

# HENRY HSU

Houston, TX 77030

+1-510-996-6837   [hsuhengjui@gmail.com](mailto:hsuhengjui@gmail.com)   [linkedin.com/in/yarikama](https://www.linkedin.com/in/yarikama)   [github.com/yarikama](https://github.com/yarikama)

## SUMMARY

A former GenAI Team Lead experience in building production GenAI platforms, promoted from Intern to Team Lead within a year with doubling the B2B client base. I 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

M.C.S., Computer Science

Aug. 2025 – Dec. 2026 (Expected)

Houston, TX

### National Yang Ming Chiao Tung University (NYCU) - Top 3 Taiwan University

B.S., Industrial Engineering and Management (GPA: 4.07/4.30)

Minor: Computer Science (Domain GPA: 4.13/4.30)

Sep. 2020 – Jun. 2024

Hsinchu, Taiwan

## WORK EXPERIENCE

### GenAI Team Lead

MaiAgent Co., Ltd. – Award-Winning B2B GenAI Startup

Sep. 2024 – Aug. 2025

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 RAG (Retrieval-Augmented Generation) architecture with async framework and dual-database synchronization (RDB/Vector DB), enabling tag-based filtering and multi-knowledge base management; scaled 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, fixing bugs and adding features in integrations with **AWS Bedrock**, **Claude**, **Elasticsearch**, **Cohere**, **OpenAI**, **MCP Client**, **PostgreSQL**, 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

## PROJECTS

### Agentic Hybrid RAG | Graph RAG (Neo4j, Cypher), Vector RAG (Milvus), Multi-Agent (LangGraph)

Feb. 2024 – Aug. 2024

- Implemented a query classification agentic AI system that picks a global or local approach to request slicing from Graph RAG, Vector RAG (hybrid embedding search) or a Hybrid (Graph & Vector) RAG path to answer the contextual or detail-oriented queries
- Used ML clustering algorithms (K-means and DBSCAN) to recognize pivotal nodes of the vector database, and built a low-cost Graph RAG; saved 35% of semantic graph indexing token cost, with only 2.56% worse on average in RAGAS metrics
- 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

- Enabled user tuning of LLM settings and inspection of chain-of-thought from ReAct Agent and fetched chunks by Streamlit
- Used 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)
- 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

- Trained and compared different models to classify fitness motions in videos, including squats, bench presses, and deadlifts by 33 human skeleton keypoints extracted by MediaPipe and body angles derived from the keypoints
- Used ML/DL models (LSTM and Random Forest) to analyze the data, 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:** C/C++, Python, TypeScript | **Databases:** SQL (PostgreSQL, SQLite), NoSQL (Neo4j, Milvus, Elasticsearch, Chroma)
- **Generative AI:** LangChain, LangGraph, LlamaIndex, MCP | **Backend:** Django, FastAPI | **Machine Learning:** PyTorch, scikit-learn
- **AWS:** Bedrock, EC2, S3, RDS, ElastiCache | **Agile, Scrum** | **Unix:** Linux, FreeBSD | **Tools:** git, Docker, Shell Script (Bash)