O. FELIX AMORUWA

famoruwa@berkeley.edu | 909-731-9011 | felixamoruwa.info

Forward Deployed Engineer and Technical Product Leader with 12+ years in production AI systems — from hand-coding BPTT in C++ (2004) to building a full RLHF/DPO/GRPO post-training workbench benchmarking TRL, VeRL, OpenRLHF, and NeMo RL across Apple Silicon (MPS) and CUDA today. Hands-on expertise in post-training pipelines (PPO, GRPO, DPO, RLHF, SFT), inference optimization, and open-source LLM deployment at scale. Scaled production inference infrastructure to 675M+ engagements and 50K TPS with sub-25ms TP99 at Intuit; NeurIPS published researcher in neural architectures.

AI/ML RESEARCH & PROJECTS

RL Workbench — Post-Training RL Platform (2026)

- Built 3-phase post-training workbench covering the full RLHF/DPO/GRPO pipeline: Reward Lab for designing and A/B testing reward functions (RLVR, learned, hybrid) across 4 datasets (GSM8K, MATH, HumanEval, UltraFeedback), Playground for real TRL-powered GRPO/DPO training with live SSE metric streaming on Apple Silicon (MPS) and CUDA, and Arena for head-to-head framework benchmarking (TRL, VeRL, OpenRLHF, NeMo RL) with GPU passthrough in Docker containers.
- Implemented 12 RL algorithms (PPO, GRPO, DAPO, REINFORCE, REINFORCE++, RLOO, DPO, SimPO, IPO, KTO, ORPO, SPPO) with algorithm-specific metric profiles, cross-tab workflow lineage tracking, and standardized throughput/memory/convergence benchmarking across frameworks — directly informing framework selection and quantization strategy for production deployments.

aeval — AI Model Evaluation Platform (2025–2026)

- Built local-first LLM evaluation platform with 5 core eval types (factuality, reasoning, instruction-following, safety, code generation), adversarial safety testing with refusal detection, and data contamination detection via SHA-256 hashing; stack: FastAPI orchestrator, TimescaleDB, Redis job queue, Next.js dashboard, Ollama.
- Statistical rigor: bootstrap confidence intervals, Welch's t-test, Cohen's d effect size, saturation detection; CI/CD integration with regression detection and automated safety gates for open-source model evaluation pipelines.

AutoEval — Automated Visual Evaluation for Robot Model Training (2025)

- Built automated visual evaluation system reducing model evaluation cycles from 72 hours to ~4 minutes; multimodal AI (Claude/GPT-4V) performs spatial reasoning on prediction frames, generating structured PASS/FAIL reports with confidence scores — zero-integration architecture captures from any visualization tool via screen capture.

BRAIN — Protein Structure Prediction ML Platform (UC Berkeley 2004; rewritten 2026)

- Built production ML platform in PyTorch with 5 neural architectures (feedforward, GRU, Transformer, ESM-2, multi-task), MLflow experiment tracking, Optuna HPO, and FastAPI serving — 823 automated tests, Docker orchestration (6 containers); spans 413 to 8B parameters (19M-fold scale increase from original 2004 C++ BPTT implementation).
- NeurIPS 2014 accepted paper on artificial neural networks for protein secondary structure prediction.

Deep Learning Education Platform (2025–2026)

- Developed 60 interactive demos across 20 chapters covering Goodfellow's Deep Learning textbook — transformers, GANs, diffusion models, autoencoders, optimization.

PROFESSIONAL EXPERIENCE

STREAMIO AI | Oakland, CA

Founder & CEO — Production AI Platform (macOS, Linux, iOS & Web) 09/24 – Present

- Integrated Claude MCP SDK for real-time multimodal AI analysis of screen captures; built OpenClaw multi-agent orchestration framework with gateway protocol, subagent delegation, and session management — enabling coordinated LLM agent workflows across multiple industry verticals.
- Engineered real-time HLS livestreaming pipeline: multi-stream canvas compositing (up to 9 concurrent sources at 30fps), FFmpeg WebM-to-MPEG-TS transcoding, and WebSocket communication layer; shipped cross-platform desktop builds (macOS DMG with code signing/notarization, Linux .deb) with native ScreenCaptureKit integration via Swift.
- Built end-to-end auth and payments pipeline using Kinde OAuth 2.0, Stripe subscriptions, and Electron SafeStorage; integrated ElevenLabs TTS/STT with 6 auto-classified voice profiles and SHA-256 audio caching.

FINTELLECT AI | Oakland, CA

Founder & CEO — AI Financial Platform (iOS/Android/Web) 09/24 – Present

- Architected RAG retrieval pipeline with ChromaDB vector store, multi-provider LLM orchestration (Claude, GPT-4, Gemini) with fallback routing, structured output validation, and token budget optimization — demonstrating hands-on model landscape awareness across open-source and proprietary LLMs.
- Built domain-specific conversational AI agents scoped to distinct financial focal points; integrated LLM models for automated trade analysis and AI-powered macroeconomic charting at scale.

INTUIT | Mountain View, CA

Staff Product Manager — Developer Frameworks & Platform Infrastructure 06/21 – 09/24

- Scaled production inference infrastructure to 675M+ engagements in FY23 across QuickBooks, TurboTax, Mint, Mailchimp, and Credit Karma; optimized throughput from 6K to 50K TPS via rSocket migration supporting ~1.5M concurrent connections with sub-25ms TP99 — directly analogous to hitting throughput and latency targets in production AI deployments.
- Delivered ICE Self-Service platform (DevPortal, GitOps config, ICE Playground), reducing developer onboarding from 2–3 weeks to minutes in pre-prod and <24 hours for production, while mitigating $1M+ in projected opex growth; achieved 275% YoY growth in ICE engagements.
- Extended Java and Python SDK Starter Kits with scaffolding templates, build configurations (Gradle/Maven), testing frameworks, and CI/CD integration — empowering developers to go from zero to production-ready microservice in minutes.
- Conducted enterprise-wide Service Language Assessment across 9 languages, analyzing usage data and developer feedback to inform strategic investment decisions presented to CTO; worked closely with telemetry and usage data (SQL, BigQuery) to prioritize developer pain points across ~20 mobile apps and 30+ product SKUs.

- Implemented ICE Presence in async chat, generating $480K/month in additional invoicing; deployed Background-to-Foreground Messaging on iOS/Android with <100ms latency. DeveloperWeek 2022 speaker.

SPLUNK INC. | San Francisco, CA

Senior Product Manager — Search Orchestration 01/19 – 01/21

- Led query performance optimization initiative achieving up to 10x performance improvements for enterprise beta customer (Assurance) by building mirrored topology for benchmark testing — directly applicable to winning critical POCs and benchmarks.
- Owned Search Service (Go microservices), Search Catalog (PostgreSQL metadata service), and Splunk Processing Language (SPL/SPL2); delivered Scheduler Service end-to-end in ~4 months, demoed at Splunk .conf19. Splunk .conf18 Speaker.

KAISER PERMANENTE | Pleasanton, CA

SOA Technical Product Manager 08/12 – 12/18

- Led enterprise Splunk Logging-as-a-Service (1.7 TB daily volume, 200+ internal customers); built Redis and XC10 caching layer across the enterprise to address scalability, fault tolerance, and data redundancy at scale.

IBM | Pittsburgh, PA / Foster City, CA / RTP, NC

Software Engineer — Business Intelligence Products 07/05 – 03/10

- Resolved multiple $1–5M high-severity customer escalations for enterprise BI software; led cross-functional root cause analysis and customer-facing resolution, improving case resolution time by 20% across the Support Organization.

BANK OF AMERICA MERRILL LYNCH | Charlotte, NC

Tech MBA Summer Associate — Global Corporate & Investment Banking T&O 06/11 – 08/11

- Developed enhanced estimation model using Monte Carlo simulation to optimize DMAIC phase distribution across a $494M portfolio in Global Technology & Operations.

EDUCATION

CARNEGIE MELLON UNIVERSITY, SILICON VALLEY
Master of Science — Software Management Mountain View, CA (08/16)
CARNEGIE MELLON, TEPPER SCHOOL OF BUSINESS
Master of Business Administration — Concentrations: Finance, Quantitative Analysis, Economics Pittsburgh, PA (03/12)
UC BERKELEY, COLLEGE OF ENGINEERING & HAAS SCHOOL OF BUSINESS
Bachelor of Science — Computational Engineering Science Berkeley, CA (12/05)
Certificate — Corporate Finance for Financial Engineering

TEACHING EXPERIENCE

DE ANZA COLLEGE | Cupertino, CA

Adjunct Faculty — Computer Information Systems Department 04/18 – Present

- CIS 35A/B, 36A/B — Java Programming; CIS 64C/64G — Introduction to PL/SQL; CIS 44F — Introduction to Data Analytics; CIS 64E — Fundamentals of Large Scale Cloud Computing (AWS, GCP); CIS 102 — Ethical Hacking; CIS 104 — Digital Forensics; CIS 95F — Managing Cloud Projects.

ADDITIONAL INFORMATION

- FAA Commercial Drone License; FAA Private Pilot License (intended CFI)
- O'Reilly & OER Publications: Data Analytics and Managing Cloud Projects (forthcoming Summer 2026)
- First Tee Assistant Golf Coach — San Francisco Chapter (2012–Present), Pittsburgh Chapter (2011–2012)
- Interests: Classical & Jazz Saxophone (Alto/Tenor/Baritone), Triathlon, Golf, Intramural Basketball & Soccer