小明

性别 · 年龄 · 籍贯

电话 · 意向岗位 · 邮箱

教育背景

xx 大学, 计算机科学与技术, *硕士* 2024.09 - 2027.06

xx 大学, 信息与计算科学, *学士* 2020.09 - 2024.06

技术特长

- PyTorch、Python、CUDA
- 大模型训练：熟悉 DeepSpeed 分布式训练框架，掌握 LoRA 微调方法，具备混合精度训练及显存优化实践经验
- 推理优化：使用 vLLM 和 Triton 构建高性能推理服务，开展过并发测试与性能调优
- NLP 技术栈：深入理解 Transformer 架构，实践过 Prompt Engineering 与 RAG 相关技术，熟悉对话系统构建与生成模型评估方法

项目经历

大规模中文情感分析系统, 角色: 负责人 2024.10 - 2025.03

- 负责搭建电商和社交媒体场景的中文情感分析系统。
- 基于 RoBERTa-wwm 训练情感分类模型，并加入领域词表提升效果。
- 尝试使用 Qwen 进行 prompt 和 LoRA 微调，以改进长文本与隐含情绪的识别能力。
- 将大模型能力蒸馏到小模型以提升推理速度，并将系统部署在 Triton 与 vLLM 上完成基础测试。
- 最终模型效果与速度较原方案均有提升。

智能客服意图识别与多轮对话系统, 角色: 核心成员 2025.04 - 2025.09

- 参与线上教育平台的智能客服系统开发，目标是实现多轮对话、意图识别和任务执行。
- 参与意图识别和知识问答部分的设计，尝试加入 RAG 来提升问答表现。
- 使用 LLaMA 的微调方法改进指令执行效果，使用工具优化显存和训练速度。
- 系统加入对话记忆和拒答策略，整体多轮对话的成功率和意图识别表现明显提升，并在实际平台稳定运行。

荣誉证书

- 2024 年获全国人工智能创新竞赛一等奖
- 2025 年获研究生国家奖学金