Xingfan Xia

8502 134th Ct NE, Redmond, WA 98052
xingfanxia@gmail.com • +1 (507) 403-1689 • https://ax0x.ai

PROFILE

Startup CTO and full-stack engineer shipping 95% of production code via **Agentic Coding** (Claude Code, Codex – 3B+ tokens burned). Build custom QA evaluation pipelines and agent orchestration patterns. Seeking to push the frontier of AI-native software engineering.

WORK EXPERIENCE

Compute Labs, Remote, Redmond, WA

- Co-Founder and CTO — Aug 2024 – Present
  - **Architecting multi-agent AI systems for GPU infrastructure underwriting, building LLM-powered automation that reduced deal evaluation from 2 weeks to under 10 minutes.**
  - Architected multi-model AI orchestration layer (Claude, Gemini, OpenAI) powering 11 parallel domain-expert sub-agents for automated deal evaluation, with three-layer pipeline: Gemini extraction, Claude Opus analysis, and multi-model ensemble for report generation.
  - Built state machine workflow engine orchestrating AI agent pipelines processing $1.3B+ in GPU deals, with prompt experimentation system for continuous model optimization.
  - Built GPU market intelligence platform scraping 3,200+ prices from 45+ providers using multi-model LLM extraction (GPT/Claude/Gemini), with 4-stage data pipeline including adaptive outlier detection and financial depreciation modeling.

Amazon Web Service – Athena, Redmond, WA

- Senior Software Engineer - Data Infrastructure — Apr 2022 – Jun 2024
  - **Building petabyte-scale data infrastructure for AI/ML analytics, enabling high-performance distributed query processing for machine learning workloads at AWS scale.**
  - Architected CTAS feature for Iceberg Tables enabling efficient data pipeline workflows for ML training data preparation. Released at AWS RE:INVENT 2022 with millions of queries executed since launch.
  - Optimized Hive JSON Serde library for data ingestion, improving query performance by 30% and reducing processing failures by 60%.

Apple, Cupertino, CA

- Machine Learning Engineer - Siri AI Platform — Aug 2020 – Apr 2022
  - **Building on-device ML models and AI-powered features for Siri proactive assistant across iOS, macOS, watchOS, and tvOS.**
  - Implemented on-device user profiling framework with multiple ML models predicting user interests and preferences, serving as upstream signal provider for Apple Maps and Apple News recommendation systems.
  - Overhauled topic prediction model using transformer-based BERT architecture, increasing prediction accuracy by 14% and user engagement by 11%. Optimized inference pipeline to reduce on-device latency by 12%.

Airbnb, San Francisco, CA

- Machine Learning Engineer - Trust & Safety — Jan 2019 – Aug 2020
  - **Building ML-powered fraud detection and trust & safety systems using XGBoost, logistic regression, and real-time inference engines.**
  - Developed ML models (XGBoost and logistic regression) for fake review detection, cutting fraudulent reviews to 1% of peak volume and preventing $2M in losses. Deployed with optimized inference mechanism improving performance by 30%.
  - Architected real-time ML inference microservice with Kafka streams and Redis caching for sub-100ms risk signals across multiple downstream consumers.

EDUCATION

Carleton College, Northfield, Minnesota

- B.A. in Computer Science — Graduated Jun 2018

TECHNICAL SKILLS

- Programming Languages: Proficiency in Python, Java, Golang, TypeScript, Scala, Objective-C
- AI & Machine Learning: LLM APIs (Claude, GPT, Gemini), Multi-Agent Systems, Prompt Engineering, PyTorch, Scikit-learn, XGBoost, BERT, On-Device ML, MLOps, Claude Code, Cursor
- Infrastructure & Data: FastAPI, Next.js, React, GCP Cloud Run, Spark, Iceberg, Kafka, Airflow, AWS (Athena, S3, Lambda, EC2)