Henry Hsu

Houston, TX 77030
+1-510-996-6837 hsuhengjui@gmail.com linkedin.com/in/yarikama github.com/yarikama

SUMMARY

Former Generative AI Team Lead with experience building production Generative AI platforms that doubled the B2B client base and grew the user base to 20K. Built key product features including Agentic AI Systems, Advanced RAG, and Backend infrastructure, leading to two major industry awards. Now pursuing a Master of Computer Science at Rice University.

EDUCATION

Rice University - Top 20 U.S. University **Aug. 2025 – Dec. 2026 (Expected)**
M.C.S., Computer Science *Houston, TX*

National Yang Ming Chiao Tung University (NYCU) - Top 3 Taiwan University **Sep. 2020 – Jun. 2024**
Minor: Computer Science (Domain GPA: 4.13/4.30) *Hsinchu, Taiwan*
B.S., Industrial Engineering and Management (GPA: 4.07/4.30)

WORK EXPERIENCE

Generative AI Team Lead **Sep. 2024 – Aug. 2025**
MaiAgent Co., Ltd. – Award-Winning B2B GenAI Startup *Taipei, Taiwan*

- **Building AI Agents –** Upgraded chatbots into AI agents by integrating memory systems, tool APIs, and MCP Client, resulting in **120%** partner growth (**CTBC Bank**, **MSI**, **HPE**, **iGroup**) and **567%** user expansion (**3K** to **20K** users) while cutting LLM token usage by **67%+**
- **Scaling Advanced RAG –** Refactored advanced RAG system with async framework and dual-database synchronization (RDB/Vector DB), scaling the system from **3M** to **20M+** text chunks while maintaining API response times and **doubling** indexing performance
- **Contributing to Open Source –** Contributed **14** merged pull requests to the **LlamaIndex** open-source project (**15K+ stars**), fixing bugs and adding features in integrations with **AWS Bedrock**, **Claude**, **Elasticsearch**, **Cohere**, **OpenAI**, **MCP Client**, and **AI Agent Workflow**
- **Testing Framework –** Introduced pytest testing framework for unit and E2E coverage to prevent production regressions, achieving **67%** test coverage from **zero** baseline, which reduced production hotfixes by **90%** initially and sustaining a **50%** reduction long term
- **Enterprise Knowledge System:** Built a semantic tagging and permission management multi-knowledge base system for B2B clients

PROJECTS

Agentic Hybrid RAG | *Graph RAG (Neo4j, Cypher), Vector RAG (Milvus), Multi-Agent (LangGraph)* **Feb. 2024 – Aug. 2024**

- **Self-Routing Agentic RAG:** Implemented a query classification agentic AI system that picks a global or local approach to request slicing from Graph RAG, Vector RAG or a Hybrid (Graph & Vector) RAG path to answer the contextual or detail-oriented queries
- **ML Clustering:** Applied ML clustering algorithms (K-means and DBSCAN) in recognizing pivotal nodes from vector database to build a low-cost Graph RAG; saved **35%** of indexing token cost, with only 2.56% worse in RAGAS metrics
- **Advanced Hybrid RAG:** Enhanced vector RAG hit rate using data retrieved from Graph RAG with HyDE, balancing both global and detailed information for more accurate responses. Achieved an MAR@10 of **88.2%** on multi-hop datasets

GenAI Tutoring System with Multi-Agent RAG | *LangChain, CrewAI, Streamlit, Chroma* **Feb. 2024 – Jul. 2024**

- **Multi-Agent Framework:** Utilized **CrewAI** to assign roles to LLMs for collaborative problem-solving, assisting students with problem-solving, exam detection, concept analysis, and compared open-source LLMs (Llama 3) and commercial LLMs (GPT-4)
- **Interactive UI:** Enabled user tuning of LLM settings and inspection of chain-of-thought from ReAct Agent and citations by **Streamlit**
- **Competition Achievement: Top 3** of 50 teams in the Computer Science AI workshop, winning the Outstanding Award

AI Fitness Motion Classification | *Keras, scikit-learn, MediaPipe, OpenCV, NumPy* **Feb. 2024 – Mar. 2024**

- **Video Data Extraction:** Extracted temporal 33 human skeleton keypoints and body angles from 24 videos by **MediaPipe**
- **ML/DL Model Training:** Trained and compared ML/DL models (LSTM & Random Forest) to classify fitness motions in videos, including squats, bench presses, and deadlifts by 33 human skeleton keypoints, achieving an accuracy of **80.52%** & **88.31%**, respectively

AWARDS

Presidential Hackathon Winner (2024) - Urban noise detection with LLM-powered structured analysis
23rd Golden Peak Award (2025) - Outstanding Commercial Product, MaiAgent AI Platform
Two-Time Dean's List Recipient - Top 5% Academic Performance, NYCU
Atona Case Competition Finalist (Top 1%) - National enterprise transformation competition
AI Workshop Outstanding Award (Top 3/50) - Multi-Agent RAG tutoring system, NYCU CS

SKILLS

- **Languages:** Python, C/C++, TypeScript | **Databases:** SQL (PostgreSQL, SQLite), NoSQL (Neo4j, Milvus, Elasticsearch, Chroma)
- **GenAI:** LlamaIndex, LangChain, LangGraph, MCP | **Backend:** Django, FastAPI, Nginx | **Frontend:** HTML, CSS, JavaScript, React
- **Tools:** git, Docker, Shell Script | **Methodologies:** Agile, Scrum | **AWS:** Bedrock, EC2, S3, RDS, ElastiCache | **AI/ML:** PyTorch, scikit-learn