**南京科远智慧科技集团股份有限公司**

**基于大模型的EmpoworX智能问答系统开发项目**

**产品试用总结报告**

**（编制时间：2025年7月23日）**

编 制： 靳鸿祥

审 核： 顾坤

批 准： 周松明

**目 录**

[1 概述 2](#_Toc22469)

[2 项目产品试用情况 2](#_Toc11411)

[3 验收部门意见 2](#_Toc15399)

[4 经验与教训 2](#_Toc21407)

[5 持续改进的内容及改进计划 2](#_Toc10761)

# **概述**

智能问答产品的试点背景如下：

1. **客户方情况**

平台开发部的在线文档系统试用，包括小科办公、大数据、empoworx平台、业务平台，具体是将各个平台的在线文档转为markdown文件，从而上传到智能问答系统，从而实现知识录入和知识问答。

化工数字化安全中心的完整系统试用，具体是将化工数字化安全中心的文件上传到系统，从而实现知识录入和知识问答。

1. **试用环境**

智能问答产品的环境部署在10.18.30网络环境下，具体地址是外网访问地址，从而保证平台开发部的在线文档能够正常调用智能问答系统接口。

1. **应用场景**

在线文档的问答助手、化工数字化中心的知识助手、常见问题的诊断助手。

1. **产品试用整体质量情况**

智能问答产品整体试用没有出现阻塞性问题，出现较多优化问题，部分问题是技术本身限制，部分问题是硬件资源问题。

# **项目产品试用情况**

**试用目标：**

（1）文档平台的文档录入到知识库；

1. 提供相关接口给平台开发部进行试用；
2. 按照平台开发部的试用需求定制开发功能；
3. 化工安全规范的文档录入到知识库；
4. 提供完整系统给化工安全数字中心部门进行试用；

**验收标准：**

1. 提供相应接口，按照平台开发部试用需求，定制开发完成，满足平台开发部的日常使用需求，系统无明显异常。
2. 提供完整系统给化工安全数字中心部门进行试用，将化工安全规范的文档录入到知识库、系统无明显异常。

**试用周期：**2025年5月20日到2025年7月23日

**试用现场：**平台开发部文档平台、化工安全数字中心部门

**产品试用整体情况：**

1. 平台部门目前试用问题6个，已解决4个，其中有2个作为遗留问题；  
   2、化工数字化中心目前试用问题10个，已解决8个，其中有2个作为遗留问题;

**遗留问题整体情况：**

1、系统无法回答大局观和综述性问题：该问题来自于RAG本身技术问题，目前通过生成目录知识块，部分解决上述问题，但是整体RAG就是将文档知识切片，所以针对一些大局观性问题，就比较难以回答。

2、系统无法回答isEmpty函数问题：该问题来自原始文档是“isEmpty”，不是“isEmpty函数”，同技术本身有问题，目前是采用关键词和语义向量检索知识块，那么“函数”必定会作为一个关键词，所以导致检索到的知识块有误，后续考虑将ElasticSearch数据库替换为专门的向量数据库，直接采用纯语义向量进行匹配，依赖于嵌入向量的嵌入效果。

3、回答无法兼容上下文问题：该问题来自于检索系统无法支持上下文，大模型本身支持上下文功能，但是检索支持不了上下文功能，所以目前业界普遍做法就是问题重写，根据上下文对话历史，将最新用户问题改写成完整的用户问题，从而执行后续的RAG流程，但是目前显卡资源受限，如果在现有流程增加问题重写，回答速度显著下降，后续考虑增加专门用于问题重写的小模型。

4、回答不出来但是检索出知识块问题：该问题由于检索系统本身并不支持判断问题和知识块的绝对相关性，此外，大模型判断问题和知识块的相关性输出格式又不固定，没有办法根据大模型输出的“知识块未找到您想要的答案！”隐藏知识来源段落，后续考虑增加流程来判断知识来源段落是否显示。

# **验收部门意见**

化工数字化部门整体评级：

智能问答产品整体试用情况，基本能满足企业自建规章制度知识库的快速检索使用需求，目前产品存在的遗留问题需持续改进。

对比市面上的友商产品，希望开发团队能尽快推进后续改进计划，进一步提升产品性能，增强我们产品的市场竞争力。

最后，提出两点建议：

1、建议将语音输入识别功能纳入APP应用端优化提升规划；

2、建议新增新旧版本文档对比分析功能，对修订内容进行标识和总结。

平台部门整体评级：

智能问答系统解决了公司的文档库的规模大、交付及其他开发人员查找文件资料的难题；该智能问答系统，能够很好从用户查询的问题中提取关键字并进行查询、过滤和回答，甚至连相关文件中的信息都可以匹配查找，这些传统查询很难做的，整体试用从UI，交互，响应速度方面来看也是不错的。

# **经验与教训**

本项目在试用推进过程中遇到了一系列挑战，这些问题不仅揭示了项目管理中的薄弱环节，也为未来的项目提供了宝贵的经验与教训。以下是对整个试用期间问题的总结及相应的经验教训：

1. **外网环境搭建工作耗时：**针对平台部门文档平台试用，因为平台部门文档服务挂载在10.18.30网络下服务器，同办公网络不互通，所以需要首先ERP申请三个虚拟服务器、然后将装载4090服务器挂载到10.18.30网络下、最后开放智能问答系统服务地址，保证其能够被外网访问到。
2. **外网环境操作限制较多：**在平台部门进行文档平台试用过程中，将其部署到外网环境时遇到了诸多问题，比如限制文件上传大小在30MB、所有请求不能携带内网地址，这些问题的排查既耗时又费力。此次经历提醒我们，在处理涉及内外网环境的情况时，应该首先确保内网环境的操作无误，然后再检查外网环境的问题。这样做可以避免由于网络问题导致不必要的长时间排查工作，提高效率和准确性。
3. **开发环境无法满足需求：**在平台部门试用阶段，将文档平台和小科办公接入智能问答系统时，由于环境的紧缺，导致项目本身试用开发测试和平台试用开发时，无法同时运行环境，导致环境需要来回切换，影响到部分开发工作。无法同时运行的原因就是测试没有较好的办法在测试环境单独部署。

# **持续改进的内容及改进计划**

在本项目的试用过程中，发现了本身技术限制的问题。针对这些问题，已经提出相应的改进措施。

1. **系统无法回答大局观和综述性问题**，改进计划就是尽可能增加一些元信息知识块，从而使得目录结构知识块作为基础信息，增强系统的回答效果。
2. **检索系统识别关键词有误的问题**，改进计划就是优化检索系统，改用轻量化的向量数据库，以向量搜索的内容为基准，在向量搜索的基础上优化文本检索方法，从而保证语义匹配。
3. **检索系统没有多轮对话检索的问题**，改进计划就是本身大模型支持多轮对话理解，但是检索系统没有办法直接支持多轮对话检索，业界内普通的做法就是基于历史对话，重写问题，保证问题的完整性，但是据我们实测发现，在显卡资源受限的情况下，回答速度受限，后续考虑用占用资源较少的技术实现。
4. **Python环境打包的问题**，改进计划就是提取出不同cuda版本的环境依赖包，针对不同cuda提供不同的Python运行包，此外，解决打包过程中无法创建ES索引的问题。
5. **深度思考的流程问题**，改进计划就是针对现有模型的不同思考类型，在模型管理页面提供不同的思考类型选择。