林子杨

📱 **电话**：13537685397 | ✉️ **邮箱**：ziyanglin1997@163.com | 💼 **领英**：www.linkedin.com/in/ziyang-lin | 💬 **微信**：zi24yang10la8 | 🌐 **个人网站**：https://ziyanglin.netlify.app | 📍 **地点**：深圳

AI软件开发工程师

个人总结

- 谢菲尔德大学计算机科学与人工智能专业本科毕业生（2022年7月本科毕业）
- 在2019-2020学年以一等学位成绩完成大二课程
- 在2020-2021学年，gap去学习更多与自然语言处理相关的技术和数学知识
- 大三在导师 Dr Anton Ragni 的指导下完成"基于深度学习的自动语音识别"相关的毕业设计项目
- 于2022年11月至2024年3月任职软件开发工程师于深圳市计量质量检测研究院技术研发中心
- 2024年3月至今担任AI软件开发工程师于深圳市智软通科技有限公司，专注于RAG技术应用开发和实时多模态模型应用开发

教育经历

英国谢菲尔德大学

计算机科学与人工智能 本科 计算机科学学院

2018年09月 - 2022年07月

英国 谢菲尔德

研究型毕业设计：

GLOM for Automatic Speech Recognition - 导师: Dr Anton Ragni (在自动语音识别领域，参考Hinton提出的GLOM模型，设计开发拥有更佳可解释性的表征学习模型，并在学术的高性能计算集群节点上进行训练；导师的最终反馈设计语可点击此处查看)

工作经历

深圳市智软通科技有限公司

AI软件开发工程师

2024年03月18日 - 至今

深圳市

- 负责基于检索增强生成（RAG）技术的智能体系统设计与开发，主要利用Milvus和Weaviate向量数据库构建高效知识检索系统
  - 设计并实现了混合检索策略，结合稀疏检索(BM25)和密集检索(向量相似度)提升检索准确率
  - 优化了文档分块策略，实现了基于语义的自适应分块，提高了检索结果的相关性
  - 实现了多级缓存机制，减少重复查询的计算开销，提升系统响应速度
- 参与高级Prompt Engineering实践，提升LLM应用效果
  - 设计并实施了基于Chain-of-Thought和ReAct框架的复杂推理提示策略
  - 开发了自动化提示优化系统，通过A/B测试评估不同提示模板的效果
  - 实现了动态提示模板库，根据用户输入和历史交互自适应选择最佳提示策略
- 成功开发并部署金融数据分析智能体，实现了对复杂金融数据的自动化分析与洞察提取
  - 设计了多阶段分析流水线，包括数据清洗、异常检测、趋势分析和预测建模
  - 实现了自定义DSL (Domain Specific Language)，使金融分析师能够通过自然语言查询复杂数据
- 设计并实现政务问答智能体，优化政务信息检索效率，提升用户查询体验
  - 构建了政务领域知识图谱，增强了实体关系理解和复杂查询处理能力
  - 实现了多轮对话管理系统，支持上下文理解和澄清问题的交互式对话
- 参与公司自研**多模态低延迟实时语音对话智能体**，集成语音识别、自然语言处理与语音合成技术
  - 优化了端到端语音生成模型至300ms
  - 实现了实时流式处理架构，支持边说边理解的交互模式
  - 设计了情感识别模块，根据用户语音情绪调整回复策略
  - 利用llama.cpp高效推理引擎部署GGUF格式的14B参数多模态语音模型，实现在**单张RTX 3090 GPU**上的低资源高性能推理，优化内存占用和计算效率
  - **全流程实时延迟500ms**，并提供工具调用和复杂提示词能力，优化部署至2xA100即可部署非量化模型 并实现基座模型的任意替换
- 参与基于vLLM部署优化与系统架构设计
  - 实现了模型量化(A8W8 A4W4)技术，在保持性能的同时显著减少模型大小
  - 实现了高效低延迟推理服务架构，支持模型并行和张量并行计算
  - 实现了动态批处理和请求排队机制，提高了系统吞吐量
  - 实现了模型热加载和缓存策略，优化了多模型切换场景下的资源利用
  - 利用llama.cpp优化本地推理性能，实现了高效的模型量化和内存管理方案
  - 基于ollama构建轻量级模型服务，支持多种开源模型的快速部署和API调用
- 参与多代CAAS平台的开发，完成其中的·Mxgent·SOP Agent 的优化
  - 参与研发公司对话管理系统，支持复杂多轮对话流程
  - 参与公司意图识别Agent的核心实现，提高了对话系统的鲁棒性和灵活性
  - 参与自研工具rpa模块的设计和实现，实现低参数模型正确调用工具的可能 并对 function calling 和 mcp 协议兼容

深圳市计量质量检测研究院

软件开发工程师 技术研发中心
2022年11月 - 2024年03月
深圳市

- 主导电动汽车充电桩检定平台网站开发项目，基于XAMPP（Apache + MySQL + PHP + PERL）技术栈，实现了检测数据的采集、存储与可视化分析功能
- 负责CT在线检测电子元器件项目的核心开发工作：
  - 设计并实现CT扫描图像数据采集与预处理流程
  - 基于百度PaddlePaddle框架进行深度学习模型选型与二次开发
  - 优化模型训练流程，提升缺陷检测准确率
  - 完成模型部署与系统集成
- 熟练应用计算机视觉相关技术，包括OpenCV库进行图像处理与分析
- 开发工业相机应用程序，整合Basler与海康威视工业相机SDK
- 参与基于Qt框架的跨平台应用程序开发
- 技术栈：C++、Python、PHP、MySQL、PaddlePaddle、OpenCV、Qt

项目经历

英国伦敦帝国理工学院暑期自然语言处理课程+项目

暑期在线研究学生
2020年06月 - 2020年08月

- 参加由伦敦帝国理工学院教授Lucia Specia指导的2020年NLP在线研究夏季项目
- 该项目旨在开发回归和分类模型，以评估编辑过的新闻标题(英文)的幽默度
- 关于我为此项目已完成的工作的具体细节，请随时前往我的个人主页或我的Github主页查看
- 教授Lucia Specia给我的推荐信链接：Reference Letter

社团和组织经历

谢菲尔德大学人工智能社团

Event Officer
2020年09月 - 2021年07月
英国 谢菲尔德

- 被选为AI社团2020-2021学年的Event Officer
- Event Officer在学期中负责主持开展和人工智能相关的学习项目(策划workshop或校友讲座)

专业知识

- **大语言模型与应用开发**：深入理解Transformer架构、注意力机制、自监督学习和微调技术
- **Prompt Engineering**：掌握高级提示工程技术，包括Few-shot Learning、Chain-of-Thought、ReAct框架和提示模板设计
- **LLM部署与优化**：精通模型量化、参数高效微调(PEFT)和推理加速技术，熟悉llama.cpp高效推理库的量化技术和内存优化方案，掌握ollama本地模型管理与部署框架的容器化应用
- **检索增强生成(RAG)**：熟悉向量检索、混合检索策略、文档分块和上下文优化技术
- **多模态学习**：理解跨模态信息融合、多模态表示学习和对齐技术
- **机器学习基础**：深度学习、强化学习、自然语言处理、计算机视觉
- **数学基础**：线性代数、多元微积分、概率论、信息论

专业技能

- **编程语言**：Python, C++, JavaScript
- **深度学习框架**：PyTorch, TensorFlow, JAX, PaddlePaddle
- **LLM工具链**：LangChain, DSPy, LlamaIndex, PEFT, Transformers
- **向量数据库**：Milvus, Weaviate, FAISS, Pinecone
- **模型部署**：ONNX Runtime, TensorRT, vLLM, DeepSpeed, Triton, llama.cpp, ollama
- **语音处理**：SpeechBrain, Whisper, ESPnet, Kaldi, FFmpeg, librosa, Wav2Vec2, HuBERT, pyAudioAnalysis
- **音频处理**：WebRTC, DTMF, VAD (Voice Activity Detection), 音频增强
- **开发工具**：Docker, Git, Linux, Ray
- **计算机视觉**：OpenCV, Pillow, torchvision
- **Web开发**：FastAPI, Flask, React

语言

普通话（母语），英语（流利）