周若骏

邮件: ruojun.zhou@duke.edu

手机: +86 18112388439　　+1 9196417116

https://github.com/Arnold-Jun

教育经历

杜克大学, 普拉特工程学院 达勒姆，美国北卡罗来纳州

电子与计算机工程工程硕士（机器学习/大数据） *08. 2024 – 05. 2026*

核心课程: C++中的计算机系统与工程基础, 编程, 数据结构和算法，机器学习，向量空间

西交利物浦大学，数学物理学院 苏州, 中国

应用数学理学学士学位 专业绩点: **3.9/4.0** 总绩点: 3.7/4.0

核心课程: 拓扑，优化论，数理统计，应用概率，泛函分析，数值分析，运筹学 *09. 2020 – 07. 2024*

荣誉/奖项：一等荣誉学位

技能

编程语言: Python, C/C++, Java， MATLAB **框架技能**: Spring AI, Langgraph4j, LangChain, DSPy, LlamaIndex, OpenCV

项目 & 实习经历

小红书（研效工程组） 上海, 中国

大模型后端研发 *04. 2025 – 08. 2025*

- **智能代码审查（Smart CR）智能体研发:**
  - **智能工作流编排**：基于**LangGraph4j**设计实现6**节点状态图架构**，开发多条件边缘路由和人机协作决策逻辑，支持任务暂停、恢复和重放操作。
  - **MCP工具生态建设**：参与开发代码审查相关**MCP工具**，实现HTTP/Stdio双协议通信机制，构建动态工具加载和容错机制。
  - **A2A协议通信**：基于**JSON-RPC**规范实现CrAgent的**A2A**通信服务端，开发**任务订阅**、**状态同步**和**SSE事件流传输**。
- **Smart CR系统研发**:
  - **对话与Prompt优化**：深度参与**智能CR一轮对话**和**prompt工程优化**，通过few-shot设计，角色设定与输出约束，使智能CR召回率提高**10%**。
  - **扫描平台构建**：基于主智能CR链路和集成SonarQube静态扫描，构建**分支级全量代码扫描自动化平台**，实现多源头和多风险级别问题透出。
  - **平台独立开发**：独立开发**Prompt管理平台**，支持多部门按场景自定义prompt模板，实现prompt版本管理，提高团队迭代效率。
  - **稳定性治理**：参与智能CR全链路的**数据埋点与告警配置**，监控请求延迟、模型响应成功率和责任链链路执行成功率，支持异常自动告警。
- **大模型适配服务开发**:
  - **推理服务构建**：深度参与构建部门统一的大模型推理后端服务，支持对接多种开源与闭源模型，实现输入输出格式标准化（OpenAI API兼容）。
  - **容错机制开发**：开发针对大模型token限流场景下的适配层容错机制，包括请求重试（指数退避）和请求排队，保障高并发下服务稳定性。

大模型私有化部署与 AI Agent 构建 昆山杜克大学, 研究助理

- **Agent框架开发**: 基于**DSPy + LlamaIndex**设计AI Agent，支持**记忆管理**、**工具检索**、**判别迭代**、**任务规划**和**流式推理输出**。
- **大模型推理优化**: 使用**LoRA**和**ZeRO-3**进行参数高效微调，增强模型对校园特定术语和规则的理解，规范和优化模型输出。
- **向量数据库检索**: 使用**Chroma + Redis**构建**RAG**系统，提高上下文理解能力，减少幻觉。
- **工具集成 & API扩展**：开让Agent能调用**邮件发送**、**Google搜索**、**计算器**、**场地预定**等函数或外部API，支持复杂任务自动化。
- **模型输出增强**：通过一个多阶段的LLM管道，使用本地模型作为隐私代理生成隐私保护的提示，并委托云端模型生成高质量响应。
- **隐私推理**：为学校部门部署**小语言模型**，通过本地加噪声和去噪的方式，让服务器端大语言模型指导小模型推理，提高文档检索效率。

保护隐私的多模态脑肿瘤分割框架 昆山杜克大学, 研究助理

- **数据处理:** 采用Dirichlet分布为每个客户端生成权值，保证客户端间数据分布的随机性和不平衡性。对标签进行下采样以进行深度监督训练。
- **模型设计:** 设计了基于Transformer的交叉模态模块并将其集成到U-Net中，使模型能够处理多模态MR图像。融合差分隐私、联邦训练范式和多模态U-Net构建联邦多模态脑肿瘤分割框架。
- **性能评估:** 在 BraTS2022 数据集上，增强肿瘤、肿瘤核心和全肿瘤的准确率分别达到 **87.5%**、**90.6%** 和**92.2%**，优于现有方法并保持隐私保护。

基于深度学习的智能电网图像识别 苏州, 中国

- **算法创新:** 提出 **GA-Kmeans** 方法，为 YOLO 获取全局最优先验锚点。
- **工作流优化:** 基于二值先验锚点、形态学处理、边缘检测和形状匹配进行绝缘子定位。使用预训练的ConvNeXt对分割后的绝缘子块进行特征提取，训练线性SVM分类器对每个图像块进行分类。
- **模型增强:** 在YOLO-v8中集成**通道注意力**和**空间注意力**，实现端到端的故障检测。改进后的YOLO-v8简化了检测工作流，将故障分类准确率提高到**96.71%**。

出版物

- “Fed-MUnet: 基于联邦学习的多模态脑肿瘤分割。” 2024年IEEE国际健康网络，应用与服务会议（HealthCom）。arXiv:2409.01020