

Xuanpu Zhang

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

EDUCATION

National University of Singapore | Singapore

Aug. 2026 - Jan. 2028

Master of Computing in Artificial Intelligence, Dissertation in Multi-Agent Security

- Conduct research on multi-agent security in Ng See Kiong's group at the NUS IDS.

McGill University | Montreal, Canada

Sept. 2022 - May 2026

Bachelor of Science in Computer Science - Artificial Intelligence; Minor in Mathematics

SELECTED PROJECTS

PPT-to-Notes Agent — Multimodal Study Platform

Jul. 2025 - Present

Founder; Sole full-stack AI product engineer

Technologies: React, TypeScript, Python, FastAPI, PostgreSQL, Celery, Redis, S3, Docker, POML

Project Summary: Process PPT / lecture recording and generate detailed notes with **grounded page / 30s video reference**

- Designed agent Harness workflow aims to generate high quality long context notes with grounded sources from lecture slides and/or lecture recording.
- Designed Agentic OCR with Gemini 2.5 Flash to convert PDF page images into structured JSON/Markdown, preserving text, image descriptions, tables, diagrams, and token metadata for downstream reasoning and synthesis.
- Built RAG over structured OCR, notes, frames, and transcripts to support understanding of 100+ page slide decks; returned grounded answers with accurate page-level references and source citation metadata.
- Deployed AWS serverless services for an AI study product serving 120 users, including 10 paid users, as of Aug. 2026.
- Built parallel, interruptible, resumable multi-model processing with per-page JSON checkpoints, corrupt-output recovery, failed-image tracking, and S3 synchronization.

HuiyuanClaw - Agent Orchestration Harness

Feb. 2026 - May 2026

Agent orchestration and developer-tools builder

Technologies: Node.js, JavaScript, WebSocket, MCP, JSONL, Git, Cloudflare Tunnel

Project Summary: Built an OpenClaw-like harness, driven by Codex / Claude Code underneath, that mimics its ability to assign tasks to sub-threads, **featuring high observability and controllable thread observation.**

- Redesigned an open-source project around explicit multi-agent orchestration, representing each delegated task as a persistent, inspectable conversation with assigned ownership and human control; made task decomposition and concurrent execution visible through a browser-based agent harness.
- Scoped MCP-spawned agent sessions to specific projects and folders, providing stable local instructions and smaller skill/tool surfaces; controlled each conversation as an explicit workspace unit to reduce unnecessary context use and keep delegated work tied to its assigned scope.
- Prototyped MCP-driven session control and cross-session chat/reply workflows in Jan. 2026, before the Codex app session control launch (May 2026) and Claude Code's standalone cross-session messaging release (Aug. 2026).
- Exposed live per-agent progress and cross-session chat/reply workflows in a browser interface, letting users inspect persistent conversation records and Git-aware changes, intervene directly, and interrupt individual runs during concurrent execution.
- Inspired by Claude Code best practices, introduced a Critic Agent that critically evaluates agent implementations across visual output, project CI/CD, and detailed user requirements, then has the agent keep running to fix the issues on its own.

JD-Tailored Resume Generation Agent

Apr. 2026 - Present

Project Summary: Built an agent that automatically generates formatted resumes from JDs, persisting users' project experience, skills, and other information to enable targeted resume generation and job screening.

- Built profiles of users' skills and past projects through in-depth follow-up questioning, creating detailed structured entries for each experience that users can manually edit and adjust.
- Generated resumes tailored to each JD with customizable, persistent formatting and prompt-adjustable writing style.
- Exported resumes as Markdown for users to manually adjust and refine.
- Used a Critic Agent to catch errors in the agent workflow, such as visual overlap, duplicated content, and results that fall short of expectations.
- Extensible to automatically screening suitable jobs from JDs and planning development for skill gaps.

alva: Home Design / Living Pain-Point Discovery Agent

Sept. 2026

Project Summary: Built a visual system that helps users design furniture and decor styles **starting from the floor plan**; used agent capabilities to gather user information and **provide customized pain-point and design-issue analysis**, culminating in **deliverable reports for both users and designers**.

- Applied Socratic questioning over a pre-built 70-question checklist, offering suggested options based on prior answers and the user profile so users make choices instead of writing open-ended answers.
- Used agent vision to model user-supplied floor plans into a foundational digital-twin home space, self-validating the result with topology and image-overlap checks and automatically fixing detected issues.
- Turned vague user requests into 3D furniture using MCP and the agent's own 3D modeling capabilities, with adjustable parameters such as color, size, and length.
- Reviewed newly added furniture against the user profile and resident needs (e.g., routines of elderly residents and children), flagging conflicts with their habits.
- Surfaced pain points users had not yet noticed, drawing on their profile and habits, a pre-built pain-point list, and agent reasoning.
- As team leader, led a 4-person team in building a collaborative production architecture with tickets as the smallest atomic unit for fast parallel development; implemented automatic ticket submission, assignment, claiming, and ownership signing. All code was produced through vibe coding.

1B LLM for Robot Agentic intention detection in Embodied AI

Jun. 2025 - Aug. 2025

Team lead; LLM/MCP system owner

Technologies: Raspberry Pi, Python, 1B LLM, MCP, ROS, STT/TTS

- Led a 3-person team building a voice-controlled robot agent on Raspberry Pi; owned the 1B LLM/MCP and ROS integration, initiating tasks within 3 seconds in internal demos.
- Designed the agent pipeline from speech-to-text and 1B LLM intent recognition to MCP routing, function-level commands, ROS execution, and text-to-speech feedback.
- Deployed and tuned a 1B LLM for intent recognition on Raspberry Pi, mapping user requests into function calls and validating behavior in simulation before physical execution.

Long-Context LLM Evaluation

Mar. 2026 - Apr. 2026

Experiment design lead; evaluation implementation owner

Technologies: Python, GPT-4o, JSON, POML, sequence alignment, precision/recall/F1

- Built a four-condition long-context evaluation that showed exact evidence-chain recovery falling to zero in the longest setting, exposing a clear GPT-4o reliability limit.
- Proposed and led the lecture-slide study, defining evidence-grounded logical-chain tasks with exact page-ID citation requirements.
- Implemented structured JSON/POML artifacts and sequence/set metrics including Needleman-Wunsch alignment, LCS recall, and precision/recall/F1.
- Built structured page/block JSON and POML prompts requiring exact source-page IDs; stored model outputs as JSON artifacts for comparison with gold evidence chains.

WORK EXPERIENCE

China Mobile | AI Training Engineer Intern

Jun. 2025 - Aug. 2025

Chinese-Thai translation model fine-tuning and evaluation

Shanghai, China

Technologies: Python, PyTorch, Qwen3-8B, LLaMAFactory, LoRA

- Fine-tuned Qwen3-8B with LLaMAFactory LoRA and curated Chinese-Thai data, raising BLEU-style scores from the teens to the twenties in project evaluations.
- Compared baseline and adapted model outputs; diagnosed and corrected recurring Buddhist-era versus Gregorian year translation errors.

ISS Tech | Software Developer Intern

Jun. 2024 - Aug. 2024

GPU-accelerated real-time video compression

Zhengzhou, China

Technologies: FFmpeg, C/C++, Linux, Bash, A100 GPU, Android Studio

- Doubled A100 real-time video compression capacity from 12 to 24 concurrent streams by tuning FFmpeg encoder parameters.
- Developed GPU-assisted streaming workflows in C/C++, Linux, and Bash, balancing compression throughput, playback responsiveness, and visual quality.

TECHNICAL COMPETENCY

Agent Systems: Multi-agent orchestration, Persistent conversation state, Scoped workspaces, Human-in-the-loop control

LLM Engineering: MCP, Agentic OCR, RAG, Prompt/context engineering, LLM evaluation

LLM Training: Fine-tuning, LoRA, PyTorch, LLaMAFactory, NLP evaluation, Offline reinforcement learning

Software Engineering: Python, TypeScript, Node.js, FastAPI, PostgreSQL, Redis, Celery, React

Systems and Delivery: Docker, Linux, Git, AWS S3, Full stack development, CI/CD, AWS Serverless, GitHub Actions, FFmpeg

Languages: Chinese (Native), English (Business), French (Beginner)