

自研搭建AI助手知识库

背景

AFE有自建自有知识库功能，但知识库的切片效果不佳，存在两个问题：

- 1、文档处理能力不足：包括word/pdf/excel文档中的图片、表格无法有效提取处理；
- 2、切片规则不佳：如下图一段文本被强硬拆分在两个不同的片段；



任务

搭建一个AI助手知识库demo，能支持上传文件，查看切片效果；可以测试召回效果，查看召回入参和回参；

目标

目标：首要做好切片，其次是召回。

支持文档类型：

pdf、doc、docx、ppt、pptx、wps、ppsx

xlsx、xls、csv、md、txt、html、json、xml、log

jpg、png、jpeg、bmp、gif

切片规则：

表格文档：默认切分/按行切分

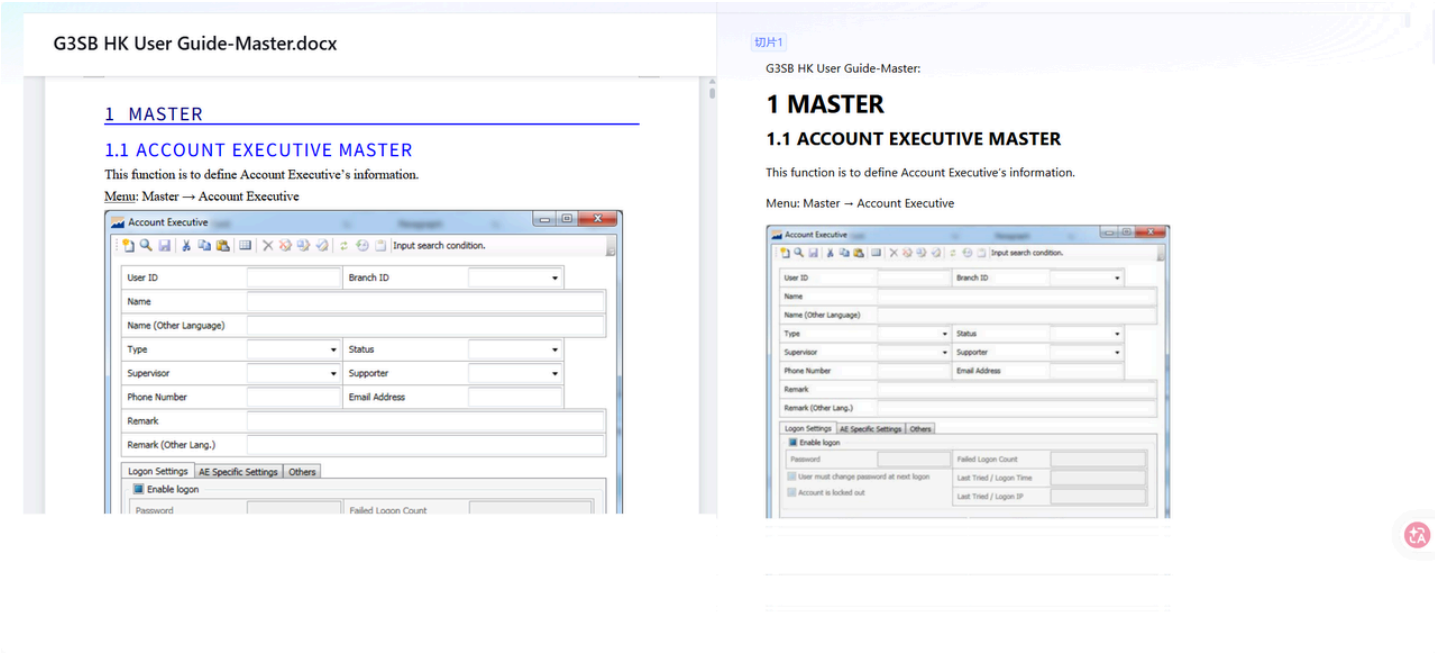
其他文档：默认切分/通用标识符切分

切片内表格：Markdown格式

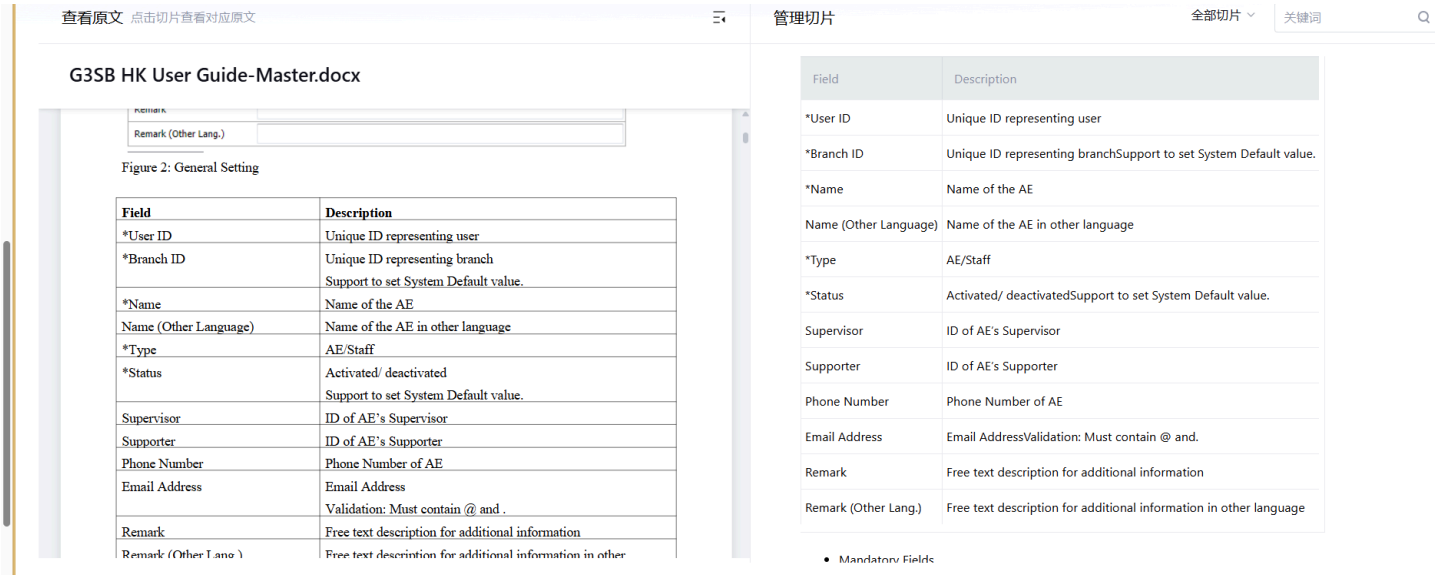
期望

期望效果（基础）：

1. 图片展示在原文本展示的问题，不会丢失或者移动到其他位置



2. 表格内容正确提取，并整理为markdown文本



3. 切片规则合理，对于word至少同一个标题下的内容在一个切片中

查看原文 点击切片查看对应原文

G35B HK User Guide-Master.docx

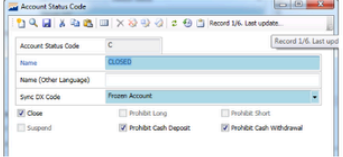
Figure 5: Others

1.2 ACCOUNT STATUS CODE MASTER

User can update, create or revoke account status on "Account Status Code Master".

Define the account status code, name and choose the action(s) by checking the checkbox(es) below, user can execute the status code when account status is being used.

Menu: Master → Account Status Code



管理切片

全部切片 关键词

Figure 5: Others

切片2

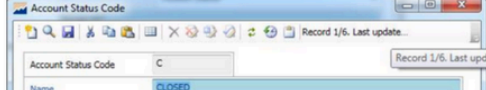
G35B HK User Guide-Master:

1.2 ACCOUNT STATUS CODE MASTER

User can update, create or revoke account status on "Account Status Code Master".

Define the account status code, name and choose the action(s) by checking the checkbox(es) below, user can execute the status code when account status is being used.

Menu: Master → Account Status Code



期望效果（高级）：

1. 对图片图形能够提取为图片，且识别文本信息转为文字

H5 i-Trade用户手册_2022_Nov_tc.pdf



H5 i-Trade用户手册_2022_Nov_tc.pdf



H5 i-Trade用户手册_2022_Nov_tc.pdf



H5 i-Trade用户手册_2022_Nov_tc.pdf



2. 对pdf水平分布的栏目文本能够识别为正常顺序

huawei_annual_report_2024_cn.pdf 35.46MB 160个切片

提交后重新处理文档 文档切片设置 原文修改

查看原文 点击切片查看对应原文

管理切片

全部切片 关键词

huawei_annual_report_2024_cn.pdf

华为是谁？

华为创立于1987年，是全球领先的ICT（信息与通信）基础设施和智能终端提供商。华为目前有20.8万员工，业务遍及170多个国家和地区，服务全球30多亿人口。我们致力于把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。

把握时代机遇，点亮浩瀚星空

共筑数智底座，拥抱全面智能化

2024年，我们见证了以人工智能技术为主要驱动力的第四次工业革命的来临。人工智能技术即将成为全球经济的重要推动力，并加速进入商用普及阶段。个人、家庭和企业都将因为人工智能发生深刻变化。通信网络的容量、联接时延和网络架构也将因为人工智能全面升级。推动各行各业走向全面智能化时代。华为将加强技术创新投入，提供高质量的产品与服务，以构建坚实的数智底座，建立行业灯塔，帮助照亮千行万业智能化升级的方向。

加强技术创新，坚持开放合作

全面智能化时代的到来，为技术创新带来了前所未有的机遇和挑战。华为扎根研究与创新，秉承开放合作原则，与产业界、学术界协同驱动技术进步和产业升级。华为坚持客户需求与科学技术双轮驱动的创新机制。过去三年，每年将公司销售收入的20%以上投入研究与开发，以保持产品与解决方案的竞争力。我们深入行业场景，牵引产品规划，持续强化在软件、硬件、芯片、终端、网络、云的战略合作与研发投入，构筑更开放易用平台，使能百模千态，努力为客户创造更大的价值。

践行以质取胜，坚守安全可信

华为坚信“质量为企业的生命”，坚持“以质取胜”。我们致力

构建万物互联的智能世界

华为是谁？

华为创立于1987年，是全球领先的ICT（信息与通信）基础设施和智能终端提供商。华为目前有20.8万员工，业务遍及170多个国家和地区，服务全球30多亿人口。我们致力于把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。

把握时代机遇，点亮浩瀚星空

共筑数智底座，拥抱全面智能化

2024年，我们见证了以人工智能技术为主要驱动力的第四次工业革命的来临。人工智能技术即将成为全球经济的重要推动力，并加速进入商用普及阶段。个人、家庭和企业都将因为人工智能发生深刻变化。通信网络的容量、联接时延和网络架构也将因为人工智能全面升级。推动各行各业走向全面智能化时代。华为将加强技术创新投入，提供高质量的产品与服务，以构建坚实的数智底座，建立行业灯塔，帮助照亮千行万业智能化升级的方向。

加强技术创新，坚持开放合作

全面智能化时代的到来，为技术创新带来了前所未有的机遇和挑战。华为扎根研究与创新，秉承开放合作原则，与产业界、学术界协同驱动技术进步和产业升级。华为坚持客户需求与科学技术双轮驱动的创新机制。过去三年，每年将公司销售收入的20%以上投入研究与开发，以保持产品与解决方案的竞争力。我们深入行业场景，牵引产品规划，持续强化在软件、硬件、芯片、终端、网络、云的战略合作与研发投入，构筑更开放易用平台，使能百模千态，努力为客户创造更大的价值。

践行以质取胜，坚守安全可信

华为坚信“质量为企业的生命”，坚持“以质取胜”。我们致力于为客户提供高质量的产品和服务。做到产品生命周期内免维护；我们将三十多年的质量管理经验延伸到所有的合作伙伴，构建全流程、端到端的质量管理体系，共同推动相关产业链的高质量发展。