Boqin Yuan

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

Summary

I'm an MSCS student at UC San Diego working at the intersection of **ML Systems** and **Agentic AI**. Previously a Machine Learning Engineer at a YC-backed startup, I build and study LLM-based agents with a focus on memory, reasoning, and scalable deployment. My work bridges research and production, turning agent ideas into reliable systems. See more at boqiny.github.io.

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*

- Conduct research with Stable Lab (Prof. Jishen Zhao) on agent memory and ML systems, designing AMA-Bench for long-horizon trajectory-based agent memory evaluation and PRO-V-R1 for reasoning-enhanced programming agents in RTL verification (DAC 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).

Selected Publications

AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications *ICLR 2026 Workshop on MemAgents*
Diagnosing Retrieval vs. Utilization Bottlenecks in LLM Agent Memory *ICLR 2026 Workshop on MemAgents*
PRO-V-R1: Reasoning Enhanced Programming Agent for RTL Verification *DAC 2026*

Skills

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