Zhirui (Raymond) Xia

AI/ML Engineer | Summer 2026: May–Sep (12+ weeks) | F-1 CPT-eligible | Open to relocation

z4xia@ucsd.edu | (+1) 857-869-9869 | https://www.linkedin.com/in/zhirui-xia/ | https://github.com/Raymondxzr

EDUCATION

University of California, San Diego — La Jolla, California
M.S. in Computer Science — *09*/2025 – **06/2027** (expected)
B.S. in Mathematics-Computer Science — 09/2021 – 03/2025

- Overall GPA: 3.91/4.00; Major **GPA: 3.94**/4.00; Provost Honors (all quarters)
- **Relevant Courses:** Machine Learning, Deep Learning, AI: Probabilistic Models, Search and Optimization, Data Analysis and Inference, Parallel Computing, LLM Security, Numerical Analysis, Linear Algebra, Statistics

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Intern — Shenzhen, China
Tencent AI Lab — 06/2024 – 09/2024

- Designed and built a real-time, **multi-stage LLM-driven** battle-report pipeline for Tencent's *Honor of Kings* (~100M DAU): ingested game engine telemetry, performed event extraction and story-driven segment aggregation, constructed coherent and personalized battle narratives, and time-budgeted commentary guided by TTS estimates.
- Built a hybrid **GraphRAG** + **vector RAG** (LangChain, LlamaIndex) to ground storyline generation on structured and unstructured game data, ensuring factual consistency and reducing hallucinations; achieved **14% higher Q&A accuracy** over fine-tuning alone.
- Distilled GPT-4 traces into Qwen2.5-7B with **SFT** and aligned with **DPO** for factuality and readability, **reducing inference cost by 35%** at comparable narrative quality.

Software Engineer Intern — Guangzhou, China
Ai4C Applied Research Institute — 06/2023 – 10/2023

- Owned the customer web portal with a headless **OrchardCore** backend behind a **GraphQL** gateway and a **React** front end, delivered sub-second P95 page loads, and shipped with **Docker** and GitHub Actions to a **Kubernetes** cluster with zero-downtime releases and reduced deployment time by **40%.**
- Built an in-portal **LLM chatbot** with **Semantic Kernel** and the **OpenAI API** that conducts short client conversations and generates an initial draft proposal, using reusable prompt templates, lightweight conversation memory, and retrieval of client profiles and policy documents.

RESEARCH & PROJECTS

Agentic Framework for Weather Science (Advisor: Dr. Rose Yu) — 09/2024 – Present
Graduate Researcher / Co-author

- Publication: current work submitted to **ICLR 2026**; earlier benchmark work (Aquilon) accepted at the **ICML 2025** World Models Workshop.
- Co-built a **unified Python execution environment** exposing core weather-science tools via clean APIs and a FastAPI code-execution service with resource pooling, acquire/release semantics, timeouts, and guarded concurrency.
- Designed **agentic workflows** that generate code, execute it, inspect intermediate outputs, and iteratively self-correct; also implemented a single-shot variant with a bounded error-fix loop for lower latency.
- Implemented a **synthetic data pipeline** that converts weather narratives into verifiable tasks by combining LLM-based claim extraction, templated question generation, and auto-generated verifier code against ERA5/WeatherBench2.

PPO for Beginners (Advisor: Dr. Sicun Gao) — 06/2023 – 09/2023

- Contributed **Pytorch** enhancements to the **1.1k+ star** PPO for Beginners repo, adding Generalized Advantage Estimation (GAE), entropy regularization, target-KL early stopping, and adaptive learning rate scheduling.
- Authored Part 4 on PPO optimization tradeoffs and a reproducible training setup, enabling newcomers to reach **~40%** average return gains on OpenAI Gym baselines.

Play Prediction with Bayesian Personalized Ranking (BPR) — 09/2023 – 12/2023

- **Ranked 1/604** in a UCSD Kaggle-style "would play" prediction challenge by building a **TensorFlow BPR recommender** with pairwise loss and negative sampling, reducing **popularity bias** with weighted negatives and score calibration, and handling **cold start** with popularity priors and side-feature warm starts.

SKILLS

- **Languages & Frameworks**: Java, Python, C/C++, C#, Go, R, MATLAB, SQL, .NET, Flask, GraphQL, React
- **AI/ML**: LLM, PyTorch, TensorFlow, Hugging Face, Agentic workflows, RAG, online serving & monitoring, A/B testing
- **Systems & Databases**: Distributed systems, Linux, Docker, Kubernetes, microservices, CI/CD, PostgreSQL, MongoDB