Boqin Yuan

San Diego, CA | b4yuan@ucsd.edu | 217-991-2180 | boqin-yuan | boqiny

Summary

UCSD MSCS student specializing in AI/ML and software engineering. Machine Learning Engineer with 1 year of industry experience and a background spanning AI research and production-grade ML systems. My expertise spans **Machine Learning, ML Systems, LLM Applications and Agentic Framework**, focused on translating research into real-world deployments.

Education

University of California, San Diego *Sep 2025 – Jun 2027*
M.S. in Computer Science *(GPA: 4.0)*

University of Illinois at Urbana-Champaign *Aug 2020 – May 2024*
B.S. in Mathematics & Computer Science; B.S. in Statistics *(GPA: 3.9)*

- Graduate with **Highest** Distinction; James Scholarship & Deans List

Work Experience

Machine Learning Engineer *San Jose, CA*
CambioML (YC S23) *July 2024 – July 2025*

- Engineered and productionized **Anyparser**, a fine-tuned 1B & 2B **vision-language model** for parsing PDFs into structured Markdown (text, tables, charts). Fully fine-tuned and post-aligned with preference data to improve robustness, achieving **higher accuracy** than GPT-4 baselines. Optimized inference with **SGLang**, delivering **8x** throughput on L4 GPUs. Deployed as a SaaS on AWS using **ECS** + **Lambda**, with a **React** frontend and **DynamoDB** + **Cognito**.
- Orchestrated the design and deployment of **Energent.ai**, a **computer-use agent (CUA)** sandbox powered by Claude that autonomously executes diverse desktop tasks. Engineered a **multi-agent system** comprising data, web, and coding agents with orchestration, **long-term memory persistence**, and state management, enabling tool integration and MCP. Leveraged **Kubernetes** to provision isolated per-user sandbox VM sessions, scaling to support **1000**+ users worldwide.

Machine Learning Engineer Intern *Jinan, China*
Inspur Group *May 2023 – Aug 2023*

- Constructed and annotated a custom volleyball dataset and trained **YOLOv7-based object detection models** for real-time AI fitness assessment. Implemented and optimized **YOLO Pose keypoint detection** for athletic movement analysis (e.g., long jump scoring), and accelerated inference with **TensorRT**, reducing latency while maintaining high accuracy in challenging outdoor environments. Integrated models into the backend system in **C++**.

Research Experience

Research Assistant *San Diego, CA*
Advised by Prof. Jishen Zhao *Sep 2025 – Present*

- Conducting research on **multi-agent systems** and **agent memory architectures**, focusing on scalable agent coordination and long-horizon memory reasoning. Co-authored **PRO-V-R1** (DAC 2026 submission) on reasoning-enhanced programming agents for RTL verification, and currently leading a **memory-agent benchmark** project submitted to **ICML 2026**.

Researcher, NCSA SPIN Program *Champaign, IL*
National Center for Supercomputing Applications *Aug 2023 – May 2024*

- Worked with Professors **Kaiyu Guan** and **Sheng Wang** on geospatial ML research, pioneering the first application of the **Prithvi-100M foundation model** (IBM-NASA) for multi-temporal crop classification in Illinois, achieving **75% mean IoU**. Developed an **auto-labeling pipeline** using Gemini 1.5 Pro to distill ResNet-50 models for crop classification and residue regression, and built explainable ML models for tillage and harvest detection from Sentinel-2 imagery, reaching **80% precision/recall**. Presented results at **EGU 2024** (EGU24-14253).

Publications

PRO-V-R1: Reasoning Enhanced Programming Agent for RTL Verification *arXiv Preprint, 2025*
L-MARS: Legal Multi-Agent Workflow with Orchestrated Reasoning and Agentic Search *arXiv Preprint, 2025*
aiXiv: A Next-Generation Open Access Ecosystem for Scientific Discovery Generated by AI Scientists *arXiv Preprint, 2025*

Skills

Languages: Python, C++/C, Java, SQL, TypeScript, R, Bash
AI/ML: PyTorch, Transformers, **RAG**, **LangGraph**, LangChain, DeepSpeed, TensorRT, vLLM, SGLang, PEFT, RLHF
Backend & Systems: FastAPI, REST APIs, LLM APIs (OpenAI, Claude, Gemini), Distributed Systems, React, Spark
Cloud & Infrastructure: AWS, GCP, Azure, **Docker**, **Kubernetes**, Terraform, CI/CD, **Git**, **Linux**