Zhiheng Jiang

310-709-5870 zhjiang@ucla.edu linkedin.com/jiangzhiheng github.com/jzh001

Education

University of California, Los Angeles Sep 2024 - June 2028
Bachelor of Science in Computer Science and Engineering (GPA: 4.00 / 4.00) *Los Angeles, CA*

- **Courses:** Data Structures and Algorithms, Software Construction, Multivariable Calculus, Differential Equations, Discrete Math
- **Awards:** 2025 Samueli Undergraduate Scholarship Recipient, Dean's Honors List
- **Work Authorization:** F1 Student Visa with STEM CPT/OPT — Eligible for **H-1B1** (faster and easier to obtain than H-1B)

Experience

Institute for Creative Technologies, University of Southern California June 2025 – Sep 2025
Visiting Academic *Playa Vista, CA*

- Led a reinforcement learning research project, partnering with U.S. DEVCOM Army Research Lab, advised by Dr. Volkan Ustun
- Designed benchmark problems for Preference-Driven Multi-Objective Reinforcement Learning with novel evaluation metrics
- Implemented an end-to-end training and evaluation pipeline with caching, multiprocessing and efficient tensor manipulation
- Designed Heterogeneous Graph Neural Network-based knowledge distillation framework for Proximal Policy Optimization (PPO)
- Authored a Python toolkit (GraphAllocBench) using Gymnasium, Stable Baselines3, PyTorch Geometric, Wandb and and PyMOO

Structures-Computer Interaction Lab at UCLA Oct 2024 – June 2025
Undergraduate Researcher *Los Angeles, CA*

- Developed digital twin simulation of springs using Large Language Models and Robot Tool Calling, advised by Prof. Khalid Jawed
- Worked with scientist from Amazon AWS to develop multi-agent, multimodal LLM-robot workflows with memory and reasoning
- Developed a LangGraph clone using Anthropic API from scratch for more customizable LLM memory for multi-agent systems
- Created LLM agentic pipelines for Physics-informed Neural Architecture Search for 2D Reduced-Order Beam Simulations
- Developed LLM workflows with OpenGL physics simulations, Intel RealSense video input (OpenCV) and Sawyer Robots (ROS)

Institute of High Performance Computing, A*STAR June 2020 – Feb 2022
Research Assistant *Singapore*

- First-author publication on music genre communities with Dr. Hoai Nguyen Huynh, cited by Spotify Research (2024)
- Performed K-means, Girvan-Newman and Shannon Entropy clustering on scale-free user-oriented networks with NetworkX
- Improved community detection algorithms on large social networks parsed with BeautifulSoup from an AJAX review database
- Performed text mining using Natural Language Tool Kit (NLTK), TF-IDF and dependency parsing to describe genre communities

Publications and Talks

[1] **Zhiheng Jiang**, Yunzhe Wang and Volkan Ustun, "GraphAllocBench: A Flexible Preference-Driven Multi-Objective Reinforcement Learning Benchmark" *Pending submission to ML Conference. Presented at SoCal CS REU Symposium 2025 (Harvey Mudd College).*

[2] Mason Zhao, **Zhiheng Jiang**, Henry Braid and M. Khalid Jawed, "LLM-Guided Model Development of Elastic Structures" *UCLA Undergraduate Research Week 2025 (Presenters: Zhiheng Jiang and Henry Braid).*

[3] **Zhiheng Jiang** and Hoai Nguyen Huynh, "Unveiling music genre structure through common-interest communities" *Social Network Analysis and Mining*, Vol. 12, No. 35 (2022).

Projects

Neural Network Diagram Generator with Agentic RAG | *LangChain, LlamaIndex, Pydantic, Docker, Gradio, Ollama, PostgreSQL* 2025

- Created a full-stack containerized Agentic RAG Gradio application to generate professional neural network model diagrams
- Designed multi-agent LLM-RAG workflows with Anthropic and Gemini API, and OpenAI gpt-oss-20B locally with Ollama
- Open-sourced application on HuggingFace Spaces, and hosted text embeddings on Supabase PostgreSQL vector database

When2Fly Scheduling Web App | *ReactJS, ExpressJS, CI/CD, PostgreSQL, Git, GitHub Actions, Vercel, Render* 2025

- Designed a full-stack web app for Software Construction Lab Final Project, to connect students for Uber ride-sharing to LAX
- Developed comprehensive backend integration tests using Jest and Supertest to validate REST API endpoints (including authentication, CRUD and time-based queries), integrating them into a CI/CD pipeline for automated testing and deployment.
- Ensured robust foreign key and data integrity constraints in a PostgreSQL environment through automated testing.

Today I Learnt AI Competition –Advanced Category Champion | *LLMs, VLMs, Finetuning, PyTorch, HuggingFace, VertexAI, Docker* 2024

- Team leader of champion team, winning 10,000 SGD (7,500 USD) cash prize competing against 60 university-level finalist teams
- Finetuned large deep learning models with high test scores, for audio (99.5%), Vision-Language Models and object detection (86.3%) and Transformer Question Answering (99.9%), with quantization, to achieve high inference speeds on a DJI robot
- Finetuned SOTA models such as YOLO, DETR, RoBERTa and OpenAI Whisper on VertexAI and Google Cloud Platform (GCP)

Technical Skills

Languages: Python, C++, MATLAB, Java, Javascript, SQL, LaTeX, Shell Scripting (Bash), YAML, JSON, HTML
Technologies: PyTorch, HuggingFace, LangGraph, LangChain, LlamaIndex, Pandas, Numpy, ReactJS, NodeJS, FastAPI, GitHub Copilot
Concepts: Large Language Models, Machine Learning, Computer Vision, Natural Language Processing, Reinforcement Learning