李婧漪

lijingyi030505@163.com
18108658857

教育背景

上海财经大学 (211 工程)，统计与数据科学院，应用统计学，硕士 2025.9 - 2027.6

中南财经政法大学 (211 工程)，统计与数学学院，应用统计学，本科 2021.9 - 2025.6

- **主修课程**：数据结构，算法设计与分析，机器学习，神经网络，数据库系统原理，时间序列分析
- GPA：3.9/4.0，校一等奖学金 (前 5%)，校优秀学生

实习经历

众安保险，算法实习生 2025.4 - 2025.8

- **Deepseek-r1+Llama Factory 微调的保险风控应验**：
  (1 **项目介绍**：为查询不同保单人的保险查询是否存在骗保风险，针对包含身份信息，保险咨询内容和理赔信息情况，基于保险理赔真实场景，根据问保人的对话实现业务侧需要的风险诊断模型。
  (2 **具体实现**：针对问答（QA）数据，采用 Llama Factory 训练框架通过低秩适应（LoRA）方法对模型进行 SFT 微调。随后，再利用包含金融咨询数据的数据集（PairWise），采用直接偏好优化（DPO）算法对模型做更进一步的优化处理，最后模型增强了对常规咨询，幻觉的误判，提升了业务侧的筛选判断速度。
- **车险定损部门智能图像增强与识别**：
  (1 **图像增亮功能上线**：针对车险部门需要，协助复现 HVI(cvpr2025,sota) 图像增亮模型并应用于车辆领域，上线夜视拍摄增亮功能，相较于之前的黑暗场景图像有显著的亮度提升。
  (2 **车辆定损识别分类**：为实现车辆损伤部位的实例分割，训练复现车损领域 Co-DETR(iccv2023,sota) 对受损最多的保险杠、车门、翼子板进行分类。

项目经历

GLM 增强的 RAG 运维问答系统设计 (CCF-AIOps 挑战赛) 2025.7 - 2025.9

- **项目介绍**：面向比赛提供的网络运维私有化文档 (含 HTML/图像/XML, 等异构数据), 构建高效 RAG 系统, 采用 GLM-4 作为 LLMs, 通过分析文档的树状结构作为知识路径来优化检索和重排, 相较于初始不使用知识路径优化和排序算法的基础模型在 NLG 指标上 (Rouge,Bleu 等) 提升了 **15%** 左右。
- **检索前优化**：通过分析文档的树状结构作为知识路径, 结合 OCR 与 GLM4V-9B 提取图像语义，利用文档树状结构与自定义设计 SentenceSplitter 进行纯文本稳定分块，增强图文表示。
- **检索中优化**：使用 Hyde 实现虚拟文档和查询关键词扩展，粗排阶段利用 BM25 稀疏检索 (文档/路径维度) 和 gte-Qwen2-7B 密集检索进行两路稀疏检索粗排, 在重排阶段利用 bge-reranker-v2-minicpm 重排模型利用倒数排序融合进行重排, 提高检索内容的可信度。
- **答案优化生成**：重排得到的 top k 文本块与图像的内容拼接，设计多种 prompt（重视 top1，思维链侧重问答模版）优化回答。

LayoutXLM 驱动的银行贷款文本自动化 OCR 解析 2025.4 – 2025.8

- **项目介绍**：针对银行月结单在业务中存在的模板差异大、图文混排复杂、易被篡改等问题，构建“文档结构识别与防伪检测”系统。项目以 PDF 格式的月结单作为输入，设计涵盖图像 OCR、表格结构提取、文本语义分类与图像防伪的多阶段处理流程，输出结构化字段（如账户、金额、交易明细）及安全校验结果。文档整体通过率 **90%+**。
- **文档要素检测与解析**：使用 logo 检测识别 (YOLO)、文字检测识别 (DBNet)、二维码检测、图像方向校正等方法，结合 OCR 与 PDF 要素解析，实现对文档中多模态要素的精准识别与提取。
- **文档结构化信息提取**：使用 LayoutXLM 模型融合文字、位置和图像三类信息进行文档实体识别，通过统一标签体系、多模板联合训练和银行分流式后处理，使其能高效适配不同银行月结单模板，输出标准化结构化结果。
- **表格线条结构检测**：通过 Cycle-CenterNet 模型以中心点及角点联合回归的方式检测单元格位置，结合业务场景调整多阈值、关键点聚合、宽度过滤、定制 NMS 等后处理参数，提取月结单中复杂的图文表格结构。
- **单元格归纳与分类**：基于 GFTE 框架，构建单元格图并引入图卷积网络（GCN）进行结构归纳，实现跨单元格和嵌套结构的自动识别。同时，引入 ALBERT 提取文本语义向量，结合空间特征增强节点表。

学术成果

- Zhao, Lufei.[1]; **Li, Jingyi.**[2]; et al. Prediction of protein secondary structure by the improved TCN-BiLSTM-MHA model with knowledge distillation. *Scientific Reports.* **2024**, 14, 16488. (JCR Q2, IF=4.6)
- Yang, Xinyi.[*]; **Li, Jingyi.**[*]; et al. Research on information leakage in time series prediction based on empirical mode decomposition. *Scientific Reports.* **2024**, 14, 28362. (JCR Q2, IF=4.6) ([*]Co-first authors)

相关技能

- 熟悉 python/C++/SQL 等语言，了解 shell 脚本工具语言
- 熟悉 Linux 开发环境及 Docker，Git 等开发工具，有大模型微调和部署经验