郭通

男 | 25岁

13283043057 | 2798815944@qq.com

教育背景

桂林电子科技大学　人工智能　硕士　GPA: 3.5　　2023-09 - 2026-06

重庆科技大学　计算机科学与技术　本科　GPA: 3.2　　2019-09 - 2023-06

校园经历： 1. 通过**CET4**，2. 大学期间获得多项奖学金，3. 竞赛获奖: 中国大学生计算机设计大赛**全国三等奖**，国家级；省级大学生创新训练计划项目(**负责人**)，省级；中国高校计算机大赛团体程序设计天梯赛**省级三等奖**，省级；第十二届蓝桥杯**省级二等奖**，省级；第十三届蓝桥杯**省级三等奖**，省级。

实习经历

流连电商　AI agent应用开发实习生　厦门　　2026-01 - 2026-03

- 参与公司AI Agent 应用与模型服务基础设施建设，**设计并落地统一的模型管理与推理框架**，支持模型注册、热加载/卸载、版本管理，为多个业务系统提供底层模型服务能力。
- 主导开发多模态发票审核 Agent，通过对比多模型效果及微调实验，基于**Qwen3-1.7B + Qwen3-VL-4B**模型使用 LangGraph 搭建多轮对话工作流，结合视觉语言模型、规则引擎与状态管理，实现发票识别、审核校验、异常协商与结果反馈的自动化闭环。
- 推动AI能力与业务系统集成，完成**部分API服务与OpenClaw及企业飞书接入**，支撑审核、通知等场景落地。

项目经历

企业级多模态发票审核智能 Agent 系统

技术栈： **OpenClaw • 飞书 • Qwen3-VL • Prompt Engineering • Lora • FastAPI**

项目描述： 面向企业采购场景中发票审核依赖人工、商品信息易错配与异常沟通成本高的问题，构建多模态发票审核智能体，实现**发票内容自动识别、结构化校验以及供应商侧异常问题的全流程自动化对话协商与审核闭环处理**。

- **模型选型及微调**：基于公司业务场景和服务承载能力，对多个开源模型及不同的模型组合策略进行对比试验和微调实验，**对比实验多家多模态模型及OCR模型（Qwen/GLM/DeepSeek/Paddle OCR）**，不同参数量（0.9B/2B/4B/8B），以及部分模型进行**lora微调**实验，最终选定Qwen3-1.7B作为对话主模型，负责意图识别及多轮对话， Qwen3-VL-4B作为发票识别及信息提取模型。
- **输入解析与意图路由决策**：在Agent入口对供应商消息进行意图识别，区分确认开票、暂时无法开票、价格异议、修改咨询与闲聊等场景。首先通过 LLM 进行意图分类，决定进入审核、协商或人工介入流程；当模型判断不稳定时，由**关键词兜底机制**接管，并结合当前状态阶段返回对应业务话术，避免误判和无效流转。
- **多模态识别与规则协同审核**：基于Qwen3-VL 对发票图片/PDF进行内容理解与字段抽取，识别商品名称、数量、单价及类目结构；同时设计**规则引擎**对商品、数量、单位等进行精确校验，实现 **LLM语义理解 + 规则精确校验**的混合审核架构，降低纯模型方案的幻觉风险。
- **LangGraph工作流与复杂状态管理**：基于 LangGraph设计发票审核 Agent 主流程(**单步执行 + 外部循环**)，实现多阶段多状态流转，并针对特殊异常情况可人工介入对话，支持多轮会话衔接与异常中断恢复。
- **缓存优化与容灾降级机制**：实现**意图识别缓存**，减少重复调用大模型带来的延迟与成本；针对大模型服务**异常、掉线**等情况，设计**关键词匹配 + 状态话术模板**的降级机制，保障审核流程连续性与系统可用性。
- **业务集成**：同时设计本地客户端与OpenClaw+飞书机器人双通道，实现多端任务处理，支持发票上传、审核结果查询、消息推送与人工协同处理，形成可落地的企业级智能审核系统。

AI 会议助手（基于LangGraph的多模态 Agent）

技术栈：**LangGraph • MCP • LLM • Prompt Engineering • RAG • PostgreSQL • Nginx**

项目描述： 面向企业会议纪要内容冗长、格式不统一与行动项易遗漏的问题，构建**多模态会议纪要智能体**，实现会议内容自动结构化与可执行输出。

- 输入解析与路由决策: 在 Agent 入口对用户输入进行**意图识别**，区分闲聊对话与会议纪要任务。首先通过 LLM 进行意图识别并给出置信判断，决定是否进入会议助手主流程；当模型判断不稳定时，启用规则兜底机制，通过关键词（如“会议”“讨论”“决定”或“hello/hi”等）进行降级路由，避免误触发会议生成或产生幻觉。
- 多工具协同执行（基于MCP）: 将**文档抽取、语音转文字与文档导出**等能力封装为**MCP**，由 Agent 根据当前状态进行组合调用，实现模型与工具解耦的执行流程。
- 上下文构建与长内容处理: 对**用户输入、ASR转写结果与文档解析内容进行统一聚合**；在内容超长场景下，由 **Agent触发内部RAG机制**，通过结合**语义与token计数的Chunk策略与滑动窗口**切分信息，结合**向量检索与重排序筛选**关键上下文，避免关键信息丢失被过度压缩。
- 会议类型驱动的生成与校验: 根据识别的会议类型选择对应的**差异化Prompt策略与结构化模板**生成会议纪要，并在生成后进行结构与可执行性校验（如行动项缺失、信息不明确等），必要时触发修复生成，降低幻觉风险。
- 状态与记忆管理: 实现**智能上下文组装与长期记忆管理**，通过**Token预算管理**与**历史记忆检索**，显著提升多轮对话质量。使用PostgreSQL持久化Agent会话、运行状态与生成产物，支持LangGraph Checkpoint，实现多轮会话与中断恢复。设计智能上下文管理工程。