小明

男 · 年龄 · 籍贯

电话 · 邮箱 · 意向岗位

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

技能特长

在学习和项目中接触过一些语言模型相关的内容，做过从模型训练到上线的流程，对常见的 NLP 做法都有一定了解，包括模型微调、对话任务以及一些生成相关的评估方法。对 Transformer 也比较熟悉，在处理长文本和情感方面做过一些 Prompt 调整的尝试，也用过一些轻量化的方法让模型变得更小更快。主要使用 PyTorch 进行开发，了解基本的训练方式，也用到过混合精度和一些训练加速的技巧，能把训练流程整理得比较清楚。在训练大模型时也尝试过分布式和一些显存优化的工具，比如 DeepSpeed 之类，对显存占用和通信的影响有过一定认识。在部署方面，用过 vLLM 和 Triton 来搭建推理服务，也做过简单的性能调试和并发测试。日常开发主要在 Linux 和 Python 环境下，能使用 Hugging Face 的工具处理模型和数据，也会做一些基础的 CUDA 性能排查。

项目经历

大规模中文情感分析系统 2024.10 - 2025.03

负责人

- 在该项目中，我负责搭建电商和社交媒体场景的中文情感分析系统
- 先基于 RoBERTa-wwm 训练情感分类模型，并加入领域词表提升效果
- 之后尝试使用 Qwen 进行 prompt 和 LoRA 微调，以改进长文本与隐含情绪的识别能力
- 为满足上线需求，我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 2025.04 - 2025.09

核心成员

- 在这个项目中，我作为核心成员参与了线上教育平台的智能客服系统开发，主要目标是实现多轮对话、意图识别和任务执行
- 我在项目中参与了意图识别和知识问答部分的设计，也尝试加入 RAG 来提升问答表现
- 模型方面使用了 LLaMA 的微调方法来改进指令执行效果
- 在训练过程中使用了一些工具优化显存和训练速度
- 系统加入了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都有明显提升，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