

符某某

171-6200-1234 | Xyu@outlook.com

2002-03 | 男 | 福州

在校生 | 大语言模型 自然语言处理 大数据开发



教育经历

西交利物浦大学（英国利物浦大学学历学位，海外前 150）	2024 年 09 月-2026 年 06 月
计算机科学（大模型推理能力方向） 研究型硕士 全日制研究生	苏州
福建江夏学院	2020 年 09 月-2024 年 06 月
数据科学与大数据技术 工学学士 全日制公办本科	福州
绩点: 87.84（专业第一，1/50），英语: 雅思 6.5（阅 7.5，听 7.0），六级 532，四级 578，优秀毕业: 面向微博的文本情感分析与可视化（94）	
奖学金: 校一等奖学金 2 次，校二等奖学金 2 次，校国境外访学奖学金 1 次，学生工作: 校英语协会主席（2021-2022，社团人数 300+）	
荣誉: 校三好学生，校优秀学生干部，校学风建设标兵，院优秀毕业生，院创新实践先进个人，院社会工作先进个人，院优秀共青团员	

实习经历

福建高岭数据科技有限公司	2024 年 06 月-2024 年 08 月
人工智能算法工程师实习生 预研组 福建省烟草公司某市公司 BI 系统	福州
· 融合 RAG 的 Text2SQL 多智能体系统: 基于 LangChain、LangGraph 和 FastAPI，开发了一套融合 RAG 的 Text2SQL 多智能体系统，应用于某市烟草 BI 系统。系统适配 GPT 系列和 Qwen 系列云端模型，同时支持 Xinference 部署的本地模型。通过专业 Prompt 模板设定不同的智能体节点，节点间通过条件边构建有向无环图。为提升系统性能，在 Text2SQL 垂直领域使用 GPT-4o 对 Qwen2-7B-Instruct 进行数据蒸馏微调，其准确性达到 80%。此外，为提高 RAG 的检索召回率，定制了一种融合表结构和表描述的文档读取器。用户通过 RAG-Agent 提出问题，RAG-Agent 检索最相关的表信息并发送给 SQL-Agent，SQL-Agent 根据上下文生成并执行 SQL 查询。系统设定重试机制，进行结果硬判断。若出现错误，则返回上一步骤；若成功，则发送结果至 Analyst-Agent 进行分析并生成报告。编写了自动化测试脚本，通过结果转置进行匹配和判断，以确保系统的准确性和稳定性。系统格式化输入和输出，并根据 SQL 查询结果生成 ECharts 数据以配合前端展示。	
· 融合 RAG 的通用 API 多智能体系统: 基于融合 RAG 的 Text2SQL 多智能体系统项目，进一步开发了一套融合 RAG 的通用 API 多智能体系统。定制了一个专业的通用 API 请求工具，支持填写任意 API 请求的 header、path、params 和 body 的固定参数、期望参数和默认参数。用户通过 RAG-Agent 提出问题，RAG-Agent 检索到最相关的 API 请求信息，包括固定参数和默认参数。API-Agent 根据用户问题，调用通用 API 工具，获取期望参数，并按固定参数、期望参数和默认参数的优先级融合参数，发出请求。API 请求成功后，响应的数据结果将交给 Analyst-Agent 进行最终分析。系统的其他功能与之前的项目类似，确保了系统的准确性和稳定性。	
北京中蓝恒量科技有限公司	2024 年 02 月-2024 年 06 月
人工智能研发工程师实习生 研发部	北京
· 百度贴吧分布式异步爬虫: 使用 Scrapy 框架开发了爬取指定贴吧所有帖子详细内容的通用爬虫。优化代理机制，降低了请求切换代理的频率，实现了以低成本采集全部数据。实现了进程监控，确保爬虫中断后可重新续爬。采购并配置多台 Linux 服务器环境，编写自动化一键部署功能，实现控制台实时返回多台服务器运行结果，加快爬虫任务的速度。	
· 二次元知识库问答助手: 使用 Scrapy 框架开发了爬取萌娘百科网站全部数据的树形爬虫。使用 Llamaindex 定制数据库读取器，将萌娘百科数据读取成 Documents 并分块，使用 bge-large-zh-v1.5 模型向量化数据并存储到 Qdrant 向量数据库中。加载 4bit 量化的 Qwen1.5-14B-Chat 模型，根据输入的 Prompt 检索相关文档，经过相似度筛选和重排序后生成回答。增加了自定义文档上传功能，上传文档后进行向量化存储。定制 System Prompt，使模型能够进行角色扮演，结合 Gradio 框架与用户进行交互。	
· 哔哩哔哩虚拟聊天主播: 使用 LangChain 设计定制表情动作控制工具，结合 OpenAI 的 Tool Calling 功能，通过 WebSockets 连接 VtubeStudio 的 API，对虚拟主播进行表情和动作控制。定制 System Prompt，使大模型能够进行角色扮演。新增持久化记忆功能，对不同用户的交互数据进行隔离，确保个性化的互动体验。实时传入 B 站弹幕数据，将弹幕数据传入阻塞队列，逐一生成响应文本。使用 GPT-SoVits 多线程合成响应音频，重采样后将音频传入阻塞队列，通过 VoiceMenter 虚拟声卡实现口型驱动和声音输出。	

研究经历

国家级大学生创新训练计划（项目编号：202213763012，优秀结项）	2022 年 05 月-2023 年 05 月
第一主持人	
· 项目名称: 基于机器学习的新冠患者住院优先级预测系统设计与实现	
· 负责内容: 数据预处理和特征工程，算法模型训练，应用接口编程	
· 项目成果: 以第一作者在 CN 期刊发表学术论文 1 篇，获计算机软件著作权 1 项	
· 论文发表: 符鑫隆,林姍,牛辉等.基于 CatBoost 的患者住院优先级预测模型[J].信息化研究,2023,49(01):43-47.	

项目经历

面向微博的文本情感分析与可视化	
· 项目功能: 零样本文本分类模型微调与推理，多线程 GPT 数据标注和 Prompt 管理，微博通用爬虫及进程控制，数据处理分析与可视化	
· 技术栈: 前端 Vue，后端 Flask，数据库 MySQL，爬虫 Scrapy，算法框架 PyTorch+HuggingFace	
基于机器学习的分布式系统故障诊断系统	
· 项目功能: 三级网络集成模型在线训练与推理，数据可视化，数据库与文件系统操作，用户认证与管理	
· 技术栈: 前端 Vue，后端 SpringBoot，数据库 MySQL，算法框架 Sklearn	
学院官网与论坛系统	
· 项目功能: 用户身份与权限管理，新闻发布与管理，论坛互动功能，个人信息管理，管理员用户管理	
· 技术栈: 前端 HTML+CSS+JS，后端 JSP，数据库 MySQL	
福建房价大数据平台	
· 项目功能: 数据采集，数据预处理与清洗，数据存储与分析，数据迁移，数据可视化	
· 技术栈: 前端 HTML+CSS+JS，后端 Flask，数据库 MySQL，爬虫 Scrapy，数据处理 MapReduce，数据仓库 Hive，数据迁移 Sqoop	

竞赛经历

第 15 届中国大学生计算机设计大赛全国三等奖	2022 年 07 月
第 13 届蓝桥杯 Python 程序设计大赛福建二等奖	2022 年 05 月
2021 年全国大学生英语竞赛二等奖	2021 年 05 月

专业技能

- 熟悉 Python 开发，了解 Java 开发，了解 Scala 开发，了解 JavaScript 开发
- 熟悉深度学习神经网络的底层原理，了解自然语言处理和大语言模型的底层原理，了解 Agent 和 RAG 的底层原理
- 熟悉大数据技术的基本理论和实际应用，包括数据采集，数据预处理，数据存储，数据挖掘，数据可视化
- 了解数据结构，计算机网络，计算机组成原理，操作系统的基础知识