文嘉懿

19875380719 | 12333207@mail.sustech.edu.cn | ku1yin.github.io

教育经历

南方科技大学 | **智能制造与机器人** | **硕士在读** 2023年09月 – 2026年06月

- 研究方向：人机交互、多模态交互、智能体设计
- 核心课程：机器人设计科学与社会价值、工业应用与实践中的设计创新

中南大学 | **土木工程** | **本科** 2019年09月 – 2023年06月

- 相关课程: C++程序设计基础、科学计算与MATLAB

项目经历

基于LangChain框架的论文查询RAG系统 - 项目负责人 2025.02-至今

- 基于**LangChain框架**构建，封装自定义工具链，远程调用**SiliConflow API**，实现本地论文内容结构化解析与摘要生成。
- 集成多轮对话与**观点验证**模块，支持自然语言查询，基于文献内容动态构建正反观点对比结构。
- 集成BM25+Faiss**混合检索**、**Rerank模型**，实现**Contextual Chunk切割**，大幅提高召回率。
- 基于本项目核心代码，目前天池云CCKS2025-人工智能领域论文复杂问题问答评测比赛top3.

基于树莓派的大模型语音情感陪伴助手 - 项目负责人、团队唯一成员 2025.04 - 至今

- 基于 **llama.cpp** 在 Raspberry Pi 5 上部署量化后的本地大语言模型，适配低功耗嵌入式场景。
- 集成 **Vosk**（ASR）与 **GPT-SoVITS**（TTS）模块，构建客制化端到端语音交互系统，实现语音输入 - LLM 处理 - 音频输出全链路。
- 设计并实现基于 Prompt Engineering 的多轮记忆与人格设定机制，具备情绪响应与长时记忆能力。

IxDL：交互设计语言智能体 - 团队主要成员 2025.03 - 至今

- 构建基于 **LangChain + RAG** 架构的智能体，结合自建 IxDL 语法知识库，建立从自然语言到交互语言的语义映射与反向转换链条。
- 使用 **GPT-4 API** 构建双向 Agent，支持自然语言到代码结构的多模态转换；设计中间表示实现 Web UI 效果预览。
- 结合 **MLLM 多模态大模型**，通过图片递归切割与语义合并，构建前端交互区域结构树，支持自动结构解析与标注。

多智能体情感交互系统（全栈开发与部署） - 项目负责人、团队唯一成员 2025.06 - 至今

- 构建支持 **多智能体记忆管理与语义交互**的情感对话系统，支持 **OpenAI/Grok/xAI** 等多大模型 API 热切换，具备高扩展性与稳定性。
- 前端采用 **Streamlit** 实现模块化组件，并基于Nginx + 云服务搭建完整交互平台，部署于腾讯云与本地并行环境。
- 统一设计 Prompt 接口标准与语料缓存机制，表达控制与情绪状态触发机制（如表情包与语音展示联动）。

“大音”：一个新的混音应用程序的开发与参与式用户研究 - 团队主要成员，文旅部重点实验室项目 2024.07 - 2025.04

- 基于**Flask框架**构建后端，集成**ffmpeg**实现音频自动切片与处理，支持实时上传与处理用户录音数据。
- 利用**UMAP**与**t-SNE**进行音频特征降维，结合**D3.js**动态渲染交互式散点图，实现音频片段的可视化浏览与连续播放。

实习经历

联想集团，SSG，技术产品经理-AI方向 2025.05-至今

- 跟踪AI前沿技术，深度调研10余款AI开发、产品提效软件，输出20+页AI研究报告并参与内部技术研讨会，为产品长期规划提供参考。
- 采用Dify框架，开发智能零售门店智能体最小可实现产品，实现基于RAG的多轮对话、跑通智能导购到自动下单全流程.

研究成果

- 一篇CCF-A会议一作在投。一项发明专利在审。

专业技能

- **编程语言:** Python（Pytorch、LangChain）、C++、C#、MATLAB
- **大模型工具:** Finetuning、llama.cpp 量化部署、Prompt 工程、RAG 系统搭建、Dify、GPT-SoVITS、Vosk
- **其他技能:** Fusion 360、Blender、Unity 3D、FMOD、Max/MSP、Wwise、Figma、Xmind、Axure、PS、AI、AU、PR

学生工作&个人荣誉

- 担任中南大学学生会心助会（朋辈心理互助会）干事及副部长，组织多项院校级活动。
- 2020年中南大学心助会优秀学生干部、南方科技大学研究生学业一等奖学金。

自我总结

- 作为一名具备大模型全栈开发经验的研究型开发者，我擅长将前沿技术与实际应用场景结合，推动 AI 系统从原型构建到工程部署的全过程落地。在多个项目中主导了大模型的训推、边缘设备部署与语料构建流程，具备扎实的 Python/C++ 编程基础与系统集成能力。跨学科背景强化了我的系统思维与快速学习能力，未来希望在 AI Infra 或大模型垂直应用方向持续深耕。