9.22 第一次会议记录

一周任务安排：

* 下载power bi（1~2人）
* 这周六9.27商学院研讨室第二次会议讨论

任务1【刘颖，李秋逸】

**①选择清晰、结构化强的文档进行 OCR 识别，输出为文本。**

* 人工筛选质量好的图片进行样本识别
* 可以通过excel截图输入提高OCR识别率

**②使用赛题提供的OCR识别器，使用pycharm进行识别，输出结果后根据实际进行调试，调整参数后进行优化**

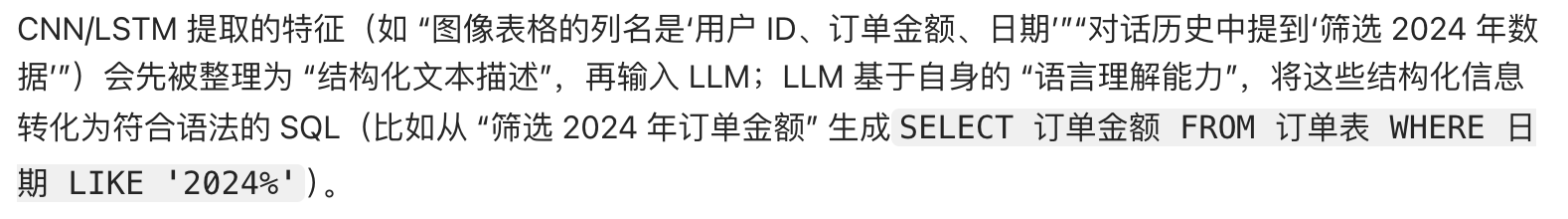
任务2【庹华艺，张泽慧】

**①在OCR识别工作完成之前，先寻找相关文本数据挑选重要特征，输出为JSON**

* JSON格式是一种轻量级的文本数据交换格式，核心目标是在不同系统（如前后端、不同编程语言）之间高效传递结构化数据。
* CNN：提取图像表单的布局特征，非文本数据转文本特征
* LSTM：处理时序数据，提取表单内文本的序列特征

**②将自然语言转化为SQL，在LLM中进行思维发散。**

* CNN/LSTM是“前置特征特区工具”，“转换为SQL”的核心环节通常由LLM完成。通常是CNN/LSTM 提取特征→将特征整合为 LLM 可理解的输入→由 LLM 完成“文本→SQL”的转换
* LLM的核心优势是处理文本序列，但对非文本数据或复杂时序文本的直接理解能力较弱，需要CNN/LSTM做前置转换。



任务3【梁紫荃，刘泽辉】：

①查找问题3“混合架构部署包&成本对比报告”相关文献，对相关模型形成初步了解

②大模型环境布置

③协助任务一、二