郭志阳

出生年月: 1998.06.14　　电话: 17798545843
邮箱：17798545843@163.com　　意向岗位: 大模型算法工程师

教育背景

2023.09-2024.11　　纽卡斯尔大学　　通信与信号处理|硕士

- **主修课程：智能信号处理，信息论，信号调制与处理，机器学习**
- **毕业设计：基于深度学习的 16-QAM-MIMO-OFDM 联合信道估计系统**

2017.09-2022.06　　南京邮电大学　　材料物理|本科

专业技能

- **1.大模型的训练及微调**
  - 熟悉主流**大语言模型架构**以及各种工业界常用的预训练模型包括各类 **Bert-like** 模型及其下游任务应用，具备**模型选型与架构对比**分析能力。
  - 具备完整的大模型训练与对齐实践经验，包括**预训练，监督微调，**以及 **RLHF/DPO** 的落地经验。近年来研究方向为大语言模型的底层实现与应用，深度掌握相关开发与部署的全流程技术栈，同时掌握自然语言处理主流技术。
  - 熟悉低资源下训练模型时常用的 **peft，LoRA，QLoRA，模型量化，混合精度训练技术**，以及对模型输出的后处理。
  - 熟悉以 **DeepSpeed** 为主的分布式训练框架，熟悉数据质量管理与数据配比对性能的影响，有经验缓解微调中出现指令丢失以及灾难性遗忘的问题

- **2.模型的推理与优化**
  - 熟悉**模型压缩**技术，掌握模型的**蒸馏，量化**等方法，能根据业务需求对模型的大小与性能进行权衡与调整，提升在不同硬件环境下的可部署性。
  - 具备模型**推理优化**经验，可以使用 **tensorRT**，**ONNX Runtime** 对模型的推理过程进行加速，以满足高并发场景下以及一些在线任务的 QPS 要求。

- **3.检索增强与知识库（RAG)**
  - 熟练使用 **milvus，faiss** 等向量数据库，可以完成 embedding 模型的微调与训练以及多路混合召回等任务。
  - 掌握**意图识别，query 改写，多路召回，粗排精排，**以及 **RAG 提示词工程**等工作，对 RAG 架构设计，性能优化，与落地应用有深刻理解。
  - 拥有基于 **neo4j 图数据**库搭建 **GraphDB** 并与大模型结合实现检索增强生成的经验，提升模型在专业垂直领域的回答准确性与可控性。

- **4.Agent 开发及应用**
  - 熟悉基于 langchain 框架的智能 Agent 开发流程，包括**工具调用，记忆机制，链式调用**等模块的设计与实现，能够灵活构建面向特定业务场景的智能助手。
  - 具有将 langchain 与大模型，embedding 模型，向量数据库，图数据库，缓存数据库深度结合的经验。

工作经历

2024.10-2025.08　　临床医疗辅助问答系统　　杭州明阳计算机技术有限公司

- **项目简介：**

该项目主要基于 **qwen2.5-72B** 大模型，为当地某公立医院构建智能问答系统。在充分融合医院现有信息系统的基础上，项目实现了 PC 端与移动端的多端协同，为普通患者与医疗专业人员提供便捷，智能的医疗咨询入口。在技术路线方面，总体上采用”**大模型对齐优化 + 检索增强（RAG）**”的架构设计，兼顾医学专业性，用户可理解性与合规性，为智慧医疗落地提供了有效解决方案。

- **工作内容**

1．搭建前置文本纠错模块：从医疗专用语料（如专业药品，疾病学名）中抽取核心实体，同时通过对比学习的思想训练出适用于医疗垂类的 SimCSE 模型用于专业名词纠错，同时利用合规用户 QA 语料，训练通用文本纠错模型，实现医疗垂类与通用场景的双重覆盖。

2．构建训练数据：针对 **RLHF 三个阶段**对于数据的不同需求，分别设计并构造了对应的数据集，确保模型在微调与对其过程中实现稳定收敛与持续性能提升。通过高质量数据构建，有效提升了模型在医疗场景下的回答准确性和安全性，保证了 RLHF 优化过程的顺利实施。

3．参与模型微调：基于 qwen2.5-72B 的基座模型，在 **DeepSpeed** 的框架下完成了 RLHF 三阶段训练，全流程使用 **LoRA** 微调，提升模型在医疗问答中的专业性与回答一致性。同时在 **LLama-Factory** 的框架下进行 DPO 优化，与 PPO 形成对比的同时，可以实现在资源受限的环境下的部署，确保生成准确，专业，可控，便于快速落地的医疗场景应用。

4．参与搭建 RAG 系统：基于 bce-embedding-base 模型完成对目标语料的向量化表示，并利用 milvus 向量数据库构建文本向量索引，实现快速，精确的语义检索。同时使用了 multi-query，BM25 等召回优化策略，并使用 bce-reranker-base 的重排序模型，实现了对 RAG 系统前，中，后的全流程优化。

- **实现技术：**

DeepSpeed，LLama-Factory，RLHF，LORA，Qwen2.5，RAG，milvus

2025.05-2025.08　　MedBrain 辅助诊疗 Agent　　杭州明阳计算机技术有限公司

- **项目简介：**

基于电子病历，医学文献，药品说明书，诊疗指南以及临床检验指标等多源医疗数据，采用 Langchain 构建多源数据检索框架，结合 DeepSeek-V3 大模型的时序预测与因果推理能力，开发了辅助诊疗 Agent，有效提升医生诊疗效率，缩短患者就诊时间，并改善了就医体验。

- **功能职责：**

1．**模型选型：**在架构设计中，选用 GLM-embedding3 作为词嵌入模型，DeepSeek-V3 作为核心对话大模型，结合 bce-reranker-basev1 作为重排序模型；整体上采用 API 调用的方式完成系统搭建，实现了模块化，可扩展的项目架构设计。

2．构建 RAG 系统：设计并实现了基于 **Milvus(向量数据库)与 Neo4j(图数据库)**的混合式 RAG 系统，通过**多路召回**与重排序模型，显著提升了检索信息的准确性与全面性，为大模型生成高质量，高可靠性的 prompt。

3．**优化 RAG 系统：**引入 **multi-query(多查询生成)，BM25(关键词匹配)，父文档检索器**等多元策略，构建更鲁棒性的检索流程，显著提升了检索结果的准确性。

4．**构建缓存系统：**引入 **Redis** 作为高性能缓存数据库，对频繁访问的查询结果与向量片段进行缓存，显著降低对底层数据库的访问压力，大幅降低 tokens 的消耗，提升系统响应速度的同时，节省约 30%的成本。

- **实现技术：**

LangChain，DeepSeek，embedding3，bce-reranker，Milvus，Neo4j，multi-query，BM25，Redis

项目经历

2024.02-2024.03　　基于知识图谱的医疗问答系统

- **项目简介：**该项目旨在构造一个基于知识图谱的医疗问答系统，不同于向量数据库的 RAG 方式，采用”**neo4j 图数据库+大模型对话生成**“的技术路线，给用户提供医疗知识问答，临床诊断辅助，医学常识科普等多类应用场景。
- **工作内容**

1．知识图谱构建：利用 **neo4j** 独立搭建医疗知识图谱，涵盖疾病，症状，药品，检查等 **500 种实体和 60 种关系。**实现医疗知识的结构化表示**。**

2．问题分类与答案生成：基于**规则派**将任务拆分成三个子任务（**问题分类，问题解析，答案搜索**），通过问题分类对输入问题进行类别判定，结合问题解析模块进行意图和识别，最后再通过知识图谱搜索，生成准确可靠的回答。

3．**意图识别模块优化：**之前通过关键字匹配来实现的意图识别尽管速度快，准确率高，但是召回率低，泛化性能差。这里通过引入 **BERT** 来进行多分类任务，实现了对意图识别模块的优化，增强了问答系统的整体鲁棒性。

自我评价

- 具备良好的英语听说读写能力，**能够熟练阅读与理解英文技术文档**，学术论文及开源社区资料，并可进行流畅的日常交流与技术沟通。
- 自学能力强，能**够在短时间内上手完全未接触过的新业务**，热爱新知识，热爱 AI 领域，会主动学习。
- 接触面广，兴趣爱好丰富，**能很好的与周围同事融成一片**，快速融入陌生环境。
- 抗压能力强，能多线程进行一系列工作，乐于挑战，**热爱工作，可接受加班和高强度工作**