华南理工大学
South China University of Technology

面向电网技术文档的智能信息提取与分析研究

答辩人 贾钰杰

指导老师 吴庆耀 教授

答辩时间 2025 年 06 月 27 日

总体情况

研究内容和考核指标差异表

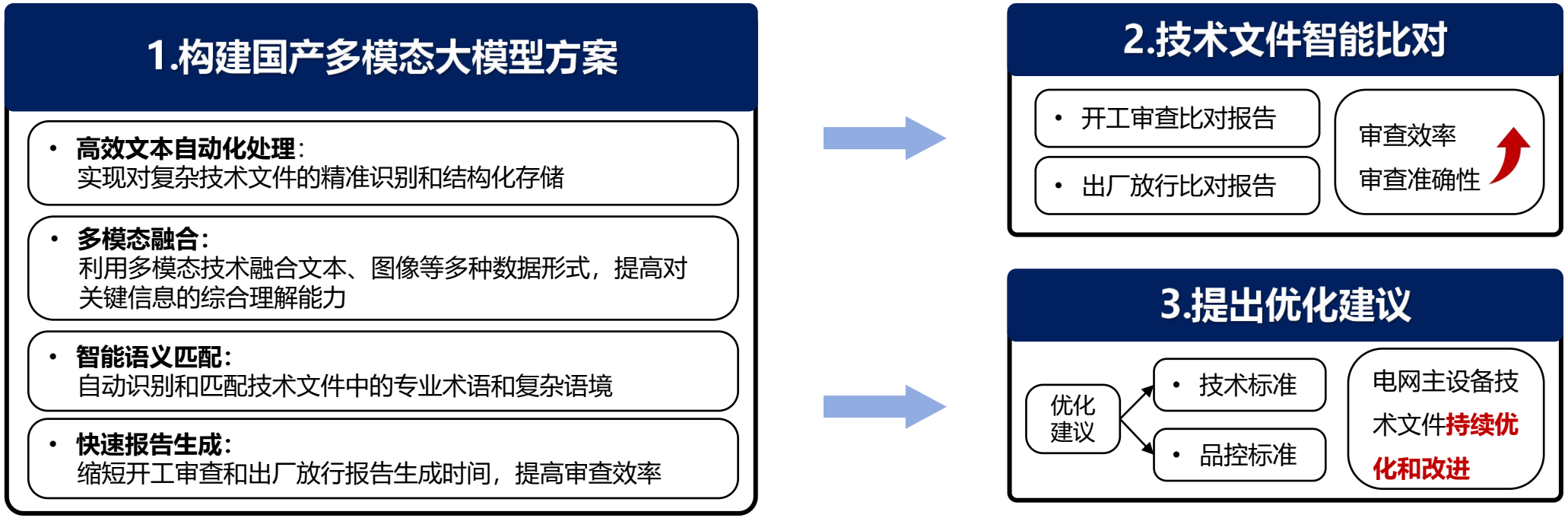
序号	类型	申报指南要求	项目单位响应情况	自评情况
1	主要研究内容	（1）研究和开发适用于电网主设备技术文件提取关键技术参数的智能算法和模型,形成结构化通用化技术标准架构模板。 （2）通过研究多模态内容识别技术及探索多模态对齐策略，实现开工审查文件内容的智能识别与判断，生成开工审查比对报告。 （3）通过研究多模态内容识别技术及探索多模态对齐策略，实现出厂放行相关文件内容的智能识别与判断，生成出厂放行比对报。 （4）通过对电网主设备技术文件的智能分析，为技术标准，品控标准提供优化建议。	1、通过【任务一：针对电力技术规范文档理解构建国产多模态大模型方案。】【任务二：针对技术规范文档的大模型微调任务设计】解决（1）研究和开发适用于电网主设备技术文件提取关键技术参数的智能算法和模型,形成结构化通用化技术标准架构模板”的内容。 2、通过【任务三：研发智能审查与比对报告生成应用。】解决“（2）通过研究多模态内容识别技术及探索多模态对齐策略，实现开工审查文件内容的智能识别与判断，生成开工审查比对报告”、“（3）通过研究多模态内容识别技术及探索多模态对齐策略，实现出厂放行相关文件内容的智能识别与判断，生成出厂放行比对报告”及“（4）通过对电网主设备技术文件的智能分析，为技术标准，品控标准提供优化建议”的内容。	完全满足指南
2	预期目标	（1）建立一套适用于电网主设备技术文件的多模态大数据模型智能算法和模型，形成结构化通用化技术标准架构模板。 （2）通过智能识别和语义匹配技术，自动化开展开工审查和出厂审查的技术文件智能比对，生成规范的开工审查比对报告、出厂放行比对报告，提高审查效率，减少人工干预和错误，提高审查准确性。 （3）通过智能分析技术，提出技术标准、品控标准提出优化建议，提升整体质量管理水平，确保电网主设备技术文件的持续优化和改进，助力供应链的数字化建设。	（1）建立一套适用于电网主设备技术文件的多模态大数据模型智能算法和模型，形成结构化通用化技术标准架构模板。 （2）通过智能识别和语义匹配技术，自动化开展开工审查和出厂审查的技术文件智能比对，生成规范的开工审查比对报告、出厂放行比对报告，提高审查效率，减少人工干预和错误，提高审查准确性。 （3）通过智能分析技术，提出技术标准、品控标准提出优化建议，提升整体质量管理水平，确保电网主设备技术文件的持续优化和改进，助力供应链的数字化建设。	完全满足指南

研究目标

（1）建立一套适用于电网主设备技术文件的多模态大数据模型智能算法和模型，形成结构化通用化技术标准架构模板。

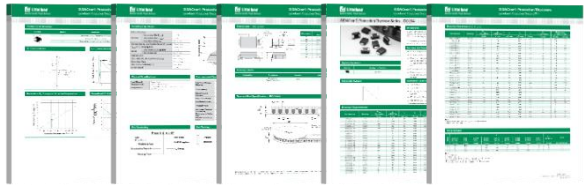
（2）通过智能识别和语义匹配技术，自动化开展开工审查和出厂审查的技术文件智能比对。

（3）通过智能分析技术，对技术标准、品控标准提出优化建议，提升整体质量管理水平。



关键挑战

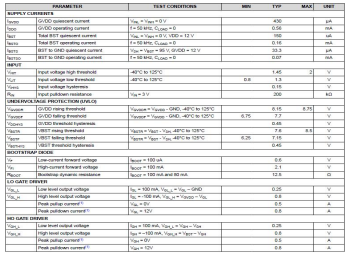
挑战一：文本结构化处理效率低



技术规范文档

耗费时间长
识别不精确

↓ 人工文本识别

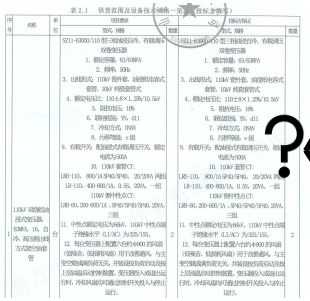


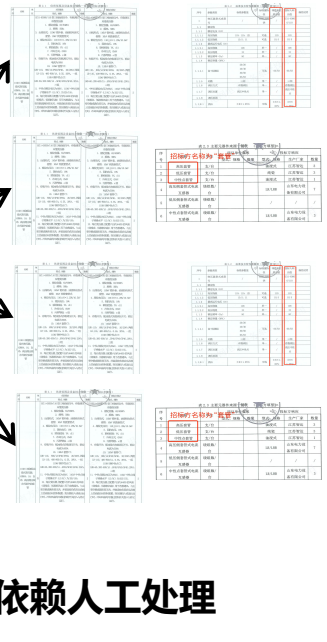
公司在技术文件的文本识别方面**处理效率较低**且对复杂技术文件的**准确率不足**

挑战二：语义识别与匹配难

模型库模板

待识别技术文档





模板匹配

↓

依赖人工处理

公司在语义识别和匹配方面**依赖人工处理，存在效率低下和易出错**的问题

挑战三：智能审查报告生成耗时长

电网管理平台

↑

依赖人工在电网管理平台上对项目进行比对

公司在开工和出厂审查过程中，报告生成主要**依赖人工比对，耗时长且易出错**

研究内容：基于大模型指令优化的指标匹配

研究内容

基于多级结构匹配技术的规范文档指标数据集选取

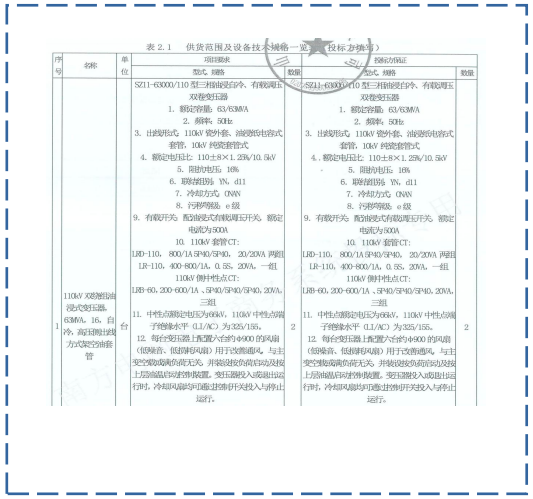
技术路线

识别模版格式

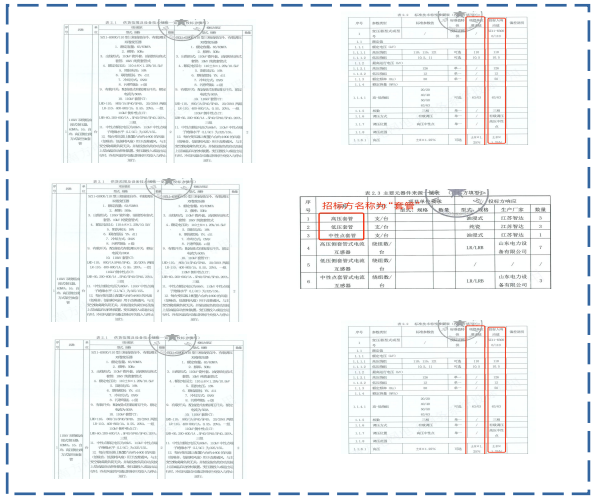
基于多级结构、语义匹配

特定模版数据选取

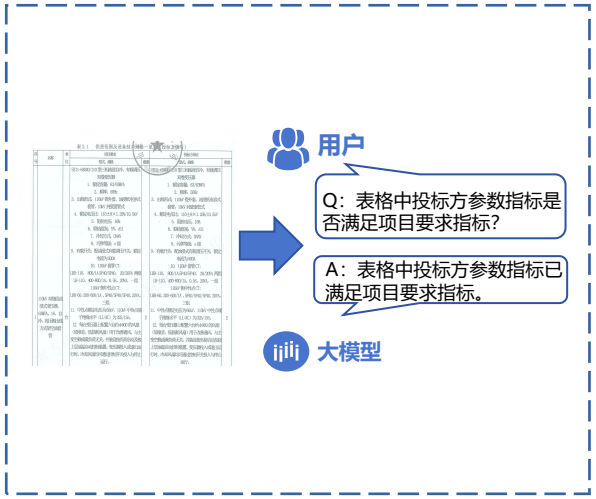
待识别技术文档



技术文档模版库



特定模版指标匹配数据集



抽取待测文档的结构和视觉特征进行文档的结构化识别

基于多级结构匹配技术进行模版匹配

选取特定模式的数据

研究内容:多模态大模型高效微调和持续学习

研究内容

针对技术规范文档的大模型高效微调任务设计（Parameter-Efficient Fine-Tuning）

技术路线

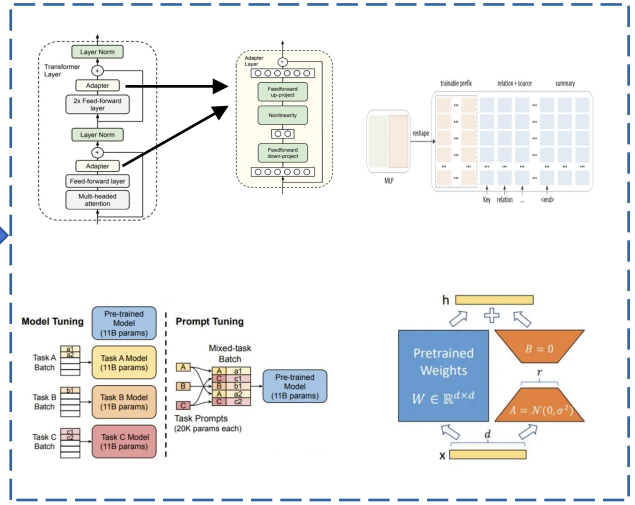
领域专用数据集构建

文档专用数据集采集

采集、标注用于微调的专用数据

大模型高效微调

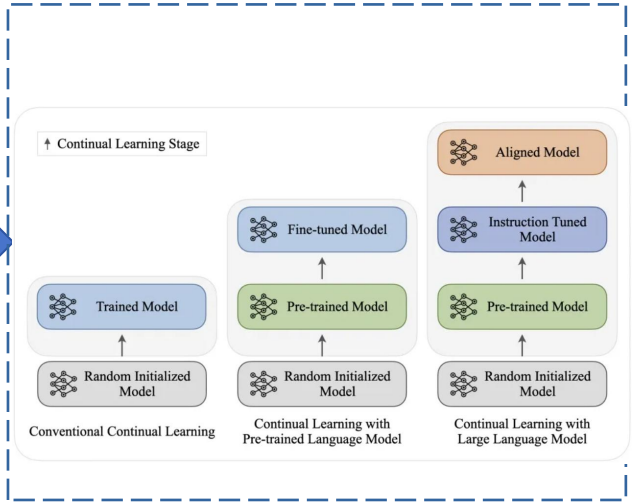
微调策略优化



选取合适的微调策略进行少量数据微调

多模态大模型持续学习

增量学习持续学习

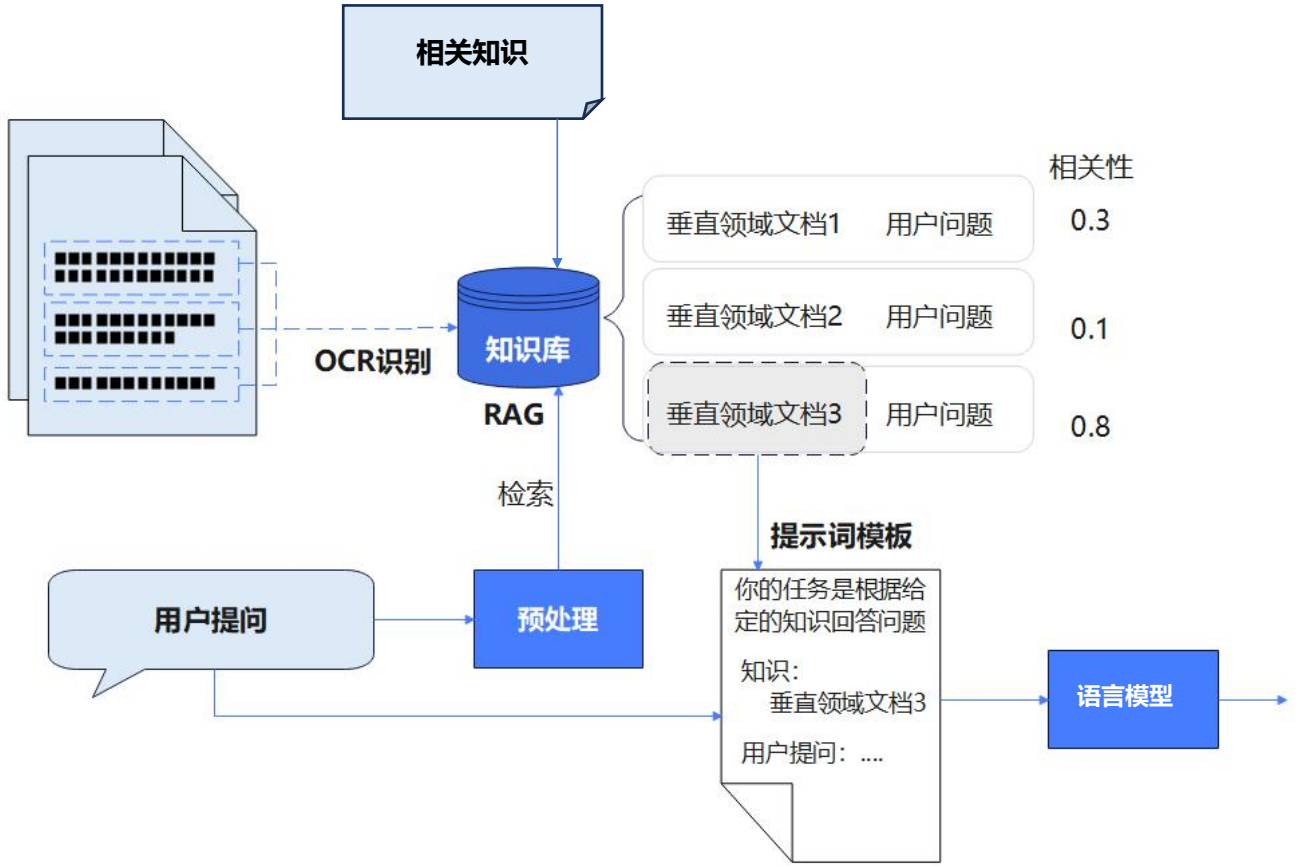


动态适应新知识并逐步更新自身知识

解决方案：基于大语言模型的本地知识库

一、整体设计

1. 使用专用OCR识别模型提取输入内容，从文档中的非结构化的文本与非标准结构的表格中识别与监造数据相关的知识，并转换为结构化的形式。
2. 利用OCR识别技术提取文档中的信息作为知识库；并把行业相关知识作为补充
3. 使用提示词(Prompt)工程和检索增强生成(RAG)技术，设计针对数据监造的专用提示词，从知识库中检索相关内容，使大模型得以结合相关内容对输入的数据进行审查。



- **零训练成本**：直接利用预训练大模型，节省时间和算力；
- **灵活性高**：仅需调整知识库和提示词即可适应文档和监造数据需求的变化；
- **效率提升**：通过知识库检索快速定位关键信息，减少大模型幻觉出现的概率。

阶段一：基于Deepseek的模型微调

模型微调：使用少量样本微调大模型

- 人工设计少量问答对，利用微调方法(Lora等)和大模型微调框架(Unsloth、Peft等)将Deepseek-R1模型微调为某领域的专家，强化大模型对需要审核的内容定位和审核推理能力，提升审查精度。根据审核内容定位和审核任务的逻辑难度，需要的样本从数百到数万不定。

- **领域精准性：**微调后模型对特定领域术语和逻辑的理解更深入。
- **可扩展性：**支持迭代优化，通过增加样本持续提升效果。

阶段二：基于Deepseek的本地知识库

一、文档内容识别

- 使用专用OCR识别模型提取输入内容，从文档中的非结构化的文本与非标准结构的表格中识别需要审查的各项指标，并转换为结构化的形式。

PDF文档

序号	参数类型	单位	正常使用条件	特殊使用条件	项目需求值或表述	投标人保证值	备注（须说明本工程适用的是正常使用条件或是特殊使用条件）
1	环境温度						
1.1	最高日温度	℃	45		45	45	正常使用条件
1.2	最低日温度	℃	-25		-25	-25	正常使用条件
1.3	最大日温差	℃	30		30	30	正常使用条件
1.4	年平均温度	℃	25		25	25	正常使用条件
2	海拔	m	≤1000m		≤1000m	≤1000m	正常使用条件
3	太阳辐射强度	W/cm2	0.1		0.1	0.1	正常使用条件
4	污秽等级			e级	e级	e级	特殊使用条件
5	覆冰厚度	mm	20		20	20	正常使用条件
6	风速	m/s	35		35	35	正常使用条件

OCR识别

HTML代码

序号	参数类型	单位	正常使用条件	特殊使用条件	项目需求值或表述	投标人保证值	备注（须说明本工程适用的是正常使用条件或是特殊使用条件）
1	环境温度						
1.1	最高日温度	℃	45		45	45	正常使用条件
1.2	最低日温度	℃	-25		-25	-25	正常使用条件
1.3	最大日温差	℃	30		30	30	正常使用条件
1.4	年平均温度	℃	25		25	25	正常使用条件
2	海拔	m	≤1000m		≤1000m	≤1000m	正常使用条件
3	太阳辐射强度	W/cm2	0.1		0.1	0.1	正常使用条件
4	污秽等级			e级	e级	e级	特殊使用条件
5	覆冰厚度	mm	20		20	20	正常使用条件
6	风速	m/s	35		35	35	正常使用条件

```
tr>
<td colspan="1" rowspan="1">1 </td>
<td colspan="7" rowspan="1">环境温度
</td>
</tr><tr>
<td colspan="1" rowspan="1">1.1 </td>
<td colspan="1" rowspan="1">最高日温度
</td>
<td colspan="1" rowspan="1">℃</td>
<td colspan="1" rowspan="1">45 </td>
<td colspan="1" rowspan="1"></td>
<td colspan="1" rowspan="1">45 </td>
<td colspan="1" rowspan="1">45 </td>
<td colspan="1" rowspan="1">正常使用条件
</td>
```

表格数据后处理

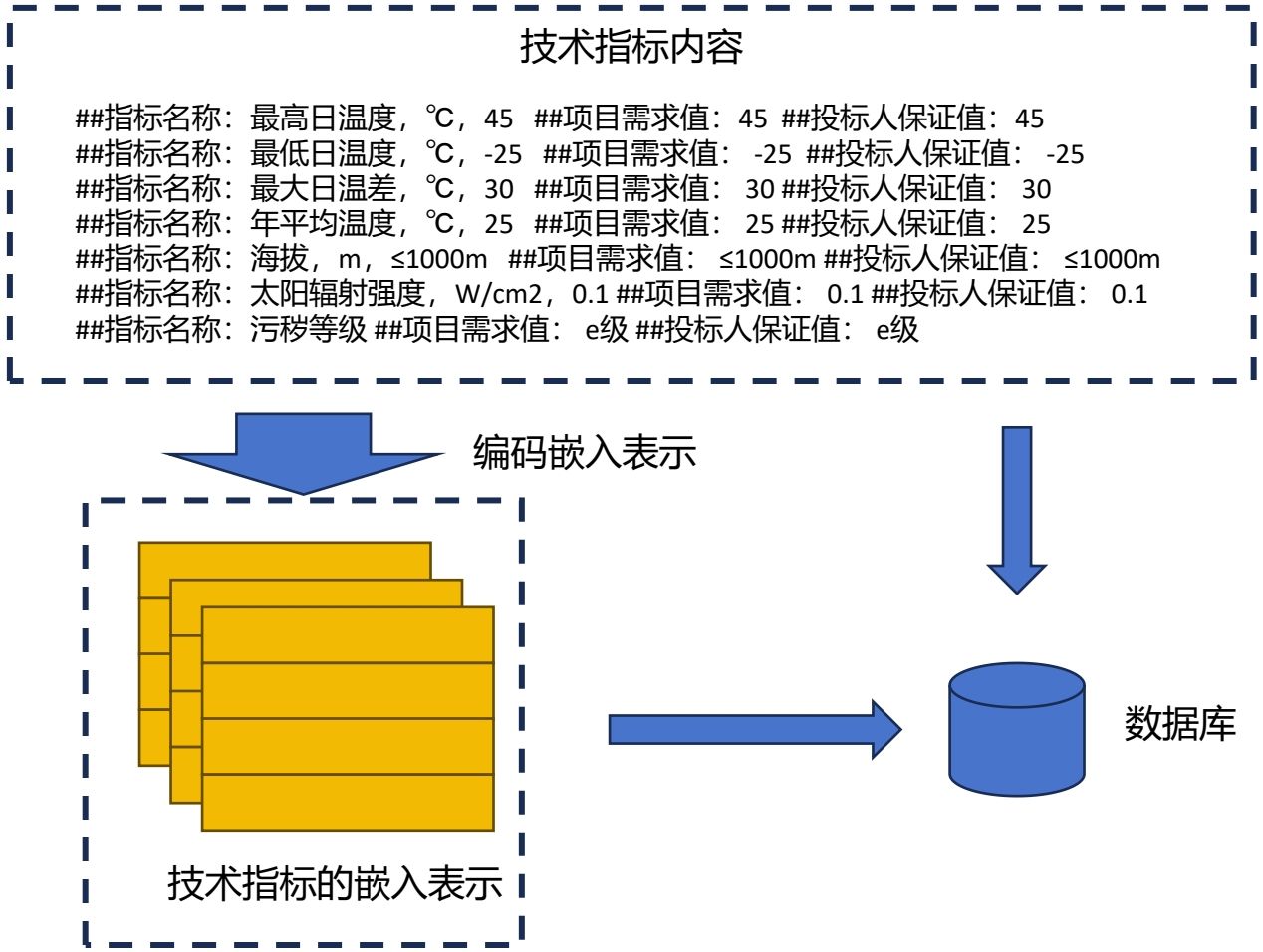
具体技术指标内容

##指标名称：最高日温度，℃，45 ##项目需求值：45 ##投标人保证值：45
##指标名称：最低日温度，℃，-25 ##项目需求值： -25 ##投标人保证值： -25
##指标名称：最大日温差，℃，30 ##项目需求值： 30 ##投标人保证值： 30
##指标名称：年平均温度，℃，25 ##项目需求值： 25 ##投标人保证值： 25
##指标名称：海拔，m，≤1000m ##项目需求值： ≤1000m ##投标人保证值： ≤1000m
##指标名称：太阳辐射强度，W/cm2，0.1 ##项目需求值： 0.1 ##投标人保证值： 0.1
##指标名称：污秽等级 ##项目需求值：_e级 ##投标人保证值：_e级

阶段二：基于Deepseek的本地知识库

二、构造本地知识库

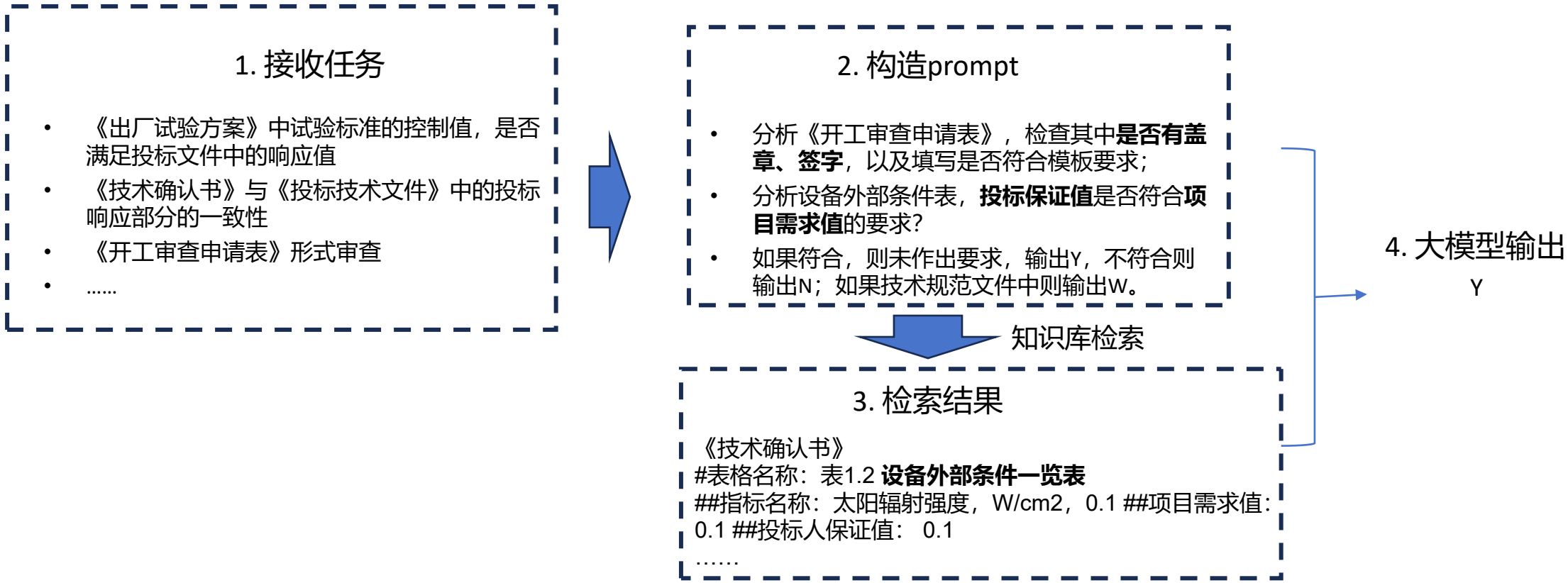
- 将上一步提取文档中的信息作为知识库；知识库中包含具体的技术指标条目及其经过大模型编码得到的嵌入表示；
- 将各条目及其对应的嵌入表示存储在数据库中，作为知识库，供后续进行检索。



阶段二：基于Deepseek的本地知识库

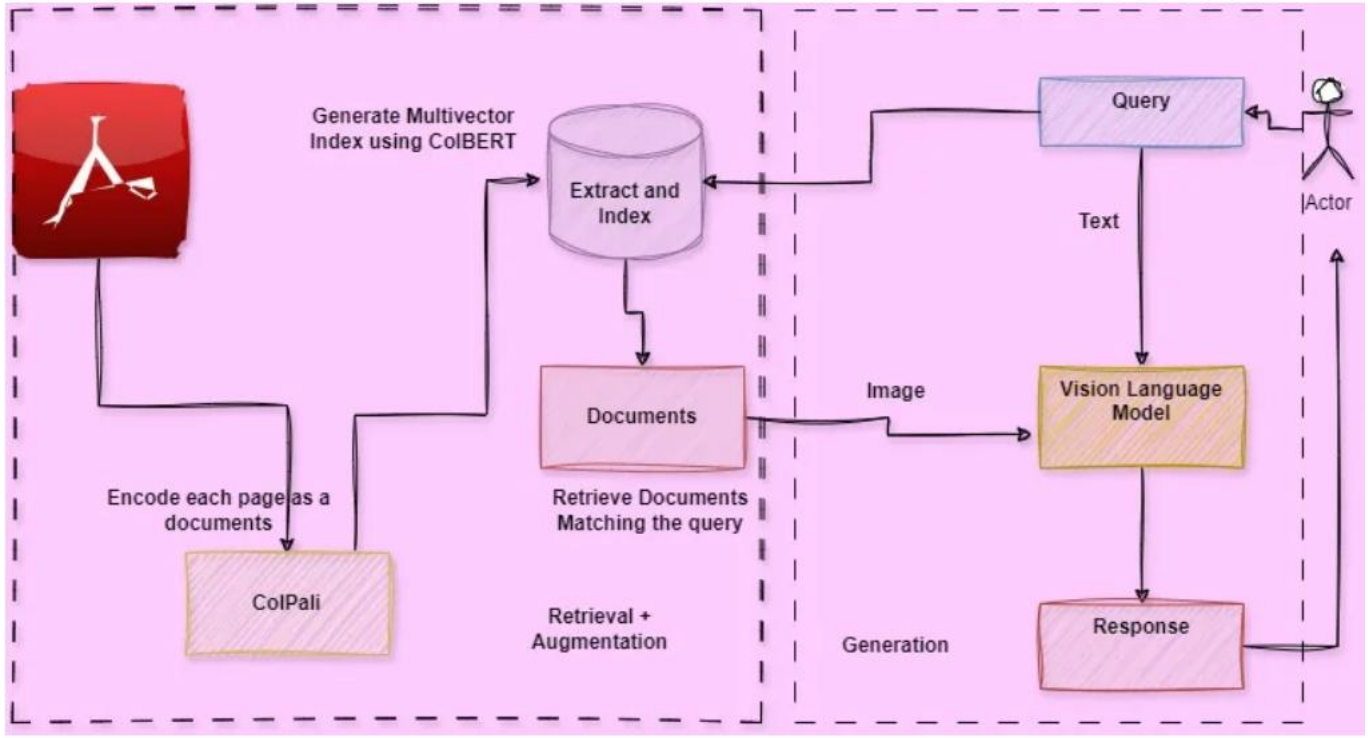
三、构造提示词，引导大模型进行预测

- 当需要进行审查时，根据**不同审查任务**构造prompt；将该prompt编码成嵌入表示，与知识库中的嵌入表示进行相似度计算，把**相似度较高的待审核指标**与多轮对话prompt同时交给大模型进行审查，得到指标是否符合要求的结果。



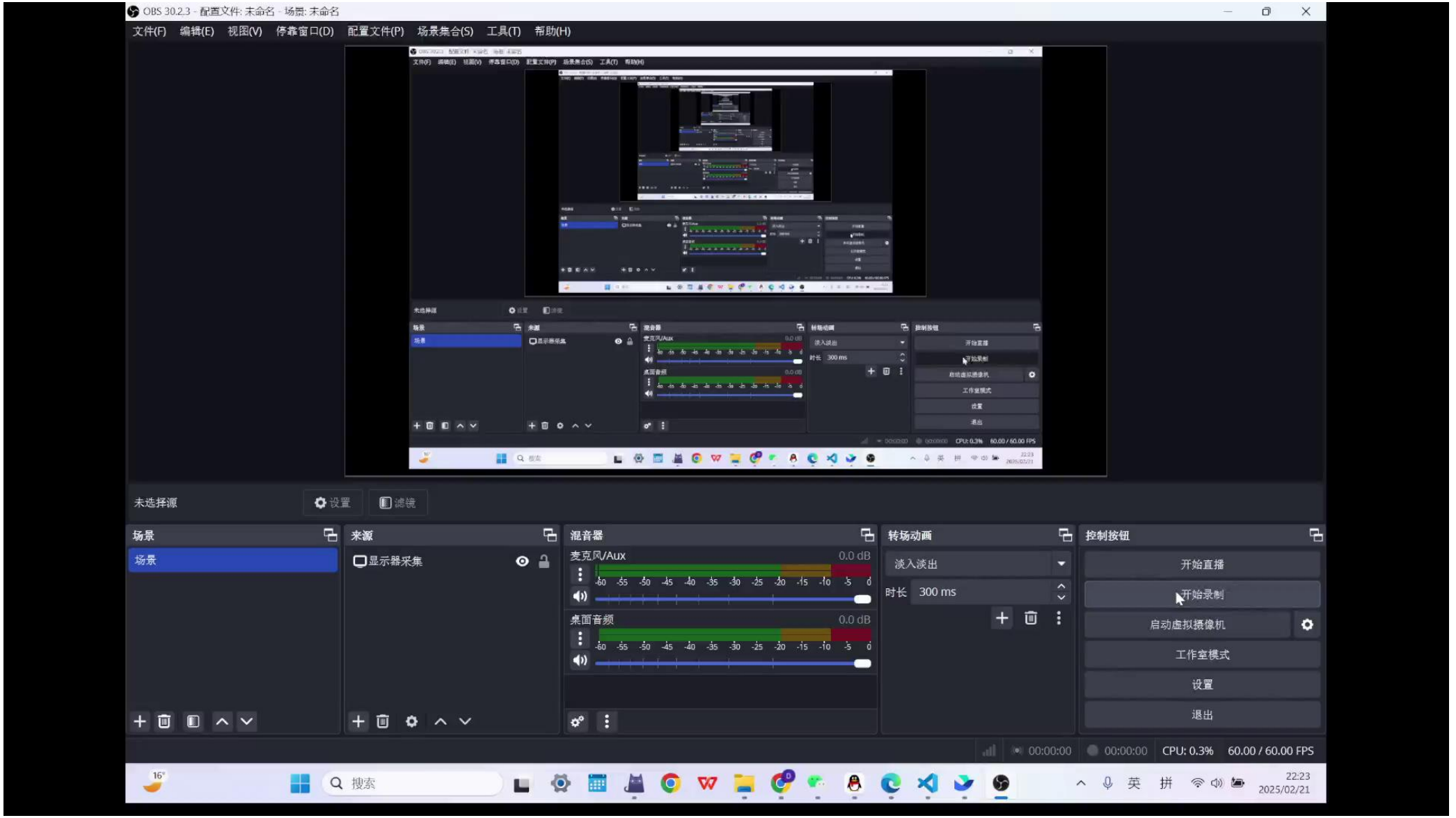
我的实践

前期探索：
1、试了一下用ocr+本地部署deepseek+本地知识库的这个流程 因为本地算力有限部署的deepseek-7B/30B的模型达不到预想的效果 且之前找的ocr模型识别效果不行
2、改成先用deepseek满血版的API（支持多模态输入）+本地知识库测试效果
3、最终决定使用ColPali + Qwen2.5 vl 7b来做



多模态RAG（检索增强生成）系统
1.文档编码和存储：
•使用ColPali对输入文档进行编码，生成嵌入向量
•将这些嵌入向量存储在向量数据库中建立索引
1.查询处理流程：
•用户查询也用ColPali进行编码
•用查询的嵌入向量从向量数据库中检索相似的文档片段(可能包含文本和图像)
•RAG系统将检索到的内容与用户查询结合,扩充上下文
•使用视觉语言模型(VLM)Qwen2.5_VL,根据增强后的上下文生成响应

Demo展示


OBS 30.2.3 - 配置文件: 未命名 - 场景: 未命名
文件(F) 编辑(E) 视图(V) 停靠窗口(D) 配置文件(P) 场景集合(S) 工具(T) 帮助(H)
未选择源
设置
滤镜
场景
场景
来源
显示器采集
混音器
麦克风/Aux 0.0 dB
桌面音频 0.0 dB
转场动画
淡入淡出
时长 300 ms
控制按钮
开始直播
开始录制
启动虚拟摄像机
工作室模式
设置
退出
00:00:00
00:00:00
CPU: 0.3%
60.00 / 60.00 FPS
搜索
英
拼
22:23
2025/02/21

感谢各位专家
敬请批评指正