企业搜索服务

产品介绍

文档版本 01

发布日期 2025-08-11

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 什么是企业搜索服务 KooSearch

什么是企业搜索服务 KooSearch

华为云企业搜索服务KooSearch提供开箱即用的企业级RAG服务，导入非结构化或者结构化业务数据，内置业界效果突出的搜索模型、高性能CSS向量数据库，多种LLM灵活对接，帮助企业客户快速构建AI搜索和文档问答服务。搜索增强大模型，数据来源于搜索，解决大模型幻觉问题，问答结果更可靠、安全。

CSS向量数据库是基于云搜索服务（Cloud Search Service，简称CSS）的Elasticsearch集群提供的具有向量检索功能的搜索引擎。CSS服务是一个集成Elasticsearch、OpenSearch等引擎、且完全托管的在线分布式搜索服务，为用户提供结构化、非结构化文本、以及基于AI向量的多条件检索、统计、报表。为KooSearch的企业级RAG服务提供向量检索的能力。

说明

仅“香港”和“新加坡”区域支持开通和使用KooSearch服务。KooSearch是公测阶段，如果有试用需求，请提**工单**申请权限。

产品功能

- **开箱即用知识问答**

  KooSearch提供了企业级RAG的服务：

  - 支持上传的文档格式
    为.doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, .md。
  - 提供文档解析拆分服务，针对pdf、影印件、图片、表格提供OCR增强的功能。
  - 支持多种拆分方式：自动拆分、层次拆分、长度拆分和自定义规则拆分。
  - 支持全文检索、向量检索和混合检索多种搜索能力。
  - 支持FAQ等结构化文档问答。
  - 支持标签和目录管理。

- **AI搜索**

  联网增强服务是专为AI大模型设计的帮助大模型应用快速获取全网实时信息的服务。AI搜索在此基础上提供大模型内容总结、搜索规划增强能力。

- **灵活配置**

- KooSearch提供模型管理能力，用户可以在模型管理页面配置符合OpenAI标准规范的LLM，供问答时选择。
- KooSearch提供提示词管理能力，用户可以在提示词管理页面管理常用的提示词，供问答时选择。

产品架构

图 1-1 KooSearch 产品架构

KooSearch产品结构请参考**图1 KooSearch产品架构**。KooSearch可以实现知识问答，AI搜索等功能：

- 支持本地知识库：将本地的文档经过智能解析、切分、向量化后存入到CSS的向量数据库中，经过检索排序召回TopK的文档给大模型使用。
- 支持AI搜索：通过对接联网增强服务增加了搜索规划功能专为大模型使用，帮助大模型应用快速获取全网实时信息。

访问方式

KooSearch提供了Web化的服务管理平台，即基于HTTPS请求的API（Application Programming Interface）管理方式和管理控制台方式。

- API方式

  如果用户需要将公有云平台上的KooSearch集成到第三方系统，用于二次开发，请使用API方式访问KooSearch。
- 控制台方式

  可视化操作，请使用管理控制台方式访问KooSearch。

  如果用户已注册公有云，可直接登录管理控制台，在从主页选择“云搜索服务”。

  如果未注册，请参见**注册华为云并实名认证**。

2 产品优势

KooSearch主要有以下特点与显著优势。

开箱即用

上传文档后即可基于文档进行问答，支持页面模型管理灵活切换LLM模型。

高精度

内置高精度的文本Embedding模型，中文Embedding评测在C_MTEB榜单排名突出。同时内置重排、搜索规划服务，支持关键词检索、向量检索、混合检索等多种检索方式，基于深度文档理解，能够从各类复杂格式的非结构化数据中自动识别文档的布局，包括标题、段落、换行、页眉、页脚等。

支持文本切片，文本切片既可以按照目录结构进行切分，也可以按照自定义规则切分，文本切片过程可视化，支持手动调整，从而保证了端到端的准确率。

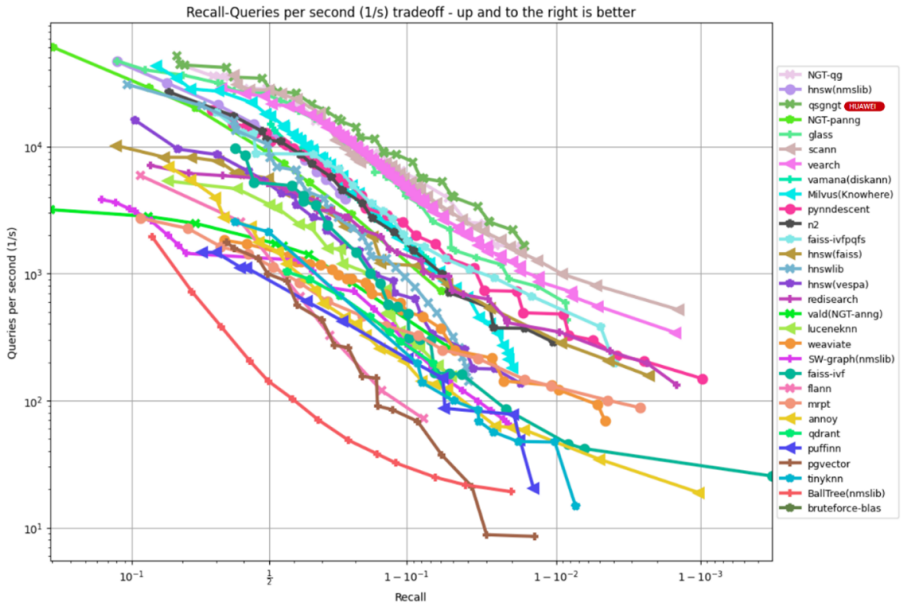
图 2-1 C_MTEB 榜单排名突出

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

高性能

Koosearch采用CSS向量数据库。CSS向量数据库ann-benchmarks打榜排名靠前，相同性能精度更好，相同精度性能更高；支持Flat、Graph、IVF、IVF_Graph、PQ等多种索引，同时支持Elasticsearch生态。

图 2-2 ann-benchmarks 打榜排名靠前



安全

支持物理多租、租户隔离、全托管服务，支持权限管理，支持对接LDAP，支持知识库级别的权限隔离，独享资源更安全。

3 应用场景

KooSearch可以帮助企业搭建企业级RAG，提升用户获取企业私有知识的效率，同时提供联网增强功能，提供互联网公域知识问答能力。主要用于智能客服应用、数字人、数字员工、AI助手、AI搜索等知识问答场景。

企业级 RAG

海量的行业数据、复杂的专业知识、各个业务系统的数据以及多种类型的文档（例如：word、pdf、Excel、ppt、图片等格式），容易导致知识碎片化，难以统一管理和调用，传统的知识检索准确率低，获取知识效率低。KooSearch企业RAG方案，通过将多样化的业务文档导入知识库，精准的知识检索，通过大模型推理能力进行相关知识推理问答，支持知识溯源，给客户提供精准可靠的信息，通过一轮或者多轮对话的方式快速获取知识，提升获取知识的效率。同时提供联网增强功能实现互联网公域知识问答，帮助客户快速构建AI搜索应用。

4 安全

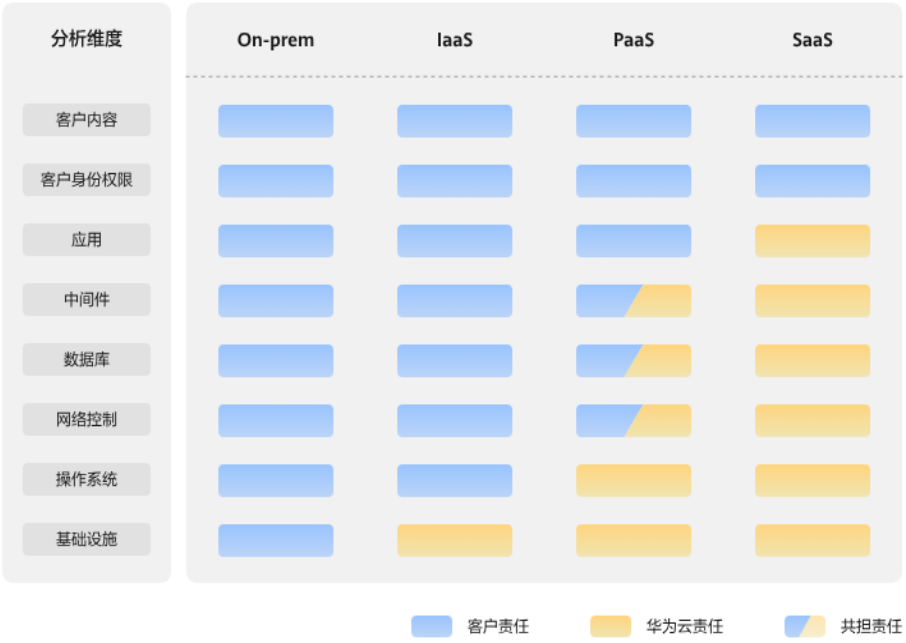
4.1 责任共担

华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

与传统的本地数据中心相比，云计算的运营方和使用方分离，提供了更好的灵活性和控制力，有效降低了客户的运营负担。正因如此，云的安全性无法由一方完全承担，云安全工作需要华为云与您共同努力，如图4-1所示。

- **华为云**：无论在任何云服务类别下，华为云都会承担基础设施的安全责任，包括安全性、合规性。该基础设施由华为云提供的物理数据中心（计算、存储、网络等）、虚拟化平台及云服务组成。在PaaS、SaaS场景下，华为云也会基于控制原则承担所提供服务或组件的安全配置、漏洞修复、安全防护和入侵检测等职责。
- **客户**：无论在任何云服务类别下，客户数据资产的所有权和控制权都不会转移。在未经授权的情况，华为云承诺不触碰客户数据，客户的内容数据、身份和权限都需要客户自身看护，这包括确保云上内容的合法合规，使用安全的凭证（如强口令、多因子认证）并妥善管理，同时监控内容安全事件和账号异常行为并及时响应。

图 4-1 华为云安全责任共担模型



云安全责任基于控制权，以可见、可用作为前提。在客户上云的过程中，资产（例如设备、硬件、软件、介质、虚拟机、操作系统、数据等）由客户完全控制向客户与华为云共同控制转变，这也就意味着客户需要承担的责任取决于客户所选取的云服务。如图4-1所示，客户可以基于自身的业务需求选择不同的云服务类别（例如IaaS、PaaS、SaaS服务）。不同的云服务类别中，每个组件的控制权不同，这也导致了华为云与客户的责任关系不同。

- 在On-prem场景下，由于客户享有对硬件、软件和数据等资产的全部控制权，因此客户应当对所有组件的安全性负责。
- 在IaaS场景下，客户控制着除基础设施外的所有组件，因此客户需要做好除基础设施外的所有组件的安全工作，例如应用自身的合法合规性、开发设计安全，以及相关组件（如中间件、数据库和操作系统）的漏洞修复、配置安全、安全防护方案等。
- 在PaaS场景下，客户除了对自身部署的应用负责，也要做好自身控制的中间件、数据库、网络控制的安全配置和策略工作。
- 在SaaS场景下，客户对客户内容、账号和权限具有控制权，客户需要做好自身内容的保护以及合法合规、账号和权限的配置和保护等。

4.2 身份认证与访问控制

KooSearch在CSS服务的控制台入口提供服务，身份认证和访问控制同CSS服务，CSS服务的身份认证和访问控制主要包括两个大的方面：一方面是通过统一身份认证服务（Identity and Access Management，简称IAM）实现服务资源层面的身份认证和访问控制；另一方面是由KooSearch服务复用CSS的安全集群内的身份认证和访问（具体请看《KooSearch用户指南》中权限管理）控制实现。两者是相互独立的模块。

4.3 数据保护技术

KooSearch主要从以下几个方面保障数据和业务运行安全：

- 网络隔离

  整个网络划分为2个平面，即业务平面和管理平面。两个平面采用物理隔离的方式进行部署，保证业务、管理各自网络的安全性。

  - 业务平面：主要是集群的网络平面，支持为用户提供业务通道，对外提供知识问答能力。
  - 管理平面：主要是管理控制台，用于管理云搜索服务。

- 主机安全

  提供如下安全措施：

  - 通过VPC安全组来确保VPC内主机的安全。
  - 通过网络访问控制列表（ACL），可以允许或拒绝进入和退出各个子网的网络流量。
  - 内部安全基础设施（包括网络防火墙、入侵检测和防护系统）可以监视通过IPsec VPN连接进入或退出VPC的所有网络流量。

- 数据安全

  在KooSearch服务中，向量数据库可以通过多副本、集群跨az部署、索引数据第三方（OBS）备份功能保证用户的数据安全。

4.4 审计与日志

CSS向量数据库基于Elasticsearch构建，提供向量检索功能。为KooSearch的企业级RAG服务提供向量检索的能力。其中，CSS的向量数据库复用云搜索服务的审计与日志的能力，详细介绍请参见**审计与日志**。KooSearch暂不支持审计。

4.5 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行**申请下载**合规资质证书。

图 4-2 合规证书下载

资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看**资源中心**。

图 4-3 资源中心

5 约束与限制

本节介绍KooSearch在使用过程中的约束和限制。

使用 KooSearch 限制

- 您输入的内容（包括但不限于构建知识库的数据）需要合法合规，不含有任何违反法律法规之内容。
- 配置第三方LLM服务时，请确认模型已对接Guard审核模型，确保模型输出内容安全合规。
- KooSearch的体验平台内容由人工智能模型合成，不代表平台立场或观点。请您确保输入内容和使用行为等符合适用的法律法规，积极构建、传播正向价值。本服务不对体验平台内容承担担保或保证责任。
- KooSearch的AI搜索内容由人工智能模型合成，不代表平台立场或观点。请您确保输入内容和使用行为等符合适用的法律法规，积极构建、传播正向价值。本服务不对AI搜索内容承担担保或保证责任。

购买 KooSearch 限制

- 仅“香港”和“新加坡”区域支持开通和使用KooSearch服务。KooSearch是公测阶段，如果有试用需求，请提**工单**申请权限。
- 企业用户实名认证后方可购买。

产品规格与限制

表 5-1 规格说明

资源类型	规格	说明
知识库总数	最大500个	如果您需要扩大知识库个数，可以**提交工单**申请变更CSS集群规格。
每个知识库文档总数	最大5000个文档	如果您需要扩大每个知识库文档总数，可以**提交工单**申请变更CSS集群规格。

6 与其他服务的关系

KooSearch与其他服务的关系如下所示。

表 6-1 KooSearch 与其他服务的关系

相关服务	交互功能
虚拟私有云（Virtual Private Cloud，简称VPC）	KooSearch创建在虚拟私有云（VPC）的子网内，VPC通过逻辑方式进行网络隔离，为用户的集群提供安全、隔离的网络环境。详细请参考**虚拟私有云用户指南**。
弹性云服务器（Elastic Cloud Server，简称ECS）	KooSearch使用CSS向量数据库集群时，集群中每个节点为一台弹性云服务器（ECS）。创建集群时将自动创建弹性云服务器作为节点。
云硬盘（Elastic Volume Service，简称EVS）	KooSearch使用CSS向量数据库集群时，集群使用云硬盘（EVS）存储索引数据，创建集群时，将自动创建云硬盘用于集群存储。
对象存储服务（Object Storage Service，简称OBS）	KooSearch使用CSS向量数据库集群时，集群快照存储在对象存储服务（OBS）的桶中，详细请参考**对象存储服务用户指南**。
统一身份认证服务（Identity and Access Management，简称IAM）	KooSearch使用统一身份认证服务（IAM）进行鉴权。详细请参考**统一身份认证服务用户指南**。
云监控服务（Cloud Eye）	KooSearch使用云监控服务实时监测向量数据库的指标信息，保障服务正常运行。详细请参考**云监控服务用户指南**。
云审计服务（Cloud Trace Service，简称CTS）	云审计服务（CTS）可以记录与向量数据库相关的操作事件，便于日后的查询、审计和回溯。详细请参考**云审计服务用户指南**。
文字识别（OCR）	KooSearch使用OCR的智能文档解析服务，对pdf、图片进行版式识别解析。

相关服务	交互功能
AI开发平台（ModelArts）	KooSearch使用昇腾资源部署的Embedding模型、Rerank模型、搜索规划模型依赖ModelArts的平台。
大模型即服务平台（MaaS）	KooSearch的模型管理支持配置大模型即服务平台MaaS上的昇腾云开源大模型。
盘古大模型（PanguLargeModels）	KooSearch的模型管理支持配置盘古大模型（PanguLargeModels）。
API网关（APIG）	KooSearch的API发布到公网访问，依赖API网关APIG。

7 基本概念

RAG

RAG（检索增强生成）是一种通过检索外部知识库来扩展模型知识范围的技术，利用非训练数据（如实时信息或内部文档）增强生成内容的准确性和时效性。其核心是通过向量检索获取相关知识片段输入大模型，从而减少生成幻觉并提供更可靠的响应。

Embedding 模型

Embedding模型的核心功能是将文本（如单词、短语或句子）转化为稠密的向量（即N维数组），从而在向量空间中建立可计算的语义表示。这种向量化表示能够通过空间距离反映语义相似性（例如"鸟"和"鸽子"的向量更接近），并支持下游任务如相似匹配和语义推理。

Rerank 模型

Rerank（重排序）模型是用于对初筛结果进行精细化排序的组件，它通过深度语义匹配优化检索结果的顺序。其核心作用是对召回阶段（如基于Embedding的向量检索）返回的Top K候选结果进行二次打分，结合上下文语义选出最相关的子集，显著提升最终排序的准确性。

搜索规划

包含2个功能，分别为多轮Query改写和意图分类。

- 多轮Query改写是指利用大型语言模型将用户当前查询与对话历史上下文结合，生成更完整、意图明确的查询的过程，同时支持复杂的Query原始问题拆解为多个问题
- 意图分类是指利用大型语言模型识别用户查询的意图。