

# HENRY (HENG JUI) 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)

## EDUCATION

### Rice University

M.C.S., Computer Science

Aug. 2025 – Present

Houston, TX

### National Yang Ming Chiao Tung University (NYCU)

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

Sep. 2020 – Jun. 2024

Hsinchu, Taiwan

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

## WORK EXPERIENCE

### AI Engineer *MaiAgent Co., Ltd. – Generative AI Startup*

Sep. 2024 – Aug. 2025

- Transformed chatbots into agents by integrating memory systems, tool APIs, and MCP Client, resulting in **120%** partner growth (MSI, HPE, iGroup) and **567%** user expansion (**3K** to **20K** users) while optimizing LLM token usage to **67%+** efficiency
- Refactored and optimized RAG architecture with async framework and dual-database synchronization (RDB/Vector DB), enabling tag-based filtering and multi-knowledge base management, which scaled the system from **3M** to **20M+** files while maintaining API response times and doubling indexing performance
- Introduced testing framework using pytest for unit and E2E coverage, achieving 67% test coverage from zero baseline, which reduced production hotfixes by 90% initially and sustaining a 50% reduction long term
- Contributed 14 merged pull requests to the LlamaIndex open-source project, fixing bugs and adding features in integrations with Bedrock Converse, Elasticsearch, Cohere, OpenAI, MCP Client, PostgreSQL, and Agent Workflow, enhancing the community ecosystem

## PROJECTS

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

Feb. 2024 – Jul. 2024

- Implemented a query classification agentic (GPT-4o-mini) system that will pick a global or local approach to request slicing from Graph RAG, Vector RAG (hybrid embedding search) or a Hybrid (Graph & Vector) RAG path
- Used K-means and DBSCAN to recognize pivotal nodes of the vector database, building a low-cost Graph RAG. Saved 35% of build token cost, with only 2.56% worse on average in RAGAS
- 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

### Problem-Solving System with Multi-Agent RAG | *LangChain, CrewAI, Streamlit, Chroma*

Feb. 2024 – Jul. 2024

- Utilized LangChain and Chroma to apply a multi-agentic RAG framework for high school student question answering
- Enabled Streamlit-based user tuning of LLM settings and inspection of chain-of-thought and fetched chunks
- Used CrewAI to assign roles to LLMs for collaborative problem-solving, helping out 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

### Fitness Motion Classification | *Keras, scikit-learn, MediaPipe, OpenCV, NumPy*

Feb. 2024 – Mar. 2024

- Used LSTM and Random Forest to analyze the data, achieving an accuracy of 80.52% and 88.31%, respectively

### Chat Bar - An Online MUD Game with Group Chatting Utility | *C/C++ (SFML), SQL*

Oct. 2023 – Jan. 2024

- Developed a real-time multiplayer role-playing chat game using socket programming with server & client architecture
- Implemented login, registration, and ranking features for user management with SHA-256 hashing, JSON, and MySQL

## 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

C/C++, Python | SQL (PostgreSQL), NoSQL (Neo4j, Milvus, Elasticsearch) | Unix (Linux, FreeBSD), Git, Docker, Shell Script (Bash) |  
Generative AI (LangChain, LangGraph, LlamaIndex, MCP) | Backend (Django, DRF, Celery) | Machine Learning (PyTorch, scikit-learn) | AWS  
(Bedrock, EC2, S3, RDS, ElastiCache) | Agile, Scrum