

O. FELIX AMORUWA

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

Technical product leader and former software engineer with 12+ years building developer-facing platforms and AI products at scale — from shipping production-ready AI solutions as a 0-to-1 founder to scaling a developer inference platform to 675M+ engagements and sub-25ms TP99 at Intuit. Hands-on LLM inference and multi-agent orchestration experience (OpenClaw, RAG pipelines, multi-provider LLM routing) with deep familiarity across the inference stack. NeurIPS-published ML researcher; benchmarked GRPO/DPO across TRL, VeRL, OpenRLHF, and NeMo RL. CMU MBA.

PROFESSIONAL EXPERIENCE

STREAMIO AI | Oakland, CA

Founder & CEO — *streamio.ai* (macOS, Linux, iOS & Web) 09/24 – Present

- Implemented OpenClaw multi-agent orchestration framework with gateway protocol, subagent delegation, profile management, and session switching — enabling coordinated AI agent workflows across real estate, insurance, health/dental, and financial markets verticals; directly analogous to agentic computation use cases Cerebras' inference platform unlocks.
- Built production-ready AI platform integrating Claude (MCP SDK) for real-time contextual analysis, Pine Script generation, and automated CMA reports via Redfin/Zillow APIs — owning the full 0-to-1 product journey from architecture through GTM.
- Engineered latency-sensitive real-time HLS livestreaming pipeline: multi-stream canvas compositing (up to 9 concurrent sources at 30fps), FFmpeg transcoding, and WebSocket communication layer — production system designed around speed as a core differentiator.
- Led customer discovery interviews and iterative product refinement, translating user feedback into prioritized roadmap decisions across AI, streaming, and payments features.

FINTELLECT AI | Oakland, CA

Founder & CEO — *fintellectai.io* (iOS/Android/Web) 09/24 – Present

- Architected multi-provider LLM inference orchestration (Claude, GPT-4, Gemini) with fallback routing, RAG retrieval pipeline (ChromaDB), structured output validation, and token budget optimization — hands-on experience with the inference stack tradeoffs Cerebras is designed to eliminate.
- Built domain-specific conversational AI agents scoped to distinct financial verticals, delivering context-aware advisory interactions at scale; executed go-to-market through App Store launch following customer discovery and influencer partnerships.

INTUIT | Mountain View, CA

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

- Achieved 275% YoY growth in ICE platform engagements, scaling to 675M+ in FY23 across QuickBooks, TurboTax, Mint, Mailchimp, and Credit Karma; scaled throughput from 6K to 50K TPS via rSocket migration supporting ~1.5M concurrent connections with sub-25ms TP99 — enterprise-scale, latency-critical platform ownership.
- Extended Java and Python SDK Starter Kits with scaffolding templates, build configurations, testing frameworks, and CI/CD integration — empowering developers to go from zero to production-ready microservice in minutes on a developer-facing platform.
- 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.
- Conducted enterprise-wide Service Language Assessment across 9 languages, analyzing usage data and developer feedback to inform strategic investment decisions presented to CTO; built Asterias, a declarative asset lifecycle management platform with GraphQL API.
- 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 for enterprise beta customer, building mirrored topology for benchmark testing and achieving up to 10x performance improvements in Splunk Cloud Services search — mirrors Cerebras' 10x inference speed value proposition.
- Owned Search Service (Go microservices), Search Catalog (PostgreSQL), and SPL/SPL2 — delivered Scheduler Service end-to-end in ~4 months; designed RICE-based prioritization framework across 3 microservice backlogs balancing Fortune 500 and third-party developer requirements. Splunk .conf18/.conf19 speaker.

KAISER PERMANENTE | Pleasanton, CA

SOA Technical Product Manager 08/12 – 12/18

- Led development and enterprise rollout of Splunk Logging-as-a-Service (1.7 TB daily volume, 200+ internal enterprise customers) and ITSI Application Monitoring-as-a-Service; built Redis caching capability addressing scalability and fault tolerance at datacenter 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 Monte Carlo simulation model to optimize DMAIC phase distribution across a \$494M portfolio in Global Technology & Operations.

AI/ML RESEARCH & PROJECTS

RL Workbench — Post-Training RL Platform (2026)

- Built 3-phase RL post-training workbench benchmarking GRPO/DPO across TRL, VeRL, OpenRLHF, and NeMo RL — with live SSE metric streaming, GPU Docker passthrough, and standardized throughput/memory/convergence benchmarking; directly maps to Cerebras' inference stack evaluation needs.
- Implemented 12 RL algorithms (PPO, GRPO, DAPO, REINFORCE, REINFORCE++, RLOO, DPO, SimPO, IPO, KTO, ORPO, SPPO) with algorithm-specific metric profiles and cross-framework lineage tracking across 4 datasets (GSM8K, MATH, HumanEval, UltraFeedback).

aeval — AI Model Evaluation Platform (2025–2026)

- Built local-first LLM evaluation platform with 5 eval types (factuality, reasoning, instruction-following, safety, code generation), adversarial safety testing, bootstrap confidence intervals, and automated regression detection — stack: FastAPI, TimescaleDB, Redis, Ollama, Next.js.

- BRAIN — Protein Structure Prediction ML Platform** (UC Berkeley 2004; rewritten 2026)
- NeurIPS 2014 published. Production PyTorch platform with 5 neural architectures (feedforward, GRU, Transformer, ESM-2, multi-task), MLflow, Optuna HPO, FastAPI serving — spanning 413 to 8B parameters (19M-fold scale); original 2004 system hand-coded in C++ with custom BPTT.
- AutoEval — Automated Visual Evaluation for Robot Model Training** (2025)
- Repurposed StreamIO's screen capture + multimodal AI pipeline to score robot model outputs against natural-language rubrics — reducing evaluation cycles from 72 hours to ~4 minutes via zero-integration architecture capturing from any visualization tool.

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
- Java Programming (CIS 35A/B, 36A/B), PL/SQL, Data Analytics, Large Scale Cloud Computing (AWS, GCP), Ethical Hacking, Digital Forensics, 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