1月14日：

现在决定，将页面展示部分用web做展示，含有点击事件部分做web交互

Qwen调度方面，分为c，java，和py三块，java部分用于去调用py来拉到数据，模型的调度部分我想着是先使用onnx做即可，完成骁龙平台加速后再考虑其他。至于c部分，额还有待思考做什么·

先从token部分开始对接吧？今天简单梳理了一下，prompt还挺复杂的

3月24日

我靠，居然鸽了这么久？qwen的整个流程算是跑通了，也梳理完成了llm的整个piplines流，transformer感觉还是太过庞大了，摘要完的代码都比摘要笔记少，准备开始找合适的ocr识别框架，来为我提取图片中的数学公式