

d.run Token 工厂效能平台 · 云原生 AI

DaoCloud AI 应用手册

d.run 智能问答（AI 应用）使用指南

文档名称	DaoCloud AI 应用手册
产品名称	d.run 智能问答（AI 应用）
所属平台	d.run Token 工厂效能平台 · 云原生 AI
产品定位	基于 RAG（检索增强生成）的智能问答与语料库管理平台
内容范围	产品介绍、功能特性、应用中心、语料库、数据分析、我的反馈、进阶配置
整理来源	https://docs.daocloud.io/tf/dak/
更新时间	2026 年 9 月

本手册基于 DaoCloud 官方文档中心「AI 应用」全站页面整理，供实施与运维人员参考。

目录

一、AI 应用介绍	3
二、功能特性.....	4
三、应用中心.....	5
四、语料库	14
五、数据分析.....	23
六、我的反馈.....	32
七、进阶配置.....	34

一、AI 应用介绍

（一）什么是 AI 应用

d.run 智能问答（AI 应用）是基于 **RAG（Retrieval-Augmented Generation，检索增强生成）** 技术的智能对话产品。它会先将您的专属语料库进行 **向量化处理**，并通过 **相似度搜索** 精准检索与问题相关的内容，再由 **大语言模型（LLM）** 生成自然流畅的答案。这种方式不仅显著减少了模型的“幻觉”现象，还能在保障 **隐私数据安全** 的同时，为您提供更精准、更个性化的回复。

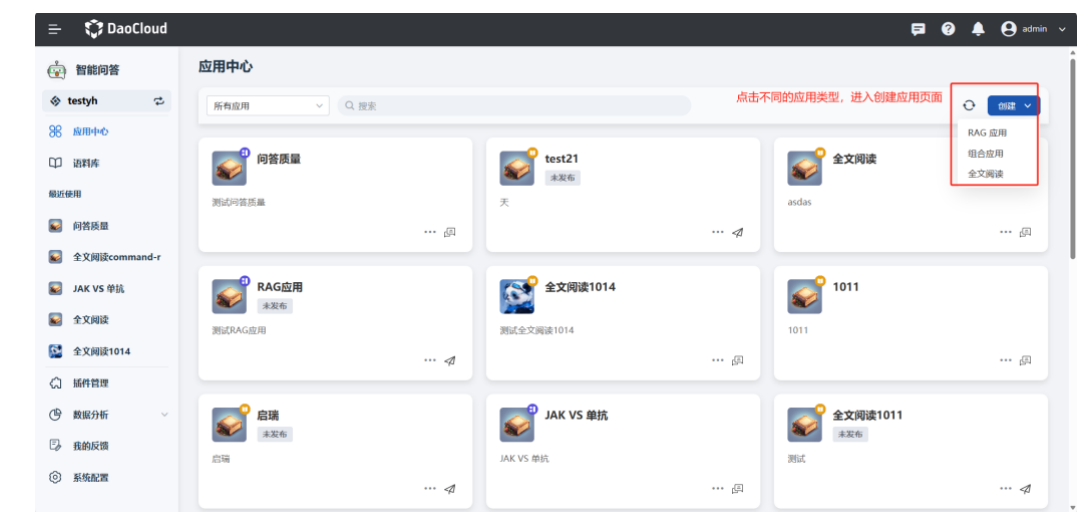


图 1 AI 应用中心

（二）平台能为您做什么

- **创建应用与插件：**将智能问答功能灵活集成到不同业务场景。
- **构建专属语料库：**导入文档、FAQ、业务数据等，让 AI 拥有您的专属知识。
- **实时优化：**通过对话日志与数据分析，持续改进语料与问答质量。
- **提升用户体验：**让回答更贴近实际需求，减少无效信息，增强交互价值。

无论是在线客服、知识管理，还是内部知识检索，d.run 智能问答都能帮助您快速构建 **高效、可信、可控** 的 AI 知识服务体系。

二、功能特性

AI 应用智能问答的功能特性按模块梳理如下。

（一）对话功能

功能	描述
上下文感知智能问答	关联应用语料库，实现具备上下文感知的智能问答功能
用户互动优化	支持用户对 AI 回答进行 评价、复制、重新生成、删除对话或提交反馈 ，提升用户体验

（二）应用中心

功能	描述
应用生命周期管理	提供从 创建、部署、监控到维护 的全生命周期管理
环境隔离	支持工作空间的绑定或解绑，实现 环境隔离 ，保障应用安全
模型服务集成	集成 GLM、Llama 等模型服务 ，提升 RAG（检索增强生成）应用的效率
多样化配置选项	提供 AI 配置、语料库关联、检索策略 等选项，优化 AI 回答质量
应用密钥支持	支持生成 应用密钥 ，用于 OpenAPI 对话的安全访问
应用链接分享	允许通过 链接分享应用 ，并支持 H5 页面展示以便于移动端使用
对话过程查看	支持查看完整的 对话过程 ，提升用户体验和问题排查效率
组合应用支持	支持 组合应用 进行意图识别，增强复杂场景处理能力
语言翻译支持	支持用户输入内容的 翻译 ，提高语料匹配的准确性
多语言回答选择	允许用户选择 AI 回答的 语言 ，满足多语言场景的需求
全文阅读应用支持	新增全文阅读应用，支持展示 图片内容
数字与化学公式支持	支持 数学公式和化学方程式 的显示，提升学术与科研场景的使用体验

（三）语料库管理

功能	描述
多种导入方式	支持 标准导入 、 格式化导入 、 手动导入 、 图文导入 等多种方式管理语料库
智能分片	支持按分割符或大小 自动分片 ，并可通过插件自定义分片逻辑
数据安全与隔离	支持设置语料库的访问级别，确保 数据安全与隔离
文件上传预览	在文件上传时，新增 分片信息预览 功能
便捷的数据导出	支持 CSV 和 Excel 格式的语料数据导出，方便后续分析与处理
OpenAPI 功能增强	支持 文件上传 、 分片处理 及 向量化 功能

（四）数据分析

功能	描述
关键指标分析	提供 问答质量 、 次数 、 分片质量 、 命中率 等关键指标的分析，帮助优化问答服务
用户反馈处理	收集并处理用户的 反馈信息 ，持续改进问答服务，提升满意度
反馈追踪与优化	追踪用户反馈 ，确保问题及时解决，不断优化用户体验

三、应用中心

「应用中心」是创建、发布和管理智能问答应用的入口。应用把语料库、模型和业务场景组合在一起，是使用 AI 应用的起点。

（一）创建应用

创建应用是将语料库、模型和业务场景结合起来，实现智能问答和数据检索的重要步骤。通过创建应用，您可以快速搭建专属的智能助手，满足具体的业务需求。

1. 创建 RAG 应用

1. 在 **应用中心** 页面中，点击 **创建** 按钮，选择创建 RAG 应用。

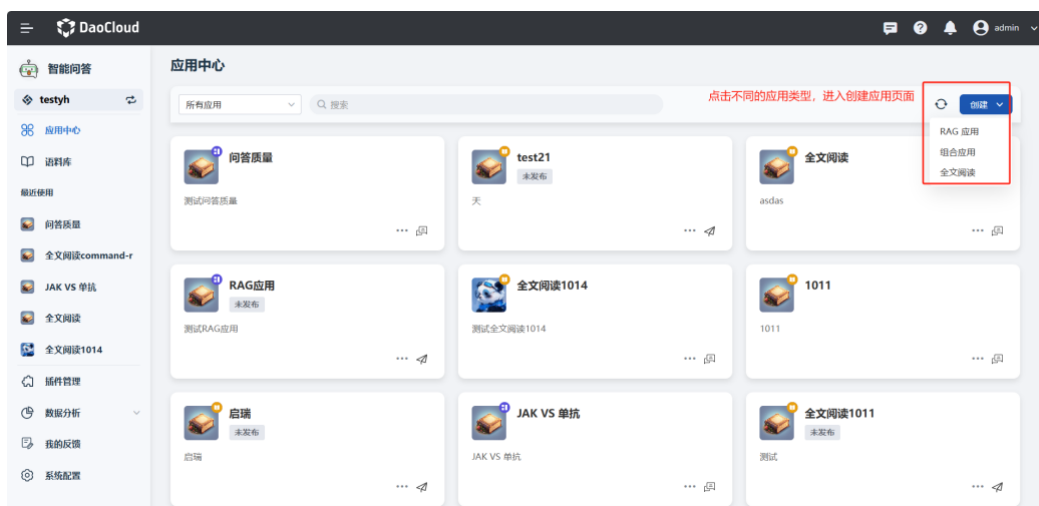


图 2 应用中心的创建入口

2. 填写应用基础信息：

- **上传应用图标：**选择 jpg、jpeg 或 png 格式的图片，确保文件大小不超过 10MB。
- **输入应用名称：**限制在 20 个字符以内。
- **填写应用简介：**简介应用，不超过 100 个字符。
- **应用调试：**选择开启。开启后，对话界面将显示对话的处理过程。

3. 完成 AI 配置：

- **选择大语言模型服务：**决定是使用本地模型服务还是在线模型服务（例如 Azure Open AI 或文心一言）。
- **设置随机度：**调整模型回答的创造性 / 发散性程度。
- **重复惩罚：**减少生成的文本中重复的内容，默认值为 1，值越大，越不会生成重复的内容。
- **向量化模型：**应用选择的向量化模型及向量化模型服务。
- **系统提示词：**指定模型所扮演的角色、它的能力，以及您希望它在跟您对话或者帮助您生成内容时的一些规则。
- **命中语料库时的模版：**为模型提供检索和回答问题的模板。
- **未命中语料库时的模版：**提供模型在检索不到相似语料时的通用提示词。

命中语料库时的模版：当检索到相似语料的时候，会使用语料库提示词模板。模板内容包含通过相似度搜索得到的 **知识块 {corpus_search_content}**、用户的输入 **{user_inputs_content}**。这些变量将被替换成对应的文本，发送给大模型，进行问答。

未命中语料库时的模版：当检索不到相似语料的时候，会使用嵌入提示词模板。嵌入的提示词拼接在用户的问题之前，作为通用的约定提示来引导应用模型进行回答输出。

命中语料库时的模版示例：

做一个知识问答游戏：

1. 回答内容必须在 "{corpus_search_content}" 中。

2. 如果问题在所提供的资料信息内无法找到, 您会回答:"抱歉, 资料库已知的信息中, 没有找到您需要的结果, 请尝试修改引用的资料库或使用模型自有能力回答。"

{user_inputs_content}

4. 关联语料库:

- **选择向量化模型:** 确定用于问题向量化的模型。
- **选择语料库:** 从已有的语料库中选择一个或多个用于模型检索。

5. 设置检索策略:

- **检索预处理:** 支持自定义插件进行问答预处理。
- **检索知识块数:** 决定提供给模型的知识块数量。
- **相似度:** 设置知识块匹配的严格程度。
- **重排序:** 启用或禁用重排序模型以改进结果排序。
- **仅语料回答:** 开启后, 模型只会按照匹配到的语料内容回答。
- **图文模式:** 控制是否以及如何输出图文内容。
- **图文分片最大数:** 确定在一个对话中最多输出多少图片。
- **图文相似度:** 设置图片匹配的严格程度。

6. 设置记忆策略:

- **最近聊天记忆轮数:** 设置对话历史的轮数或回合数。

7. 保存并发布应用:

- 点击 **保存** 按钮, 应用设置完成, 应用会保存成草稿状态。
- 点击 **发布** 按钮, 应用将会被发布到应用中心, 用户可以在应用中心中找到并使用您的应用。

请确保按照上述步骤仔细填写和配置应用的各个方面, 以确保应用能够准确和高效地响应用户的需求。

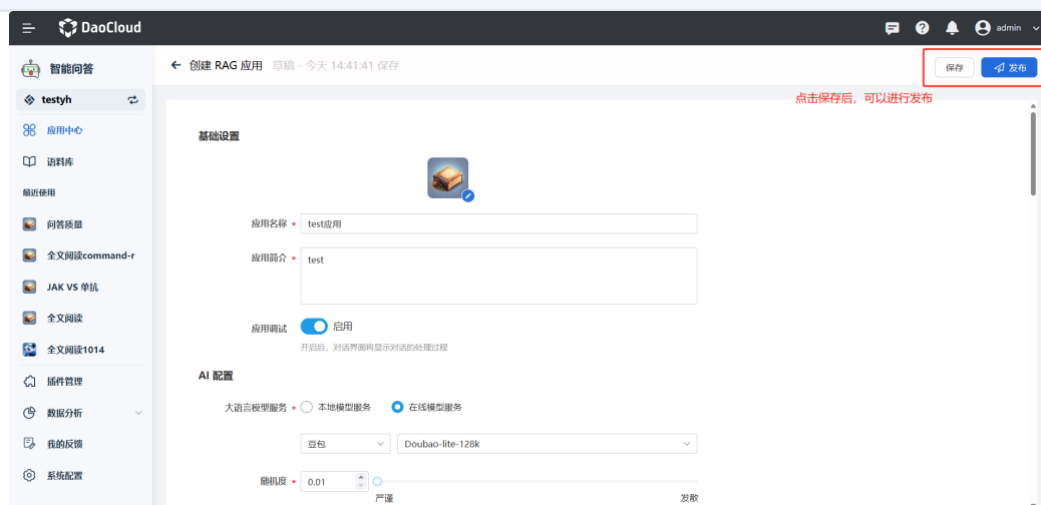


图 3 发布应用

2. 创建全文阅读应用

1. 在 **应用中心** 页面中，点击 **创建** 按钮，选择创建 **全文阅读** 应用。

2. 填写应用基础信息：

- **上传应用图标：**选择 jpg、jpeg 或 png 格式的图片，确保文件大小不超过 10MB。
- **输入应用名称：**限制在 20 个字符以内。
- **填写应用简介：**简介应用，不超过 100 个字符。

3. 完成 AI 配置：

- **选择大语言模型服务：**决定是使用本地模型服务还是在线模型服务（例如 Azure Open AI 或文心一言）。
- **设置随机度：**调整模型回答的创造性 / 发散性程度。
- **提示词：**为模型提供检索和回答问题的模板。

4. 保存并发布应用：

- 点击 **保存** 按钮，应用设置完成，应用会保存成草稿状态。
- 点击 **发布** 按钮，应用将会被发布到应用中心，用户可以在应用中心中找到并使用您的应用。

请确保按照上述步骤仔细填写和配置应用的各个方面，以确保应用能够准确和高效地响应用户的需求。



图 4 创建全文阅读应用

3. 创建组合应用

1. 在 **应用中心** 页面中，点击 **创建** 按钮，选择创建 **组合应用**。

2. 填写应用基础信息：

- **上传应用图标：**选择 jpg、jpeg 或 png 格式的图片，确保文件大小不超过 10MB。

- **输入应用名称：**限制在 20 个字符以内。
- **填写应用简介：**简介应用，不超过 100 个字符。
- **应用调试：**选择开启。开启后，对话界面将显示对话的处理过程。

3. 完成 AI 配置：

- **选择大语言模型服务：**决定是使用本地模型服务还是在线模型服务（例如 Azure Open AI 或文心一言）。
- **应用选择提示词：**为模型提供检索和回答问题的模板。

提示词中需要设置不同应用的 ID，应用 ID 从应用配置下的应用中复制获取。

4. 完成应用配置：

- **选择应用：**选择应用中心下可以正常对话的应用。
- **复制 ID：**用于填入提示词中，实现模型对话时选择不同应用。

5. 保存并发布应用：

- 点击 **保存** 按钮，应用设置完成，应用会保存成草稿状态。
- 点击 **发布** 按钮，应用将会被发布到应用中心，用户可以在应用中心中找到并使用您的应用。

请确保按照上述步骤仔细填写和配置应用的各个方面，以确保应用能够准确和高效地响应用户的需求。



图 5 创建组合应用

(二) 应用对话说明

对话 是自 ChatGPT 以来最流行、最直观的信息获取方式。在 d.run 平台上，一旦应用发布，用户即可立即使用对话功能。通过关联语料库，您可以在对话中自由提问，系统将基于丰富的语料提供精准的回答。同时，您可以随时查看历史对话记录，方便回顾和跟进交

流内容。导航栏还特别展示了最近使用的对话，帮助您快速切换和继续未完成的交流，极大提升使用效率和体验。

1. 日常对话

RAG 应用

1. 在左侧导航栏，点击 **应用中心**，选择已发布的 RAG 类型的应用，点击右下角的对话图标。

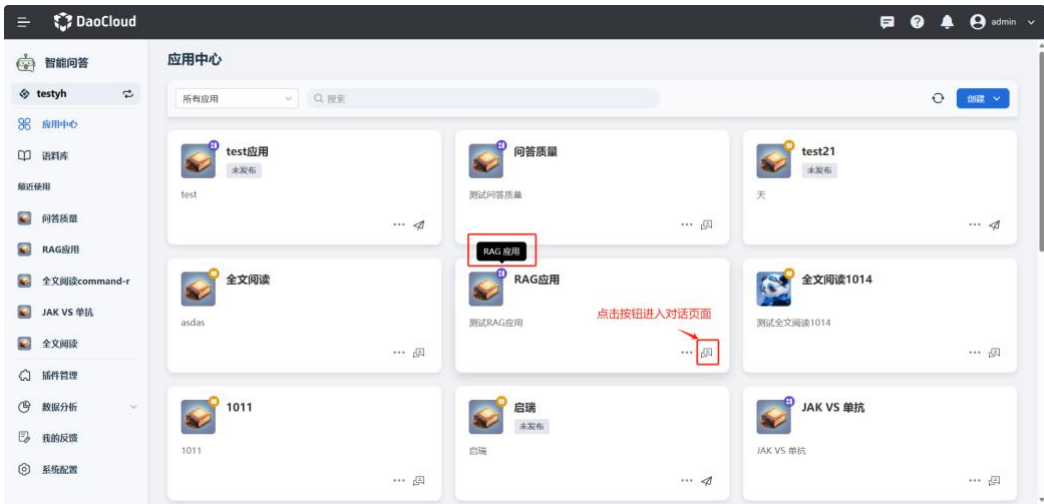


图 6 点击对话图标（RAG 应用）

2. 在对话框中输入问题，点击 **发送**，或敲击回车键，进行对话。

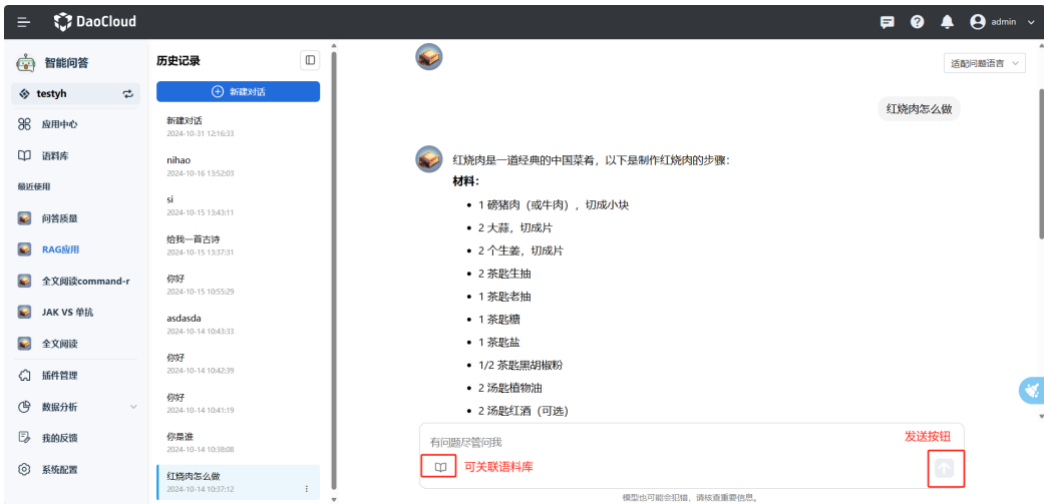


图 7 RAG 应用对话

全文阅读应用

1. 在左侧导航栏，点击 **应用中心**，选择已发布的 **全文阅读** 类型的应用，点击右下角的对话图标。

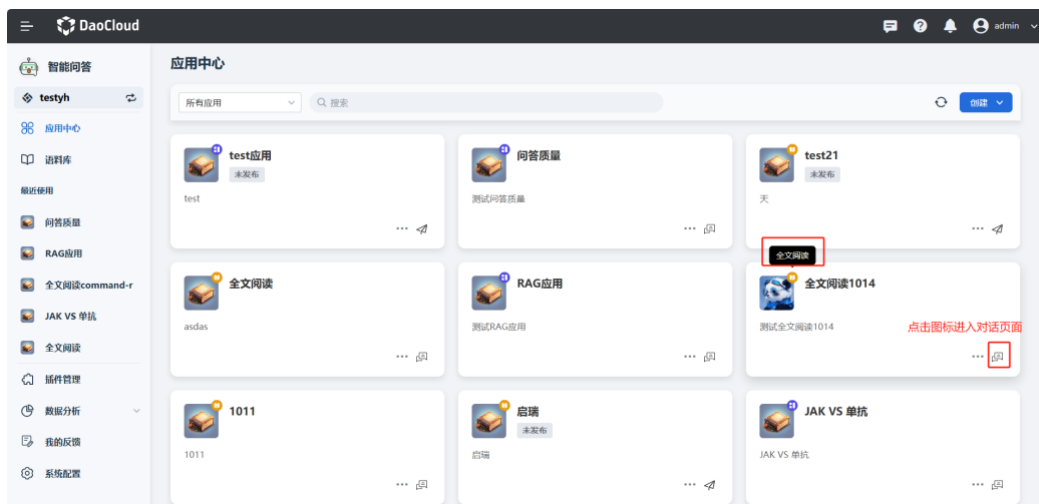


图 8 点击对话图标（全文阅读应用）

2. 在对话框右下角上传文件，根据文件内容进行提问对话。点击 **发送**，或敲击回车键，进行对话。

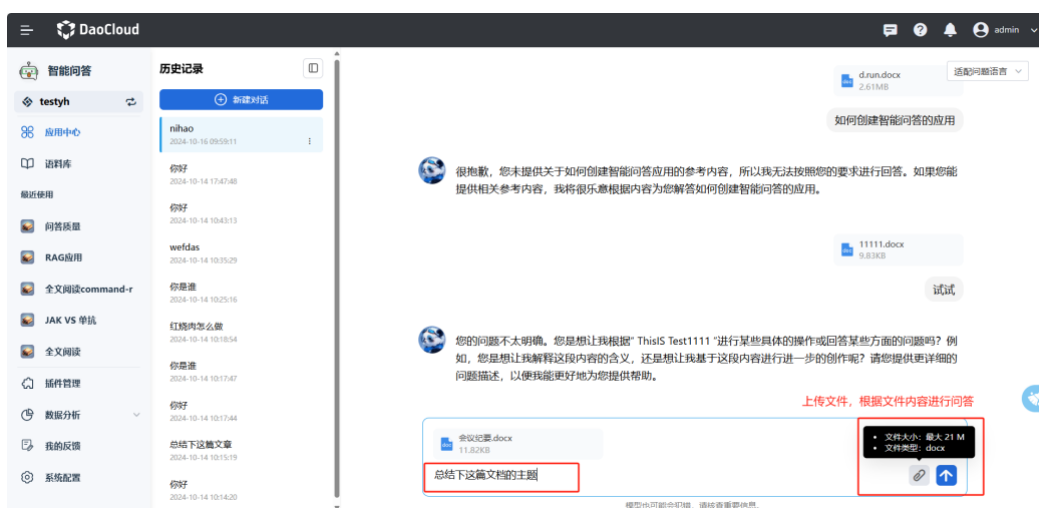


图 9 全文阅读应用对话

组合应用

1. 在左侧导航栏，点击 **应用中心**，选择已发布的 **组合应用** 类型的应用，点击右下角的对话图标。

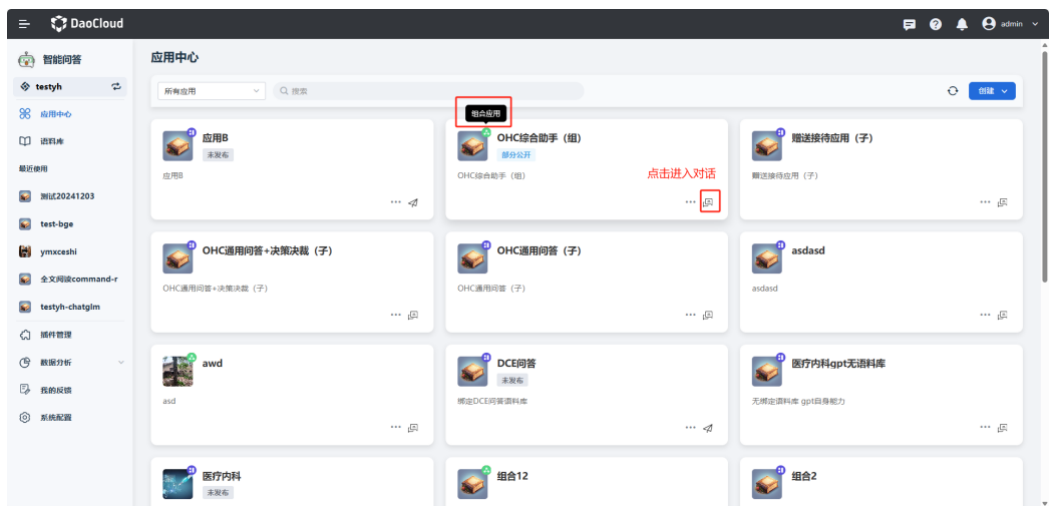


图 10 点击对话图标（组合应用）

2. 在对话框中输入问题，点击 **发送**，或敲击回车键，进行对话。

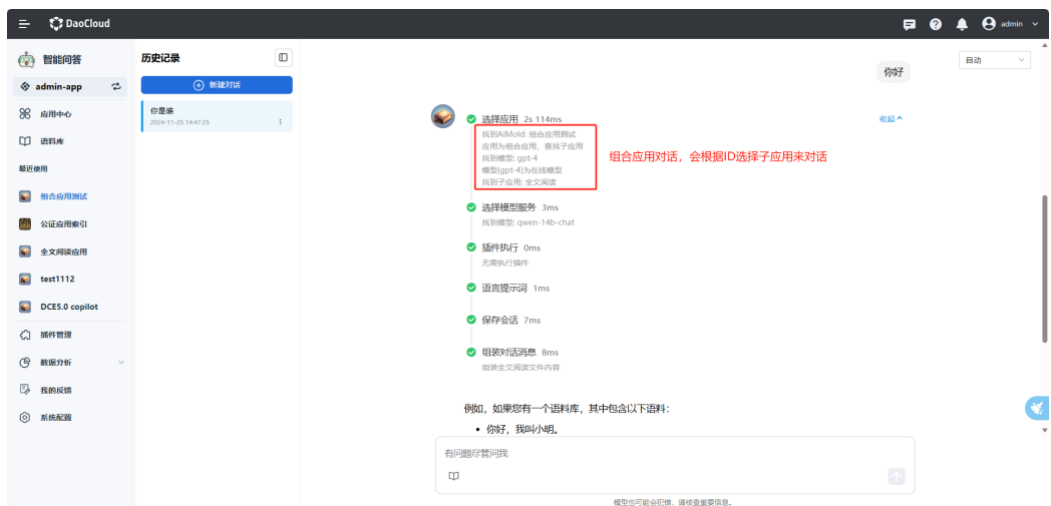


图 11 组合应用对话

2. 对话管理



图 12 对话管理

- **置顶、重命名和删除：**在历史记录窗格中，点击某一条对话右侧的 ，可以置顶、重命名和删除对话。
- **清空：**在对话页面左上角，点击  图标，可以 **清空** 对话。
- **关联语料库：**在输入框左下角，点击  图标，可以关联语料库。语料库的选择：**聊天应用** 会在选中的语料库中去匹配最相似的语料作为参考，回答您的问题。显示操作成功表示已经选中语料库，也可以取消选中某个或某些语料库。
- **停止对话：**提问后，点击输入框右侧的图标，可以在回答过程中停止对话，让助手停止输出内容。

如果无法选择和更改关联的语料库，只能添加语料库，说明应用的向量化模型和语料库的向量化模型不一致。应用的向量化模型和语料库的向量化模型一致才可以使用。

3. 几个实用图标

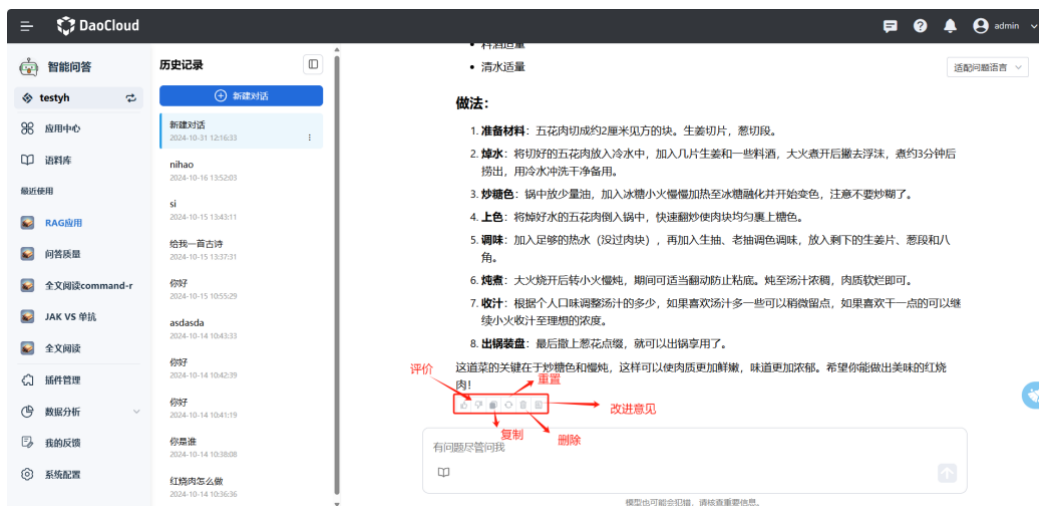


图 13 对话中常用的操作图标

- **评价：**可以点赞  或点踩  某条回答，取决于您对回答内容是否满意。
- **复制：**可以复制某条回答。
- **重置：**可以重置、重新生成某一条回答。
- **删除：**点击垃圾桶图标，可以删除某条回答。此后在聊天助手上下文对话时，将不会对删除的对话有记忆。
- **改进意见：**点击某条回答的最后一个图标，可以提交反馈，根据聊天助手的回答结果好坏，在反馈中提出反馈意见。

关于随机度：对于重置的内容，可以通过管理员调节模型的 **随机度** 来控制聊天助手多次回答的一致性。若随机度高，聊天助手多次回答的结果可能会有不一样的答案；如果要求回答的准确度高，可以降低随机度，这样聊天助手每次生成的结果会接近一致。

四、语料库

语料库是 AI 应用的知识来源。AI 应用通过向量化与相似度检索，从语料库中找出与问题最相关的内容，再交给大模型生成回答。因此，语料库的构建质量直接决定了问答的准确程度。

（一）创建语料库

语料库是指收集整理的大量文本数据集合，用于训练、测试和优化自然语言处理模型。通过构建高质量的语料库，可以帮助大模型更好地理解用户输入的问题，提高问答准确率和智能化水平。在 d.run 中，可以很方便地创建语料库。

1. 在 **我的语料** 页面中，点击 **创建语料** 按钮。
2. 参考下列要求填写语料库基本信息，并点击 **下一步**：
 - **语料库名称**：名称包含大小写字母、数字和符号（限制 20 个字符）。
 - **向量化模型服务**：可选择 **bge-large-zh** 和 **bge-large-en**。
 - **可见范围**：可选择公开 / 私有 / 指定工作空间。
 - **简介**：简要描述语料库中的内容信息，可包含中英文、数字（限制 100 个字符）。

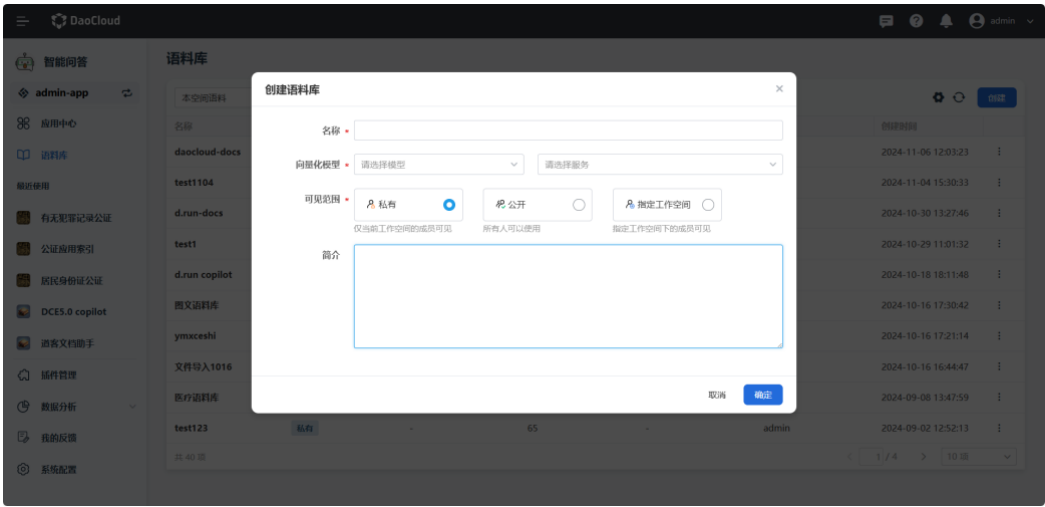


图 14 创建语料库

（二）管理语料库内容

语料库管理是确保数据规范、有序和高效利用的关键环节。通过有效管理语料库，可以方便地维护、更新和导出文本数据，从而保障模型训练和应用的质量与效果。

点击 **语料库** 创建时间旁的 **⋮** 按钮进入语料库整理页面。

- 查看语料库中文件名称、向量化状态、文件分片数以及创建时间。
- 点击操作中的 **导出** 按钮，可以将某一文件导出。

- 点击操作中的 **删除** 按钮，可以将文件在语料库中删除。

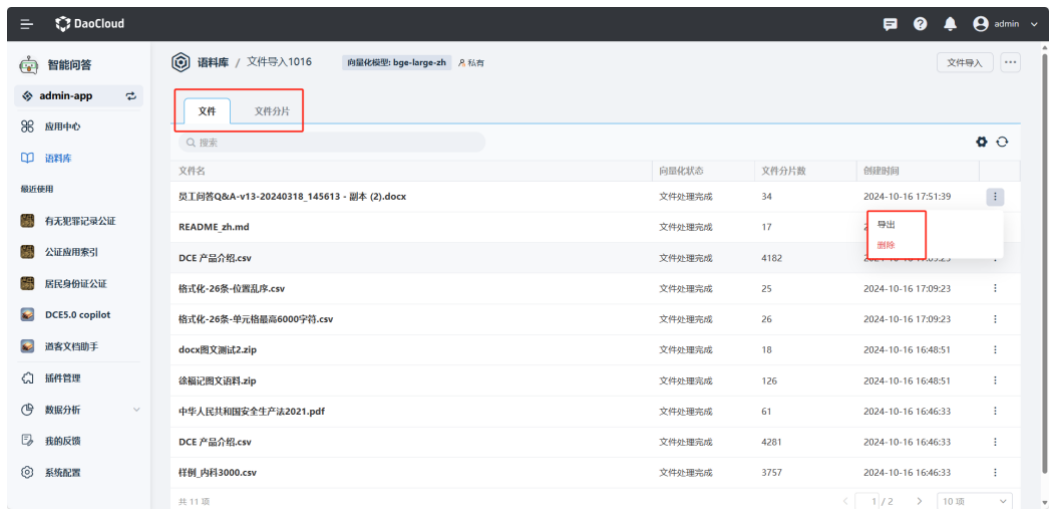


图 15 语料库文件列表

- 点击上方的 **文件分片** 进入文件分片详情，可以查看文件的所有分块信息。
- 输入文件分片描述来搜索具体的分片。其中文本相似度用于衡量搜索内容和语料库中分块的相似度。这里用【欧氏距离】作为相似度计算指标，故相似度越小，则距离越接近，文本越相似。
- 在文件分片中点击某一分片的 **!** 按钮，即可编辑语料分片内容或删除某一分片。

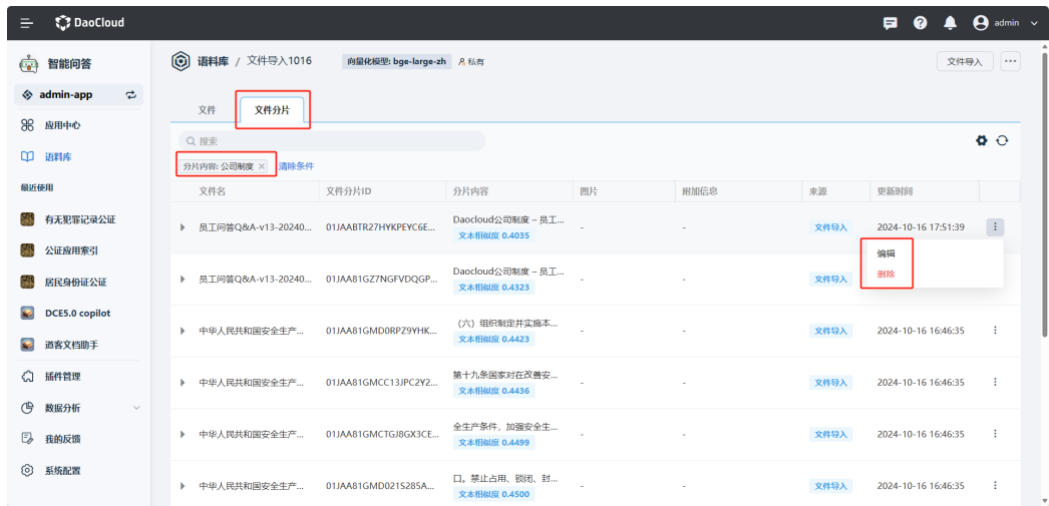


图 16 文件分片详情

(三) 语料导入

语料库支持多种导入方式，您可以根据数据来源和格式选择最合适的一种。

1. 文件导入

文件导入是将本地或其他来源的文本文件上传到语料库的过程。通过导入文件，可以丰富语料库的数据来源，提升模型的覆盖范围和理解能力，从而提高系统的整体表现。

1. 点击语料库旁的  按钮，选择 文件导入 方式。

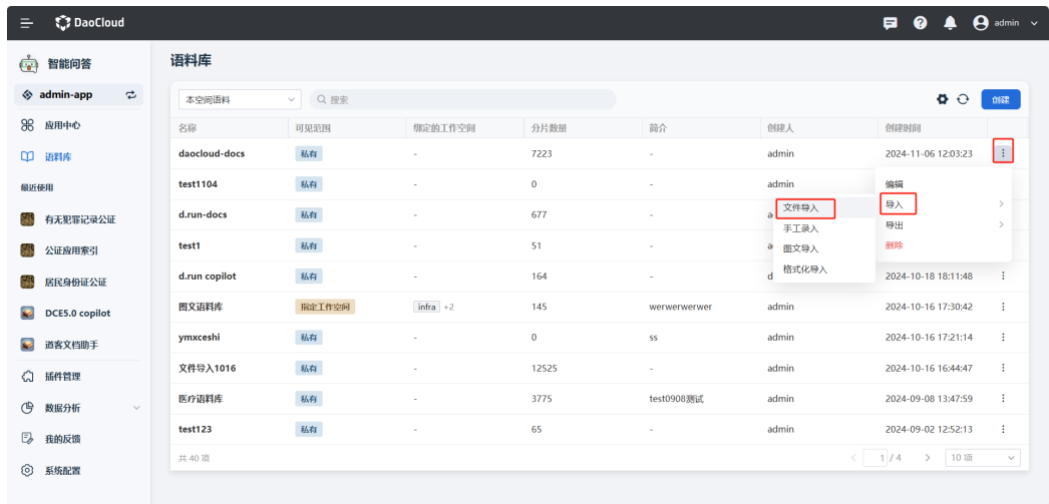


图 17 选择文件导入方式

2. 导入数据：点击 上传文件，并选择文件分片的处理方式：标准处理、自定义处理（即插件处理，请到插件接入处查看）。

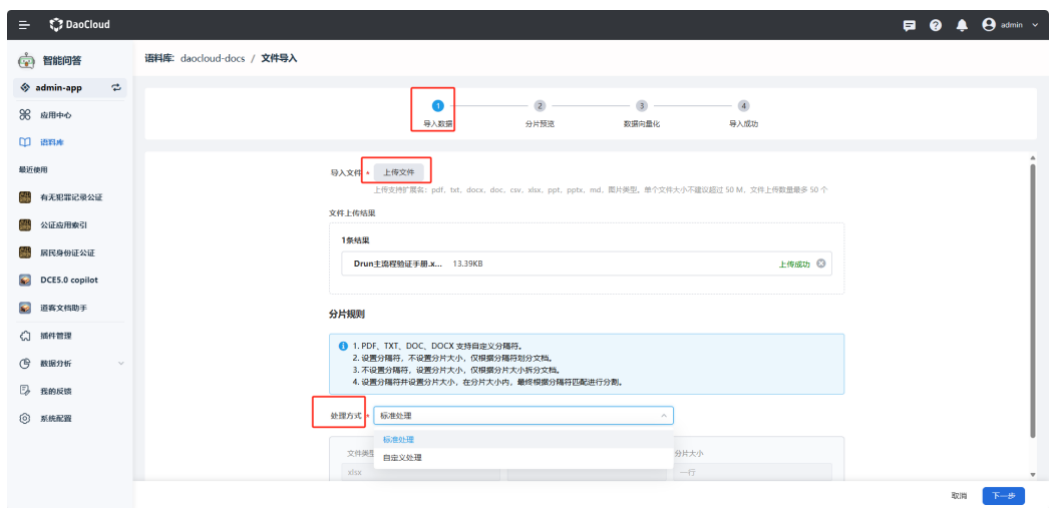


图 18 上传文件并选择分片处理方式

目前支持 pdf、txt、docx、doc、csv、xlsx 等格式，单个文件的大小不建议超过 50M，文件上传数量限制为 50 个。

分片规则之标准处理：

1. PDF、TXT、DOC、DOCX 支持自定义分隔符；
2. 设置分隔符，不设置分片大小，仅根据分隔符划分文档；
3. 不设置分隔符，设置分片大小，仅根据分片大小拆分文档；
4. 设置分隔符并设置分片大小，在分片大小内，最终根据分隔符匹配进行分割。

3. 分片预览：预览分片是否正确，如果不正确可以回到上一步修改分片规则或文件内容。

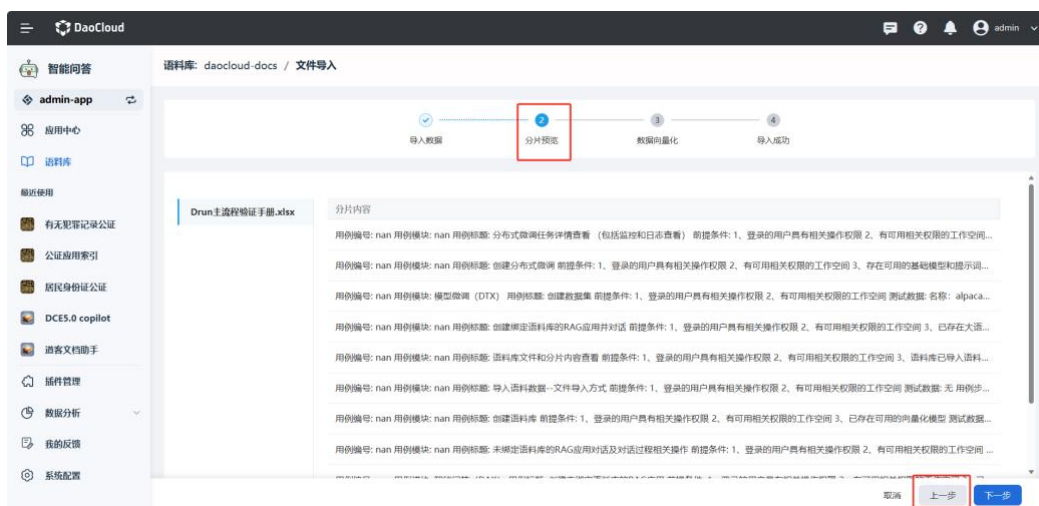


图 19 分片预览

4. 数据向量化：查看文件分片数量、重复分片数量、本次导入分片数以及向量化状态。当向量化处理成功后，点击 **下一步**。

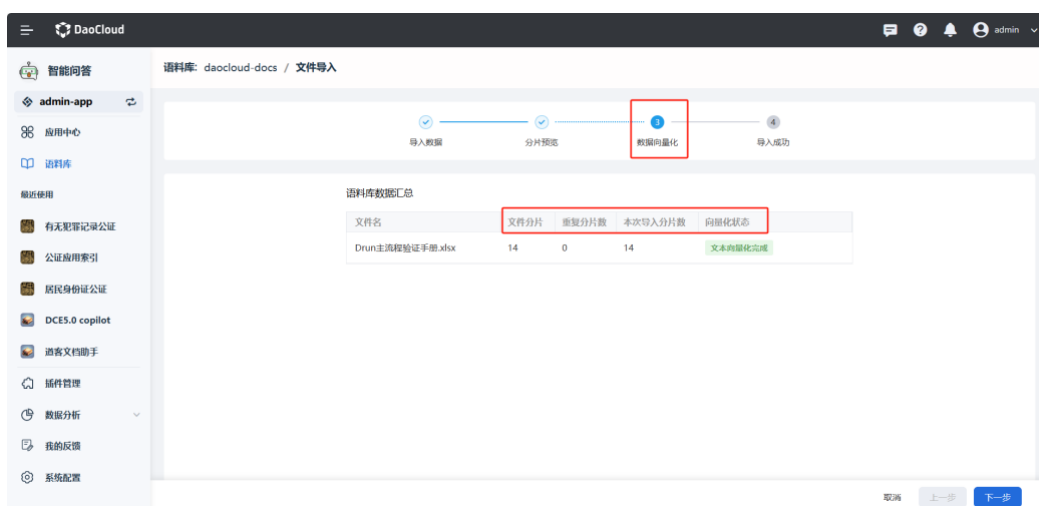


图 20 数据向量化

5. 待文件状态为文件处理完成后，点击 **确定。**

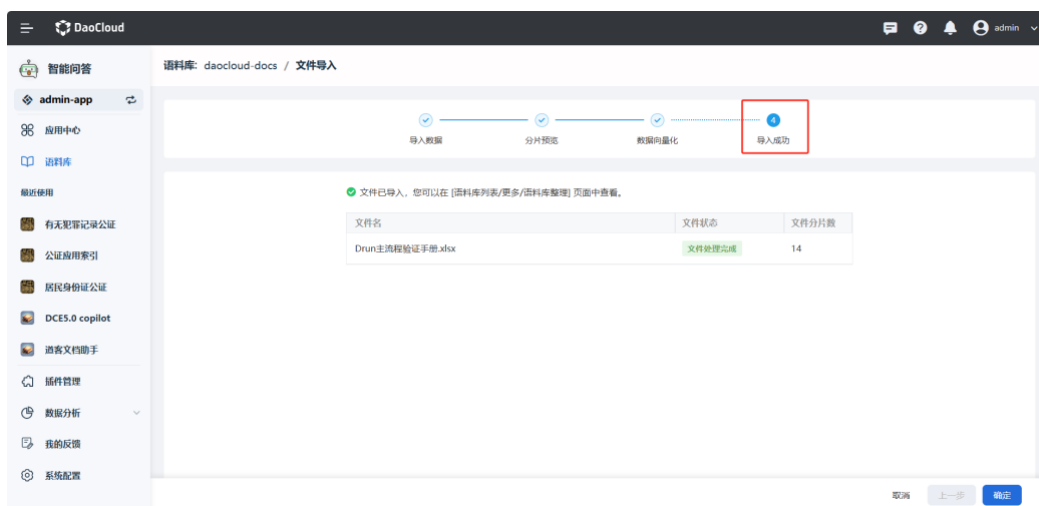


图 21 文件处理完成

2. 手工录入

1. 在语料导入时可以选择使用 **手工录入** 的方式导入分片。
2. 点击 **手工录入** 后，会弹出 **新增文件分片** 弹框。

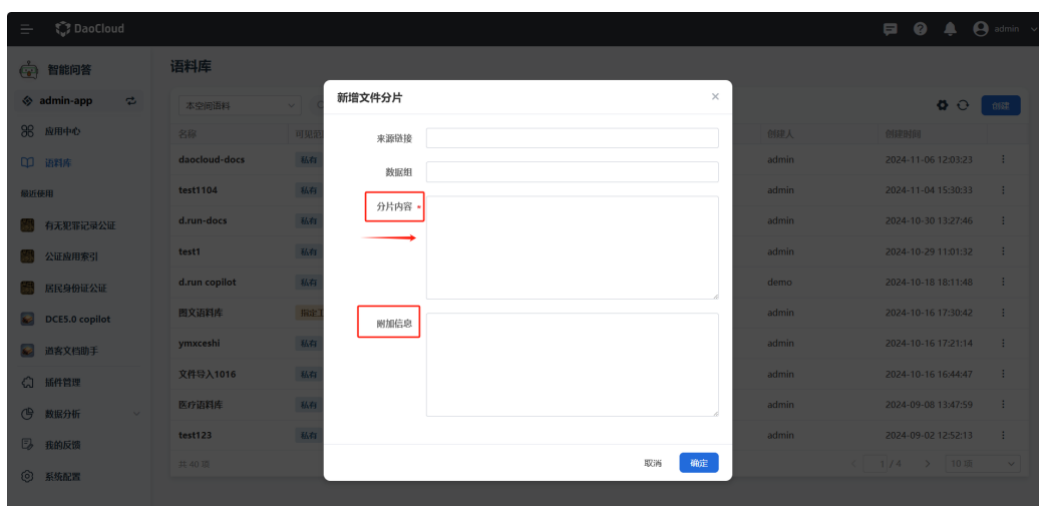


图 22 手工录入新增文件分片

在箭头位置录入信息，如对信息有备注，可在附加信息中录入。

3. 图文导入

在导入图文前，需要将导入的语料进行处理后再导入（目前仅支持 **Word** 和 **Excel** 的图文处理）。

预处理 Docx 文档

1. 直接支持带图文的 **Docx** 文档按照约定的字符长度分割。

2 或 3 个 AE 堆叠实现 SAE,并用 Softmax 对最终的 EEG 特征 进行分类.另一种 AE 的变[降噪自编码器(denoising autoencoder, DAE)^[25] 通过向输入层加入噪声干扰,训练网络能够对含噪数据进行编码、解码,从而提高隐层特征的表达 能力.Xu 等人利用堆栈 DAE (stacked DAE, SDAE)进行特征学习并利用 Softmax 分类此外,Li 等人网利用变分 AE (variational AE, VAE)无监督地学习 EEG 信号的低维表示.

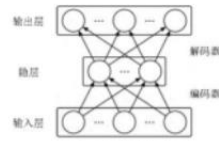


图5 自编码器框架

深度信念网络(deep belief networks, DBN)中)是一种概率生成模型.它由若干层受限玻尔兹曼机(restricted Boltzmann machines, RBM)组成.其深度结构能够有效学习高层特征表示.一个 RBM 只有两层神经元:可视层和隐层,分别用于输入 数据和特征检测,两层之间双向连接.层内神经元之间无连接.将多个 RBM 堆叠构成 DBN,其中底层 RBM 的输出作为 高层 RBM 的输入,DBN 就可用于无监督学习,也可用于有监督学习.如图 6 所示.通常采用含有两个隐层的 DBN 进行 EEG 情绪识别^[21].其训练主要包括 3 个步骤:首先,对每层进行无监督的预训练.其次,将 RBM 展开为编码器和解码器, 利用反向传播算法对各层进行无监督的微调,学习使得输入和重构向量尽可能相似的网络权重和偏置;最后,在顶层 RBM 的可视层加入表示类别的神经元,反向传播更新参数,实现有监督的微调.然而 DBN 的隐层缺少约束,使得其很难挖掘 EEG 电极间和频段间的相关信息.Chao 等人^[22]引入胶质细胞链(glia chain),利用 DBN-GCs 网络学习隐层神经元的关联.胶 质细胞能够调节隐层神经元的激活概率,并向相邻胶质细胞传递信号.此外,文章利用集成学习,将特征子集分别送入 5 个 DBN-GCs,拼接各模块的隐层输出,作为判别 RBM 的输入.

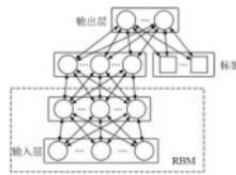


图6 深度信念网络框架

3.2 卷积神经网络与循环神经网络

CNN^[23] 是一种前馈神经网络,利用反向传播算法进行训练.其中一个典型层包含 3 级:卷积、非线性变换、池化. 卷积运算具有局部连接、权值共享,等变表示的特性.通常在同一卷积层内使用多个不同的卷积核以学习不同特征.非线性 变换,也称为激活函数,能够给模型加入非线性因素,提高表达能力.目前大部分 CNN 采用线性整流(rectified linear unit, ReLU)函数^[24]作为激活函数.池化是一种过拟合的方法,使用某一位置周围的总体特征代替网络在该位置上的输出.常用的 池化函数包括最大池化、平均池化等.现有研究设计或应用不同的 CNN 结构进行 EEG 情绪识别.Wang 等人^[25]将 LeNet 和 ResNet^[26]这两种典型网络架构分别用于情绪识别.Cimtay 等人^[27]在预训练的 InceptionResNetV2 模型的基础上增加 5

图 23 按字符长度分割

2. 也支持手工用 `<split>` 标签，提前规划好文档分割段落。

|
开始标签

`<split>`
→
中间的图文内容，为一个分片

2 语音情感的认知神经科学研究进展
→

2.1 情感的神经机制
→

情感产生的脑机理研究经历了一个较长的过程。受到神经解剖学、神经生理与认知心理学等相关科学发展的影响。思想家和科学家对情绪奥秘的探讨可以追溯到古代的臆测和神秘主义。直到文艺复兴以后，如霍布斯(Hobbes)、洛克(Locke)、笛卡尔(Descartes)等带有唯物主义者色彩的哲学家才把知觉、思维、知识、情绪等和神经与脑的活动联系了起来。1872年达尔文(Darwin)在《人和动物的表情》一书里论述了情绪的生物学基础，强调了环境对情绪行为的作用，形成了情绪生理心理学的雏形。其后的詹姆斯(James)提出了最早的情绪生理-心理理论，为探讨情绪的性质指出了一条必由之路。James-Lang 理论(1885年)即情绪外周理论，强调情绪的产生是植物神经系统活动的产物。1912年，Mills首次提出了情感的大脑半球假说，右脑更多地决定了人的空间感、抽象思维、音乐感与艺术性。1931年，Cannon提出了情绪的丘脑学说，认为丘脑对情绪调节起着重要作用。随后，Papez提出了 Papez 环路理论，认为下丘脑是情绪表达中心，边缘系统是情绪体验部位。但当时这一回路并没得到科学证实。Maclean 于 1952 年提出了情绪脑的概念，划分了较为精细的情绪相关脑区网络，得到研究者的广泛认同(如图 6 所示)。

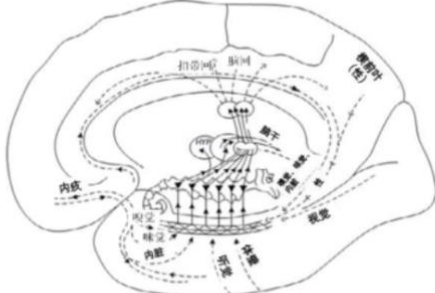


Fig.6 MacLean's limbic system theory 图 6 MacLean 提出的边缘系统理论 [28]

`</split>`
→
结束标签

图 24 手动 split 分割

对于 Docx 文档中的图片信息，整理的时候请直接粘贴到文档（不要使用形状或者文本框包裹图片），以免程序无法检测从而遗漏图片的处理。

预处理 xlsx 文档

xlsx 文件需要符合固定的模板格式，Q 为问题，A 为答案。

序号 Seq	信息1 Info1	信息2 Info2	信息3 Info3	问题 Question	答案 Answer	关于 About
1	纳米技术			纳米技术如何影响材料科学？	【摘要】 纳米技术在材料科学领域产生了深远的影响。通过在纳米尺度操纵材料，科学家们能够创造出具有独特性能的新材料。例如，纳米材料在强度、重量、导电性和导热性方面展现出与传统材料无法比拟的特性。纳米颗粒被用于制造更轻、更坚固的材料，广泛应用于航空航天、建筑和汽车工业。此外，纳米技术在医疗领域也显示出巨大潜力，如用于药物输送和疾病诊断的纳米颗粒。尽管纳米技术的发展带来了巨大的经济和社会效益，但它也引发了环境和健康方面的担忧。需要进一步的研究和监督。	
2	生物技术			生物技术在医疗领域的最新应用有哪些？	【摘要】 生物技术在医疗领域的应用正在快速发展，带来了许多创新的治疗方法。基因编辑技术，如 CRISPR-Cas9，已被用于治疗遗传性疾病，并在癌症治疗中显示出潜力。再生医学也在进步，干细胞疗法和组织工程技术正致力于修复或替换受损的组织和器官。生物技术在个性化医疗中扮演着关键角色，通过分析患者的基因组信息，医生可以制定更精准的治疗方案。此外，合成生物学的进步为设计新药物和疫苗提供了新工具，特别是在应对新出现的传染病时。	

图 25 xlsx 模板格式

对于 xlsx 文档，请按照模板要求整理，插图请尽量放在一个单元格中，尽量不要横跨几个单元格放置。

生成图文语料

1. 登录环境：<https://console.d.run/ai-tools/lab>，密码为 `aitools`。

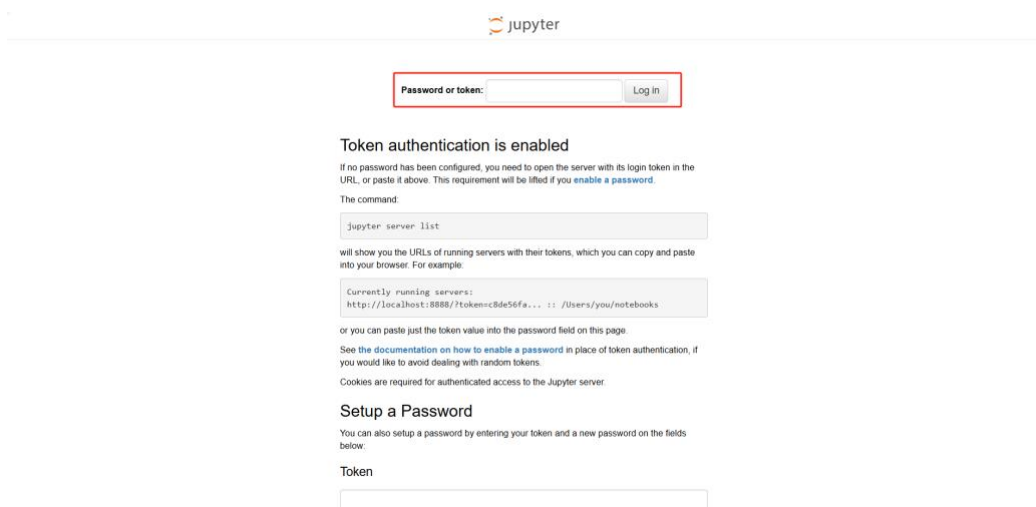


图 26 登录图文语料处理环境

2. 上传语料文件：进入目录 `/app/corpus_processing/input`，将语料文件上传到此目录下。

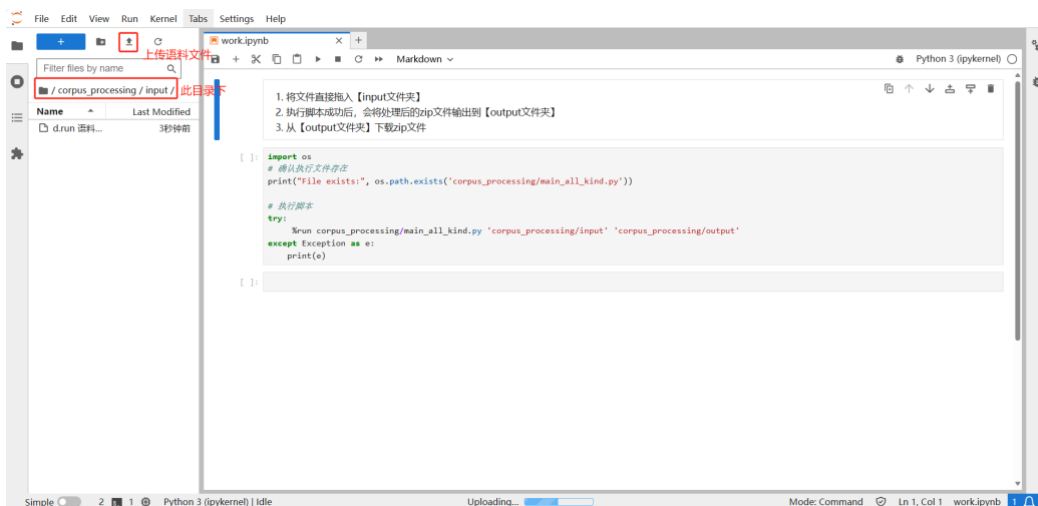


图 27 上传语料文件

3. 点击运行代码。

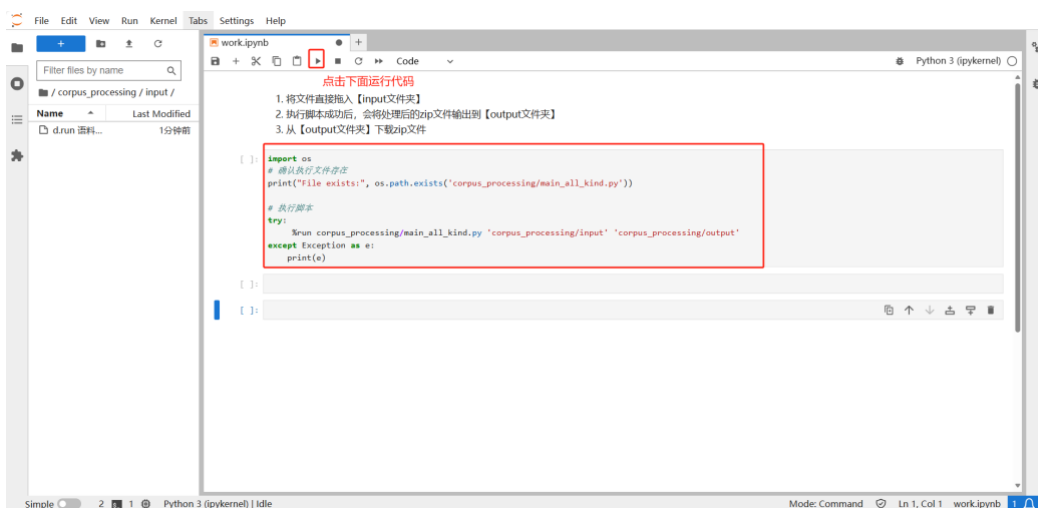


图 28 运行代码

4. 下载生成的图文语料文件：进入目录 /app/corpus_processing/output，下载 zip 文件。

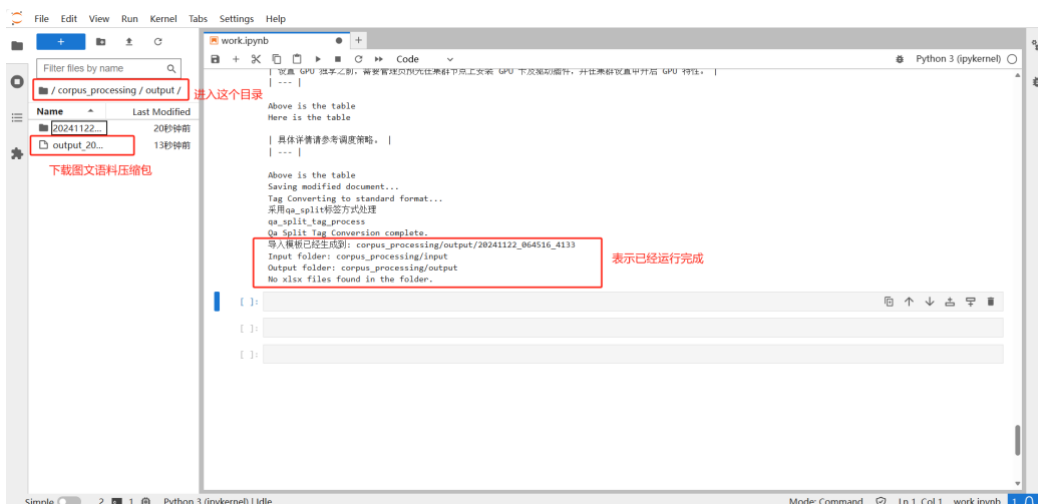


图 29 下载生成结果

5. 清理环境：清理输入和输出文件，以及清理运行日志文件。

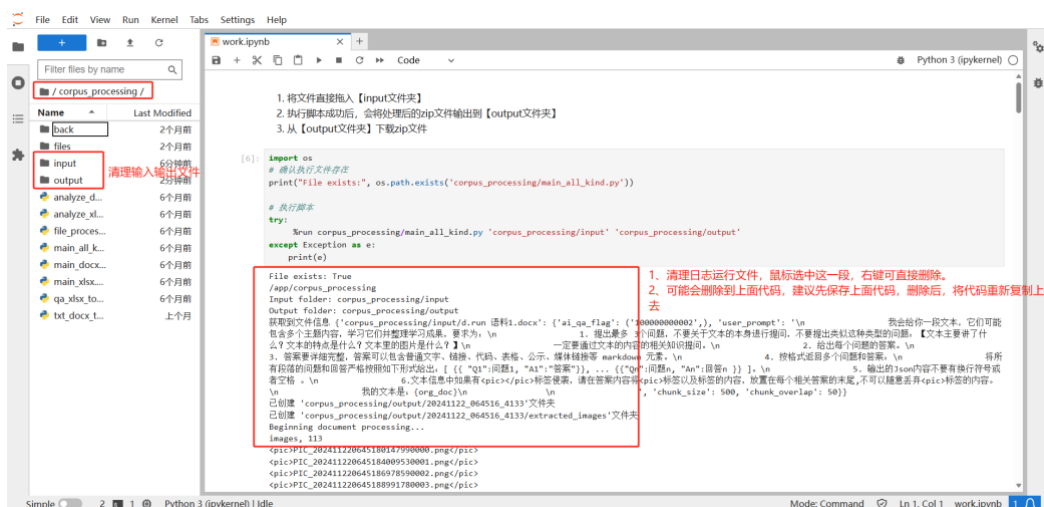


图 30 清理环境

该环境为公共环境，私有语料文件操作后，建议执行清理环境操作。

导入下载的文件

1. 点击 语料导入 → 图文导入。

2. 将处理好的文件上传，并进行向量化，等待处理成功。

4. 格式化导入

1. 在导入语料时选择 格式化导入。

2. 上传格式化文件，将数据向量化，这与上传数据的过程相同。

格式化导入 目前只支持 csv、xlsx 文件，并且要求文件内容格式如下。如果是其他类型的文件，请正常使用上传数据的方式导入。

A	B	C	D
1	seq	orgDoc	extDoc
一、调整范围部分 1、2023年8月31日前我行已发放的和已签订合同但未发放的首套住房商业性个人住房贷款。 2、2023年8月31日前贷款发放或签订合同时套数性质为二套及以上，但借款人实际住房情况已符合所在城市首套住房标准的其他存量住房商业性个人住房贷款。具体情况包含： 1 根据《中国人民银行 国家金融监督管理总局关于降低存量首套住房贷款利率有关事项的通知》（银发〔2023〕125号）要求，对于2023年8月31日前发放或签订合同时套数性质为二套及以上，但借款人实际住房情况已符合所在城市首套住房标准的其他存量住房商业性个人住房贷款，具体操作如下： (1) 房屋购买时家庭在当地没有存量住房，但因当地政策采取“认房又认贷”政策导致按照二套房贷款利率标准办理，现在当地执行“认房不认贷”政策的； (2) 房屋购买时不是家庭在当地唯一存量住房，但后期通过交易等方式出售了其他存量住房，本住房成为家庭在当地唯一存量住房且当地执行“认房不认贷”政策的； (3) 其他满足所在城市首套住房利率标准的存量住房商业性个人住房贷款。			

图 31 格式化导入的文件格式

五、数据分析

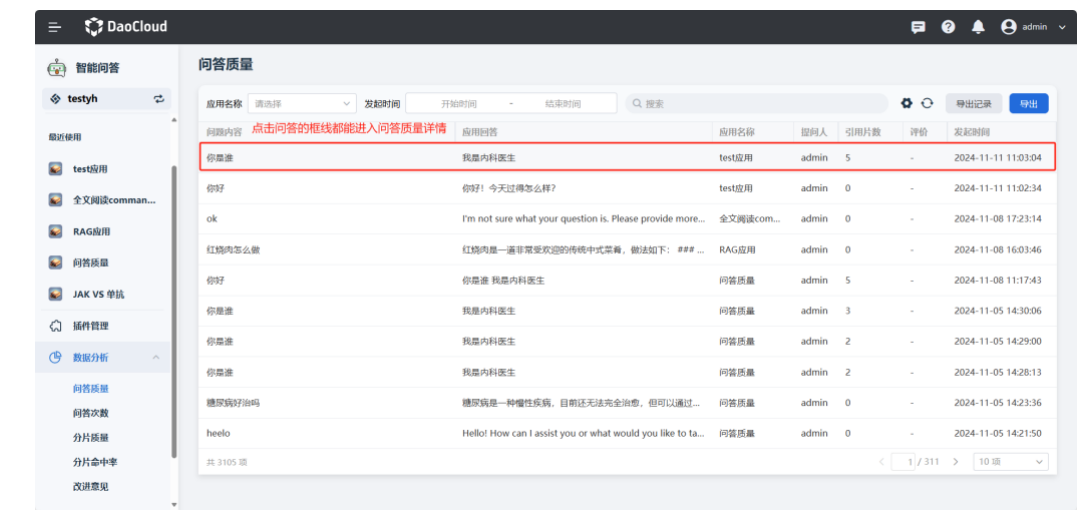
「数据分析」用于观察 AI 应用的运行质量和语料使用情况，包括问答质量、问答次数、分片质量、分片命中率，以及用户提交的改进意见。这些数据是持续优化语料和问答效果的依据。

（一）问答质量

d.run 提供了全面的问答质量反馈统计功能，帮助您深入了解用户对问答结果的满意度和反馈情况。系统不仅自动收集和汇总用户的质量评分和意见，还通过直观的报表形式展示数据趋势和关键指标，便于快速分析和发现潜在问题。此外，您还可以将这些反馈数据导出为 Excel（.xlsx）格式，方便进行离线分析、共享与归档，助力持续优化模型性能和提升用户体验。

查看问答质量详情

1. 在 **数据分析** 栏中点击 **问答质量**，找到需要查看的问答，点击该问答记录可进入详情页面。



应用名称	请选择	发起时间	开始时间	结束时间	搜索	导出记录	导出
问题内容	点击问答的横线都能进入问答质量详情	应用回答	应用名称	提问者	引用片数	评价	发起时间
你是谁	我是内科医生	test应用	admin	5	-	-	2024-11-11 11:03:04
你好	你好！今天过得怎么样？	test应用	admin	0	-	-	2024-11-11 11:02:34
ok	I'm not sure what your question is. Please provide more...	全文阅读com...	admin	0	-	-	2024-11-08 17:23:14
红烧肉怎么做	红烧肉是一道非常受欢迎的传统中式菜肴，做法如下：### ...	RAG应用	admin	0	-	-	2024-11-08 16:03:46
你好	你是谁 我是内科医生	问答质量	admin	5	-	-	2024-11-08 11:17:43
你是谁	我是内科医生	问答质量	admin	3	-	-	2024-11-05 14:30:06
你是谁	我是内科医生	问答质量	admin	2	-	-	2024-11-05 14:29:00
你是谁	我是内科医生	问答质量	admin	2	-	-	2024-11-05 14:28:13
糖尿病好治吗	糖尿病是一种慢性疾病，目前还无法完全治愈，但可以通过...	问答质量	admin	0	-	-	2024-11-05 14:23:36
hello	Hello! How can I assist you or what would you like to ta...	问答质量	admin	0	-	-	2024-11-05 14:21:50

图 32 问答质量列表

2. 在该页面可以查看以下内容：

- 用户提问内容，包括提问用户名称。
- 应用回答：应用针对用户问题产生的答案。
- 该问题使用的提示词。

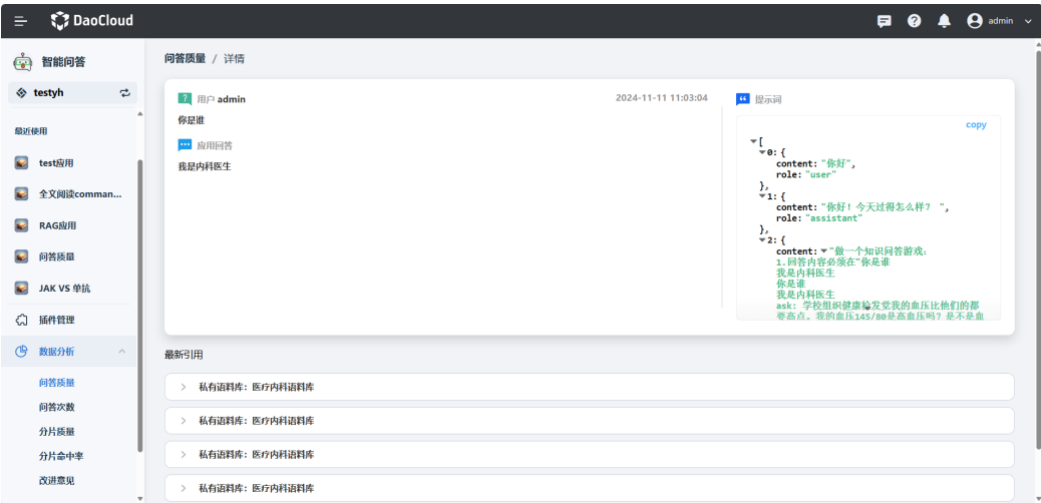


图 33 问答质量详情

3. 当有新的问答生成，可点击右上角 **刷新** 按钮查看最新问答详情。

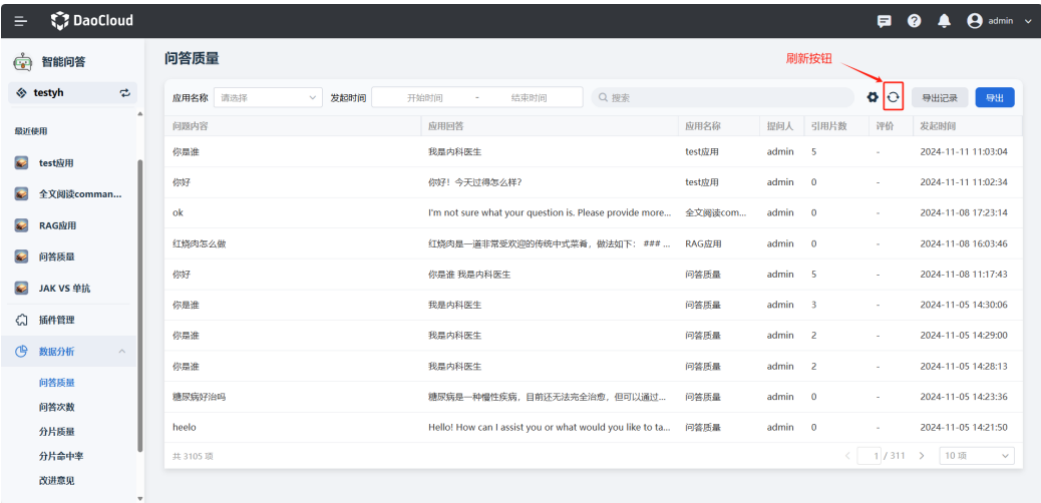


图 34 刷新问答质量

导出问答质量

您可以将该工作空间内所有用户对问答质量的反馈汇总成表格导出。

1. 在 **问答质量** 页面点击右上角 **导出** 按钮。

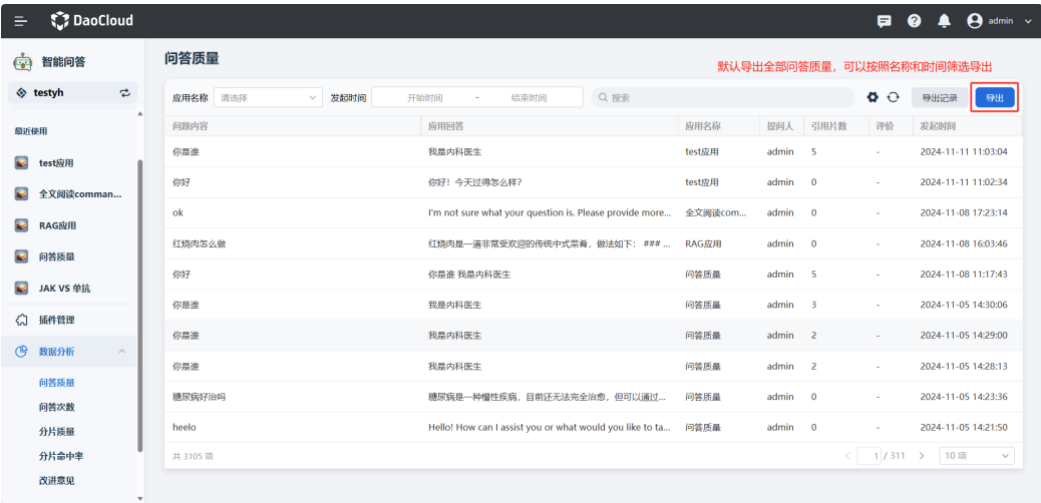


图 35 导出问答质量

2. 点击 导出记录，查看导出的问答质量记录，选择需要下载的记录，导出为 Excel 文件。

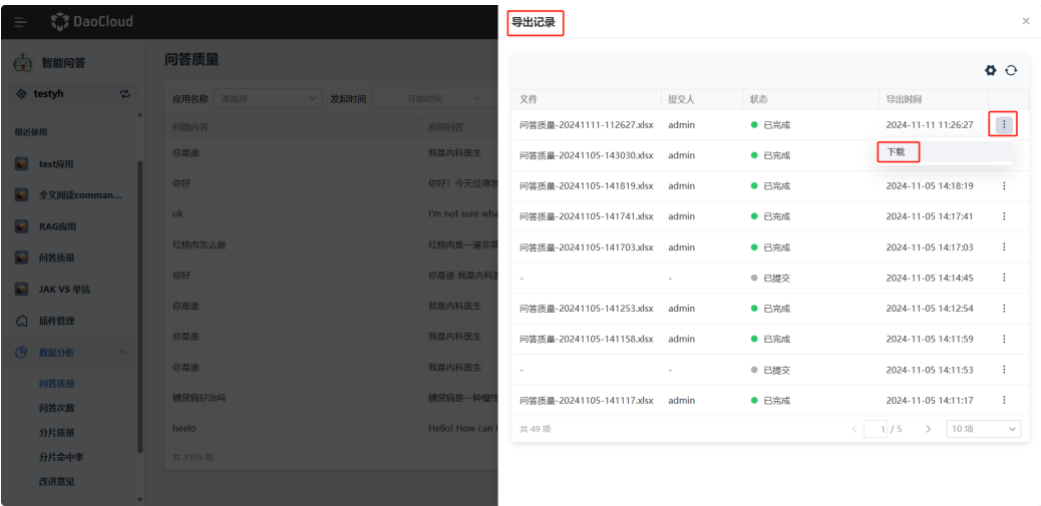


图 36 导出记录

（二）问答次数

d.run 支持记录在该工作空间创建的应用问答的次数。当有用户进行了问答，记录会保存在创建该应用的工作空间下，点击右上角 刷新 按钮查看最新问答次数数据。



图 37 刷新问答次数

您可以将该工作空间内所有用户的问答次数汇总成表格导出。

1. 在 **问答次数** 页点击右上角 **导出** 按钮。

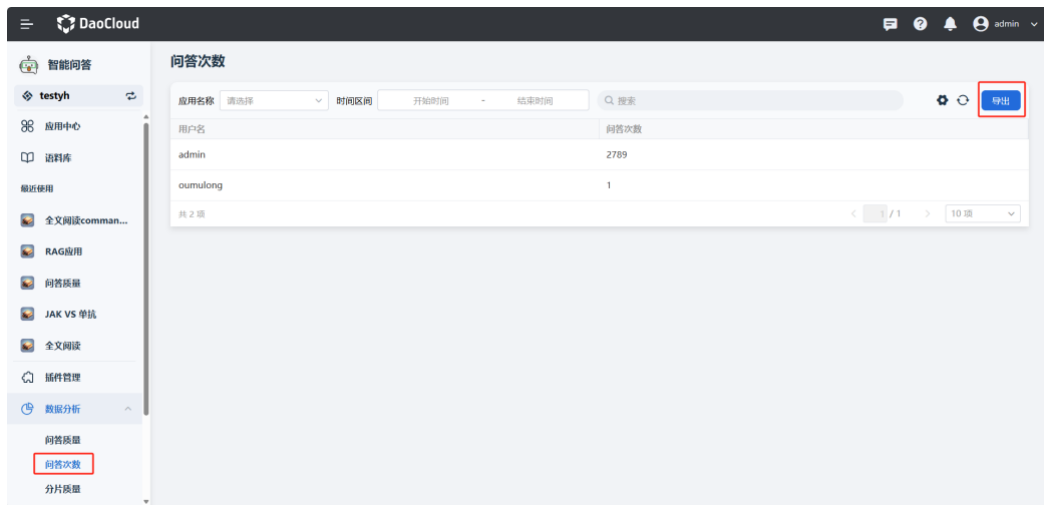


图 38 导出问答次数

2. 将该工作空间下所有用户的问答次数汇总成 **xlsx** 文件并下载。

(三) 分片质量

分片是指将较大的数据切分为多个片段，便于训练模型查找和命中。**d.run** 支持查看分片的质量。具体步骤如下：

1. 在 **数据分析** 栏中点击 **分片质量**，通过 **搜索** 找到您关注的分片，点击该分片可进入详情页面，查看该分片的详细内容。

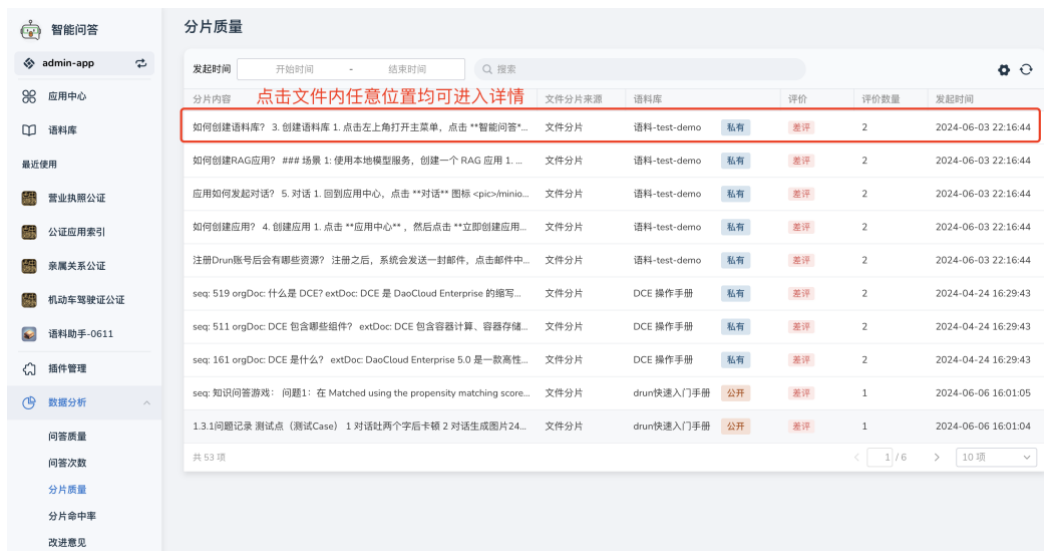


图 39 分片质量列表

2. 在该页面可以查看以下内容：

- **语料库**：分片属于哪个语料库。
- **更新时间**：该分片文件的更新时间。
- **分片 ID**：分片的唯一识别码。
- **分片内容**：切片后该分片的具体内容。
- **附加信息**：与该分片有关的分片附加内容。

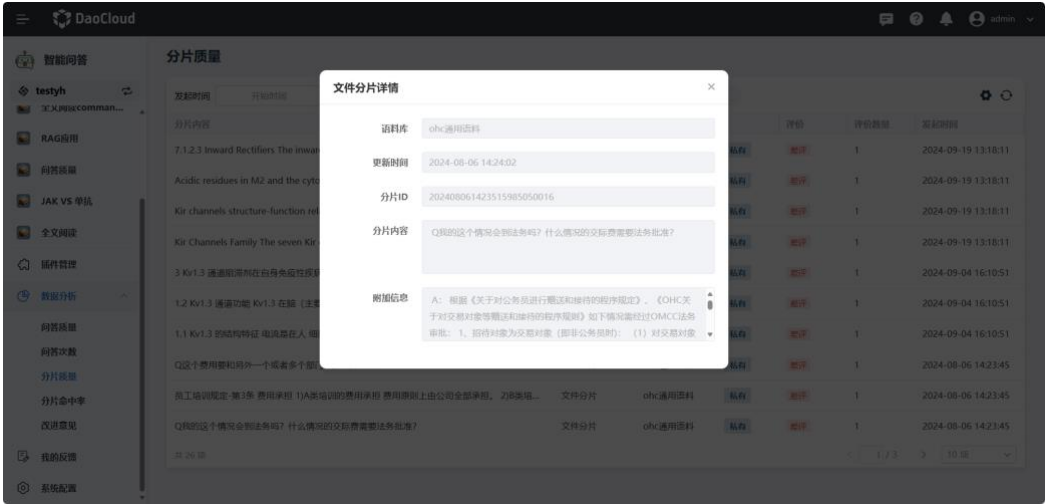


图 40 分片质量详情

3. 当有新的分片文件被评价，可点击右上角 **刷新** 按钮查看最新分片文件。

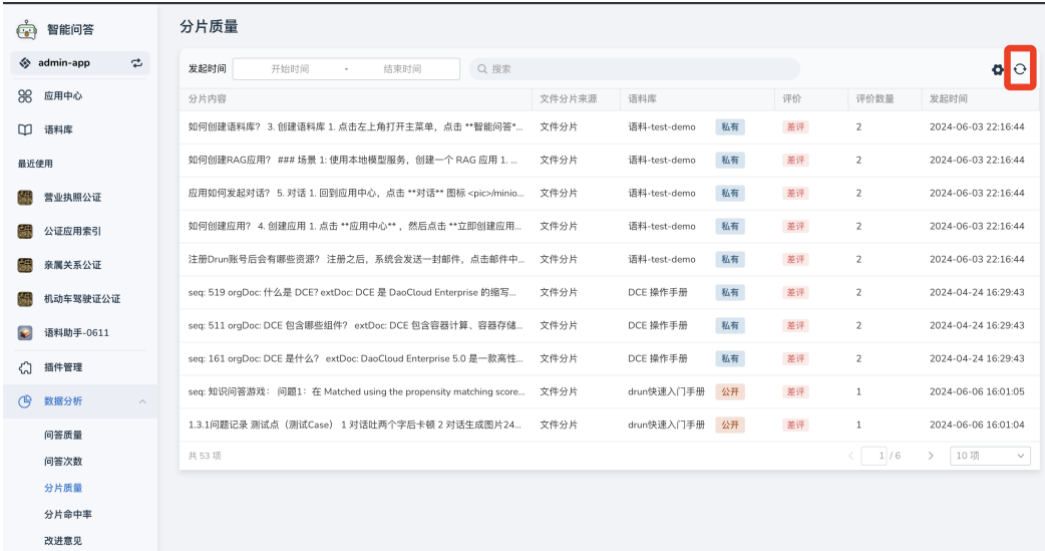


图 41 刷新分片质量

（四）分片命中率

d.run 统计了用户对于分片数据的使用情况，将分片按命中次数排序，支持查看分片内容并将分片的使用情况导出到 **xlsx** 中。

查看分片命中率详情

1. 在 **数据分析** 栏中点击 **分片命中率**，通过 **搜索** 找到您关注的分片，点击该分片可进入详情页面，查看该分片被引用的详细情况。



图 42 分片命中率列表

2. 在该页面可以查看以下内容：

- **分片的基本信息**：包括文本分片 ID 以及具体的分片内容。
- **详细信息**：该分片被问题命中的问题内容、应用名称、提问人以及提问时间。

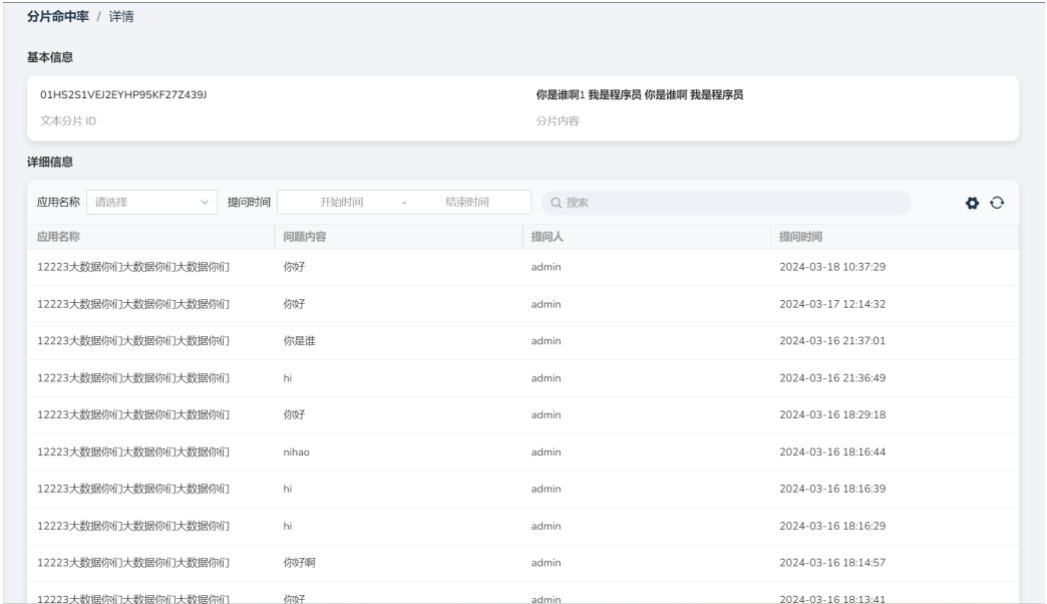


图 43 分片命中率详情

3. 当有新的问答生成，可点击右上角 **刷新** 按钮查看最新分片命中率结果。



图 44 刷新分片命中率

分片命中率详情中也可 **刷新** 查看最新结果。

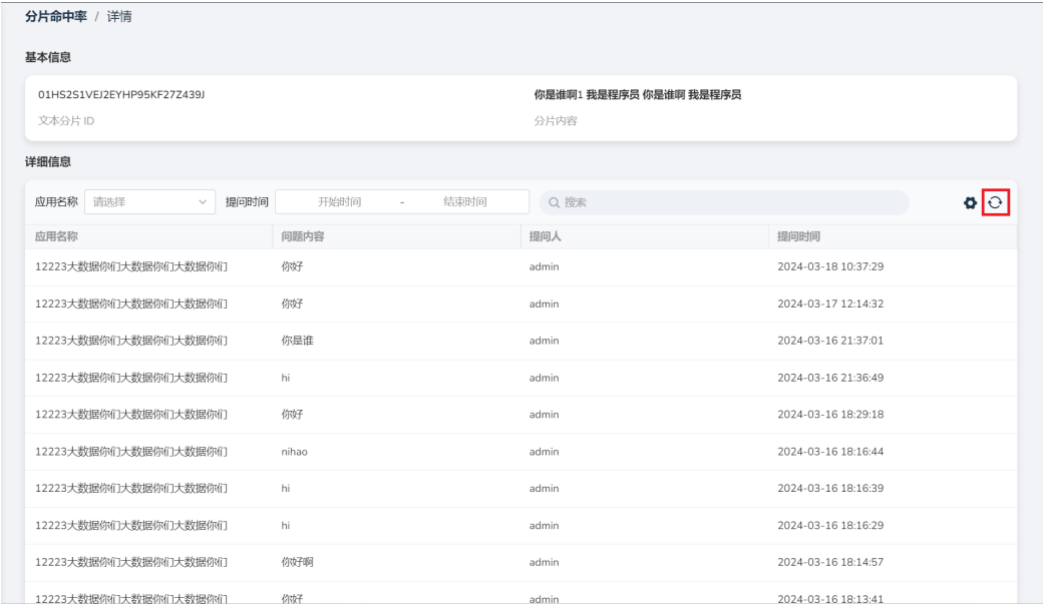


图 45 刷新详情结果

导出分片命中率

1. 在 **分片命中率** 页点击右上角 **导出** 按钮。



图 46 导出分片命中率

2. 导出在该工作空间创建的分片，将这些分片使用情况的反馈汇总成 **xlsx** 文件并下载。

（五）改进意见

d.run 提供了用户提交反馈的入口，管理员可在此处看到在这个工作空间的应用在对话上有哪些问题需要改进。

处理反馈

1. 在 **数据分析** 栏中点击 **改进意见**，找到需要处理的反馈，点击右侧的 **⋮** 按钮。
2. 在弹出菜单中选择 **处理反馈**（在范围内点击工单信息可进入详情界面）。

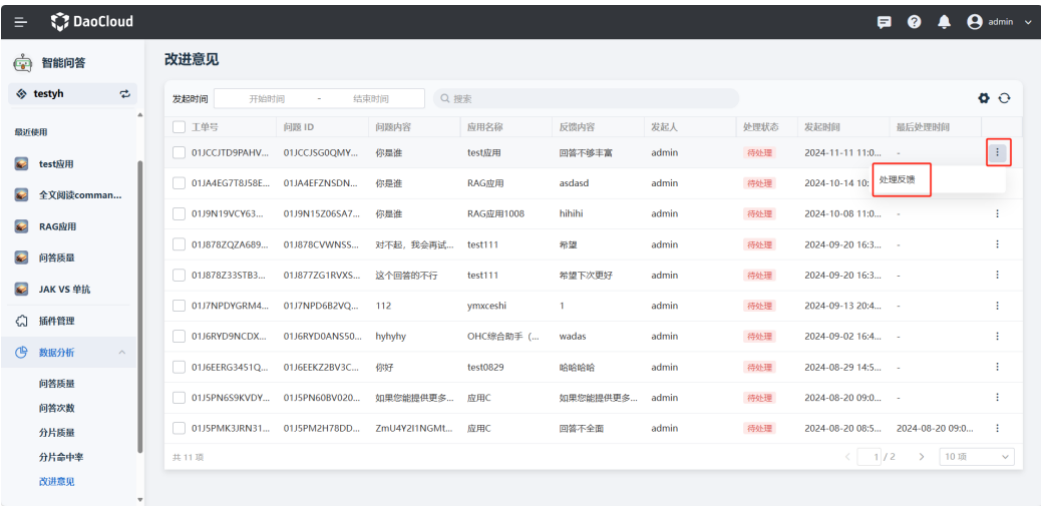


图 47 处理反馈入口

3. 在该页面可以查看以下内容：

- 用户对问答的反馈内容。
- 问答的相关信息：助手名称、模型名称、引用条数、处理时间以及问答详情。

- 最新引用：可查看回答所有引用的分片详情。
- 工单处理的详情。

工单号: 01HS2XGPQ2WZDE9FGZ6E822E9D 待处理

用户反馈

用户反馈内容

你好

相关信息

助手名称 12223大数据你们大数据你们大数据你们

模型名称 -

引用条数 5

处理时间 -

用户 admin

hi

2024-03-16 13:59:51

提示词

[[{"content":"hi","role":"user"}]]

应用回答内容

Hello! How can I help you today?

图 48 问答相关信息

工单号: 01HS2XGPQ2WZDE9FGZ6E822E9D 待处理

最新引用

私有语料库:

来源 手工录入

分片ID 01HS2HSIDRR14CNFN0NVGBS1P9

文件分片内容 seq: 38
orgDoc: 在办理个人外汇业务中, 哪些客户是关注客户?

关联文本分片

> 私有语料库:

> 私有语料库:

> 私有语料库:

取消

确定

图 49 最新引用

4. 点击工单状态后的 待处理，可将状态调整为 处理中 或 处理完成，并填写下方的 处理描述。

5. 点击右下角 确定 即可修改反馈处理状态。

工单处理

工单状态 待处理

处理描述 待处理
处理中
处理完成

取消

确定

图 50 修改反馈处理状态

31/38

6. 当有新的改进意见提交，可点击右上角 **刷新** 按钮查看最新的改进意见。

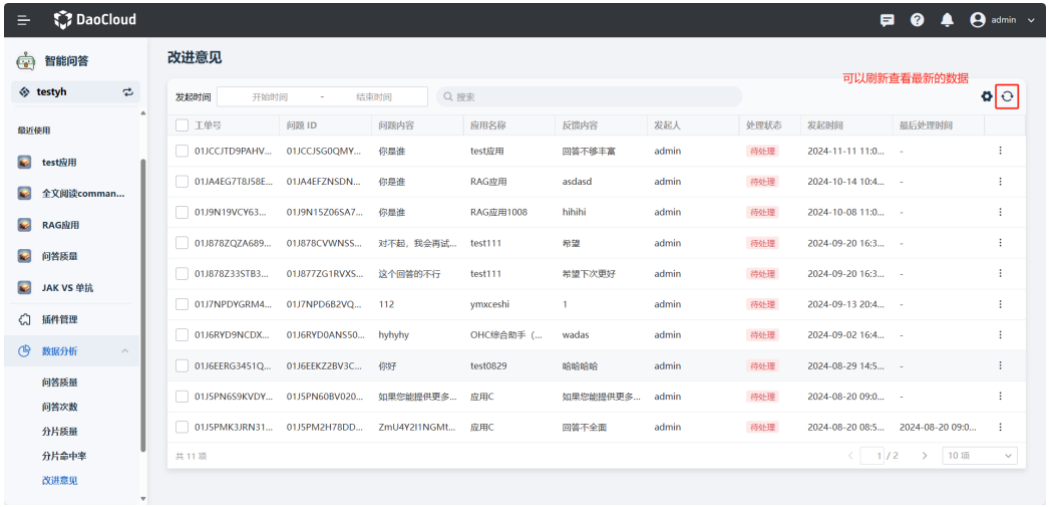


图 51 刷新改进意见

导出反馈

将所有问答的详细内容以及问答反馈汇总成表格导出。

- 1. 点击 **问答反馈** 左方，选取需要导出的反馈，点击右上角 **导出** 按键。
- 2. 将问答内容汇总为 **xlsx** 文件并下载。



图 52 导出反馈

六、我的反馈

d.run 记录了当前用户发出的反馈信息，以方便查看后续该反馈得到的回馈。

「我的反馈」功能用于集中记录并管理当前用户提交的所有反馈信息，方便您在后续跟进查看每条反馈的处理进度与最终结果。无论是对问答内容的准确性、模型的表现，还是系

统功能的建议与问题，您都可以通过该功能提交反馈，并随时回溯历史记录。这不仅有助于平台及时优化 AI 模型与知识库内容，还能保障您的意见被有效采纳和回应，从而持续提升整体使用体验。

(一) 查看反馈详情

1. 点击 **我的反馈**，可通过搜索找到需要查看的反馈，点击范围内任意位置均可进入详情。

DaoCloud

admin

我的反馈

发起时间

开始时间 - 结束时间

搜索

工单号

点击任意位置即可进入详情

问题内容

应用名称

反馈内容

处理状态

发起时间

最后处理时间

01JCCJTD9PAHV95TM9J9CSBC2	你靠谱	test应用	回答不够丰富	处理完成	2024-11-11 11:03:33	2024-11-11 13:55:41
01JAE6G78J58EE4N058F79ZHP	你靠谱	RAG应用	asdasd	待处理	2024-10-14 10:42:47	-
01J9N19VCY63Q052HSZX841HX0	你靠谱	RAG应用1008	hihihi	待处理	2024-10-08 11:03:29	-
01J876ZQZA689QSEY48DA6H3V	对不起，我会再试一次。七龙珠是一部经...	test111	希望	待处理	2024-09-20 16:32:43	-
01J876Z33TB3M37VFTKVK41GE	这个回答的不行	test111	希望下次更好	待处理	2024-09-20 16:32:22	-
01J7NPDYGRM42T4C5M0475P5HD	112	ymacshi	1	待处理	2024-09-13 20:41:20	-
01J6RYD9NCDXVW1M1NA6H12TT	hyhyhy	OHC综合助手（超）	wadas	待处理	2024-09-02 16:42:49	-
01J6EERG3451QK7QEZ3WEWNHRWZ	你好	test0629	哈哈哈哈	待处理	2024-08-29 14:56:54	-
01J5PN659KVDYV6Y0ETYSQK7Z7	如果你能提供更多的背景信息或明确您的...	应用C	如果您能提供更多的...	待处理	2024-08-20 09:07:48	-
01J5PMKJRN31AN90HBDVGMZ6Y	ZmU4Y21fNGMtyNjZ00YyZlTkZjYl...	应用C	回答不全面	待处理	2024-08-20 08:57:03	2024-08-20 09:06:39

共 11 页

< 1 / 2 > 10 页

图 53 我的反馈列表

2. 在该页面可以查看以下内容：

- **用户反馈：**用户本人发出的反馈内容。
- **相关信息：**助手名称、模型名称、引用条数、处理时间以及问答详情。
- **处理意见：**管理员处理反馈后发出的意见。
- **历史引用与最新引用。**

工单号: 01HS2YR1TCN7YC2JR20WP0M6A 处理完成

用户反馈

用户反馈内容

你好

相关信息

助手名称 12223大数据你们大数据你们大数据你们

模型名称 -

引用条数 5

处理时间 2024-03-16 14:26:21

用户 admin

hi

2024-03-16 14:21:20

提示词

["content":"hi","role":"user"]

应用回答内容

Hello! How can I help you today?

处理意见

办好

图 54 反馈详情

历史引用

> 私有语料库:

> 私有语料库:

> 私有语料库:

> 私有语料库:

> 私有语料库:

最新引用

> 私有语料库:

> 私有语料库:

> 私有语料库:

> 私有语料库:

> 私有语料库:

图 55 引用详情

七、进阶配置

本章介绍两项平台级的进阶能力：定义全局默认参数的 **系统配置**，以及用于自定义分片逻辑的 **分片插件**。两者均由平台管理员在工作空间或平台级别完成配置。

（一）系统配置

系统配置用于定义平台中与智能问答应用相关的全局默认参数，包括聊天记忆轮数、语料库提示词模板、嵌入提示词模板等。这些配置会影响应用在运行时的行为与输出效果，并在创建新应用时作为默认设置自动应用，帮助您快速构建一致且可控的对话体验。

系统配置中的参数将会作为创建应用时的默认参数。

DaoCloud | d.run

系统配置

聊天记忆轮数

10

语料库提示词模板

做一个知识问答游戏：
1. 回答内容必须在“{corpus_search_content}”中。
2. 如果问题在所提供的资料信息内无法找到，你会回答“抱歉，资料库已知信息中，没有找到您需要的结果，请尝试修改引用的资料”

嵌入提示词模板

{user_inputs_content}

保存

图 56 系统配置

1. 聊天记忆轮数

指在一个对话系统中，用于追踪和管理对话历史的轮数或回合数。每当用户与系统进行一次交互，对话的轮数就会增加一次。此处指用户与聊天应用进行交互时，应用会在一个小记忆窗口中保持一段时间的对话历史，记忆轮数大小即保持时间的长短。

- 默认值为 10。
- 最低为 0。

2. 语料库提示词模板

该模板在应用关联使用模板时启用，模板内容包含通过相似度搜索得到的知识块、用户的输入和嵌入提示词模板。

3. 嵌入提示词模板

嵌入的提示词拼接在用户的问题之前，作为通用的约定提示来引导应用模型输出针对问题的回答。

（二）分片插件实现与集成指南

本节介绍如何实现一个分片插件，并将其集成到系统中。

例如，您可以部署一个插件，该插件将接收 Markdown 格式的文本，然后将其按照标题进行分片，最后将插件添加到系统。现在就可以在语料库中导入 Markdown 格式的文本，并选择该插件进行分片。

1. 实现分片插件

设计插件功能

在设计插件功能时，需要确定以下几点：

- **输入格式：**插件将接收什么类型的文本作为输入？在本示例中，我们将接收 Markdown 格式的文本作为输入。
- **分片标准：**根据什么标准将文本进行分片？在本例中，我们将根据 Markdown 文档的标题进行分片。

实现插件功能

编写代码实现插件功能。下面是一个示例代码的框架：

```
from markdown_splitter import MarkdownSplitter

def import_markdown_file(file_path):
    # 读取 Markdown 文件内容
    with open(file_path, 'r', encoding='utf-8') as file:
        markdown_text = file.read()
```

```
# 使用插件进行分片
splitter = MarkdownSplitter()
splitted_text = splitter.split_by_heading(markdown_text)

# 处理分片后的文本
# ...
```

将分片功能暴露成 HTTP 服务，接口如下。

请求信息

- **HTTP 方法：**POST
- **请求 URL：**[/embeddings/FileSplitAndLocalEmbedDoc](#)
- **支持格式：**JSON

请求体

请求体应包含待处理文档的内容。文档应采用如下 JSON 格式：

```
{
  "fileGetPath": "http://xxx/file.md",
  "params": {
    "key1": "value1",
    "key2": "value2"
  }
}
```

名称	类型	必填	描述
fileGetPath	string	是	文件路径
params	object	否	自定义参数，key 为 string，value 为 string

响应信息

- **状态码：**200 OK
- **内容类型：**JSON

响应体

响应体将包含处理后的嵌入结果，格式如下：

```
{
  "content": [
    {
      "orgDoc": "h1: 流沙之战\n 作者: Micky Neilson\n\n 正午的太阳坚定地凝视着希利苏\n 斯的流沙, ... 暴力。",
      "pageIndex": 0
    }
  ],
  "resultCode": "2008000000001",
  "resultMessage": "success"
}
```

名称	类型	必选	中文名 / 说明
----	----	----	----------

content	object	true	分片列表
orgDoc	string	true	分片内容
pageIndex	integer	true	分片页码
resultCode	string	true	状态码，200800000001 成功，200800000002 失败
resultMessage	string	true	状态描述 success / failed

测试插件功能

在实现插件功能后，务必进行测试以确保其正常运行。可以编写单元测试或手动测试来验证插件的正确性。

2. 将插件集成到系统中

在系统中添加插件接口，使得语料库模块可以调用插件的功能。

添加插件接口

1. 在 **插件管理** 页面中，点击 **创建** 按钮。

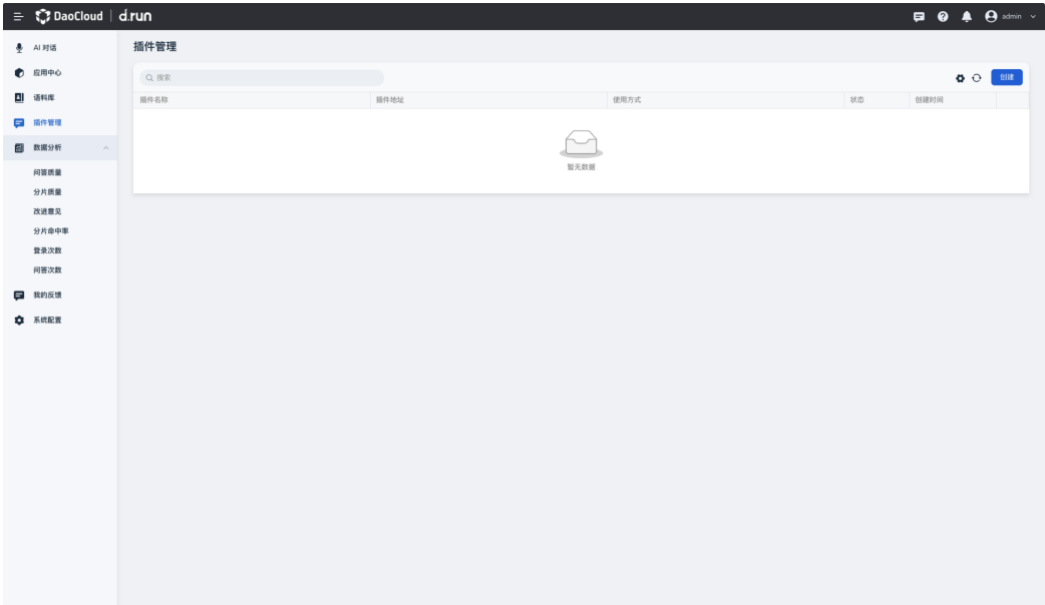


图 57 创建插件

2. 参考下列要求填写插件信息，并点击 **确定**：

- **插件名称**：名称可以包含大小写字母、数字、连字符（“-”）、中文，最长 63 个字符。
- **插件地址**：插件的地址应以 <https://> 或者 <http://> 开头。
- **文件类型**：插件处理的文件类型，当前支持的文件类型包括 pdf、txt、docx、doc、csv、xlsx、ppt、md。
- **使用方式**：插件的功能介绍，可以包括分片的方法和其他功能说明。

- **参数列表：**键值对，描述在分片的时候需要填写对应参数的值，例如 key1:value1, key2:value2。

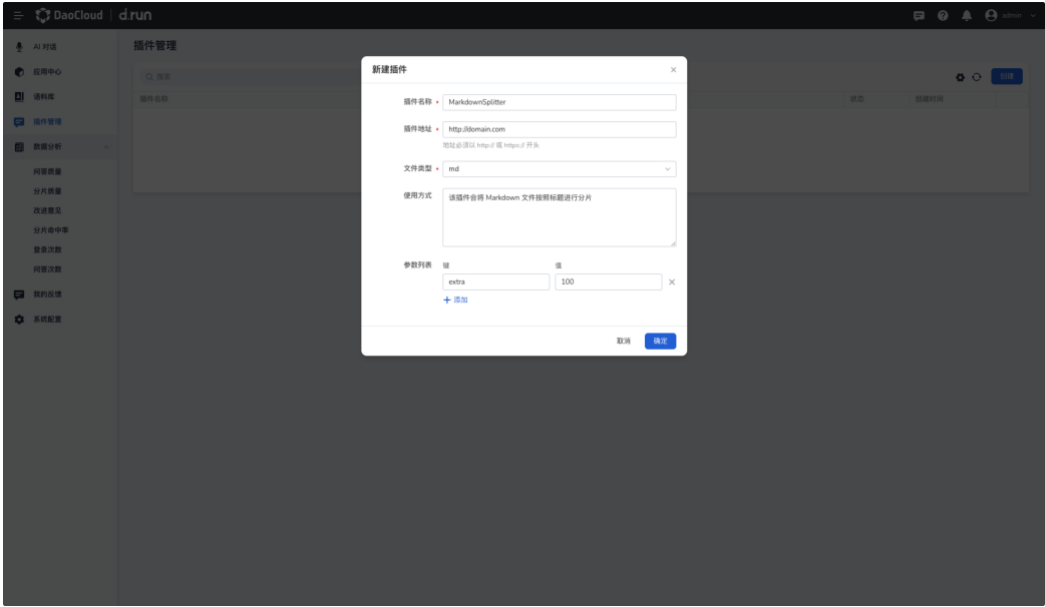


图 58 填写插件信息

填写正确信息并点击确定后，页面会提示创建成功。

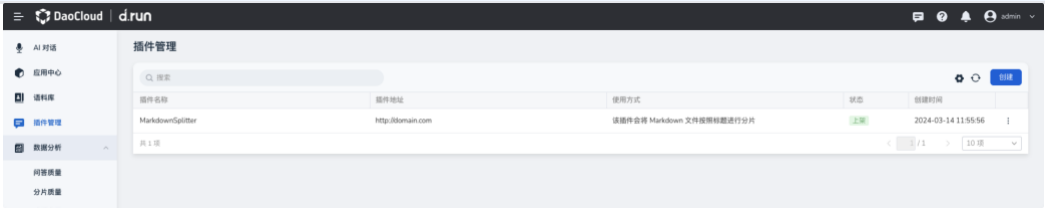


图 59 创建插件成功

使用插件

在本例中，可以在语料库导入模块中调用插件功能，以在导入 Markdown 类型文件时进行分片。