Yuheng Li

(323) 998-3656 | liyuheng0830@gmail.com | https://github.com/Yuhengli77

EDUCATION

University of California, San Diego — **San Diego, CA**
Master of Science in Computer Science — *Expected Dec 2026*
University of California, Los Angeles — **Los Angeles, CA**
Bachelor of Science in Mathematics of Computation — *June 2025*

SKILLS

Programming Languages: Python, C, C++, SQL, Java, MATLAB, Bash
Frameworks & Tools: PyTorch, LangChain, LangGraph, scikit-learn, Pandas, Git, Fast API, AWS
Data & Storage: MySQL, PostgreSQL, MongoDB, Faiss
Other: Large Language Model, NLP, AI Agent, Recommender System, LLM Distributed Training & Serving

EXPERIENCE

Advance.AI — **Singapore**
Algorithm Intern (Applied Machine Learning) — *Jun 2025 – Sep 2025*

- Designed and implemented an **OCR + LLM** pipeline in **Python** to automatically annotate ID document layouts across any country or format, integrating post-processing, and **Prompt Engineering** to ensure high labeling accuracy, cutting annotation time from **1-2days** to under **1 hour**.
- Developed and trained layout models in **PyTorch** using **generated dataset**, improving recognition accuracy to **99.1%** across diverse ID documents, reducing maintenance costs, and eliminating manual review.
- Iteratively optimized the **end-to-end pipeline**, including data collection, cleaning, model training and evaluation, enabling the company to achieve **100% self-developed** algorithms for its major client's ID document processing.

Goldstate Securities Co., Ltd. — **Shenzhen, China**
Data Scientist Intern — *Jul 2024 – Sep 2024*

- Developed custom stock volume and price indicators using **Python** and **Pandas** to detect abnormal trading behaviors and uncover market patterns, improving the accuracy of short-term analysis.
- Built Python modules with Pandas and **Seaborn** to process and visualize large-scale trading data, enabling dynamic stock screening based on volume-price relationships and enhancing market insight generation.
- Designed an interactive GUI tool allowing users to upload portfolio data, configure take-profit and stop-loss thresholds, and receive action recommendations through **LLM integration** and automated daily reporting.

PROJECTS

Two-Stage Sequential Recommender System | *Machine Learning, Deep Learning* — *Oct 2025 - Present*

- Designed and implemented a **two-stage recommendation** framework using KuaiSAR dataset, integrating a **DSSM-based recall** stage with a **Transformer ranking** stage to balance system efficiency and precision.
- Developed and benchmarked **SASRec ranking model** to capture long-range sequential dependencies in user behavior, achieving a **100%** improvement in **Hit Rate@50** compared to ItemCF and NeuMF baselines.
- Engineered robust training data using **negative sampling** techniques and addressed **cold-start challenges** by incorporating advanced feature construction, effectively mitigating popularity bias.

LLM-Driven Generative Engine Optimization | *UCSD Research Project* — *Sep 2025 – Dec 2025*

- Designed and implemented a **multi-agent** optimization workflow using **Gemini** and **LangChain** to iteratively optimize content, enhancing its **visibility and citation** likelihood in Generative Engine responses.
- Architected an automated Analyst-Editor agent loop that diagnoses content weaknesses based on 6 key metrics and synthesizes targeted editing strategies to improve retrieval ranking.
- Bulit a comprehensive **evaluation framework** using **LLM-as-a-judge** to benchmark optimization success, tracking metrics across 5 iterations to ensure content integrity while maximizing visibility.

RAG-powered Real Estate Search Assistant | *LLM & RAG System* — *Jun 2025 – Sep 2025*

- Developed an **RAG**-powered real estate assistant with **LangGraph** that reduced search effort by implementing a two-stage retrieval process, ensuring users always get relevant recommendations even without exact matches.
- Implemented robust tool-use capabilities that convert natural language requirements into **SQL filters & vector search** queries for property retrieval and applied **prompt engineering** to handle edge cases.
- Designed a multi-turn conversational **agent workflow** that correctly identifies user intent, invokes tools, and maintains context, delivering a seamless and intuitive interaction experience.