Arin Gadre

aringadre@gmail.com | LinkedIn | GitHub | Portfolio

Education

University of California, Santa Cruz — Santa Cruz, CA

B.A. in Network and Digital Technology — *Expected June 2025*

- Relevant Coursework: Applied ML: Deep Learning, Intro to Software Engineering, Computer System Design, Computer Networks, Compiler Design, Advanced Programming in C++, Data Structures & Algorithms, Logic Design, Natural Language Processing

Experience

Co-op Intern - LLM Benchmarking Project — Jan 2025 – June 2025

Nutanix — *Remote*

- Developed Flask-based ML inference system for CPU-accelerated LLMs (Mistral-7B, LLaMA) with Prometheus metrics pipeline
- Implemented production-grade server infrastructure with Waitress WSGI, supporting concurrent model inference and batch processing
- Built Intel IPEX optimizations with bfloat16 precision and thread management, improving inference efficiency for 7B+ parameter models
- Created RESTful API endpoints with robust error handling and logging for model inference and metrics collection
- Integrated LlamaCpp for optimized CPU inference, achieving 2x throughput improvement through GGML quantization and kernel optimizations
- Designed Grafana dashboards visualizing Prometheus metrics for real-time monitoring of model latency, throughput, and resource utilization

Software Engineer Intern — Sept 2023 – Present

The Difference LLC — *Remote*

- Architected Flutter mobile app components with optimized state management using Provider, streamlining 15+ interactive features
- Integrated various third-party APIs to enhance application functionality and boost user engagement
- Developed new back-end functions in Laravel and performed DB migrations, expanding service capabilities
- Implemented debouncing techniques to optimize search functionality, reducing API response times by 60%
- Created comprehensive API documentation using Swagger, improving development workflow and reducing onboarding time by 50%
- Implemented automated CI/CD pipeline using GitHub Actions, reducing deployment time from 45 to 12 minutes

Projects

BingeFlix | *React, Node.js, MongoDB, ChatGPT API* | Live Demo — Spring 2024

- Developed a unified streaming search platform aggregating movies, sports, and anime from various services
- Led team as Scrum Master through SCRUM processes, managing sprints and maintaining documentation
- Integrated streaming APIs for real-time availability and ChatGPT API for personalized recommendations

Real-Time Drawing Board | *Django, Django Channels, Redis, React* | Live Demo — Spring 2024

- Engineered real-time collaborative drawing platform using Django Channels and WebSocket for multi-users
- Implemented Redis channel layer and ASGI interface with Daphne server for efficient WebSocket communication
- Deployed containerized backend on Railway.app and frontend on Vercel with CI/CD pipeline

Dino Run Game | *SystemVerilog, FPGA* | Demo Video — Fall 2024

- Designed and implemented a dinosaur runner game using SystemVerilog with finite state machines
- Programmed VGA rendering for pixel-perfect graphics on FPGA with real-time obstacle spawning
- Developed modular game logic for score tracking, collision detection, and difficulty adjustments

Technical Skills

Languages: Python, Java, C/C++, JavaScript, TypeScript, Dart, PHP, Verilog, SQL
Frameworks & Tools: React, Laravel, Flutter, PyTorch, TensorFlow, Keras, Tailwind, Node.js, Express, MongoDB, MySQL, Grafana, Prometheus, Flask
Development: SCRUM, Agile, API Integration, Unit Testing, Debugging Tools, Git, Android Studio, Android SDK, NLTK