

Junfan Zhu

📍 Santa Clara, CA, 95051 ✉ junfanzhu98@gmail.com 📞 773-236-6422 [in junfan-zhu](#) [🌐 junfanz1 \(520+★\)](#)

AI Researcher, LLM/ML Engineer. | CS, Math, Finance: Versatile Problem-Solving Blend. | Traverler, Marathoner.
I design cutting-edge LLM systems pipelines, focusing on scalable architectures and multimodal agentic reasoning.

Experience

AutoX

San Jose, CA

AI/Machine Learning Engineer – LLM, VLM/VLA, Agents, Reasoning

Jul 2025 – Present

· Advancing autonomous intelligence by building next-gen reasoning agents and multimodal foundation models.

Societe Generale Corporate & Investment Banking

Chicago, IL

Quantitative Research Analyst – Machine Learning

Mar 2021 – Dec 2024

Summer Intern – Data Scientist

Jun 2020 – Aug 2020

- Led AI initiatives developing reflection-enabled Quant Agent, fine-tuned via QLoRA and contrastive learning, built market digital twins for robustness, boosting retrieval accuracy by 25%, recognized as “Hero of the Week”.
- Pioneered multimodal Agentic HybridRAG pipeline with CrewAI, optimized distilled LLM with adaptive retrieval, enabling alt-data flywheel and boosting E2E system efficiency by 25%, cutting hallucination by 27%.
- Designed GRPO-style reward system with ROUGE/BERTScore evaluation and LLM-as-a-judge, aligning generation with safety guardrails, enhancing factual grounding, reducing False Positive by 35%.
- Architected scalable small-big LLM inference pipeline, streamlined across teams with cross-functional impact.
- Spearheaded hybrid batch-stream pipeline for high-freq volatility prediction, with interpretable ML (XgBoost, VAE, ensemble, Kolmogorov-Arnold Networks), drift detection & retraining automation, lifting accuracy 15%.
- Optimized large-scale real-time portfolio analytics infra, driving +\$2.6M annual revenues, 20% efficiency boost.
- Championed mentorship and collaboration among 50+ teammates across researchers, PMs and Leads globally.

AI Projects & Open Source Contributions

Multi-Agent Orchestration – LangGraph Reflection Researcher

[LangGraph Github](#), [RAG](#)

- Engineered LangGraph-based multi-agent system with self-reflection and retrieval-grounded alignment; integrated LangSmith trace for reasoning introspection, cutting hallucination 40% with iterative expert routing.

DeepSeek R-1 Mixture-of-Experts (MoE) in PyTorch

[MoE Github](#)

- Implemented scalable 8-expert MoE model with top-k routing, expert load balancing, and capacity-aware gating; enabled parallel sparse activation and DeepSeek-R1-style distributed training scalability.

MCP Multi-Server – Interoperable Agent2Agent LangGraph AI System

[MCP Github](#)

DeepSeek V-3 Multi-Head Latent Attention (MLA) in PyTorch

[MLA Github](#)

Cursor Full-Stack Vibe Engineering – E2E Micro-SaaS App with LangChain

[Cursor Github](#)

Crypto Blockchain Full-Stack Decentralized Platform (Flask, React)

[Blockchain Github](#)

2025 NVIDIA GTC Conference – Technical & Industrial Insight

[GTC Github \(160+★\)](#)

Education

Georgia Institute of Technology

Atlanta, GA

M.S. Computer Science (Computer Systems, Machine Learning)

Jan 2021 – May 2023

University of Chicago

Chicago, IL

M.S. Financial Mathematics (Statistics, Computational Finance)

Sept 2019 – Dec 2020

Dalian University of Technology

Dalian, China

B.Econ. Finance, Minor Mathematics | Outstanding Thesis (1%)

Sept 2015 – Jun 2019

Awards, Certificates & Skills

- 2nd Prize, Finalist, Asia Supercomputer Challenge (ASC). | C++, CUDA, Kubernetes, Docker, Parallelism.
- Meritorious Winner, International Mathematical Contest in Modeling (MCM). | MATLAB, Python, R, SQL.
- Top 10 Algo Trader, Rotman International Trading Competition (RITC). | Scikit-learn, Pandas, Pydantic.
- [Mt Kilimanjaro Climbing \(5895m\)](#). [Azure AI Engineer](#). [SocGen Cloud Expert](#). [Ray](#). [AI Infra](#). [Private Pilot](#).