# 工作汇报

1. 自我介绍 学历背景、研究方向、入职时间、实习内容 1页ppt
2. 工作总体内容 PLM智能问答（问答对、文档、多轮对话）1页ppt
3. 问答对 （总体研究 、 挑战和解决方法）2页ppt
4. 文档 （总体、 ）3-4页ppt 文档链接、tfs链接
5. 优化想法 1页

## 工作背景/内容

**项目名称：PLM智能问答项目**

**项目背景：**现有的PLM问题解答，前一版自动回答的效果不好，目前工作流程主要由后台工作人员接收用户提问，人工检索相关内容并汇总，选择最符合的结果回复提问，包括问答对、截图、表格、文本等数据。总体业务逻辑为，对于能够由标准QA知识库解答的内容直接由问答对解答，超出部分由解决方案文档的相关知识解答。

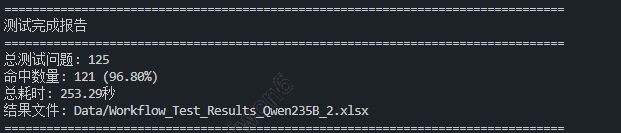
|  |  |
| --- | --- |
| 工作任务 | 工作目标 |
| 调研基于问答对的算法方案，完成开发上线 | 产出效果测评文档、代码；上线后业务测试准确率≥95% |
| 基于LightRAG框架的基于解决方案文档的算法模块 | 产出模块代码，相关接口，用户文档，提升文档解析准确率 |
| 调研KAG框架/结合业务需要搭建适配版 | 完成技术调研报告及框架搭建方案 |

## 时间安排与当前进度

|  |  |  |
| --- | --- | --- |
| 工作任务 | 计划时间 | 当前进度 |
| 问答对算法模块 | 7月-7月底 | ✅ 已完成（7/22上线） |
| LightRAG框架算法模块 | 8月上旬-8月底 | 框架构建、测试中 |
| KAG框架调研 | 8月中下旬 | 技术调研进行中 |

## 实现效果与量化结果

1. 基于问答对的算法方案：
   1. **准确率**：由业务方问答数据衍生的测试集命中率达96.8%（目标≥95%），测试数据包含：口语化、同义表达、错别字、简化表达、关键词组合、混合型方面的测试类型。
   2. **响应效率**：代码上线并集成在MaaS平台上，通过MaaS平台提供统一接口，测试平均运行时间在800ms-1200ms
   3. **上线范围**：提供通用可泛化的重排序与智能判断模块，针对协定好的json格式{“query”，“documents”}数据均可提供模块功能支持。



1. 基于LightRAG框架的基于解决方案文档的算法模块方案：
   1. **文档解析**：针对解决方案文档进行自动化word到pdf的转化，通过MinerU工具对PDF文档进行解析，可以获得Markdown格式、HTML格式的文本内容，以及对于图片和表格解析得到的图片内容。
   2. **标签索引**：鉴于解决方案文档的层次结构，通过解析目录的方式，针对chunk添加“文档名称-一级标题-二级标题”的元数据标签，用以后续主题匹配与内容检索。
2. KAG框架方案：
   1. 基于KAG框架的生成效果，由于其融合知识图谱与大型语言模型的混合推理架构，通过逻辑符号引导的推理流程实现专业领域的高精度问答和决策支持，可以提高专业性场景的适配度，通过严格事实性保障减少“幻觉”问题。

## 实现方案与技术挑战

1. 基于问答对的算法方案：
   1. 实现方案：

方案流程为:

|  |
| --- |
| | Json输入 | -> | 并行双路重排序 | -> | 结果聚合 | -> | 分数清洗&去重 | ->  | RRF融合排序 | -> | Qwen3-235B非思考模型判断 | -> | 清洗结果并输出 | |

原方案流程为：处于业务方通用性考虑，上线部分为重排序&判断的模块部分。

|  |
| --- |
| | 数据增强 | -> | QA知识库向量化 | -> | TF-IDF结合BERT语义的混合检索 | ->  | 重排&判断（现方案模块）| -> | 输出结果 | |

通过BGE-M3的多语言能力，使中英文混合Query的匹配准确性提高18%。

RRF融合并清洗后的候选问题质量提升，使最终LLM匹配环节的无效调用减少，单次平均耗时减少30%左右。

通过结果清洗和反向索引，避免无效输出，提升答案可靠性。

* 1. 技术挑战：
     1. Query语种形式多样：通过双路重排序异构互补，Qwen3-Reranker是轻量级的中文优化模型，侧重于语义连贯性理解；BGE-M3是多语言泛化模型，侧重于细粒度的跨语言匹配。通过双模型重排序与RRF融合，聚合了不同模型的排序偏好，BGE-M3的多语言能力弥补了Qwen3在非中文场景的局限性，并通过Qwen3的中文语义能力加强同义表达的问题的匹配能力。
     2. 相似度阈值局限性：由于基于QA的回答具有高匹配度的要求，需保证用户询问和匹配问题核心诉求一致，避免答非所问。单纯的通过相似度阈值筛选得到的效果不佳，引入非思考的Qwen3-235B语言模型进行判断。
     3. 大模型误判模糊语义：增强提示词，增加严格的完全匹配定义，降低模型温度设置，避免扩散思维、跳出定义等，提高答案可信度和一致性。
  2. 后续优化：探索小模型蒸馏+负样本微调提升模型对于模糊语义的判断能力，降低误判率与延迟。

1. 基于LightRAG框架的算法方案：
   1. 实现方案：通过LightRAG框架构建整体RAG系统，针对文档解析部分采用MinerU工具进行解析。

整体工作流程为：

|  |
| --- |
| | 微调BERT或小参数大模型（例如Qwen-1.8B）对Query分类（简单：单事实检索/复杂：多跳推理）| -> | 简单查询：向量检索库/复杂查询：知识图谱推理 | ->  | 证据聚合与重排序（去噪后，向量结果：RRF重排序/图谱结果：路径置信度打分）| ->  | 答案生成（轻量回答：直接抽取答案/需推理生成：LLM生成）| |

* 1. 技术挑战：
     1. 文档解析乱码：采用MinerU对文档进行解析，因为中文编码问题会产生乱码，影响结果匹配。通过预处理文档与编码方式匹配MinerU编码工具，解决乱码问题。
     2. 表格嵌图抽取分辨率低：针对表格数据抽取出的图片内容因为嵌套关系导致内层图片的分辨率低。考虑通过针对表内图片的预处理和保存解决，针对表格数据先抽取其中图片内容并保存。
     3. 表格嵌图解析错误：针对表格中包含图片的数据解析出的HTML格式文本内容存在错误的将图片内容解析成文本的情况，导致数据乱序和异常。通过抽取图片的同时在原地打上标签，将多元表格数据转化为单文本表格。

1. 基于KAG框架的算法方案：

目前整体还处在路线调研和探索阶段，可以预见的技术挑战有：

* + 1. 非结构化文本（如PDF、网页）到图谱三元组的**高精度自动抽取**

采用联合抽取模型代替流水线模型

* + 1. **多元异构数据的模式对齐**