Yuheng Li

+1 323 998 3656 | liyuheng0830@gmail.com | LinkedIn | GitHub

Education

University of California, San Diego — La Jolla, 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 — *Jun. 2025*

Skills

Programming Languages: Python, C/C++, Java, MATLAB, Bash
Frameworks & Tools: PyTorch, HuggingFace, LangGraph, scikit-learn, Git, FastAPI, Docker, AWS
Data & Storage: MySQL, PostgreSQL, MongoDB, Faiss
Other: Large Language Models, Natural Language Processing, AI Agents, Recommender Systems

Experience

Advance.AI — Singapore
Machine Learning Engineer Intern — *Jun. 2025 – Sep. 2025*

- Addressed the scarcity of field-level annotations for ID documents by designing an **LLM-driven** pipeline that converts OCR outputs into structured training data, scaling labeled samples to **50K+** with accuracy exceeding **98%**.
- Fine-tuned a **multimodal** LayoutLM in **PyTorch** by integrating visual patch embeddings with textual signals; leveraged image context to resolve OCR ambiguities, increasing field-level **F1-score by 11%**.
- Scaled the pipeline to diverse document layouts by adapting prompts for new formats and implementing post-processing quality filters to validate LLM's outputs, expanding layout coverage by ∼**50%** at ∼**95%** less annotation cost.
- Containerized the optimal model as a **FastAPI** service using **Docker**; conducted comprehensive integration testing to validate model robustness and system stability for real-time identity verification.

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

- Developed a **LSTM** model to predict market trends, synthesizing features from pricing and fundamental indicators to achieve a **13%** reduction in RMSE compared to previous baselines.
- Designed a walk-forward **validation** workflow to validate algorithm performance across 2 years of data, implementing automated evaluation scripts to verify model stability and robustness before production deployment.
- Built an LLM-driven automated reporting pipeline to analyze portfolio holdings, integrating market data with static company profiles to generate daily strategy reports, reducing manual effort by ∼**90%** in pilot testing.

Research

LLM-Driven Generative Engine Optimization | *Prof. Yiying Zhang* — *Oct. 2025 – Jan. 2026*

- Co-authored **SourceBench:** Can AI Answers Reference Quality Web Sources? ***(ICML** 2026 Under Review)* [arXiv]
- Implemented a workflow using Gemini and **LangChain** to iteratively refine web content, enhancing visibility and citation likelihood in Generative Engine responses.
- Designed an evaluation framework using **LLM-as-a-judge** to track optimization effects, achieving a ∼**12%** uplift in ranking metrics while maintaining >**0.95** content integrity.

Projects

Two-Stage Sequential Recommender System | PyTorch, SASRec, DSSM — *Oct. 2025 – Dec. 2025*

- Designed and implemented a two-stage recommendation framework using the KuaiSAR dataset, integrating a **DSSM-based** recall stage with a **Transformer** ranking stage to balance system efficiency and precision.
- Developed and benchmarked a **SASRec** ranking model to capture long-range sequential dependencies in user behavior, achieving a **2**× improvement in **Hit Rate@50** compared to ItemCF and NeuMF baselines.

RAG-powered Real Estate Search Assistant | LangGraph, RAG, SQL, Vector Search — *Jun. 2025 – Sep. 2025*

- Developed a **RAG**-powered real estate assistant with **LangGraph** that reduced search effort by implementing a **two-stage** retrieval process, ensuring users receive relevant recommendations even without exact matches.
- Implemented robust **tool-use** capabilities that convert natural language queries into **SQL** filters and **vector** search queries for property retrieval, leveraging prompt engineering to identify user intent and invoke tools.