Zhiheng Jiang

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

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:** Machine Learning, Linear Algebra, Probability and Statistics, Multivariable Calculus, Data Structures and Algorithms, Operating Systems, Digital Design
- **Awards:** 2025 UCLA Samueli Undergraduate Scholarship Recipient, Dean's Honors List

Experience

UCLA Computational Machine Learning Lab Oct 2025 – Present
Undergraduate Researcher *Los Angeles, CA*

- Working under Prof. Cho-Jui Hsieh to explore methods for LLM compression and denoising with Singular Value Decomposition
- Benchmarked layer-adaptive SVD-based LLM model compression approaches on factual+logical reasoning and safety alignment
- Devised novel finetuning-free LLM layer editing method which outperformed existing SVD-based denoising approaches
- Performed subspace clustering of LLM weight matrices using gradient-based heuristics to improve post-training SVD reduction

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

- First author in reinforcement learning paper, 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
- Designed Heterogeneous Graph Neural Network with flexible preference conditioning for Proximal Policy Optimization (PPO)
- Authored a Python toolkit (GraphAllocBench) using Gymnasium, Stable Baselines3, PyTorch Geometric, Wandb and PyMOO
- Pending review at top Machine Learning conference, presented at SoCal CS REU Symposium (Harvey Mudd College)

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
- Refined numerical optimization methods using PyTorch Differential Equations (torchdiffeq) for reduced-order spring simulations
- Trained Physics-Informed Neural Networks with Autograd and Discrete Differential Geometry to predict spring elastic energy
- Worked with scientist from Amazon AWS to develop LangGraph clone from scratch to manage LLM multi-agent LLM memory
- Developed LLM tool wrappers for 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 journal publication on large social network analysis 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

[1] **Zhiheng Jiang**, Yunzhe Wang, Ryan Marr, Ellen Novoseller, Benjamin T. Files and Volkan Ustun, "GraphAllocBench: A Flexible Benchmark for Preference-Conditioned Multi-Objective Policy Learning," *arXiv preprint arXiv:2601.20753*, 2026.

[2] **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

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