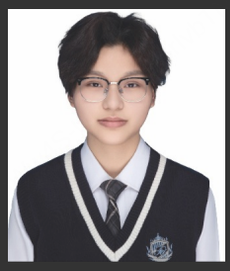
陈子芮



性　　别：女
电　　话：13564546051
求岗职位：算法工程师
联系邮箱：13564546051@163.com
到岗时间：随时到岗
期望薪资：面谈

个人信息

2018-09 ~ 2022-07　　**武昌理工学院（本科）**　　**计算机科学与技术**

工作经历

2022-06 ~ 至今　　**宁波博登智能科技有限公司**　　**NLP-算法工程师**

1. 主要负责文本数据处理与建模工作，包括数据清洗、特征工程和深度学习模型构建。
2. 参与核心项目的算法研发与优化，运用BERT、Transformer等先进模型解决实际业务问题，并通过模型调优等技术持续提升算法性能。
3. 主导完成了多个NLP项目的落地应用，从算法选型、模型训练到最终部署全流程参与，确保算法解决方案在实际业务场景中的有效实施。

专业技术

2022-08 ~ 至今

编程语言与环境: Python（精通）、开发环境Windows、Linux、熟练驾驭多类云平台的部署工作
数据处理与分析: numpy、panda
机器学习算法: 线性回归、逻辑回归、SVM、 K-means、KNN等
深度学习技术

- 深度学习框架与算法：TensorFlow、Pythorch、CNN、RNN、LSTM
- 预训练模型微调：FFT（全参量微调）、PEFT-LoRA （参数高效微调）

开发框架与工具熟练掌握：RAG、LangChain、AutoGen、CrewAI、Gradio、Streamlit、Chainlit、Neo4j
多模态生成与处理：Stable Diffusion、Video Diffusion：熟悉生成式AI，适用于图像和视频生成任务
开源大模型：

- 熟悉 **DeepSeek**、**Qwen**、**Mistral**、**Llama**、**ChatGLM**，能够利用开源模型进行开发和微调。

项目经验

2025-02 ~ 至今　　**化工专用问答引擎**

技术栈：Autogen，微软Graphrag，Ollama，chainlit
项目简介：在化学研究与工程实践中，工程师和研究人员常面临从实验报告、专利文献等海量私有文档中快速获取精确信息的难题，传统方式效率低且易漏关键数据。为此，我们整合自然语言处理与智能检索技术，打造高效精准的化工专用问答引擎，可智能解析化学术语、定位核心知识点，显著提升信息获取效率与准确性，助力科研创新与工程实践优化

项目流程：

1.数据处理层

(1),利用 Ollama 框架完成 Mistral-7B 大模型的本地化部署后，基于化学领域数据集，通过 LoRA 轻量化微调策略对模型进行领域适配优化，结合微软 GraphRAG 技术，实现数据的语义解析与三元组提取并进行社区聚类，识别具有相似特征的实体群体，挖掘潜在的关联模式，neo4j可视化。

(2), mxbai-embed-large-v1对文本进行语义编码，构建高维语义表征体系。

2.智能决策层

AutoGen框架构建多智能体协同系统，部署Qwen2.5作为核心推理模型

路由Agent： 通过动态意图理解机制解析用户提问语义，基于问题类型判别模型，决策执行本地知识库的范围检索，或触发网页信息获取通道

检索Agent：结合知识图谱的关系遍历与向量相似度计算，从向量库中快速定位相关信息

3. 交互应用层

基于Chainlit框架构建自然语言交互层，实现面向用户的对话式交互接口与后端智能决策系统的无缝集成

2024-06 ~ 2024-12 **医学生物百科全书**

技术栈：BERN2工具，Langchain，NLM数据库，OCR（光学字符识别）

项目背景：在生物医学领域，海量文献、研究成果与临床数据呈指数级增长，分散于各类学术数据库、专业文献及公开网站中。由于数据格式多样、语义表述复杂，传统检索方式难以快速精准获取有效信息，极大制约了科研人员、医疗从业者的知识获取效率，也为大众健康科普带来挑战。本项目旨在构建生物医学智能百科，通过整合多源数据，运用先进的数据处理与自然语言处理技术，实现知识的高效提取、结构化存储与智能检索，打破生物医学知识壁垒，提升知识传播与应用效能

项目流程：

1.数据采集：从网络多渠道爬取生物医学相关数据，涵盖学术论文、科普文章、临床指南等，获取原始知识素材

2.数据预处理：利用 OCR 技术对采集数据进行处理，识别文本中的命名实体；同时，移除文本中的章节标题、图描述等非核心内容，减少冗余信息干扰；最后将文本分割成句子，为后续处理做准备。

3.知识提取：借助 BERN2生物医学命名实体识别（NER）与标准化（NEN）工具，从文本中提取生物医学实体、关系及属性，生成三元组数据；结合外部数据库NLM（美国国家医学图书馆），对提取的概念进行标准化映射与补充，确保知识准确性与完整性

4.知识图谱构建：将提取的三元组数据，通过合适的数据建模与存储技术，构建生物医学知识图谱，以结构化形式存储知识，直观展现实体间关系。

5.智能交互实现：基于 Langchain 框架，建立用户问题与知识图谱的交互桥梁。当用户提出问题时，系统将问题解析为 Cypher 语句，在知识图谱中进行查找匹配，获取相关知识；再通过 OpenAI 的 GPT3.5 模型，将检索到的结构化知识转化为自然语言，反馈给用户，实现智能问答交互

2023-10 ~ 2024-04 **会议智录**

技术栈：WhisperX，PyAudio

项目背景：在医疗数字化与智慧医疗加速发展的当下，医院科室会诊、学术研讨、医患沟通等场景，对信息记录的准确性和及时性要求极高。传统人工记录方式弊端明显：医疗术语专业复杂，人工记录易出错，影响诊疗与病例归档；多科室协作、远程会诊时，信息量大，人工整理耗时，导致信息传递滞后。国际医疗合作增多，多语言会议场景也加剧了信息记录难度。本项目专为医疗场景定制，可实现多语言语音实时精准转录，高效识别专业医疗词汇，帮助医护人员快速获取信息，减少误差。其将语音转化为结构化文本，便于病例书写与纪要整理，提升信息处理效率，降低人工成本，为医疗协同与精准诊疗提供支持，推动医疗数字化转型。

项目流程：

1.模型微调：针对医疗领域专业术语识别挑战，我们通过大规模医疗语音语料库（涵盖临床诊疗、病例讨论、学术会议等场景），对 WhisperX 模型进行领域适应性训练。采用渐进式学习策略优化声学模型参数，重点强化对医学专有名词的特征提取能力，有效提升专业术语的识别准确率，确保转录内容的专业性与完整性。

2.实时捕捉音频：基于 PyAudio 库实现音频信号的实时捕获与流式处理，通过配置音频输入设备参数（采样率、通道数、位深度），构建低延迟的音频数据流采集通道，持续获取麦克风或会议系统输出的实时语音信号，为后续降噪处理与语音识别提供连续的原始音频数据

3.音频转录：WhisperX 执行自动语音识别（ASR），生成原始转录文本，并通过说话人分割（Speaker Diarization）技术识别不同说话人，自动标记每位说话者的发言内容及其对应的时间戳。

4.转录结果推送至用户：当语音识别系统完成音频转文本处理后，自动将生成的完整转录文本及说话人分段信息，通过系统消息、邮件或企业协作平台等渠道推送给指定用户，支持实时查看与历史记录回溯，确保会议信息及时触达相关人员

2023-01 ~ 2023-07 **金融市场情绪智能分析系统**

技术栈：Streamlit，Lora微调，Huggingfase

项目背景：在金融行业快速发展和数据驱动的背景下，市场情绪分析已成为投资决策、风险管理和市场预测的核心工具。通过分析新闻、社交媒体、财报等文本数据，提取积极、消极或中性情绪，可以为金融机构提供实时、数据驱动的洞察。然而，通用大语言模型在金融领域的专业性不足，难以精准捕捉金融文本的语义和情绪，尤其在需要处理中文金融文本的场景下。为此我们通过微调构建一个企业级的FinGPT模型，对市场情绪分析

项目流程：

1.数据处理：从权威来源和公开数据集收集金融文本数据，从Hugging Face拉取ChatGLM-6B模型及其分词器，对文本进行精准分词和格式化处理。预处理步骤包括去除HTML标签、停用词和标点符号，确保数据清洁。对于中文金融文本，结合ChatGLM-6B分词器和Jieba进行高效分词，保留语义完整性，并将处理后的数据存储为结构化格式，便于后续模型训练和分析

2.模型微调与训练：设置LoRA参数，利用金融情感数据集进行监督微调。为减少计算资源消耗，实施8-bit量化以降低显存需求并加快训练速度，最终保存微调后的模型

3.模型部署与市场情绪分析：利用Streamlit构建交互式Web应用程序，用于展示金融文本的情感分析结果，提供直观的用户界面。集成Plotly库生成可视化图表，如饼图展示正面、负面和中性情感的比例分布，折线图呈现情感趋势随时间的变化，增强数据解读的清晰度和吸引力

2022-06 ~ 2023-11 **维启中文智识者**

技术栈：HMM，Viterbi

项目背景：在自然语言处理领域，中文命名实体识别（NER）面临分词歧义、未登录词识别等独特挑战，现有深度学习方案（如 BERT）虽精度高但计算成本大，难以适配教育场景教学需求与小型应用轻量化部署要求。经典的 Viterbi 算法结合隐马尔可夫模型（HMM）具备模型结构清晰、计算复杂度低的优势，既能通过概率图模型直观展现 NER 核心原理，适合 NLP 入门教学中的算法拆解与实践；又能以轻量级架构满足中小企业文档标注、垂直领域信息提取等小型应用场景需求。当前学界对先进模型研究较多，但缺乏针对教育场景与轻量级应用的中文 NER 系统整合方案，本项目通过优化经典算法流程，填补这一应用空白

项目流程：

1.数据处理与准备:

从公开语料库（如人民日报标注语料）和网络文本收集中文数据，清洗噪声与特殊字符,对文本进行分词和实体标注,最后划分训练集、验证集和测试集，确定模型参数初始值，搭建训练环境

2.模型构建与训练:

构建隐马尔可夫模型,实现 Viterbi 算法，计算文本序列中最优实体标签路径,使用训练集训练模型，通过验证集调优参数（转移概率、发射概率），以准确率、召回率、F1 值评估性能。

3.系统集成与部署:

将训练好的模型集成到 NER 系统，用测试集验证不同文本类型和长度的识别准确性，修复问题,部署系统至服务器或本地环境，应用于教育教学、小型文本分析等场景，持续收集反馈优化。

自我评价

本人工作态度认真负责，具备优秀的职业素养和团队协作精神。面对任务，我积极主动，注重细节，确保高效完成目标。我善于沟通，乐于倾听意见，快速适应环境并提出创新解决方案。作为NLP算法工程师，我精通Python，熟练运用PyTorch等框架及预训练模型微调，并掌握LangChain、Gradio等开发工具，擅长处理NLP任务及生成式AI开发。我尊重领导指导，虚心学习，不断提升能力，以为团队和公司创造更大价值为己任。期望在未来工作中，发挥优势，与团队共成长，为公司发展贡献力量