

KYLE YIN XU

Irvine, CA ◇ kyleyinxu@gmail.com ◇ (858) 922-8296 ◇ github.com/KYinXu ◇ linkedin.com/in/kyleyinxu ◇ kyinxu.com

EDUCATION

University of California, Irvine *B.S. Computer Science '27, Intelligent Systems Specialization (GPA: 3.83)* Sep 2023 - Present
Courses: Automata Theory, Data Structures, Algorithms, Operating Systems, Networks, Computer Architecture, Machine Learning, Data Mining, Search Systems, Databases & Web Apps, Transformers & LLMs, Probabilistic & Graphical Models

PROFESSIONAL EXPERIENCE

Forge Rewards (YC W24) — Palo Alto, CA / New York City, NY, *Software Engineer* Jul 2025 - Present

- Deploy 20+ Flutter/React Native DSP apps cross-platform with Docker, reaching top 100 App Store and 600K+ users.
- Build scalable APIs using Firestore as a ground-truth datastore to sync 100k daily orders with POS/payment systems.
- Developing agentic CRM dashboard; cut GraphQL query latency by 500ms and led database ETL pipeline migrations.

Variantz — Singapore, *Software Engineer Intern* May 2025 - Aug 2025

- Optimized React/Vite performance via code-splitting, reducing bundle size by 85% and loading pages 200ms faster.
- Restructured PostgreSQL to 3NF schema with query indexing, cutting API response latency by 50% across 30+ routes.
- Automated CI/CD pipelines and led Agile sprints, halving deployment lead time and boosting development velocity.

RapDevPro — San Diego, CA, *Web & Application Development Intern* Dec 2022 - Feb 2024

- Developed data-driven analytics platform with PHP for real-time moderation of 10,000+ posts through 15+ key metrics.
- Spearheaded Flutter development of social media app with 8 engineers, accelerating dual-platform launch by 8 weeks.
- Architected a unified mobile and web API layer, reducing redundant backend logic by 70% via centralized cloud sync.

PROJECTS

TransformerLens (Contributor), *Python, PyTorch, JLens, Transformers, Hugging Face* Jul 2026

- Developed integration tooling between JLens and HF for modern LLM architectures, lowering VRAM overhead 25%.
- Engineered Gated DeltaNet layer integrations using a modular bridge pattern, refactoring 2 cores of legacy codebase.

NVIDIA Nemotron Reasoning Challenge, *Kaggle, LangChain, Ollama, LoRA, QLoRA, PEFT* Apr 2026

- Architected agentic knowledge distillation pipeline to fine-tune Nemotron:30b model on complex math/logic puzzles.
- Utilized LoRA/QLoRA on synthetic datasets to optimize reasoning, boosting target eval accuracy from 0.35 to 0.61.

NYT Connections Solver, *Python, PyTorch, Hugging Face, DeBERTa, Llama, LoRA, CUDA, scikit-learn* Mar 2026

- Fine-tuned DeBERTa for complex wordplay resolution, increasing accuracy on custom semantic benchmarks by 200%.
- Architected CUDA-accelerated fine-tuning pipelines, outperforming baseline models 4x with <1 average swap error.

Symptoms Searcher, *Flutter, Python, Go, MongoDB, Kaggle* Feb 2026

- Designed a polyglot microservice backend using Go and Python to ingest and normalize high-volume health datasets.
- Constructed concurrent data pipeline to populate MongoDB index, optimizing for low-latency search on 100k records.

Fabflix Media Discovery Platform, *React, Node.js, Java, MySQL, Apache Spark, JMeter, MongoDB* Oct 2025

- Devised idempotent, multi-threaded ETL pipeline migrating 120K MySQL rows to MongoDB for 3x higher throughput.
- Conducted JMeter load-testing and system benchmarks to optimize NoSQL indexes and reduce search latency by 40%.

RESEARCH

Biointelligence Lab — Irvine, CA, *ML Research Assistant* Feb 2026 - Present

- Engineering a Graph Attention Network (GAT) for AMP classification, achieving 98% AUC via structural embeddings.
- Refining clustered cross-validation to mitigate homology-based data leakage and ensure generalization over all AMPs.
- Benchmarking performance of GAT, GCN, and EGNN model architectures to optimize accuracy and inference speed.

TECHNICAL SKILLS

Languages: Python, JavaScript, TypeScript, Java, Go (Golang), C++, C, SQL, Dart, PHP

Frameworks: React Native, Node.js, Next.js, Angular, Express, Flutter, Flask, GraphQL, REST

Machine Learning: PyTorch, CUDA, scikit-learn, LoRA, PEFT, LangChain, Agentic Harness, Transformers

Systems/DevOps: Docker, CI/CD, Google Cloud Platform (GCP), Firebase

Databases: MongoDB, PostgreSQL, MySQL, Firestore, JDBC, Cloudflare R2, AWS S3, Redis