KooSearch

Product Overview

Issue	01
Date	2025-08-12

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions

HUAWEI and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 What Is KooSearch?.......... 1

2 Advantages.......... 4

3 Use Cases.......... 6

4 Security.......... 7

4.1 Shared Responsibilities.......... 7

4.2 User Authentication and Access Control.......... 9

4.3 Data Protection.......... 9

4.4 Auditing and Logging.......... 9

4.5 Certificates.......... 10

5 Constraints.......... 12

6 Related Services.......... 14

7 Concepts.......... 16

1 What Is KooSearch?

Overview

KooSearch offers an enterprise-grade, out-of-the-box retrieval augmented generation (RAG) service. By integrating enterprise knowledge bases consisting of both structure and unstructured data, an industry-leading search model, a high-performance CSS vector database, and a variety of large language models (LLMs), KooSearch allows you to quickly build your own AI search and document Q&A applications. As a promising way to mitigate LLM hallucination, RAG helps LLMs generate more accurate, reliable, and secure answers.

An Elasticsearch-based vector database hosted by Cloud Search Service (CSS) powers a vector search engine. CSS is a fully managed, distributed search service based on open-source Elasticsearch and OpenSearch. You can use it for structured and unstructured data search, and enable vector-based composite search, statistical analysis, and reporting. It provides vector search capabilities for KooSearch.

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**.

Highlights

- **Out-of-the-box Q&A service**

  KooSearch provides an enterprise-grade RAG service that:

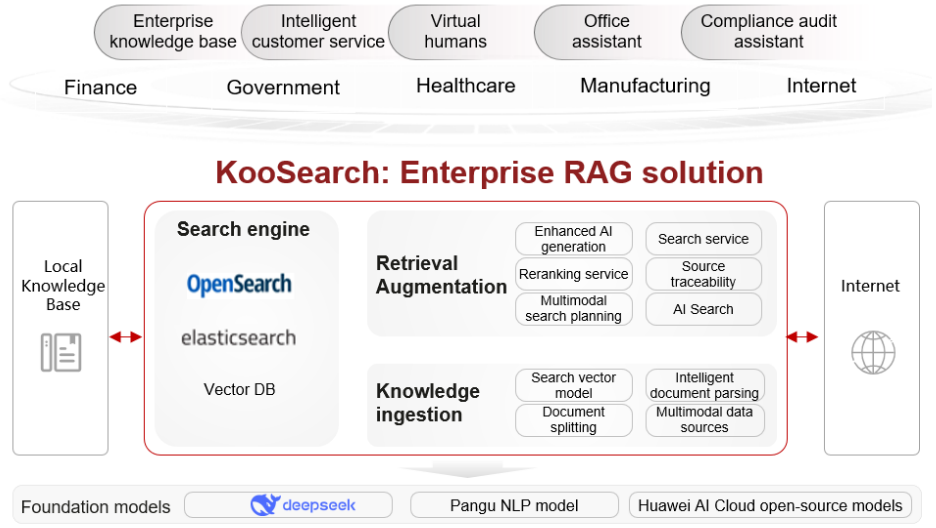
  - Supports a variety of document formats, including .doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, and .md.
  - Provides document parsing and splitting services and enhanced OCR for PDFs, scanned copies, images, and tables.
  - Supports multiple splitting methods: automatic splitting, hierarchical splitting, length-based splitting, and custom rules-based splitting.
  - Supports full-text search, vector search, and hybrid search.
  - Supports Q&A on structured documents, such as FAQs.
  - Supports tag and directory management.

- **AI search**

  An enhanced web search service gives LLMs access to up-to-date information available on the Internet. On top of this, AI search integrates LLMs and search planning to deliver a smart search experience.
- **Flexible configuration**
  - On KooSearch's model management page, you can choose any LLM that complies with OpenAI standards to provide the Q&A service.
  - KooSearch provides a prompt management function, where you can manage frequently used prompts for Q&A.

Product Architecture

Figure 1-1 KooSearch architecture



KooSearch supports practical use cases such as knowledge Q&A and AI search.

- KooSearch supports local knowledge bases. After intelligent parsing and splitting, local documents are vectorized and stored in the CSS vector database. The top-K documents are retrieved and ranked, and returned to an LLM, which then summarizes these results to generate answers.
- KooSearch supports AI search. It integrates an enhanced web search service, coupled with search planning, giving LLMs access to up-to-date information on the Internet and enhancing their performance.

Access Methods

KooSearch provides HTTPS-based APIs as well as a web-based console, which you can use to easily access the service.

- Using APIs

  If you want to integrate KooSearch into a third-party system for secondary development, use the APIs to access KooSearch.
- Web-based console

  For visualized operations, use the KooSearch web console.

If you have already registered on the public cloud platform, log in to the cloud service management console and click **Cloud Search Service**.

If you have not registered, **Signing Up for a HUAWEI ID and Enabling Huawei Cloud Services**.

2 Advantages

KooSearch has the following advantages:

Out-of-the-box usability

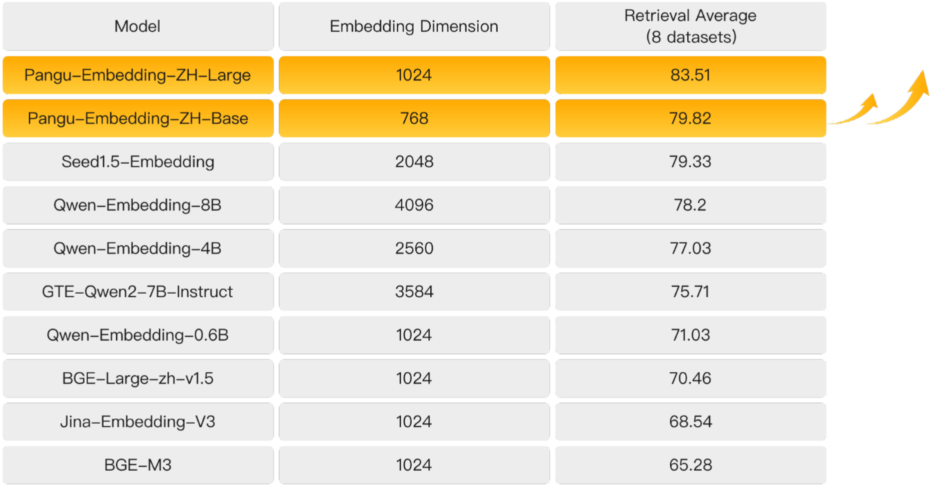
Document Q&A can be made ready immediately after documents are uploaded. You can switch between different LLMs on the model management page.

High accuracy

KooSearch uses a high-accuracy text embedding model, which has a high ranking on C-MTEB (Chinese Massive Text Embedding Benchmark). It also provides built-in reranking and search planning services; and supports multiple search methods, such as keyword-based search, vector search, and hybrid search. Based on deep document understanding, KooSearch can accurately identify the layout of documents from complex unstructured data, including titles, paragraphs, line breaks, headers, and footers.

KooSearch supports flexible text slicing—It can slice documents based on their directory structure or custom rules. The whole slicing process is visualized and can be manually adjusted to ensure end-to-end accuracy.

Figure 2-1 High ranking on C_MTEB

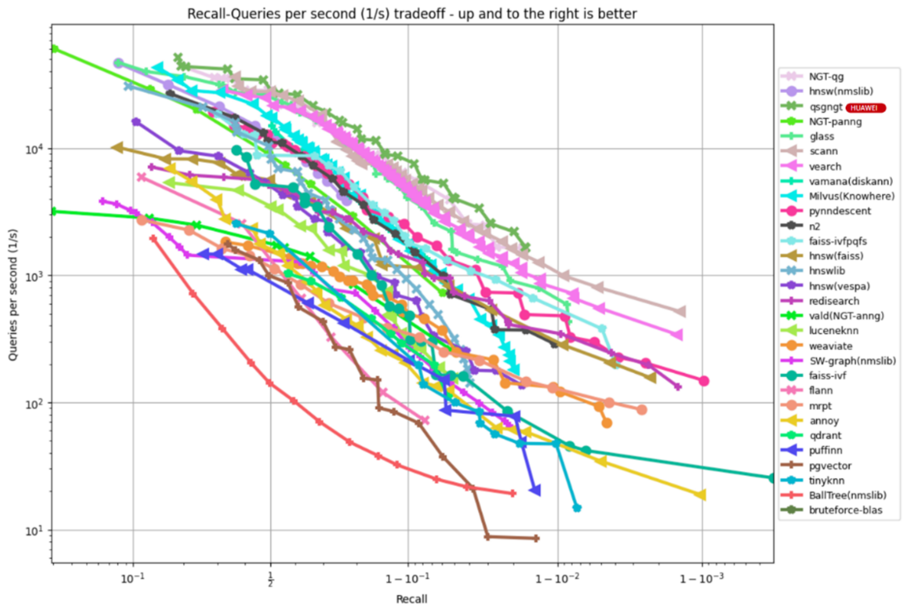


Model	Embedding Dimension	Retrieval Average (8 datasets)
Pangu-Embedding-ZH-Large	1024	83.51
Pangu-Embedding-ZH-Base	768	79.82
Seed1.5-Embedding	2048	79.33
Qwen-Embedding-8B	4096	78.2
Qwen-Embedding-4B	2560	77.03
GTE-Qwen2-7B-Instruct	3584	75.71
Qwen-Embedding-0.6B	1024	71.03
BGE-Large-zh-v1.5	1024	70.46
Jina-Embedding-V3	1024	68.54
BGE-M3	1024	65.28

High Performance

The KooSearch uses a CSS vector database, which has a high ranking on ANN-Benchmarks. This database supports multiple types of indexes, such as Flat, Graph, IVF, IVF_Graph, and PQ, and is fully compatible with the Elasticsearch ecosystem. It achieves a perfect balance between performance and precision.

Figure 2-2 High ranking on ANN-Benchmarks



Security

KooSearch supports physical multi-tenancy for secure tenant isolation; fully managed services; fine-grained permission management, including LDAP integration; and knowledge base isolation.

3 Use Cases

KooSearch can be used to quickly build enterprise-grade Retrieval-Augmented Generation (RAG) and AI search solutions. These solutions enhance efficiency in accessing private-domain knowledge, and by enabling Internet access, they enable an intelligent Q&A service to generate answers based on the latest public-domain knowledge from the Internet. KooSearch can power knowledge Q&A capabilities for a wide range of intelligent applications, such as intelligent customer service, virtual humans, digital employees, AI assistants, and AI search.

Enterprise-Grade RAG

There is a high fragmentation of knowledge, which can be attributed to the following factors: vast amounts of data from different domains, highly specialized, domain-specific knowledge, and disparate service systems storing documents in various formats (such as Word, PDF, Excel, PPT, and images). This fragmentation often results in inefficient knowledge management and access and low search accuracy. With the KooSearch Enterprise RAG solution, you can store all kinds of documents in a unified knowledge base to enable accurate knowledge search, and an LLM, with its reasoning capabilities, can organize and summarize the search results to generate reliable answers to user queries. (The answers contain links to the source documents.) Users can quickly and accurately find the knowledge they are looking for through a single-turn or multi-turn dialog with AI. Additionally, the RAG solution offers Internet access, allowing answers to be generated based on the latest public-domain knowledge from the Internet.

4 Security

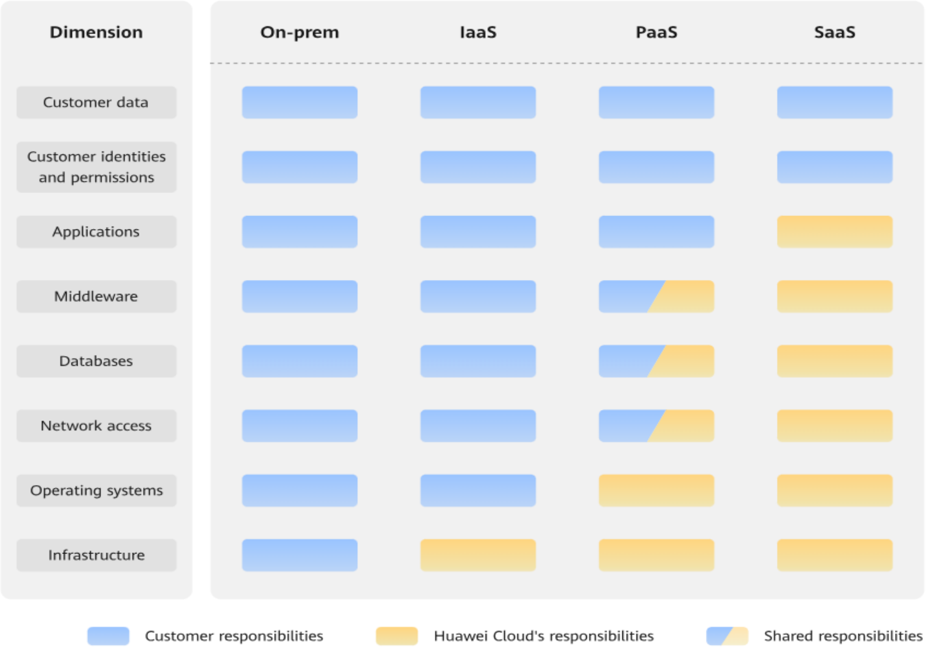
4.1 Shared Responsibilities

Huawei guarantees that its commitment to cyber security will never be outweighed by the consideration of commercial interests. To cope with emerging cloud security challenges and pervasive cloud security threats and attacks, Huawei Cloud builds a comprehensive cloud service security assurance system for different regions and industries based on Huawei's unique software and hardware advantages, laws, regulations, industry standards, and security ecosystem.

Unlike traditional on-premises data centers, cloud computing separates operators from users. This approach not only enhances flexibility and control for users but also greatly reduces their operational workload. For this reason, cloud security cannot be fully ensured by one party. Cloud security requires joint efforts of Huawei Cloud and you, as shown in **Figure 4-1**.

- **Huawei Cloud**: Huawei Cloud is responsible for infrastructure security, including security and compliance, regardless of cloud service categories. The infrastructure consists of physical data centers, which house compute, storage, and network resources, virtualization platforms, and cloud services Huawei Cloud provides for you. In PaaS and SaaS scenarios, Huawei Cloud is responsible for security settings, vulnerability remediation, security controls, and detecting any intrusions into the network where your services or Huawei Cloud components are deployed.
- **Customer**: As our customer, your ownership of and control over your data assets will not be transferred under any cloud service category. Without your explicit authorization, Huawei Cloud will not use or monetize your data, but you are responsible for protecting your data and managing identities and access. This includes ensuring the legal compliance of your data on the cloud, using secure credentials (such as strong passwords and multi-factor authentication), and properly managing those credentials, as well as monitoring and managing content security, looking out for abnormal account behavior, and responding to it, when discovered, in a timely manner.

Figure 4-1 Huawei Cloud shared security responsibility model



Dimension	On-prem	IaaS	PaaS	SaaS
Customer data	🟦	🟦	🟦	🟦
Customer identities and permissions	🟦	🟦	🟦	🟦
Applications	🟦	🟦	🟦	🟨
Middleware	🟦	🟦	◪	🟨
Databases	🟦	🟦	◪	🟨
Network access	🟦	🟦	◪	🟨
Operating systems	🟦	🟦	🟨	🟨
Infrastructure	🟦	🟨	🟨	🟨

🟦 Customer responsibilities 🟨 Huawei Cloud's responsibilities ◪ Shared responsibilities

Cloud security responsibilities are determined by control, visibility, and availability. When you migrate services to the cloud, assets, such as devices, hardware, software, media, VMs, OSs, and data, are controlled by both you and Huawei Cloud. This means that your responsibilities depend on the cloud services you select. As shown in **Figure 4-1**, customers can select different cloud service types (such as IaaS, PaaS, and SaaS services) based on their service requirements. As control over components varies across different cloud service categories, the responsibilities are shared differently.

- In on-premises scenarios, customers have full control over assets such as hardware, software, and data, so tenants are responsible for the security of all components.
- In IaaS scenarios, customers have control over all components except the underlying infrastructure. So, customers are responsible for securing these components. This includes ensuring the legal compliance of the applications, maintaining development and design security, and managing vulnerability remediation, configuration security, and security controls for related components such as middleware, databases, and operating systems.
- In PaaS scenarios, customers are responsible for the applications they deploy, as well as the security settings and policies of the middleware, database, and network access under their control.
- In SaaS scenarios, customers have control over their content, accounts, and permissions. They need to protect their content, and properly configure and protect their accounts and permissions in compliance with laws and regulations.

4.2 User Authentication and Access Control

KooSearch is accessible through the CSS console. It reuses CSS's user authentication and access control mechanisms, which consist of two aspects: Identity and Access Management (IAM) for resource-level user authentication and access control; the cluster security mode and the associated user authentication and access control mechanisms (see the "Permission Management" part in *KooSearch User Guide*). These two aspects are independent of each other.

4.3 Data Protection

KooSearch employs the following methods to protect data and service security:

- Network isolation

  The entire network is divided into two planes: service plane and management plane. The two planes are isolated physically to ensure the security of the service and management networks.

  - Service plane: It is the cluster's network plane. It provides service channels and knowledge Q&A capabilities for users.
  - Management plane: It provides a management console that you can use to manage CSS.
- Host security

  This includes the following security measures:

  - VPCs and security group rules can be configured to enhance host security.
  - Network access control lists (ACLs) enable granular control of inbound and outbound traffic to regulate data flow across your network perimeter.
  - The internal security infrastructure (including network firewalls, intrusion detection system, and protection system) monitors all network traffic that enters or exits the VPC through an IPsec VPN.
- Data security

  The vector database used by KooSearch uses multiple replicas, cross-AZ cluster deployment, and a third-party (OBS) backup solution to ensure data security.

4.4 Auditing and Logging

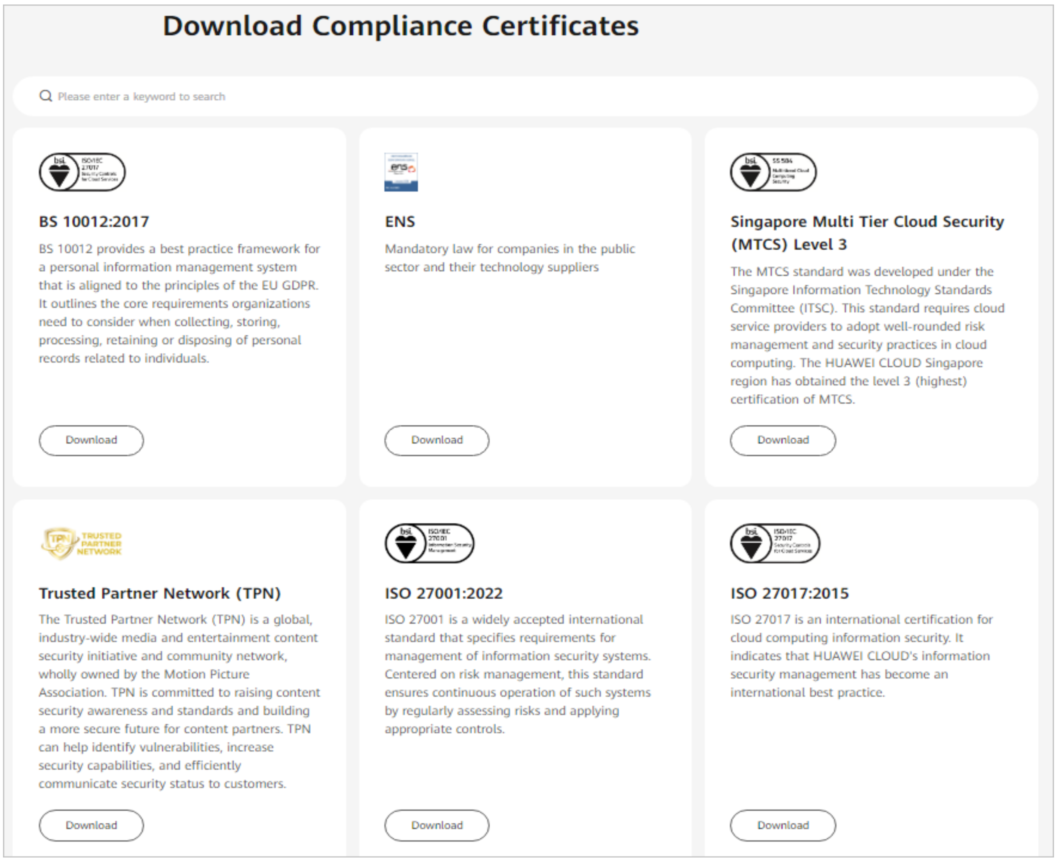
An Elasticsearch-based CSS vector database provides vector search capabilities for KooSearch, an enterprise-grade RAG service. This vector database reuses CSS's auditing and logging capabilities. For details, see **Audit and Logs**. KooSearch does not provide auditing capabilities.

4.5 Certificates

Compliance Certificates

Huawei Cloud services and platforms have obtained various security and compliance certifications from authoritative organizations, such as International Organization for Standardization (ISO). You can **download** them from the console.

Figure 4-2 Downloading compliance certificates



Download Compliance Certificates

Please enter a keyword to search

BS 10012:2017

BS 10012 provides a best practice framework for a personal information management system that is aligned to the principles of the EU GDPR. It outlines the core requirements organizations need to consider when collecting, storing, processing, retaining or disposing of personal records related to individuals.

Download

ENS

Mandatory law for companies in the public sector and their technology suppliers

Download

Singapore Multi Tier Cloud Security (MTCS) Level 3

The MTCS standard was developed under the Singapore Information Technology Standards Committee (ITSC). This standard requires cloud service providers to adopt well-rounded risk management and security practices in cloud computing. The HUAWEI CLOUD Singapore region has obtained the level 3 (highest) certification of MTCS.

Download

Trusted Partner Network (TPN)

The Trusted Partner Network (TPN) is a global, industry-wide media and entertainment content security initiative and community network, wholly owned by the Motion Picture Association. TPN is committed to raising content security awareness and standards and building a more secure future for content partners. TPN can help identify vulnerabilities, increase security capabilities, and efficiently communicate security status to customers.

Download

ISO 27001:2022

ISO 27001 is a widely accepted international standard that specifies requirements for management of information security systems. Centered on risk management, this standard ensures continuous operation of such systems by regularly assessing risks and applying appropriate controls.

Download

ISO 27017:2015

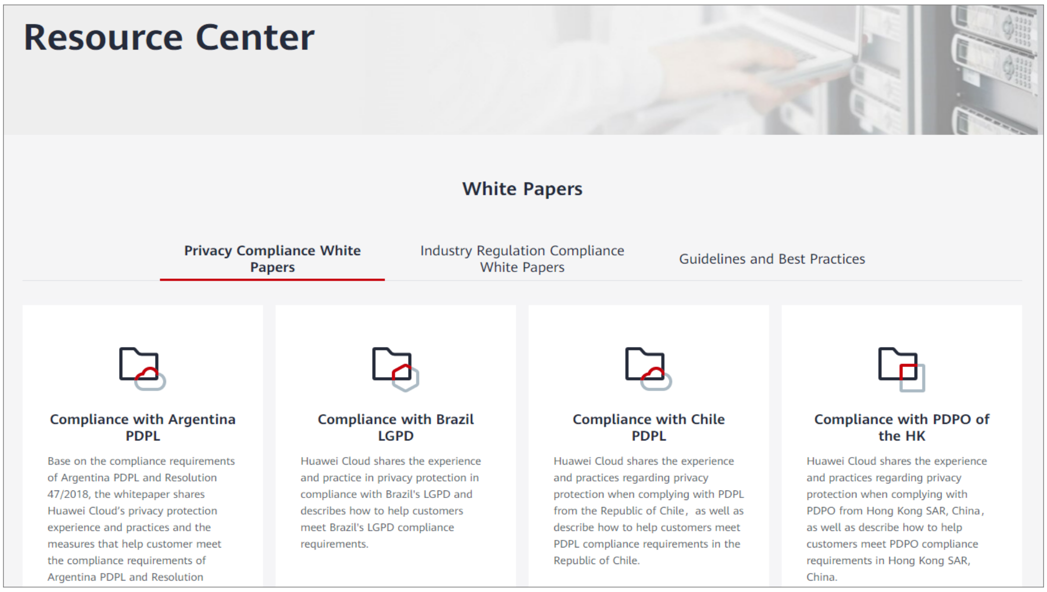
ISO 27017 is an international certification for cloud computing information security. It indicates that HUAWEI CLOUD's information security management has become an international best practice.

Download

Resource Center

Huawei Cloud also provides the following resources to help users meet compliance requirements. For details, see **Resource Center**.

Figure 4-3 Resource center



Resource Center

White Papers

Privacy Compliance White Papers | Industry Regulation Compliance White Papers | Guidelines and Best Practices

Compliance with Argentina PDPL

Base on the compliance requirements of Argentina PDPL and Resolution 47/2018, the whitepaper shares Huawei Cloud's privacy protection experience and practices and the measures that help customer meet the compliance requirements of Argentina PDPL and Resolution

Compliance with Brazil LGPD

Huawei Cloud shares the experience and practice in privacy protection in compliance with Brazil's LGPD and describes how to help customers meet Brazil's LGPD compliance requirements.

Compliance with Chile PDPL

Huawei Cloud shares the experience and practices regarding privacy protection when complying with PDPL from the Republic of Chile，as well as describe how to help customers meet PDPL compliance requirements in the Republic of Chile.

Compliance with PDPO of the HK

Huawei Cloud shares the experience and practices regarding privacy protection when complying with PDPO from Hong Kong SAR, China，as well as describe how to help customers meet PDPO compliance requirements in Hong Kong SAR, China.

5 Constraints

This topic describes notes and constraints on using KooSearch.

Constraints on Using KooSearch

- All input data provided to KooSearch, including but not limited to the data used to build knowledge bases, must fully comply with applicable laws and regulations.
- When using a third-party LLM, make sure the Pangu Guard model is used for content moderation, ensuring output security and compliance.
- Any output on the KooSearch experience platform is AI generated and does not represent the official positions or views of the service provider. Make sure your use of the service complies with applicable laws and regulations and positive societal values. The service provider makes no representations or warranties regarding any content on the KooSearch experience platform.
- Any output of the KooSearch AI search service is AI generated and does not represent the official positions or views of the service provider. Make sure your use of the service complies with applicable laws and regulations and positive societal values. The service provider makes no representations or warranties regarding any content associated with the KooSearch AI search service.

KooSearch Availability

- 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**.
- To subscribe to KooSearch, you must first complete real-name authentication as an enterprise user.

Product Specifications

Table 5-1 Specifications

Resource Type	Specifications	Description
Total number of knowledge bases	Maximum: 500	To use more knowledge bases, submit a service ticket to change CSS cluster node specifications. For details, see **Submitting a Service Request**.
Total number of documents per knowledge base	Maximum: 5,000	To include more documents in each knowledge base, submit a service ticket to change CSS cluster node specifications. For details, see **Submitting a Service Request**.

6 Related Services

This topic lists services that KooSearch depends on or is related to.

Table 6-1 Relationships between KooSearch and other services

Service	Description
Virtual Private Cloud (VPC)	KooSearch services are created in the subnets of a VPC. VPCs provide a secure, isolated, and logical network environment for your clusters. For details, see **Virtual Private Cloud User Guide**.
Elastic Cloud Server (ECS)	In the cluster that runs the CSS vector database used by KooSearch, each node is an ECS. When you create a cluster, ECSs are automatically created.
Elastic Volume Service (EVS)	The CSS vector database used by KooSearch uses EVS to store index data. When you create a cluster, EVS disks are automatically created.
Object Storage Service (OBS)	The CSS vector database used by KooSearch uses OBS to store cluster snapshots. For details, see **Object Storage Service User Guide**.
Identity and Access Management (IAM)	IAM authenticates access to KooSearch. For details, see **Identity and Access Management User Guide**.
Cloud Eye	KooSearch uses Cloud Eye to monitor the metrics of the CSS vector database in real time. For details, see **Cloud Eye User Guide**.
Cloud Trace Service (CTS)	With Cloud Trace Service (CTS), you can record operations associated with the vector database for query, auditing, and backtracking later. For details, see **Cloud Trace Service Guide**.
Optical Character Recognition (OCR)	KooSearch uses the intelligent document parsing service enabled by OCR to parse PDF documents and images.

Service	Description
ModelArts (AI development platform)	The embedding, reranking, and search planning models used by KooSearch are deployed using ModelArts and Ascend compute resources.
Model as a Service (MaaS)	On the KooSearch model management page, users can choose open-source models compatible with Ascend MaaS to use with KooSearch services.
Pangu Large Models (PLM)	On the model management page, users can choose Pangu models to use with KooSearch services.
API Gateway (APIG)	KooSearch APIs must be published on APIG before they can be made accessible from the Internet.

7 Concepts

RAG

Retrieval-augmented generation (RAG) is a technique that enhances large language models (LLMs) by integrating them with external knowledge bases. RAG leverages non-training data (such as up-to-date information and internal documents) to enhance the relevance and accuracy of AI-generated responses. At its core, RAG is about supplying LLMs with reliable knowledge retrieved through vector search to mitigate LLM hallucinations, enabling them to generate more reliable, knowledge-backed outputs.

Embedding model

An embedding model transforms text (such as words, phrases, or sentences) into dense vector representations (N-dimensional arrays) in a continuous vector space where semantic similarity can be measured through spatial distances. The model can support downstream tasks such as similarity search and semantic reasoning.

Reranking model

A reranking model reranks an initial result set by performing deep semantic matching, thus enhancing search result relevance. It operates by rescoring the top-k results retrieved during the recall phase (for example, through an embeddings-based vector search), and returning a refined subset of the most semantically relevant results.

Search planning

Search planning consists of two parts: multi-turn query rewriting and intent classification.

- With multi-turn query rewriting, an LLM rewrites user queries based on the chat history and generates new queries with clearer intents. It can also break down complex query questions into multiple simpler questions.
- Intent classification refers LLMs' ability to accurately identify users' query intents.