Xianhan Huang

@ xianhanh@usc.edu | LinkedIn | GitHub | Los Angeles, United States

Education

University of Southern California — Los Angeles, United States
M.S. in Electrical and Computer Engineering; ***GPA 3.6/4.0*** — Sep 2025 – May 2027

- **Coursework:** Data Structures and Algorithms, Web Technologies, LLMs, NLP, Deep Learning, Machine Learning

University College Dublin, Beijing University of Technology — Beijing, China
B.S. in Internet of Things Engineering; ***GPA 3.6/4.0*** — Sep 2021 – Jun 2025

- **Coursework:** Object-Oriented Design, Databases, Operating Systems, Cloud Computing

Skills

Languages: Python, Java, C++, JavaScript, TypeScript, SQL, HTML/CSS, C#

Tools & Frameworks: React.js, Node.js, FastAPI, vLLM, Flask, REST APIs, PostgreSQL, MySQL, Git, Linux, Docker, Kubernetes, AWS, GitLab CI, GitHub Actions, Prometheus, Grafana, PyTorch, XGBoost

Methods & Applications: OOD, Microservices, CI/CD, Unit & Integration Testing, Agentic AI, RAG, Tool Use, SSE, Vector Search, LLM Evaluation, Model Serving, Cloud Infrastructure, Data Visualization

Professional Experience

Nokia — Dallas Fort Worth, United States
Cloud AI-RAN R&D Engineering Intern — Jun 2026 – Aug 2026

- Built and deployed a React-based ML dashboard as 4 microservices on Kubernetes with REST APIs for authentication, real-time ML inference, LLM-powered scaling guidance, and Grafana observability.
- Built an end-to-end GPU telemetry pipeline with Prometheus and NVIDIA DCGM on a Kubernetes cluster, enabling real-time monitoring in Cloud RAN infrastructure.
- Implemented 305 unit and integration tests with pytest, debugged application workflows on Linux, and automated deployment through CI/CD; benchmarked and tuned XGBoost and N-HiTS, reducing MAE by over 90%.

Institute of Software, Chinese Academy of Sciences — Beijing, China
Computer Vision Research Intern — Jul 2023 – Apr 2024

- Engineered an attention mechanism for Masked Autoencoder in PyTorch, improving reconstruction MSE by 3% and strengthening encoder feature extraction capability under limited visibility.
- Built a Pandas/NumPy analysis pipeline linking texture complexity to reconstruction error, guiding model optimization; first-author paper at ICASSP 2024.

Project Experience

Full-Stack Document Q&A Agent with Tool-Use — Aug 2026 – Present

- Built a FastAPI backend that streams a multi-round tool-use agent in real time over Server-Sent Events, answering questions over 5,049 Qasper QA pairs with a self-hosted LLM served via vLLM.
- Chose BM25 over hybrid retrieval for 12× lower latency with no significant F1 loss.
- Built the RLVR/GRPO post-training pipeline and trained Qwen2.5 models on a SLURM GPU cluster, improving answer F1 by 65% and raising the answered rate from 62% to 98%.

Conversational Movie Recommendation Agent — Aug 2026 – Present

- Led a 6-member team building a conversational recommendation agent that retrieves 50 of 21K+ movies via vector search and recommends 10 with explanations, storing conversation history in an 8-table PostgreSQL schema.
- Built a FastAPI service and React/TypeScript UI; added admission control after Locust load tests exposed DB connection-pool exhaustion, keeping the API responsive at 300 users (190 req/s, excl. LLM).

TrojanPlanner: AI Course Planner (AI@USC Hackathon) — Nov 2026

- Led a 3-member team using Python and Google OR-Tools to build an AI course planner that turns student preferences into schedules, with 14 ms p95 schedule computation time across 1,000 test scenarios.
- Built an AI assistant with Node.js/Express and the Claude API to convert natural language into scheduling rules, and extended the planner to offer distinct schedules, with optimal top picks in all 500 feasible tests.