered by an LLM directly. For questions with unidentified intents, they will first be searched in the knowledge base. Then, the LLM summarizes the results to generate answers.
Refuse Certain Questions	When enabled, you can set **Response When Refusing a Question**. If no answer is found for a question, this preset response is returned.

Parameter	Description
General Prompt	● Use scenarios: non-RAG. In non-RAG scenarios, there are no search processes. The generative AI model generates answers directly. ● Elements: The prompt must contain the question, task instructions, and other requirements. ● Usage: The prompt can be customized. If not specified, the default prompt is used. Refer to the format of the default prompt when you write a custom prompt.
Custom Prompt for Question Generation	You are an expert in question extraction. Please summarize and generate up to {0} high-quality questions based on the content of the document text provided below. The requirements are as follows: (1) The generated questions should be answerable based on the provided document text. (2) Present the questions in a conversational and personalized manner, suitable for a knowledge base Q&A format. (3) Avoid revealing that your answer is based on some reference material. (4) Make sure the questions are diverse in terms of the knowledge points they cover. (5) Avoid overly simple questions; maintain high quality in the generated questions. Document text: {1} Note: {0} and {1} are placeholders in a fixed sequence. The retrieved document content will be filled to the location indicated by {1}. The format is as follows: [Document name]: {title1} [Document content]: {content1} [Document name]: {title2} [Document content]: {content2} ...... The number of questions generated will be filled to the location indicated by {0}.
Custom Prompt for Answer Generation	You are an expert in question extraction. Please summarize and generate up to {0} high-quality questions based on the content of the document text provided below. The requirements are as follows: (1) The generated questions should be answerable based on the provided document text. (2) Present the questions in a conversational and personalized manner, suitable for a knowledge base Q&A format. (3) Avoid revealing that your answer is based on some reference material. (4) Make sure the questions are diverse in terms of the knowledge points they cover. (5) Avoid overly simple questions; maintain high quality in the generated questions. Document text: {1} Note: {0} and {1} are placeholders in a fixed sequence. The retrieved document content will be filled to the location indicated by {1}. The format is as follows: [Document name]: {title} [Document content]: {content} The number of questions generated will be filled to the location indicated by {0} before answers are generated.

Table 3-16 Model Settings

Parameter	Description
Text diversity (top_p)	Controls the diversity of the generated text by changing how the model selects tokens for output. A higher value means a wider choice of tokens and hence a higher text diversity. The default value is **0.1**.
Maximum new tokens in the output (max_tokens)	Maximum output tokens generated by the model. With a larger value, the reply can be longer and maybe more comprehensive. With a smaller value, the reply is more brief. The default value is 2048. **NOTE** If you select NLP Model - Ascend Cloud to provide Q&A, you are advised to set this parameter to 512.
Diversity of non-RAG model's output (temperature)	A higher temperature increases the randomness and diversity in the output of a non-RAG model. The default value is 0.6.
Diversity of RAG model's output (temperature)	A higher temperature increases the randomness and diversity in the output of a RAG model. The default value is 0.3.
Text repetition penalty (presence_penalty)	Penalizes tokens that already appear in the generated text. The higher this value, the more diversified words and phrases in the generated text. The default value is **0**.

2. Click **OK**.

3.5 Experiencing KooSearch Document Search

With a knowledge base that contains the necessary knowledge in it, you can try the Document Search service on the KooSearch Experience Platform.

Prerequisites

- KooSearch has been enabled.
- A knowledge base has been created, and knowledge has been uploaded to it.
- The knowledge base that you plan to use to experience the Document Q&A service is **Enabled**.

Accessing the KooSearch Console

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Selecting a Knowledge Base

1. On the KooSearch console, choose **Experience Platform** from the left navigation pane.
2. Click ☰ in the upper right corner. In the displayed **Sources** dialog box, select a knowledge base and click **OK**. You can select a single knowledge base or select multiple ones after toggling on ● Multi-choice.

   KooSearch will search for answers in response to your questions in the selected knowledge base.

Configuring Search Settings

1. On the **Experience Platform** page, click ⚙ in the upper right corner. On the **Configure** page, configure the search settings. For details, see **Configuring Q&A Settings**.
2. Click **OK**.

Experiencing Document Search

1. In the upper right corner of the Experience Platform, click **Search**.
2. Enter a question in the text box and click 🔍.

   Click a search result to check more details. Click **Full text** to check the original document.

   NOTE

   The system currently cannot properly display content from a DOCX document that has multiple columns.

3.6 Experiencing AI Search

Besides knowledge bases, AI search is equipped with a deep thinking model and an Internet search model. They allow KooSearch to provide an intelligent search service that integrates both the parametric knowledge of an LLM and up-to-date public information available on the Internet to deliver a best-in-class AI search experience.

- Deep Thinking: A deep thinking model simulates the deep thinking process of humans to solve complex problems or make well-informed decisions.
- Internet Search: An Internet search model searches for information relevant to user questions on the Internet using a search engine.

Prerequisites

- KooSearch has been enabled.
- On the model management page, you have configured Enhanced Web Search Service or Web Search Engine Service and an NLP model that supports deep thinking. For details, see **Table 3-1**.
- You have configured Enhanced Web Search Service or Web Search Engine Service and Deep Thinking Model for the knowledge base you plan to use. For details, see **Table 3-8**.
- You have configured General Prompt for AI search for the knowledge base you plan to use.
- The knowledge base that you plan to use to experience the AI Search service is **Enabled**.

Accessing the KooSearch Console

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Select a document Q&A service created earlier, and click **Q&A** in the **Operation** column to switch to the KooSearch console.

Selecting a Knowledge Base

1. In the left navigation pane on the KooSearch console, choose **AI Search**. The AI Search page is displayed.
2. Click ⇌ in the upper right corner. In the displayed **Sources** dialog box, select a knowledge base and click **OK**. You can select a single knowledge base or select multiple ones after toggling on ● Multi-choice. KooSearch will search for answers to your questions in the selected knowledge base.

Configuring Search Settings

1. On the **AI Search** page, click ⚙ in the upper right corner. On the **Configure** page, configure the search settings. For details, see **Configuring Q&A Settings**.
2. Click **OK**.

Experiencing AI Search

1. Enter a question in the text box, and click **Think Deeper** or **Search** above to try to get more comprehensive answers.
2. Click 🔍 to start a search.

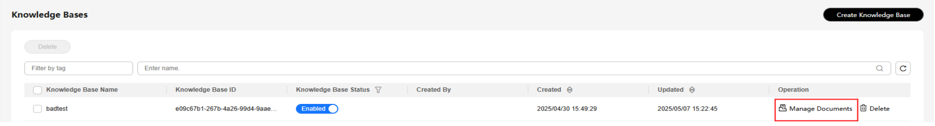
3.7 Managing KooSearch Knowledge Bases

You can view, modify, enable, disable, reference, and delete existing knowledge bases.

Viewing Knowledge Base Details

1. Assess the KooSearch console.
2. Choose **Knowledge Bases** from the left navigation pane.
3. In the row that contains the target knowledge base, click **Manage Documents** in the **Operation** column. The knowledge base details page is displayed.

Figure 3-10 Viewing knowledge base details



Knowledge Bases

Knowledge Base Name	Knowledge Base ID	Knowledge Base Status	Created By	Created	Updated	Operation
□ badtest	e09c67b1-267b-4a26-99d4-9aae...	Enabled		2025/04/30 15:49:29	2025/05/07 15:22:45	Manage Documents Delete

4. On this page, you can view details about the knowledge base, enable or disable it, reference it, and configure its settings. Additionally, you can upload documents to it, as well as manage its tasks and versions.

Renaming a Knowledge Base

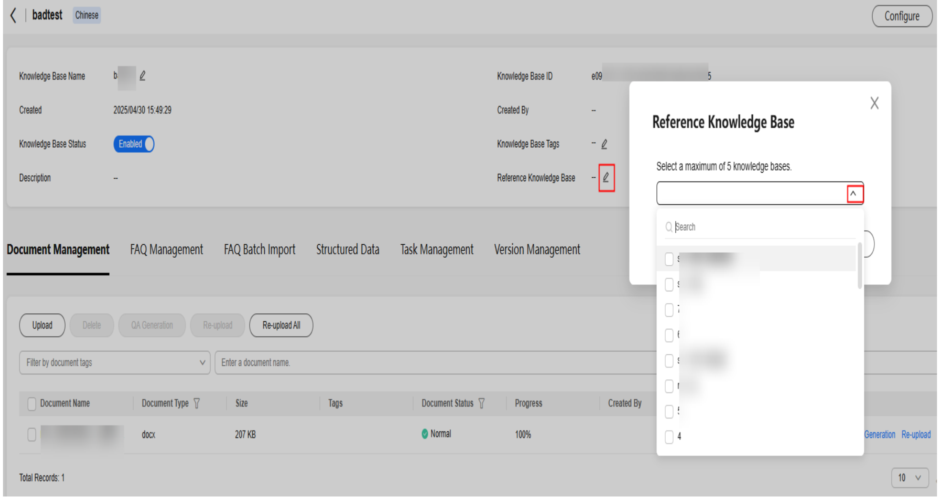
1. Assess the KooSearch console.
2. Choose **Knowledge Bases** from the left navigation pane.
3. Click ✎ next to a knowledge base name, change the name, and click **Confirm**.

Referencing a Knowledge Base

If you want to allow knowledge bases to be shared between different departments within your organization (for example, when each department maintains an independent knowledge base, yet all these knowledge bases are presented as a single knowledge base externally), you can do so through knowledge base referencing. The steps are as follows:

1. Assess the KooSearch console.
2. Choose **Knowledge Bases** from the left navigation pane.
3. Click ✎ next to **Reference Knowledge Base**, select the knowledge bases you want to reference, and click **Confirm**.

Figure 3-11 Referencing a knowledge base



badtest Chinese

Knowledge Base Name b[redacted]

Created 2025/04/30 15:49:29

Knowledge Base Status Enabled

Description --

Knowledge Base ID e09[redacted]5

Created By --

Knowledge Base Tags --

Reference Knowledge Base --

Reference Knowledge Base

Select a maximum of 5 knowledge bases.

Document Name	Document Type	Size	Tags	Document Status	Progress	Created By
[redacted]	docx	207 KB		Normal	100%	

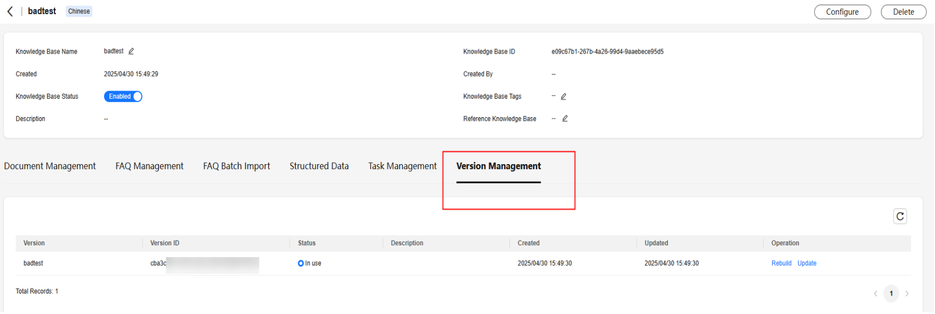
Total Records: 1

Version Management

When you create a knowledge base, an initial version of this knowledge base is automatically created. You can perform the following steps to manage its version:

1. **Assess the KooSearch console**.
2. Choose **Knowledge Bases** from the left navigation pane.
3. Click the **Version Management** tab.

   When you create a knowledge base, an initial version of this knowledge base is automatically created.



badtest Chinese

Knowledge Base Name badtest

Created 2025/04/30 15:49:29

Knowledge Base Status Enabled

Description --

Knowledge Base ID e09c67b1-267b-4a26-99d4-9aaebece95d5

Created By --

Knowledge Base Tags --

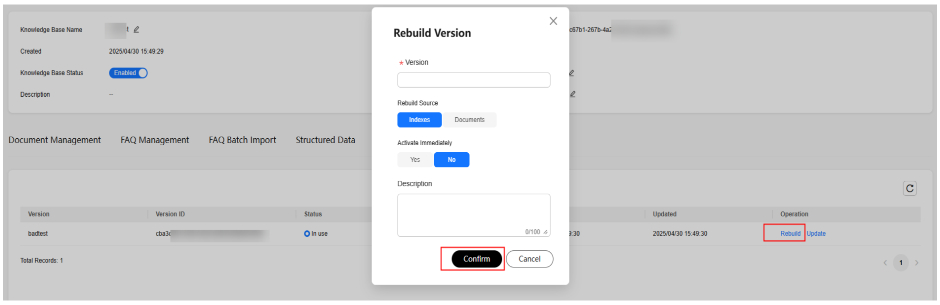
Reference Knowledge Base --

Version	Version ID	Status	Description	Created	Updated	Operation
badtest	cba3c[redacted]	In use		2025/04/30 15:49:30	2025/04/30 15:49:30	Rebuild Update

Total Records: 1

4. To create another version, click **Rebuild** in the **Operation** column, and set the necessary parameters. Then click **Confirm**.

Figure 3-12 Rebuild Version



Rebuild Version

* Version

Rebuild Source: Indexes Documents

Activate Immediately: Yes No

Description

0/100

- **Version**: Set the version name.
- **Rebuild Source**: Select **Indexes** or **Documents**.
  - **Indexes**: Rebuild a new version by reusing the existing indexes of a vector database.
  - **Documents**: Rebuild a new version from documents. When selecting **Documents**, select **Inherit** to inherit existing rules or **Latest** to use the latest rules. To reload all documents, rebuild the version by choosing **Documents** > **Latest** after configuring parsing and splitting settings for the knowledge base.
- **Activate Immediately**: whether to activate the new version immediately.
- **Description**: a brief description of the new version.

5. You can perform the following operations on a knowledge base version.

   You can perform the following operations on a version whose status is **In use**:

   - **Rebuild**: Create a new version. See **step 4**.
   - **Update**: Update the version description.

   For a version whose status is **Available**, you can perform the following operations in addition to **Rebuild** and **Update**:

   - **Disable**: When a version is no longer needed, disable it to reclaim index resources.
   - **Delete**: Delete a version that is no longer needed.
   - **Activate**: Activate an available version to change its status to **In use**. The version that was previously **In use** changes to **Available**.

   For a version whose status is **Closed**, you can perform the following operations:

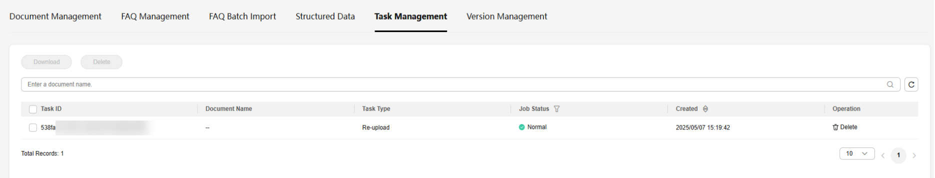
   - **Enable**: Enable a disabled version, which changes its status to **Available**.
   - **Delete**: Delete a version that is no longer needed.
   - **Update**: Update the version description.

Task Management

On the **Document Management** tab, you can create **QA Generation** and **Re-upload** tasks by clicking the corresponding buttons. You can check all these tasks on the **Task Management** tab. You can download or delete the documents generated by a QA Generation task. Yet you can only delete **Re-upload** tasks.

1. **Assess the KooSearch console**.
2. Choose **Knowledge Bases** from the left navigation pane.
3. Click the **Task Management** tab. Then select tasks. You can download or delete the documents generated by tasks.

Figure 3-13 Task Management



□ Task ID	Document Name	Task Type	Job Status	Created	Operation
□ 538fa[redacted]	--	Re-upload	● Normal	2025/05/07 15:19:42	Delete

Total Records: 1

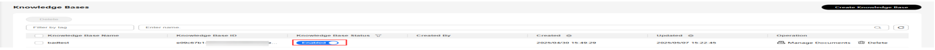
4. The downloaded documents can be uploaded on the **FAQ Batch Import** tab.

Disabling a Knowledge Base

Enabled is the default status of a knowledge base. If a knowledge base is no longer needed for Q&A or search, you can disable it.

1. [Assess the KooSearch console](#).
2. Choose **Knowledge Bases** from the left navigation pane.
3. Locate the target knowledge base. In the **Knowledge Base Status** column, change the status to **Disabled**.

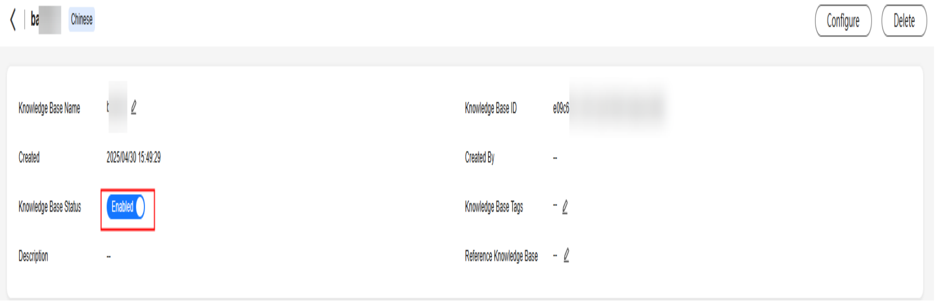
Figure 3-14 Disabling a knowledge base



	Knowledge Base Name	Knowledge Base ID	Knowledge Base Status	Created By	Created	Updated	Operation
☐	badtest	e09c67b1-[redacted]...	Enabled		2025/04/30 15:49:29	2025/05/07 15:22:45	Manage Documents Delete

Alternatively, click **Manage Documents** in the **Operation** column. On the displayed page, find **Knowledge Base Status** in the upper pane, and toggle it to **Disabled**.

Figure 3-15 Disabling a knowledge base



Knowledge Base Name t[redacted]
Knowledge Base ID e09c6[redacted]
Created 2025/04/30 15:49:29
Created By -
Knowledge Base Status Enabled
Knowledge Base Tags -
Description -
Reference Knowledge Base -

Deleting a Knowledge Base

Delete knowledge bases that you no longer need to reclaim resources.

> **NOTE**
>
> Deleting a knowledge base will also delete all its data. Exercise caution.

1. On the KooSearch console, choose **Knowledge Bases** from the left navigation pane.

   The **Knowledge Bases** page is displayed.
2. Select a knowledge base, and click **Delete** in the **Operation** column. Confirm about knowledge base deletion by entering the knowledge base name. Then click **Confirm**.
3. On the **Knowledge Bases** page, select an existing knowledge base, and click **Manage Documents** in the **Operation** column.

   The **Document Management** page is displayed.
4. Click **Delete** in the upper right corner. Confirm about knowledge base deletion by entering the knowledge base name. Then click **Confirm**.

3.8 Managing Prompts on KooSearch

Prompts are carefully crafted inputs designed to guide AI models in generating specific, high-quality outputs. They align the models' responses with user intent. Effective prompts enable AI models to generate more accurate, relevant, and coherent responses, allowing users to quickly obtain accurate information or complete specific tasks.

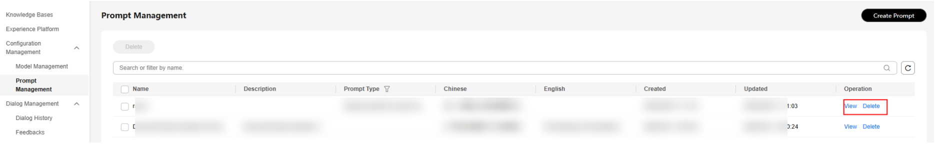
KooSearch provides a prompt management function, where you can manage frequently used prompts.

Creating Prompts

1. Assess the KooSearch console.
2. In the navigation pane on the left, choose **Configuration Management > Prompt Management**. The **Prompt Management** page is displayed.
3. Click **Create Prompt** in the upper right corner. On the displayed page, set the necessary parameters, and then click **Create**.

Parameter	Description
Prompt Name	Prompt name.
Prompt Type	The following types are supported: general prompt for search enhancement, prompt for question generation, and prompt for answer generation.
Description	A description of the prompt.
Default prompt	The service offers default prompts. If they are already good enough for you, click **One-Click Import** to import them. You can also customize prompts as needed. Supported prompt languages include Chinese, English, Arabic, Thai, Spanish, and Portuguese.

4. You can also view or delete existing prompts by clicking **View** or **Delete** in the **Operation** column. A referenced prompt cannot be deleted.



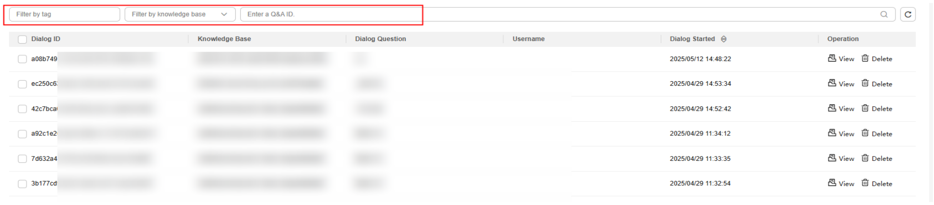
☐ Name	Description	Prompt Type	Chinese	English	Created	Updated	Operation
☐ r [redacted]		[redacted]	[redacted]		[redacted]	[redacted] 1:03	View Delete
☐ D [redacted]	[redacted]		[redacted]	[redacted]	[redacted]	[redacted] 0:24	View Delete

3.9 Managing Dialogs on KooSearch

On KooSearch, you can manage dialogs and user feedbacks that occurred on the Experience Platform.

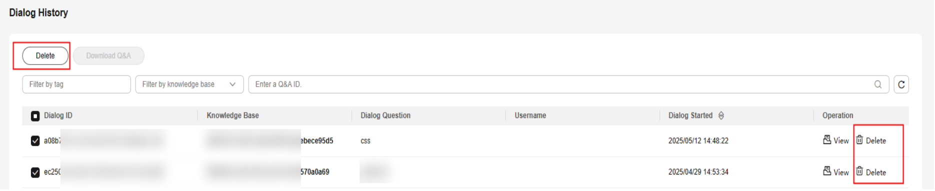
Viewing the Dialog History

1. Assess the KooSearch console.
2. In the left navigation pane, choose **Dialog Management** > **Dialog History**. On the Dialog History page, check all past dialogs.
3. Click **View** in the **Operation** column to view dialog details.
4. You can also filter dialogs by tag, knowledge base, or dialog ID.



□ Dialog ID	Knowledge Base	Dialog Question	Username	Dialog Started	Operation
□ a08b749[redacted]	[redacted]	[redacted]		2025/05/12 14:48:22	View Delete
□ ec250c6[redacted]	[redacted]	[redacted]		2025/04/29 14:53:34	View Delete
□ 42c7bca[redacted]	[redacted]	[redacted]		2025/04/29 14:52:42	View Delete
□ a92c1e2[redacted]	[redacted]	[redacted]		2025/04/29 11:34:12	View Delete
□ 7d632a4[redacted]	[redacted]	[redacted]		2025/04/29 11:33:35	View Delete
□ 3b177cd[redacted]	[redacted]	[redacted]		2025/04/29 11:32:54	View Delete

5. Click **Delete** in the **Operation** column to delete a dialog, or select one or more dialogs in the dialog list, and click **Delete** in the upper left corner.

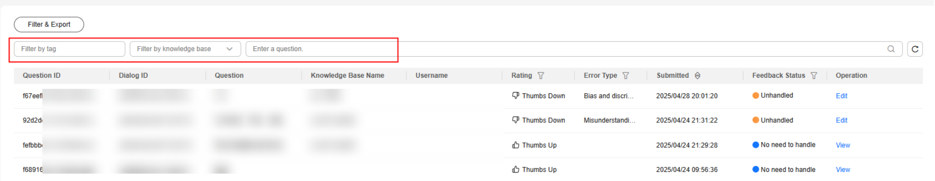


■ Dialog ID	Knowledge Base	Dialog Question	Username	Dialog Started	Operation
☑ a08b[redacted]	[redacted]ebece95d5	css		2025/05/12 14:48:22	View Delete
☑ ec25[redacted]	[redacted]570a0a69	[redacted]		2025/04/29 14:53:34	View Delete

Managing User Feedback

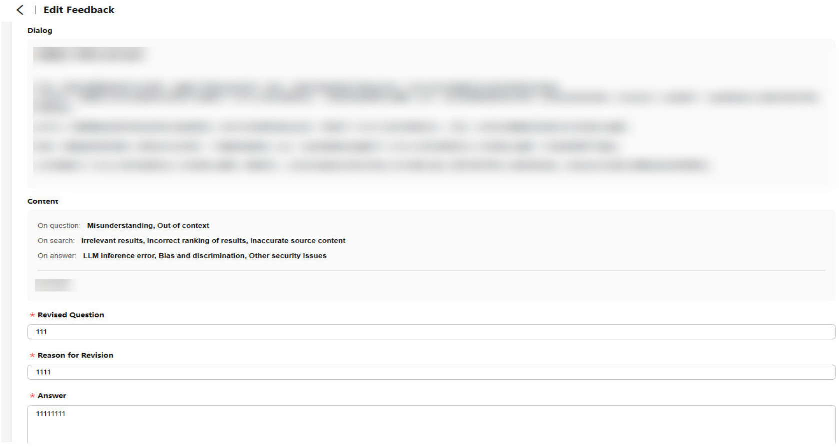
You can manage user feedbacks collected from the experience platform on this page.

1. Assess the KooSearch console.
2. In the left navigation pane, choose **Dialog Management** > **Feedbacks**.
3. On the **Feedbacks** page, filter user feedbacks by tag or relevant questions. You can also export feedbacks by setting filter criteria.



Question ID	Dialog ID	Question	Knowledge Base Name	Username	Rating	Error Type	Submitted	Feedback Status	Operation
f67eef[redacted]	[redacted]	[redacted]	[redacted]		Thumbs Down	Bias and discri...	2025/04/28 20:01:20	● Unhandled	Edit
92d2d[redacted]	[redacted]	[redacted]	[redacted]		Thumbs Down	Misunderstandi...	2025/04/24 21:31:22	● Unhandled	Edit
fefbbb[redacted]	[redacted]	[redacted]	[redacted]		Thumbs Up		2025/04/24 21:29:28	● No need to handle	View
f68916[redacted]	[redacted]	[redacted]			Thumbs Up		2025/04/24 09:56:36	● No need to handle	View

4. For a dialog whose Rating is is Thumbs Down, click **Edit** in the **Operation** column to correct the dialog.



Edit Feedback

Dialog

[redacted]

Content

On question: Misunderstanding, Out of context

On search: Irrelevant results, Incorrect ranking of results, Inaccurate source content

On answer: LLM inference error, Bias and discrimination, Other security issues

[redacted]

* Revised Question: 111

* Reason for Revision: 1111

* Answer: 11111111

- Revised Question: Enter the question.
- Reason for Revision: Find out why the user disliked the answer.
- Answer: Enter the improved answer.

Click **OK**. The **Feedbacks** page is displayed, and **Feedback Status** has changed to **Handled**.

4 Using KooSearch APIs to Provide Search and Q&A

KooSearch APIs can be published to different environments, where they can be called to access the corresponding KooSearch services.

Scenarios

After a KooSearch service is enabled, related KooSearch APIs are automatically created. On the **API Management** tab of the KooSearch service details page, you can see two types of APIs: Knowledge Bases and Document Parsing.

- Knowledge Bases: This type of APIs is used for knowledge base management, such as uploading and querying documents.
- Document Parsing: This type of APIs is used to process documents, for example, parsing documents.

Users can access KooSearch services by calling the corresponding KooSearch APIs deployed in their environments. The procedure is as follows:

1. Configure an API gateway using the APIG service: **Configuring an API Gateway**.
2. Publish KooSearch APIs on the CSS management console: **Publishing a KooSearch API**.
3. Call the published KooSearch APIs in your service environment: **Calling a Published KooSearch API**.

To change the user authentication method for a published KooSearch API, perform **Editing an API**.

To cancel a published KooSearch API, making it unavailable, perform **Taking an API Offline**.

Configuring an API Gateway

1. Create an API gateway. For details, see **Creating a Gateway**.

   NOTE

   The gateway must be in the same VPC and subnet as KooSearch.

2. Create an API group. For details, see **Creating an API Group**. An API group contains APIs used for the same service. You must create a group before creating an API.
3. Create the environment where the API can be called. For details, see **Configuring the Environment and Environment Variables**. APIs can be published in different customized environments, such as development and testing environments. RELEASE is the default environment provided by APIG.

Publishing a KooSearch API

Publish a KooSearch API in a target environment.

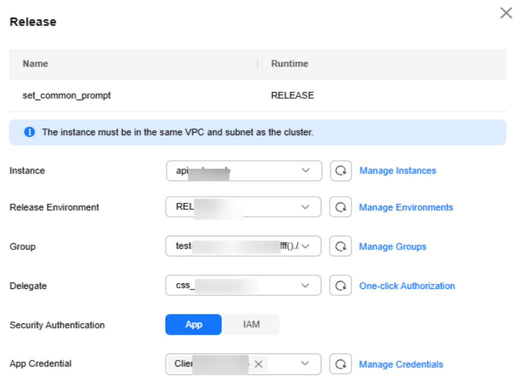
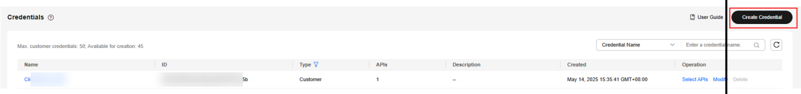
1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Click the **API Management** tab.
3. Select the API you want to publish and click **Publish** in the **Operation** column.
4. On the displayed page, configure the API gateway information.

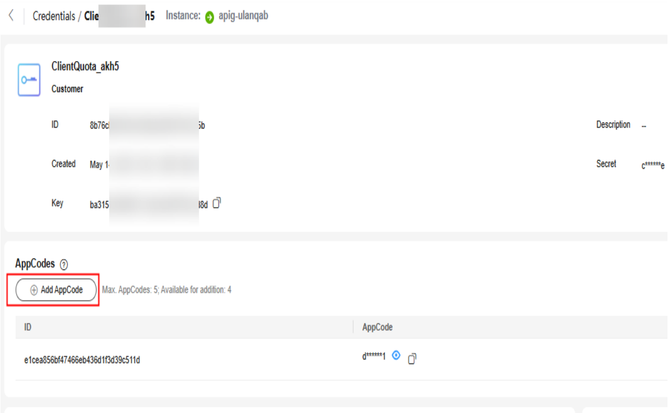
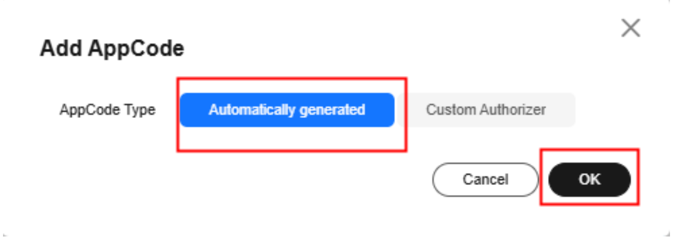
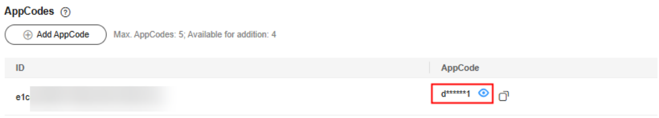
Table 4-1 Publishing a KooSearch API

Parameter	Description
Instance	To use APIG, you need to buy an instance first. To do that, click **Manage Instances** on the right. For details, see **Creating a Gateway**. **NOTE** The gateway must be in the same VPC and subnet as KooSearch.
Release Environment	An API can be called in different scenarios, such as the production environment (RELEASE) or other custom environments. For this parameter, you are advised to select **RELEASE**, which is the default online environment for official APIs. Only the APIs released in **RELEASE** can be made available for sales. For how to create a custom environment, see **Configuring the Environment and Environment Variables**.
Group	An API group contains different APIs used for the same service. You are advised to select the default group **DEFAULT**, which is automatically generated by the system. All APIs in the group can be accessed via an EIP or private IP address. For how to create a custom group, see **Creating an API Group**.

Parameter	Description

Agency Name	Select an IAM agency to grant the current account the permission to access and use APIG.
	● If you are configuring an agency for the first time, click **Automatically Create IAM Agency** to create **css-apig-agency**.
	● If there is an IAM agency automatically created earlier, you can click **One-click authorization** to have the permissions associated with the **APIG Administrator** role or the **APIG FullAccess** system policy deleted automatically, and have the following custom policies added automatically instead to implement more refined permissions control.
	`"apig:vpcChannels:*",`
	`"apig:apis:*",`
	`"apig:instances:*",`
	`"apig:envs:*",`
	`"apig:groups:*",`
	`"apig:apps:*"`
	● To use **Automatically Create IAM Agency** and **One-click authorization**, the following minimum permissions are required:
	`"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"`
	● To use an IAM agency, the following minimum permissions are required:
	`"iam:agencies:listAgencies",`
	`"iam:agencies:getAgency",`
	`"iam:permissions:listRolesForAgencyOnProject",`
	`"iam:permissions:listRolesForAgency"`

Parameter	Description
Security Authentication	Two authentication methods are available: App authentication (recommended) and IAM authentication. ● **App**: Requests for the API will be authenticated by APIG. App authentication has multiple authentication paths. AppCode authentication is recommended. – An AppCode is part of a credential used to call APIs in simple authentication mode. In this mode, the **X-Apig-AppCode** parameter (whose value is an AppCode on the credential details page) is added to the HTTP request header for quick response. APIG verifies only the AppCode, yet the request content does not need to be signed. The procedure is as follows: 1. Choose **Cloud Secret Management Service** > **Secrets**. **Figure 4-1** Cloud Secret Management Service  Release Name: set_common_prompt; Runtime: RELEASE The instance must be in the same VPC and subnet as the cluster. Instance: api[redacted] — Manage Instances Release Environment: REL[redacted] — Manage Environments Group: test[redacted]fff()./ — Manage Groups Delegate: css_[redacted] — One-click Authorization Security Authentication: App (selected) / IAM App Credential: Clie[redacted] — Manage Credentials 2. On the **Credentials** page, click **Create Credential**. On the displayed page, specify the necessary parameters. Name: credential name, which can contain 3 to 64 characters and must start with a letter. Only letters, digits, and underscores (_) are allowed. Description: credential description; length: 1 to 255 characters.  Credentials — User Guide — Create Credential Max. customer credentials: 50; Available for creation: 45 — Credential Name — Enter a credential name. Name: Cli[redacted]; ID: [redacted]5b; Type: Customer; APIs: 1; Description: --; Created: May 14, 2025 15:35:41 GMT+08:00; Operation: Select APIs, Modify, Delete 3. Click **OK**. 4. Click the name of the newly created credential to go to the credential details page. 5. Click **Add AppCode**.

Parameter	Description
	 Credentials / Clie[redacted]h5 Instance: apig-ulanqab ClientQuota_akh5 Customer ID 8b76c[redacted]5b Created May 1[redacted] Key ba315[redacted]8d Description -- Secret c******e AppCodes ⊕ Add AppCode Max. AppCodes: 5; Available for addition: 4 ID: e1cea856bf47466eb436d1f3d39c511d AppCode: d******1 6. Set **AppCode Type** to **Automatically generated**, and click **OK**.  Add AppCode AppCode Type: Automatically generated / Custom Authorizer Cancel OK 7. The generated AppCode is displayed in the list.  AppCodes ⊕ Add AppCode Max. AppCodes: 5; Available for addition: 4 ID: e1c[redacted] AppCode: d******1 After the AppCode is added, you can use it for simple authentication of API requests. When sending a request, add the **X-Apig-AppCode** parameter to the request header and omit the request signature. For example, when using curl, add the **X-Apig-AppCode** parameter to the request header and set the parameter value to the generated 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. After the credential is created, return to the Publish Credential page, select the newly created credential, and go to the next step. – Use a key pair (key and secret) for authentication. You need to create a credential on APIG to generate a credential ID and key pair. Then, bind the created credential to your API. This will allow access to your API while using App authentication. When you call an API, the client replaces the key pair in the SDK with your own key

Parameter	Description
	pair so that APIG can authenticate your identity. For how to create a credential, see **Configuring Authentication Credentials**. ● **IAM**: Requests for the API will be authenticated by IAM. NOTE If you set the authentication mode of an API to **Huawei IAM**, any APIG user can access the API, which can result in excessive charges if the API is attacked by malicious traffic.

5. Click **OK**.

   When the API status changes to Published (or Released), the API has been published to the target environment and is now callable.

Calling a Published KooSearch API

Call a published KooSearch API in your environment.

For details, see .

Editing an API

The authentication mode can be changed for a published API.

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Click the **API Management** tab.
3. Select a published API and click **Edit** in the **Operation** column.
4. On the **Edit** page, change the API's authentication mode.
   - **IAM**: Requests for the API will be authenticated by IAM.

     NOTE

     If you set the authentication mode of an API to **Huawei IAM**, any APIG user can access the API, which can result in excessive charges if the API is attacked by malicious traffic.
5. Click **OK**.

Taking an API Offline

You can remove APIs that you do not need from the environments where the APIs have been published.

NOTE

This operation will make the APIs inaccessible in these environments. Ensure that you have notified the users before performing this operation.

1. Go to the details page of a KooSearch service.
   1. Log in to the **CSS management console**.
   2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   3. Click the name of the target service to go to the service details page.
2. Click the **API Management** tab.
3. Select an API and click **Offline** in the **Operation** column.
4. On the displayed page, set **Instance**, **Offline Environment**, and **Agency Name**, and then click **OK**.
5. The API is successfully taken offline when its status changes to Not Released.

5 Upgrading a KooSearch Service

This task updates kernel patches for KooSearch clusters.

Scenarios

How It Works

Cluster nodes are upgraded one at a time. During the upgrade, you remove a node from a cluster, change the node's OS image, then mount the node's original NIC port back to reuse its old IP address. Then you initialize the node to start its processes. After this node's information is updated, you proceed to another node and repeat these steps. During the upgrade, nodes are removed and then brought back one at time, which may cause service interruptions. Therefore, perform the upgrade during off-peak hours.

Upgrade Process

Step 1 Perform the pre-upgrade check: **Pre-Upgrade Check**.

The pre-upgrade check is mostly automated. A few of the items need to be checked manually.

Step 2 Create an upgrade task and start the upgrade: **Creating an Upgrade Task**.

----End

Constraints

- A maximum of 20 clusters can be upgraded simultaneously. You are advised to perform the upgrade during off-peak hours.
- Clusters that have ongoing tasks cannot be upgraded.
- Once started, an upgrade task cannot be stopped until it succeeds or fails.

Pre-Upgrade Check

To ensure a successful upgrade, you must check the items listed in the following table before performing an upgrade.

Table 5-1 Pre-upgrade checklist

Check Item	Check Method	Description	Normal Status
Cluster status	System check	After an upgrade task is started, the system automatically checks the cluster status. If the cluster status is **Available**, the cluster can provide services properly.	The cluster status is **Available**.
Resources	System check	After an upgrade task is started, the system automatically checks resources. During the upgrade, the OS image is changed for each node. Make sure the required resources are available.	Sufficient resources are available.
Non-standard operations	Manual check	Check whether non-standard operations have been performed in the cluster. Non-standard operations refer to manual operations that are not recorded. These operations cannot be automatically passed on through the upgrade, for example, modification of system settings and return routes.	Changes caused by non-standard operations will be lost upon an upgrade. To avoid potential impact on your services, manually back up these changes before the upgrade starts.

Creating an Upgrade Task

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Click the **Upgrade** tab.
3. On the displayed page, set upgrade parameters.

Table 5-2 Upgrade parameters

Parameter	Description
Target Image	Image of the target version. After you select an image, the image name and target version details are displayed below. The supported target versions are displayed in the **Target Image** drop-down list. If the target image cannot be selected, the possible causes are as follows: ● The current cluster is already using the latest version. ● The new image is not yet available in the current region.

Agency	When a node is deleted, NICs are released. This means you need to have VPC permissions. Select an IAM agency to grant the current account the permission to access and use VPC.
	● If you are configuring an agency for the first time, click **Automatically Create IAM Agency** to create **css-upgrade-agency**.
	● If there is an IAM agency automatically created earlier, you can click **One-click authorization** to have the permissions associated with the **VPC Administrator** role or the **VPC FullAccess** system policy deleted automatically, and have the following custom policies added automatically instead to implement more refined permissions control.
	`"vpc:subnets:get",`
	`"vpc:ports:*"`
	● To use **Automatically Create IAM Agency** and **One-click authorization**, the following minimum permissions are required:
	`"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"`
	● To use an IAM agency, the following minimum permissions are required:
	`"iam:agencies:listAgencies",`
	`"iam:agencies:getAgency",`
	`"iam:permissions:listRolesForAgencyOnProject",`
	`"iam:permissions:listRolesForAgency"`

4. After setting the parameters, click **Submit**.
5. Check the upgrade task in the task list. If the task status is **Running**, you can expand the task list and click **View Progress** to view the upgrade progress.

   If the task status is **Failed**, you can retry or terminate the task.

   – Retry the task: Click **Retry** in the **Operation** column.

   – Terminate the task: Click **Terminate** in the **Operation** column.

After an upgrade task is terminated, contact technical support to handle failed items.

6 Managing KooSearch Document Q&A Service

6.1 Viewing Details About KooSearch Document Q&A Service

On the service's basic information page, you can obtain the internal IP addresses for accessing the document parsing and knowledge management services, as well as the billing mode. Additionally, you can manage services, APIs, and logs.

- **Managing services**: You can manage clusters used for the KooSearch Document Q&A service on the CSS console.
- **Managing APIs**: After a KooSearch service is enabled, a KooSearch API is automatically created. By publishing this API in different environments, you allow users to access this KooSearch service by calling this API from these environments.
- **Managing logs**: You can query KooSearch service logs to locate and diagnose issues.

Viewing Information About a KooSearch Service

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Check the basic information and configuration of a KooSearch service.

Table 6-1 Basic information

Parameter	Description
Name	Service name.

Parameter	Description
ID	Unique ID of a service, which is automatically generated by the system.
Cluster Status	Current service status.
Internal document parsing address	Internal IP address for accessing the document parsing service.
Specifications	Service specifications.
Billing Mode	The service's billing mode.
Task Status	Current task status of the service. If there is no ongoing task, **--** is displayed.
Region	Region where the service locates.
Created	When the service was created.
Internal knowledge management address	Internal IP address for accessing the knowledge management service

Table 6-2 Configuration

Parameter	Description
VPC	The VPC where the service is located.
Enterprise Project	Enterprise project to which the service belongs. You can click the project name to see more information about the enterprise project on the Project Management console.
Subnet	The subnet to which the service belongs.
Cluster Routing	KooSearch cluster route information. You can view, add, or modify routes for a cluster. For details, see **Configuring Cluster Routes for KooSearch Document Q&A Service** .
Security Group	Security group configured for the service. To modify the service's security group, click **Change Security Group** on the right.

Managing Dependent Services

KooSearch exists as a parent service for services that it depends on. For example, if KooSearch depends on an Elasticsearch vector database, the Elasticsearch cluster is a dependent service of KooSearch.

If any of the following operations is performed on the KooSearch service, it will also be performed on its dependent services in a synchronized manner: deletion, unsubscription, renewal, and change of billing mode.

You can manage dependent services for KooSearch on the CSS console.

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. On the **Dependent Services** tab, check the KooSearch service's dependent services.
3. Click **Go to Detail** in the **Operation** column to go to the Cluster Information page, where you can manage the cluster. For details, see ***Cloud Search Service User Guide***.

⚠ CAUTION

- For a dependent service of KooSearch, the following operations are disallowed: deletion, unsubscription, renewal, and change of billing mode.
- After the vector database used by KooSearch is scaled out or in, on the Model Management page of KooSearch, the connectivity of some model services may become abnormal. In this case, perform the following steps to restore connectivity:
  1. Click **Edit** on the right of the abnormal model service.
  2. Click **OK** without modifying anything.

  The connection is automatically re-established, and the status changes back to Normal.

6.2 Configuring Cluster Routes for KooSearch Document Q&A Service

Configure cluster routes to allow a KooSearch service to proactively connect to the public network or enable cross-network KooSearch API access.

Procedure

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Click **Add Route** next to **Cluster Route**.
3. In the displayed dialog box, configure the route information.

Table 6-3 Configuring cluster routing

Parameter	Description
IP Address	Enter the first 16 or 24 bits (or first two or three octets) of the remote server's IP address. For example, if the source IP address is **192.168.1.1**, enter **192.168.0.0** in the text box.
Subnet Mask	Enter the subnet mask of the IP address. ● If the first 16 bits of the IP address are provided, enter **255.255.0.0** in **Subnet Mask**. ● If the first 24 bits of the IP address are provided, enter **255.255.255.0** in **Subnet Mask**. **NOTE** The subnet mask must cover the IP network segment. That is, after the subnet mask and IP address are converted into binary values, the number of 0s at the end of the IP address must be greater than the number of 0s at the end of the subnet mask.

4. Click **OK** to complete the cluster routing configuration.
5. Click **View Route** next to **Cluster Route**. In the **View Route** dialog box, check the updated route information.

6.3 Deleting KooSearch Document Q&A Service

Delete the KooSearch Document Q&A service to release resources.

Constraints

Deleting a service also deletes its data and the dependent cluster. Exercise caution.

Deleting a Pay-per-Use Service

1. Log in to the **CSS management console**.
2. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
3. Locate the service you want to delete, and click **Delete** in the **Operation** column.
4. In the displayed dialog box, manually type in **DELETE**, and click **OK**.

Deleting a Yearly/Monthly Service

You can unsubscribe from a service billed on a yearly/monthly basis. The service will be released and all its data will be permanently deleted.

1. Log in to the **CSS management console**.
2. Locate the service you want to unsubscribe from, and click **Unsubscribe/Release** in the **Operation** column.

3. In the displayed dialog box, manually type in **YES**, and click **OK**.

   On the **Unsubscribe from Resource** page, confirm the resource information and refund amount.

4. Select a reason for unsubscription, select the acknowledgement check boxes, and click **Confirm**.

   In the displayed confirmation dialog box, click **Yes**.

   NOTE

   If the service is in **Available** state, an unsubscription order will be generated for refund, and then the service will be deleted. If the service has already expired or has been frozen, it will be directly deleted. For more information about unsubscribing from a service, see **Unsubscribing from In-Use Resources**.

7 Managing the Logs of KooSearch Document Q&A Service

Query KooSearch service logs to locate and diagnose issues.

Querying Logs

1. Go to the details page of a KooSearch service.
   a. Log in to the **CSS management console**.
   b. In the navigation pane on the left, choose **KooSearch** > **KooSearch Document Q&A**.
   c. Click the name of the target service to go to the service details page.
2. Click the **Logs** tab.
3. Search logs on the **Logs** tab.

   Select a node and then click 🔍 to search for its logs.
   - When you search for logs, the latest 10,000 logs are matched, but only a maximum of 100 logs will be displayed.
   - You can search logs by entering keywords in the search box.