

# 周凯

Tel: 15995397689

Email: [kaizhou0305@gmail.com](mailto:kaizhou0305@gmail.com)

WeChat: zk471104594

Blog: [zhoukai.space](http://zhoukai.space)



## 教育经历

吉林大学	人工智能专业硕士	吉林/长春	2025.9 - 2028.7
GPA: 3.4/4.0   院奖学金			
南京信息工程大学	数据科学与大数据技术本科	江苏/南京	2021.9 - 2025.7
GPA: 3.4/5.0   校优秀学生干部, 国家励志奖学金, 校奖学金等			

## 项目经历

### 多模态专利文档智能问答系统

**项目简介:** 基于阿里云天池竞赛数据集构建的多模态专利文档问答系统, 通过检索增强 & 大模型微调实现专利文档的智能理解与精准回答。

**技术栈:** Transformers、PyMuPDF、TRL、PEFT、LoRA、Qwen、PyTorch、Datasets、RLHF、vLLM

- 1) 多模态数据预处理:** 使用 PyMuPDF 将专利 PDF 转换为高分辨率图像 (600 DPI), 基于 GME 模型提取 3584 维向量表示, 构建问题、图像的向量数据库。
- 2) 智能检索系统:** 问题 & 图像的跨模态语义检索, 基于余弦相似度精准定位相关文档页面 (Top-K=2), 根据问题类型动态调整图像输入组合。
- 3) 模型微调:** 构建差异化训练数据集 (核心专利问答数据 + 少量问题分类辅助数据), 缓解模型过拟合风险。基于 Qwen3-VL-32B 多模态大模型, 采用 LoRA 技术完成参数高效微调, 结合梯度累积、学习率调度策略优化训练流程, 显著提升模型在专利问答任务中的响应准确率与鲁棒性。
- 4) 差异化推理策略:** 搭建问题分类模块, 精准判别纯图像位置关系 & 推理混合问题类型; 定制专属提示词, 通过差异化输入引导策略提升初答准确性。
- 5) 答案优化机制:** 设计“初答 + 风格迁移”两阶段优化流程, 基于 Few-Shot 范式构建风格示例库 (检索相似问题答案作为风格参考), 通过后处理模块提取核心表达风格, 实现答案质量与一致性的双重提升。
- 6) 性能提升:** 通过微调 & 检索优化 & 后处理, BERTScore 的 F1 从 0.6244 提升至 0.8234, EM 从 6.6% 提升至 23.27%, ROUGE-1 的 F1 从 0.2517 提升至 0.5493。

### 企业知识问答 Agent 系统

**项目简介:** 独立设计并开发企业文档问答系统, 支持上传 PDF, 构建向量索引并调用大语言模型回答问题。

**技术栈:** LangChain、Qwen、FAISS、HuggingFace、Transformers、Gradio

- 1) 使用 LangChain 框架构建 RAG 管线, 采用 Qwen-7B-Chat 实现问答生成。
- 2) 利用 FAISS 进行向量检索, 结合 HuggingFace 中文嵌入模型优化搜索准确性。
- 3) 搭建 FastAPI 接口服务及 Streamlit 前端, 支持用户实时上传、提问、查看回答与引用来源。

## 技能/证书/竞赛

**证书:** CET4、CET6

**语言:** 精通 Python 开发

**技术:** 熟悉 Transformer、BERT、QwenVL 等主流大语言模型原理与架构; 熟练运用 LoRA 完成大模型参数高效微调; 了解 DPO、PPO、GRPO 等主流 RL 对齐算法

**应用:** 熟悉 RAG 检索增强技术及主流优化方法; 掌握 Agent 开发、工具调用全流程