

Xuanpu Zhang

xuanpu.huiyuan@gmail.com | LinkedIn: xuanpu-zhang-3bb575264 | GitHub: huiyuanXP | huiyuanxp.com

教育背景

新加坡国立大学 | 新加坡

2026.08 - 2028.01

计算机硕士（人工智能方向），学位论文方向：多智能体安全

- 在新加坡国立大学数据科学研究所（IDS）Ng See Kiong 教授课题组开展多智能体安全研究。

麦吉尔大学 | 加拿大，蒙特利尔

2022.09 - 2026.05

理学学士：计算机科学（人工智能方向）；辅修数学

项目经历

PPT-to-Notes Agent — 多模态学习平台

2025.07 - 至今

创始人；独立负责 AI 产品全栈开发

技术栈：React、TypeScript、Python、FastAPI、PostgreSQL、Celery、Redis、S3、Docker、POML

项目概述：处理 PPT 与课堂录制内容，生成详细学习笔记，并提供可溯源的课件页码 / 30 秒视频片段引用

- 设计 Agent Harness 工作流，基于课件和/或课堂录制内容生成高质量、来源可溯的长上下文笔记。
- 使用 Gemini 2.5 Flash 设计 Agentic OCR，将 PDF 页面图像转为结构化 JSON/Markdown，保留文本、图像描述、表格、图示及 token 元数据，供下游推理与整合使用。
- 基于结构化 OCR、笔记、视频帧及转录文本构建 RAG，支持 100 页以上课件的理解；返回有据可查的回答，附带准确的页级引用与来源元数据。
- 为 AI 学习产品部署 AWS 无服务器服务；截至 2026 年 8 月，服务 120 名用户，其中 10 名为付费用户。
- 构建支持并行、可中断、可恢复的多模型处理流程，结合逐页 JSON 检查点、损坏输出恢复、失败图像跟踪及 S3 同步。

HuiyuanClaw - 多智能体编排框架

2026.02 - 2026.05

Agent 编排与开发者工具开发

技术栈：Node.js、JavaScript、WebSocket、MCP、JSONL、Git、Cloudflare Tunnel

项目概述：构建类 OpenClaw 的 Harness，底层由 Codex / Claude Code 驱动，模仿其将任务分配给子线程的功能，具备高可观测性与可控的线程观察能力。

- 围绕显式多智能体编排重新设计开源项目，将每项委派任务表示为持久化、可检查的会话，明确任务归属并支持人工控制；通过基于浏览器的 Agent Harness 呈现任务拆解与并发执行过程。
- 将 MCP 创建的 Agent 会话限定在特定项目和文件夹中，提供稳定的本地指令和更精简的技能/工具集；将每个会话作为明确的工作区单元进行控制，减少不必要的上下文消耗，并使委派任务始终限定在指定范围内。
- 于 2026 年 1 月实现 MCP 驱动的会话控制及跨会话消息/回复工作流原型，早于 Codex 应用会话控制功能发布（2026 年 5 月）及 Claude Code 独立跨会话消息功能发布（2026 年 8 月）。
- 在浏览器界面中实时展示各 Agent 的执行进度及跨会话消息/回复流程，用户可查看持久化会话记录与 Git 变更，在并发执行中直接干预或单独中断某项任务。
- 受 Claude Code 最佳实践启发，引入 Critic Agent，从视觉效果、项目 CI/CD、用户需求细节等维度对 Agent 的实现进行批判性评估，并使 Agent 自行持续运行修复问题。

JD 定制化简历生成 Agent

2026.04 - 至今

项目概述：构建根据 JD 自动生成格式化简历的 Agent，通过持久化用户的项目经历、技能等信息，实现针对性简历生成与职位筛选。

- 以深度追问的形式生成用户的能力与历史项目画像，为每段经历构建详细结构化条目，并支持用户手动修改调整。
- 根据 JD 生成对应简历，格式可定制、可持久化，并可通过 prompt 调整简历语言风格。
- 生成 Markdown 格式简历，供用户手动调整和优化。
- 使用 Critic Agent 检查 Agent 流程中的错误，如视觉重叠、内容重复及结果不符合预期等。
- 可扩展至根据 JD 自动筛选合适职位、自动规划能力提升方向等。

alva：房屋设计 / 居住痛点挖掘 Agent2026.09

项目概述：构建帮助用户从房屋户型图开始设计家具风格与装饰的可视化系统；利用 Agent 能力获取用户信息，据此为客户提供定制化的痛点分析、设计问题分析等，最终形成可交付的面向用户的报告与面向设计师的报告。

- 采用苏格拉底式提问法，依据预构建的 70 个问题清单向用户提问；提问时结合已回答内容与用户画像给出多个建议选项，让用户只做选择题，不做简答题。
- 利用 Agent 视觉能力对用户上传的户型图进行建模，构建基础的数字孪生家庭空间底座，并通过拓扑检查与图像重叠检查自我检验生成结果、自动修复问题。
- 根据用户提出的模糊需求，结合 MCP 与 Agent 自身的 3D 建模能力生成对应的 3D 家具，并支持颜色、大小、长度等参数调整。
- 根据用户画像与居住者需求（如老人、孩子的生活习惯等）审核新添加的家具，指出与画像习惯的冲突。
- 结合用户画像与习惯、预构建的痛点列表及 Agent 能力，提出用户尚未发现的痛点。
- 作为 Team Leader，带领 4 人小组构建团队协作开发架构，以 Ticket 为最小原子单元实现可并行的项目快速构建；实现任务自动提交、自动分配、自动接取及任务归属签名等协作功能。代码均通过 Vibe Coding 生成。

面向具身智能机器人 Agent 意图识别的 1B LLM2025.06 - 2025.08

团队负责人；负责 LLM/MCP 系统

技术栈：Raspberry Pi、Python、1B LLM、MCP、ROS、STT/TTS

- 带领 3 人团队在 Raspberry Pi 上构建语音控制机器人 Agent，负责 1B LLM/MCP 与 ROS 的集成；内部演示中可在 3 秒内启动任务。
- 设计从语音转文字、1B LLM 意图识别到 MCP 路由、函数级指令、ROS 执行及语音合成反馈的完整 Agent 流程。
- 在 Raspberry Pi 上部署并调优用于意图识别的 1B LLM，将用户请求映射为函数调用，并在实体执行前通过仿真验证行为。

长上下文大模型评测2026.03 - 2026.04

负责实验设计与评测实现

技术栈：Python、GPT-4o、JSON、POML、序列对齐、precision/recall/F1

- 构建四种条件的长上下文评测，结果显示最长设置下证据链完全恢复率降至零，揭示 GPT-4o 明确的可靠性边界。
- 提出并主导课件研究，定义基于证据的逻辑链任务，并要求精确的页码 ID 引用。
- 实现结构化 JSON/POML 产物及序列/集合评测指标，包括 Needleman-Wunsch 对齐、LCS 召回率及 precision/recall/F1。
- 构建页面/内容块级结构化 JSON 与 POML 提示词，要求输出精确的来源页码 ID；将模型输出保存为 JSON 产物，与标准证据链进行对比。

工作经历

中国移动 | AI 训练工程师实习生2025.06 - 2025.08

中泰翻译模型微调与评测中国，上海

技术栈：Python、PyTorch、Qwen3-8B、LLaMAFactory、LoRA

- 使用 LLaMAFactory LoRA 与整理后的中泰语料微调 Qwen3-8B，使项目评测中的 BLEU 类指标从十几分提升至二十几分。
- 对比基线与微调模型输出，定位并修正佛历与公历年份转换中的重复性翻译错误。

ISS Tech | 软件开发实习生2024.06 - 2024.08

GPU 加速实时视频压缩中国，郑州

技术栈：FFmpeg、C/C++、Linux、Bash、A100 GPU、Android Studio

- 调优 FFmpeg 编码器参数，将 A100 的实时视频压缩能力从 12 路提升至 24 路并发。
- 使用 C/C++、Linux 与 Bash 开发 GPU 辅助流媒体处理流程，平衡压缩吞吐量、播放响应速度与画质。

技术能力

Agent 系统：多智能体编排、持久化会话状态、限定范围的工作区、人在回路控制

大模型工程：MCP、Agentic OCR、RAG、提示词与上下文工程、大模型评测

模型训练：模型微调、LoRA、PyTorch、LLaMAFactory、NLP 评测、离线强化学习

软件工程：Python、TypeScript、Node.js、FastAPI、PostgreSQL、Redis、Celery、React

系统与交付：Docker、Linux、Git、AWS S3、全栈开发、CI/CD、AWS 无服务器架构、GitHub Actions、FFmpeg

语言：中文（母语）、英语（商务）、法语（入门）