

企业搜索服务 KooSearch

最佳实践

文档版本 01
发布日期 2025-07-17



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

华为云计算技术有限公司

地址：贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址：<https://www.huaweicloud.com/>

目 录

1 基于 KooSearch 和 DeepSeek 构建知识问答..... 1

2 基于 Dify 集成 KooSearch 与 DeepSeek 构建智能问答助手..... 7

1 基于 KooSearch 和 DeepSeek 构建知识问答

本文结合KooSearch和DeepSeek模型搭建一个基于RAG（ Retrieval Augmented Generation ）的知识问答系统。针对NLP大模型存在的知识幻觉、知识更新不及时以及缺乏企业专业知识等问题，大模型+知识库是当前普遍采用的解决方案。

KooSearch基于RAG架构，支持导入非结构化或结构化的业务数据到知识库，帮助企业客户快速、一站式地搭建企业智能问答应用。通过搜索增强大模型，数据来源于搜索，解决了大模型幻觉的问题，生成的问答结果更加可靠和安全。同时，您可以灵活配置LLM大语言模型，支持选择部署在ModelArts Studio大模型即服务平台（后续简称为MaaS）上的开源大模型，如DeepSeek模型。

应用场景

在人工智能应用中，RAG技术逐渐成为提升智能系统性能的关键手段。RAG结合了检索（Retrieval）和生成（Generation）的能力，通过检索相关知识并结合生成模型输出高质量结果。DeepSeek V3作为一款强大的开源大语言模型，特别适合用于构建RAG系统。

方案优势

- 开箱即用：上传文档后，即可进行基于文档的问答，并支持灵活的API接入。支持非结构化和结构化数据入库，提供多轮问答、可溯源及图文输出功能。
- 高精度：集成高精度文本Embedding模型，提供智能文档解析、重组及搜索规划服务，检索准确率高。
- 高性能：内置CSS向量数据库，具备高精度和高性能，支持Flat、Graph、IVF、IVF_Graph、PQ等多种索引，并兼容Elasticsearch生态。
- 安全：支持物理多租、租户隔离、全托管服务，独享资源更稳定、性能更高。
- 强大的知识检索与生成能力：集成DeepSeek模型服务，使得RAG系统能够从大规模知识库中高效检索相关信息，并生成高质量、准确的答案。

约束限制

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

前提条件

- 在云搜索服务管理控制台已创建一个KooSearch文档问答服务，请参见[开通KooSearch文档问答服务](#)。
- 已完成MaaS服务的访问授权，请参见[配置MaaS访问授权](#)。

步骤一：获取 DeepSeek 模型服务

在MaaS服务中获取DeepSeek模型服务，并记录API地址、模型名称和API Key，用于接入KooSearch。

- 登录ModelArts管理控制台，区域选择“西南-贵阳一”。
- 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
- 参考[使用ModelArts Studio的DeepSeek-R1模型框架实现对话问答](#)，在MaaS服务中获取DeepSeek模型服务。
- 在DeepSeek模型服务的调用说明页面，获取并记录“API地址”和“模型名称”，根据页面提示获取并记录API Key。

图 1-1 获取 DeepSeek 模型服务信息

调用说明



说明

- 最多支持创建5个密钥，密钥只会在新建后显示一次，请妥善保管。
- 当密钥丢失将无法找回，请新建API Key获取新的访问密钥。

步骤二：KooSearch 对接公网

通过配置NAT网关和弹性公网IP将KooSearch对接到公网，便于接入MaaS服务中的DeepSeek模型服务。

- 登录NAT网关管理控制台。

2. 在公网NAT网关列表页面，单击“购买公网NAT网关”，进入购买公网NAT网关页面。
3. 在购买公网NAT网关页面，配置NAT网关。
关键参数请参见表1-1配置，其他参数可以保持默认值或基于业务场景配置。

表 1-1 购买公网 NAT 网关

参数	说明
计费模型	建议选择“按需计费”。
区域	和KooSearch保持一个区域，选择“西南-贵阳一”。
虚拟私有云	和KooSearch选择同一个VPC。
子网	和KooSearch选择同一个子网。

4. 单击“立即购买”，确认产品规格无误后，单击“提交”，开始创建公网NAT网关。当状态为“运行中”时，表示公网NAT网关创建成功。
5. 当公网NAT网关创建成功后，在公网NAT网关列表页面，单击操作列的“设置规则”，在弹窗中单击“添加规则”。
6. 在“SNAT规则”页面，单击“添加SNAT规则”，在弹窗中配置公网NAT网关的SNAT规则。

表 1-2 添加 SNAT 规则

参数	说明
使用场景	选择“虚拟私有云”。
网段	选择“使用已有”，选择和KooSearch同一个VPC。
公网IP类型	选择“弹性公网IP”，勾选满足要求的IP。 如果没有可选的弹性公网IP，可以单击“查看弹性公网IP”跳转到弹性公网IP列表页， 申请弹性公网IP 。 弹性公网IP的建议配置如下： “计费模式”：“按需计费” “公网带宽”：“按流量计费” “带宽大小”：“20”

7. 单击“确定”完成SNAT规则的添加。

步骤三：在 KooSearch 接入 DeepSeek 模型

将MaaS服务中获取的DeepSeek模型服务接入KooSearch。

1. 登录云搜索服务管理控制台。
2. 在左侧导航栏选择“KooSearch > KooSearch文档问答”，进入KooSearch服务列表。
3. 单击KooSearch服务名称，进入服务详情页。

4. 给KooSearch配置集群路由，使KooSearch能够通过公网接入DeepSeek。
- a. 在服务详情页，单击“集群路由”后面的“添加路由”。

b. 在“添加路由”弹窗中，配置路由信息。

表 1-3 配置集群路由

参数	说明
ip地址	填写MaaS服务中DeepSeek模型服务的IP地址。 获取方式：在执行机ping DeepSeek模型服务的域名，解析后即可获得IP地址。例如获取到的DeepSeek模型服务的API地址是“https://example.example.com/v1/infers/xxxx”，则域名是“example.example.com”，在执行机执行命令 ping example.example.com ，回显如下，则表示IP地址是“xxx.xxx.xxx.xxx”。 正在 Ping example.example.com [xxx.xxx.xxx.xxx] 具有 32 字节的数据: 来自 xxx.xxx.xxx.xxx 的回复: 字节=32 时间=33ms TTL=55
子网掩码	填写“255.255.255.255”。

- c. 单击“确定”完成集群路由配置。
5. 返回KooSearch服务列表，选择服务，单击操作列的“问答”，前往KooSearch控制台。
6. 创建模型服务，接入DeepSeek。
- a. 在KooSearch左侧导航栏选择“配置管理 > 模型管理”，进入“模型管理”页面。

b. 单击右上角的“新建模型服务”，弹出新建模型服务页面。

图 1-2 新建模型服务



- c. 在新建模型服务页面，配置模型服务。

表 1-4 新建模型服务

参数	说明
模型服务名称	自定义模型服务名称，不能为空。
模型类型	选择“NLP模型-昇腾云”，表示通过昇腾云的MaaS服务提供的NLP大模型访问方式。
访问地址	填写MaaS服务中DeepSeek模型服务的“API地址”。获取方式请参见 步骤一：获取DeepSeek模型服务 。
模型描述	模型详细信息的描述。

参数	说明
昇腾云模型名称	填写MaaS服务中DeepSeek模型服务的“模型名称”，获取方式请参见 步骤一：获取DeepSeek模型服务 。
上下文长度（K）	NLP大模型在进行一次特定的推理时可以考虑的最大令牌数。可以扩展上下文以生成更全面的响应。
认证类型	勾选“自定义认证”，添加自定义请求头。“标头键”填写“Authorization”，“标头值”填写“Bearer+空格+DeepSeek的API Key”，例如获取到的API Key为“abcdef”，则此处填写“Bearer abcdef”。MaaS服务中DeepSeek模型服务的API Key获取方式请参见 步骤一：获取DeepSeek模型服务 。

- d. 单击“确定”，在弹出来的“新建模型服务声明”弹框中勾选同意，再单击“确认”创建模型服务。
- e. 当模型“连通性”显示“正常”时，表示模型对接成功。

体验知识问答

1. 创建知识库。
- a. 在KooSearch控制台的左侧导航栏选择“知识库管理”，进入知识库管理页面。
- b. 单击右上角“新建知识库”，进入新建知识库页面。
- c. 在“新建知识库”页面设置知识库信息。
在“模型设置”页签，设置“NLP模型服务”，选择上一步新建的模型服务。
其他参数可以保持默认值或基于业务场景配置。
- d. 设置好后，单击“确定”完成知识库创建。
- e. 在知识库列表，选择新建知识库，单击操作列的“文档管理”。进入知识库详情页面。
- f. 选择“文档管理”页签，单击“上传”。在上传对话框中单击“选择文档”，本地选择已提前准备好的文档。重复的文档不允许上传。

图 1-3 上传文档



- g. 单击“确认”。文档上传后，可在文档管理页签查看已上传的文档，当“文档状态”为“正常”，说明文档已上传成功。
2. 选择知识库。

- a. 在KooSearch控制台，左侧导航栏选择“体验平台”，进入体验平台页面。
- b. 单击右上角，在“引用来源”对话框勾选新建的知识库，单击“确定”。
3. 体验问答。
- KooSearch同时支持由大模型自身知识回答和通过文章内容回答。
- a. 在“体验平台”页面右上角单击“问答”，切换至问答体验页面。
- b. 在输入框中输入问题，单击，查看返回的答案。

图 1-4 体验问答



2 基于 Dify 集成 KooSearch 与 DeepSeek 构建智能问答助手

通过开源Dify平台集成企业搜索服务KooSearch的企业级知识库以及MaaS服务的DeepSeek模型，可构建具备以下能力的智能问答助手：

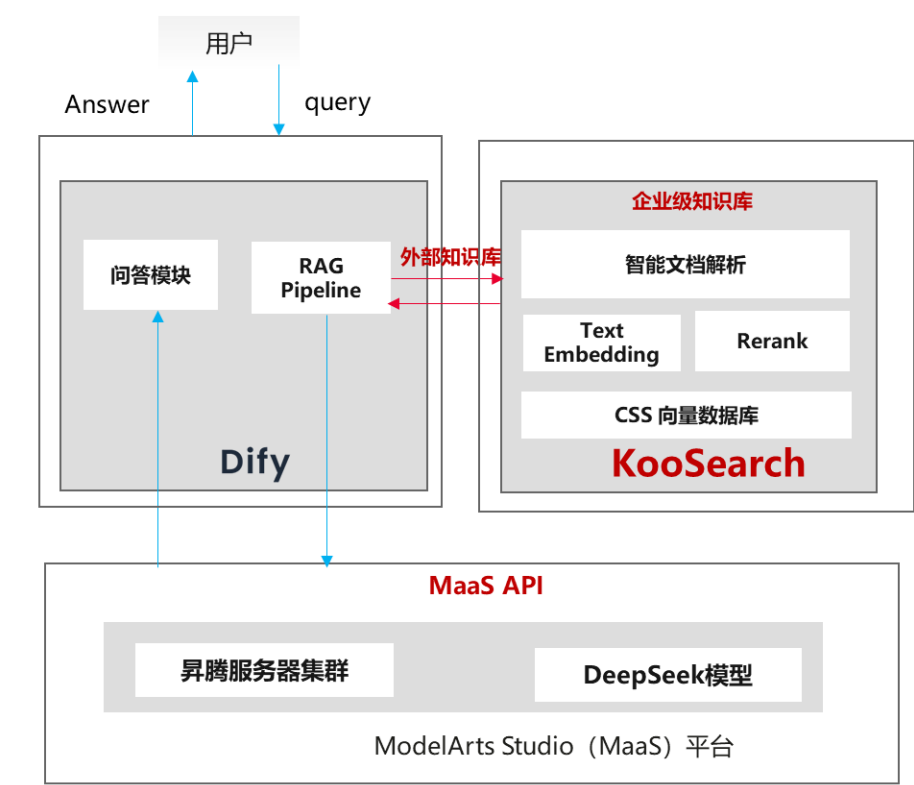
- 直接问答：调用DeepSeek模型生成回答。
- 知识增强问答：结合企业级知识库内容与DeepSeek模型输出精准答案。

应用场景

基于开源框架Dify构建，打造具备双重能力的智能对话助手。系统支持用户自由选择直接调用DeepSeek大模型进行开放式问答，或结合企业私有知识库实现精准信息检索与推理应答，适用于智能客服、知识库问答、个性化推荐等多种业务场景。KooSearch内置的OpenSearch向量数据库专注于检索优化，能够高效处理非结构化和结构化数据，而其独享版大模型服务则确保了企业数据的安全隔离和模型性能的稳定输出。DeepSeek模型的强大上下文理解和多轮对话能力进一步提升了系统的智能化水平，为企业提供了一个高效、灵活且安全的智能对话解决方案。

方案架构

图 2-1 方案架构图



1. 用户在KooSearch上传非结构化文档，经过智能文档解析和数据向量化，再存入CSS的OpenSearch向量数据库。
2. 用户通过Dify控制台提交查询请求。
3. 检索模块使用KooSearch的Embedding技术解析查询内容。
4. 在本地知识库（即CSS的OpenSearch向量数据库）中查找文档并重排序。
5. 检索到的文档经过MaaS的DeepSeek模型服务生成专业、可靠的回答。
6. 答案通过Dify控制台返回给用户。

其中，MaaS服务提供计算资源，DeepSeek模型执行复杂查询，是支持性组件。

方案优势

该方案适合需要快速构建企业级知识中枢、同时兼顾通用对话能力的组织，Dify集成KooSearch后具备如下增强优势：

- 开箱即用的**可视化配置界面**，显著降低AI应用开发门槛，企业可快速完成从数据接入到智能服务上线的全流程。
- 企业知识库高精度，KooSearch内置**高精度的文本Embedding**模型，同时内置智能文档解析、重排、搜索规划服务，提升了检索准确率。
- 企业知识库高性能，KooSearch采用**CSS向量数据库**，实现毫秒级语义检索，相同性能精度更好，相同精度性能更高。同时支持Flat、Graph、IVF、IVF_Graph、PQ等多种索引，兼容Elasticsearch生态。

- 企业知识库安全，KooSearch支持物理多租、租户隔离、全托管服务，独享资源更稳定、性能更高。

约束限制

- 仅“西南-贵阳一”区域支持开通和使用KooSearch服务。KooSearch是受限使用阶段，如果有试用需求，请提[工单](#)申请权限。
- 为确保Dify和KooSearch的正常交互，需保证二者网络互通。本案例中，二者部署于同一VPC内。

步骤一：获取 DeepSeek 模型服务

在MaaS服务中获取DeepSeek模型服务，并记录**API地址**、**模型名称**和**API Key**，用于接入KooSearch。

- 登录ModelArts管理控制台，区域选择“西南-贵阳一”。
- 在左侧导航栏中，选择“ModelArts Studio”进入ModelArts Studio大模型即服务平台。
- 参考[使用ModelArts Studio的DeepSeek-R1模型框架实现对话问答](#)，在MaaS服务中获取DeepSeek模型服务。
- 在DeepSeek模型服务的调用说明页面，获取并记录“API地址”和“模型名称”，根据页面提示获取并记录API Key。

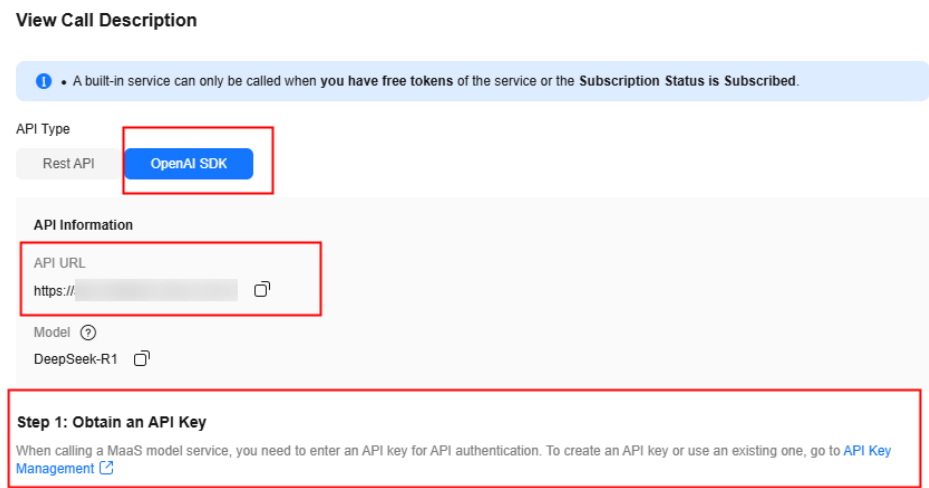
图 2-2 获取 DeepSeek 模型服务信息

调用说明



说明

- 最多支持创建5个密钥，密钥只会在新建后显示一次，请妥善保管。
- 当密钥丢失将无法找回，请新建API Key获取新的访问密钥。



步骤二：获取 KooSearch

- 获取并记录KooSearch的关键信息，用于后续配置Dify平台。
- 登录云搜索服务管理控制台，区域选择“西南-贵阳一”。
 - 在左侧导航栏中，选择“KooSearch > KooSearch文档问答”。
 - 参考[开通KooSearch文档问答服务](#)，购买KooSearch服务。必须启用“搜索模型配置”，否则会影响知识库使用。
 - 在KooSearch服务列表，单击目标KooSearch名称，进入详情页面，获取并记录“内网访问知识管理地址”。此处会获取到2个IP地址和端口号，例如“192.168.0.8:30275,192.168.0.170:30275”，任选一个IP地址和端口号即可作为KooSearch服务的IP地址和端口号。
- 当业务量大时，建议购买ELB挂载这2个地址，实现负载均衡，提升业务稳定性。操作步骤请参见[实现单个Web应用的负载均衡](#)。此时，KooSearch服务的IP地址即为ELB的服务地址，KooSearch服务的端口号即为监听器的端口信息。

图 2-3 获取 KooSearch 信息

基本信息					
名称	koo_...	产品规格	敏捷版 (KooSearch, 100, 100, 100)	区域	华北-北京-一区
ID	91b-451b-b0a6-2b...	计费模式	按需计费	创建时间	2025/03/05 09:51:21 GMT+08:00
集群状态	可用	任务状态	--	内网访问知识管理地址	192.168.0.8:30275, 192.168.0.170:30275
内网访问文档解析地址	192.168.0.8:30275, 192.168.0.170:30275				

- 在KooSearch服务列表，单击目标KooSearch操作列的“问答”进入KooSearch页面。
- 参考[创建KooSearch知识库](#)，新建知识库。
- 参考[将本地文件上传至KooSearch知识库](#)，上传文本文件。
- 在知识库管理列表，获取并记录目标知识库的“知识库ID”。

图 2-4 获取知识库 ID

知识库名称	知识库ID	知识库状态	创建人	创建时间	更新时间
Testcase_KooSea...	a1b-451b-b0a6-2b...	开启		2024/09/29 18:...	2024/09/29 18:...

步骤二：部署 Dify

在服务器上部署Dify容器，本案例以弹性云服务器ECS为例介绍操作步骤。Dify的详细安装指导请参见[Docker Compose部署](#)。

步骤1 参考[快速购买和使用Linux ECS](#)，购买ECS。

ECS的实例规则要大于或等于2U4G，虚拟私有云要和KooSearch一致，并且必须绑定弹性公网IP。

步骤2 在ECS实例上ping KooSearch服务的IP地址，验证网络连通性。KooSearch服务的IP地址请参见[步骤二：获取KooSearch](#)获取。

如果ECS和KooSearch在同一VPC下却无法ping通，则可以配置KooSearch的安全组规则，把ECS的私有IP地址添加到入方向规则。

步骤3 参考[手工部署Docker](#)，在ECS实例上安装Docker。

步骤4 克隆Dify的源代码到ECS实例上，并切换至某tag版本分支，以0.15.3版本为例。

```
git clone https://github.com/langgenius/dify.git && cd dify
git checkout -b 1.1.3 1.1.3
```

步骤5 进入docker目录，复制生成“.env”文件。

```
cd docker && cp .env.example .env
```

步骤6 执行命令，启动Dify容器。

```
docker compose up -d
```

步骤7 执行命令，确认容器是否启动成功。

```
docker ps
```

显示如下信息表示启动成功。

图 2-5 查看 Dify 容器

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS
9d08c792352c	nginx:latest	"sh -c 'cp /docker-e..."	2 days ago	Up About an hour	0.0.0.0:80->80/t
docker-nginx-1					
7d86418e447c	langgenius/dify-api:0.15.3	"/bin/bash /entrypoi..."	2 days ago	Up About an hour	5001/tcp
docker-worker-1					
a8186afc6833	langgenius/dify-api:0.15.3	"/bin/bash /entrypoi..."	2 days ago	Up About an hour	5001/tcp
docker-api-1					
29015a74b161	redis:6-alpine	"docker-entrypoint.s..."	2 days ago	Up About an hour (healthy)	6379/tcp
docker-redis-1					
d699a290f395	postgres:15-alpine	"docker-entrypoint.s..."	2 days ago	Up About an hour (healthy)	5432/tcp
docker-db-1					
520ac6aec14e	langgenius/dify-web:0.15.3	"/bin/sh ./entrypoin..."	2 days ago	Up About an hour	3000/tcp
docker-web-1					
6256e6fbbfc9	langgenius/dify-sandbox:0.2.10	"/main"	2 days ago	Up About an hour (healthy)	
docker-sandbox-1					
c67dabe27ea6	ubuntu/squid:latest	"sh -c 'cp /docker-e..."	2 days ago	Up About an hour	3128/tcp
docker-ssrf-proxy-1					

步骤8 通过ECS的弹性公网IP访问部署在ECS上的Dify平台。

在浏览器输入“http://{ECS公网IP}:80”访问Dify的开发平台。首次登录需注册管理员账号，依次填写邮箱、账号、密码。

如果是外网访问Dify平台，则还需要配置ECS的安全组规则，把外网客户端的IP地址添加到入方向规则。

----结束

步骤三：Dify 接入 KooSearch 和 DeepSeek

Dify作为一个开源的大语言模型（LLM）应用开发平台，通过配置KooSearch知识库和调用DeepSeek模型服务的API，来实现功能集成和扩展。

步骤1 在Dify平台接入KooSearch知识库。

- 1. 在Dify平台上方选择“知识库”。
- 2. 单击“连接外部知识库”。

图 2-6 进入连接外部知识库



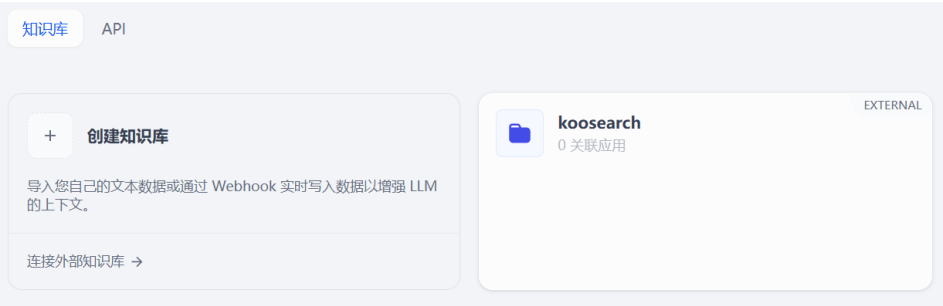
- 3. 在连接外部知识库页面配置KooSearch信息。配置说明参见表2-1。

表 2-1 添加 KooSearch

参数名称	参数值
外部知识库名称	自定义知识库名称，例如“koosearch”。
知识库描述	可选，添加KooSearch知识库的相关说明。
外部知识库API	单击加号创建外部知识库API，在弹窗中配置外部知识库API。 – “Name”：自定义KooSearch API名称，例如“koosearch_api”。 – “API Endpoint”：输入“http://{koosearch_ip}:{koosearch_port}”，koosearch_ip和koosearch_port即步骤二：获取KooSearch获取的KooSearch服务的IP地址和端口号。 – “API Key”：自定义Key，由于Dify和KooSearch在同一VPC，内网访问即可，无需API Key鉴权，此处输入任意数字和字母组成的字符串即可。
外部知识库ID	填写KooSearch知识库的ID，即步骤二：获取KooSearch获取的知识库ID。
召回设置	根据业务需要配置。

- 4. 配置完成后在知识库首页可以查看新接入的KooSearch知识库。

图 2-7 查看 KooSearch 知识库



5. 单击KooSearch知识库，进入召回测试页面，验证是否可以正常召回文本内容。如图2-8所示，回答正确，表示文档召回成功。

图 2-8 召回测试



步骤2 在Dify平台接入DeepSeek模型服务。

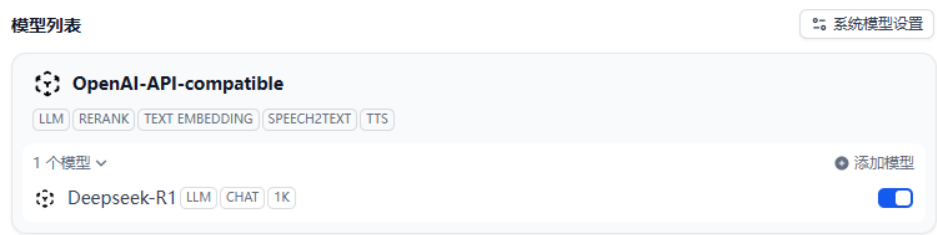
1. 单击Dify平台右上角的头像，选择“设置”中的“模型供应商”。
2. 在模型供应商页面中选择“OpenAI-API-compatible”，单击“添加模型”配置 DeepSeek信息。配置说明参见表2-2。

表 2-2 添加 DeepSeek

参数名称	参数值
模型类型	选择“LLM”。
模型名称	填写DeepSeek模型服务的模型名称，即步骤一：获取DeepSeek模型服务获取的模型名称。
API Key	填写DeepSeek模型服务的API Key，即步骤一：获取DeepSeek模型服务获取的API Key。
API endpoint URL	填写DeepSeek模型服务的Open API地址，即步骤一：获取DeepSeek模型服务获取的API地址。
最大token上限	填写“1024”。DeepSeek模型服务只支持1024。

3. 在模型列表可以查看接入的模型。

图 2-9 查看接入的 DeepSeek



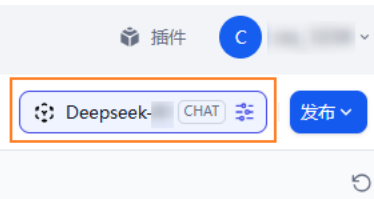
----结束

步骤四：创建智能问答助手

在Dify平台创建聊天助手，并进行调试预览，验证AI应用是否能够正常运行。

- 步骤1 在Dify平台上方选择“工作室”。
- 步骤2 单击“创建空白应用”，配置应用信息。“选择应用类型”必须选择“聊天助手”，“应用名称”可以自定义，例如chat。
- 步骤3 单击“创建”，进入应用编排页面。
- 步骤4 在编排页面右侧，确认应用已关联DeepSeek模型服务。

图 2-10 确认模型



- 步骤5 在编排页面调试智能问答助手，验证功能。

智能问答助手支持2种调试方式：直接问答和知识增强问答。

- 直接问答：聊天助手不对接知识库，直接调用DeepSeek模型生成问答。
该方式无需配置“知识库”，智能问答助手会直接使用DeepSeek回答问题，如图2-11所示。

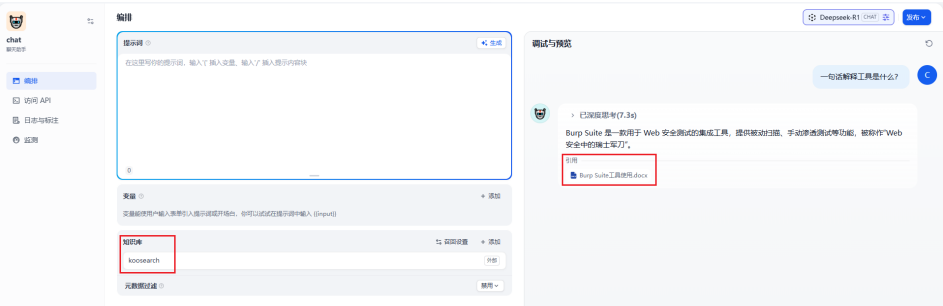
图 2-11 直接大模型问答



- 知识增强问答：聊天助手对接KooSearch知识库，并基于知识库内容，结合DeepSeek模型输出精准答案。

该方式需要先在“知识库”添加步骤1接入的KooSearch知识库，再进行问答，智能问答助手就会优先根据知识库的文档进行回答，如图2-12所示。

图 2-12 结合知识库问答



----结束