基本信息

姓　　名：　曾仔强
年　　龄：　22
学　　历：　大学本科
求职意向：　大模型开发工程师

手机：　17380236821
邮箱：　2167033233@qq.com
地址：　四川省成都市

教育背景

2021.9-2025.6　　**成都理工大学**　　**数字媒体技术**

主修课程：C 语言、深度学习、数据结构与算法、操作系统、高等数学、概率论、线性代数。

实习经历

2024.10-2025.3　　**汇人健康管理 • AI 算法实习生**

参与医学问答系统及知识图谱构建项目，负责知识图谱模块开发维护与向量库联合检索，实现医学知识高效管理；依据产品需求与技术方案，开发实体检索、关系查询功能，支撑医学问答系统多轮问答生成：参与生产环境问题排查修复，分析日志、定位知识图谱查询及问答生成问题根源并实施解决方案； 参与代码评审与技术讨论，协助完善项目技术文档，总结经验并优化代码质量与系统性能。

专业技能

基础能力：掌握 python /C 语言，熟练运用 pytorch 框架，熟悉 docker、多线程编程、异步编程、resful 编程、fastapi 接口封装，Agent、LangGraph、Dify、mcp、Function Calling

大语言模型与自然语言处理：

模型架构：熟悉 Transformer、注意力机制、BERT 等前沿架构，具备 NLP 任务建模能力

参数高效微调：熟练使用 LoRA 技术，能利用 LLaMA-Factory 实现轻量高效的大模型定制化训练

NLP 技术栈：掌握文本特征工程(TF-IDF、word2vec、Tokenizer)，支持多语言场景下的命名实体识别(NER)、本生成等任务

大模型系统开发与部署：

高性能推理：熟练使用 vLLM 进行大模型部署，实现流式响应、多轮对话与并发优化

增强技术：掌握 RAG 检索增强生成、Prompting 提示词工程等关键技术

知识库构建：使用 Milvus 向量数据库、Neo4j 图数据库构建高效知识检索系统

智能代理框架开发与部署:

框架构建：熟练使用 Dify 平台进行 Agent 框架的设计与实现，支持高效的任务编排和自动化工作流

本地化部署：本地化部署模型控制服务（MCP），优化资源管理和服务高可用性

核心技术：熟悉 Function Calling 与 ReAct 机制；LangChain、LangGraph 库，用于构建多智能体协作系统和复杂推理工作流。

项目经验

2024.10-2025.3 **智能医疗问答系统**

技术栈：DeepSeek-70B + LoRA 微调 + Hybrid RAG + Neo4j +Transformers + vLLM

知识图谱构建： 基于 CHPO 医学术语库扩展同义词，采用 PaddleNLP-UIE 实现三元组抽取，设计“预标注-人工校验-微调”闭环框架。 开发实体对齐模块，利用 SapBERT 语义相似度计算解决 15 类症状标准化问题，图谱查询效率提升 30%。

RAG 增强系统： 融合知识图谱路径检索与向量库语义检索，增强事实性与推理能力

大模型部署：基于 vLLM 异步引擎 实现流式响应（首 Token 延迟 <200ms），通过 KV Cache 共享机制支持高效多轮对话，并采用 RingAttention 技术 动态管理长上下文，显著提升大模型本地部署的吞吐量与资源利用率

个人证书

- 2024 年 蓝桥杯省级二等奖
- 2024 年 大唐杯省级二等奖
- 2024 年 成都理工大学科技立项
- CET4

自我评价

深耕大模型应用技术栈：从模型微调（LoRA）、本地部署（Transformers/vLLM）到 RAG 增强全流程落地经验。

能够运用 Dify 构建构建多智能体协作系统

具有较强的学习能力和抗压能力