

XIANG HAO

Wilmington, DE | Open to Relocation / Remote | xianghao98520@gmail.com | linkedin.com/in/xianghaosuc | github.com/Xiang-Suc | xianghao.dev

SUMMARY

Software Engineer (UC Davis MSCS) building production LLM systems — multi-model agent routing, RAG retrieval, and the async Python backends underneath them — owning features from idea conception and rapid prototyping through production deployment. Shipped a cross-platform AI assistant into iOS / Android beta and an audit-grade clinical data platform on Azure + PostgreSQL, in regulated, ML-heavy, and consumer-facing environments.

TECHNICAL SKILLS

LLM & Agents: LLM orchestration & agent pipelines, multi-model routing & fallback (OpenAI / Anthropic / DeepSeek), function calling, structured outputs & schema validation, guardrails & human-in-the-loop, SSE streaming, prompt engineering, agentic dev tooling (Claude Code, Cursor, Codex)

Retrieval & RAG: RAG architecture, embeddings, chunking & ranking strategies, hybrid retrieval, semantic search, pgvector & vector DBs, LangChain, Hugging Face

Backend & Languages: Python (FastAPI, Flask, Pandas), TypeScript, Node.js, Next.js (App Router), SQL, Java, REST API design, async processing & task queues, idempotent job design, OAuth 2.0 / JWT, RBAC, Supabase (Auth / RLS / migrations)

Data & ML: PostgreSQL (schema design, indexing & query tuning), MySQL, Redis, ETL pipelines, Pandas, PyTorch, Apache Kafka, CDISC SDTM / ADaM, D3.js

Cloud & DevOps: AWS (S3, ECS, Lambda), Azure (App Service, AD), Docker, Kubernetes, CI/CD (GitHub Actions), Prometheus / Grafana / Jaeger

Working knowledge: Snowflake, MongoDB, Django

PROFESSIONAL EXPERIENCE

Rocha Tech LLC — Software Engineer (Remote)

May 2026 – Present

- Building **Mo Paws**, a cross-platform AI pet-care assistant (React + Vite + TypeScript, Capacitor for iOS / Android, Supabase) now in closed TestFlight beta — owning auth, data model, retrieval and model routing end to end, with multi-tenant isolation enforced in PostgreSQL Row Level Security policies rather than application middleware.
- Designed a task-based **multi-model router** across 3 providers (DeepSeek for text, OpenAI for vision, Anthropic for fallback) with cost- and latency-aware selection and automatic failover; cut cold-start latency **75–80%** through knowledge-base preloading and SSE-streamed responses.
- Built a deterministic retrieval layer over a curated knowledge base — **206 breed profiles** covering the full 205-breed AKC baseline, plus **181 canine knowledge facts** — with image-based document extraction and safety guardrails, backed by a PostgreSQL schema designed and indexed for low-latency lookups on mobile clients.
- Hardened the LLM layer for production: **JSON-schema validation** on every structured output with automatic repair-and-retry on malformed responses, plus a prompt regression suite run before each prompt change to catch behavioral drift.

Studio PAV — Data & AI Engineer (Contract, Remote)

Feb 2026 – May 2026

- Shipped a GenAI classification layer (OpenAI structured outputs) that auto-categorized **10K+ transaction line items at 90% accuracy** into merchandise / benefits / engagement and de-duplicated a fragmented taxonomy, replacing manual analyst labeling.
- Designed an end-to-end Python / Pandas workflow merging multi-source Shopify + CRM records across 10 years, unifying tens of thousands of transactions into analysis-ready datasets for segmentation and ML scoring.
- Engineered a feature module computing **20+ behavioral features** per customer (RFM, AOV, recency, LTV) powering a percentile-based segmentation engine with dynamic Top 1%–30% tiers, exported as multi-sheet Excel via openpyxl.

EDA Clinical — Software Engineer (Pennsylvania, US)

Oct 2024 – Dec 2025

- Shipped an end-to-end **AI-agent pipeline** (Teams transcript → OpenAI → structured summary → auto-posted Trello cards and checklists), eliminating manual meeting documentation org-wide and **saving 5+ hours/week**.
- Built a **RAG-powered** weekly-digest dashboard (React + D3.js) over clinical datasets — document chunking and embedding-based semantic retrieval — auto-generating summaries and trend analysis and **lifting stakeholder engagement 20%**.
- Engineered an async **FastAPI** backend with hierarchical storage, custom delta versioning, idempotent long-running jobs, and RBAC / OAuth 2.0 on Azure + PostgreSQL for TB-scale regulated submission workflows, **cutting audit cycles 25%**.
- Authored a reusable protocol library standardizing health checks and CI/CD (GitHub Actions) across 5+ Dockerized / Kubernetes microservices, **cutting rollout time 30%**.
- Built a 0-to-1 ETL pipeline (**SilentNight**) simulating clinical-trial lifecycles with **100% CDISC SDTM/ADaM compliance**, and integrated the CDISC Open Rules Engine to establish end-to-end audit traceability.

HRG Institute of Research and Innovation — Software Engineer Intern (Anhui, China)

Jul 2021 – Sep 2021

- Productionized a real-time computer-vision service (PyTorch ResNet-50) on a Kafka + AWS S3 / ECS inference pipeline with Dockerized blue/green deployments and sub-second alert latency, automating **30+ hours/day** of manual safety review.

Shanxi Yunlv Tianxia Network Technology — Software Engineer Intern (Shanxi, China)

Jan 2021 – Mar 2021

- Shipped a SaaS expense / ticketing system serving **1,200+ employees** on FastAPI microservices auto-scaled via Kubernetes, with Temporal-orchestrated long-running workflows and a Prometheus / Grafana / Jaeger observability stack.

EDUCATION & CERTIFICATION

M.S. Computer Science — University of California, Davis

Sep 2022 – Jun 2024

B.S. Computer & Information Science — University of Oregon

Mar 2019 – Jun 2022

AWS Certified Cloud Practitioner — Amazon Web Services

2025