鲍凡

意向岗位：AI 工程师

出生日期：1960.05

籍贯：湖北省襄樊市

工作年限：3 年

电话：15201401579

邮箱：3dpme@live.com

兴趣爱好

编程、看电影、音乐

教育背景

清华大学　　计算机科学与技术硕士（人工智能方向）　　2021.09–2024.06

工作经历

某头部 AI 科技公司　　AI 工程师（实习）　　2022.07 – 2023.07

- 参与 AI 中台开发，主导 3 个企业级 RAG 项目落地，客户留存率提升 25%
- 推动 LangChain 与内部工具链整合，团队开发效率提升 30%

项目经历

基于 RAG 的智能问答系统（独立开发）　　2023.03 – 2024.02

1. 使用 LangChain 构建检索流程，集成 Milvus 向量数据库，优化 Embedding 模型（BAAI/bge），将响应准确率从 68%提升至 89%
2. 设计动态 Prompt 模板，结合用户上下文实现多轮对话，推理延迟降低 40%（GPU Triton 部署）

多模态文档处理系统（团队核心开发）　　2022.06 – 2023.01

1. 搭建文档解析后端，集成 OCR（PaddleOCR）与 NLP 模型（LayoutLM），实现 PDF/表格结构化提取（F1=92%）
2. 基于 LlamaIndex 构建索引，支持语义检索与自动摘要生成，效率提升 50%
3. 技术方案入选公司国产化替代白皮书

大模型垂直领域微调（主导）　　2023.09 – 2023.12

1. 使用 LoRA 对 Llama2-7B 进行医疗领域微调，通过领域语料增强与奖励模型（PPO）优化，评测指标提升 35%
2. 部署于华为昇腾 910 集群，推理成本降低 60%

专业技能

- AI 开发工具：LangChain、LlamaIndex、OpenAI/Claude API，熟悉 Prompt Engineering 最佳实践
- 算法与框架：PyTorch/TensorFlow，精通 RAG、微调（LoRA）、向量数据库（Milvus/Pinecone）
- 工程能力：AI Agent 系统设计、Docker/Kubernetes 部署、AWS/Azure 云服务、国产化环境适配（华为昇腾）
- 模型优化：设计模型评估框架（准确率/延迟/鲁棒性）、构建监控系统（Prometheus/Grafana）
- 语言：英语 CET-6（流利技术文档读写），熟练使用 Markdown/Sphinx 编写技术文档