Xinkai Chen

412-708-2025 | xinkaichen1997@gmail.com | LinkedIn | GitHub | Website

Machine Learning Engineer with **5+ years of experience** building and shipping production ML systems — spanning LLMs, Transformers-based NLP, and gradient boosted trees — driving **millions** in annual savings and earning a US patent.

Experience

Data Scientist — Chicago, IL
Elevance Health, Inc. (formerly Anthem Inc.) — *Mar 2021 – Present*

Built and owned production ML and LLM systems to automate clinical authorization decisions across 30+ medical guidelines and multiple business units, saving tens of thousands of physician and nurse hours annually.

- Core contributor to the MLOps ecosystem — built pip-installable packages using **UV**, **Poetry**, and **JFrog**, and integrated **GitLab CI/CD** pipelines for linting, testing and deployment, reducing release cycle time by **50%**
- Designed a **DAG-based LLM** pipeline for therapy visit recommendations with swappable nodes, Pydantic-enforced outputs, and robust retry logic, achieving **90%** accuracy and **∼$200K/month** in savings
- Implemented hybrid **RAG** with section-aware chunking, BM25 sparse and OpenAI embedding dense retrieval via Qdrant, and cross-encoder reranking to enhance search quality, reducing context token usage by **40%**
- Trained **LightGBM** models across 15+ medical guidelines, tuned decision thresholds for target FP rates and applied SHAP for clinical explainability — enabling **30%** automatic case approvals and **∼$100K** monthly savings
- Built clinical NLP pipelines using **UmlsBERT/CODER** embeddings and **spaCy/Stanza** to extract structured medical entities from unstructured clinical notes, powering downstream authorization models at **85% precision**
- **Led a pod of 6 data scientists**, building a scalable tree-based codebase that could be rapidly adapted across clinical guidelines — driving **25+** model deployments in five months with minimal per-use-case overhead
- **Patented** a novel system for authorization automation using Artificial Intelligence (US 12293835 B1)

Machine Learning Consultant — Pittsburgh, PA
Hamiltonian Systems, Inc. — *Jun 2020 – Dec 2020*

- Designed a multithreaded **Flask API** to orchestrate ML model training and inference pipelines with async task management and cloud deployment via REST endpoints — reducing manual engineering workload **by 90%**
- **Led a team of 5** to build inventory lifecycle prediction models using **Random Forest** and **XGBoost** on 440K+ transactions cleaned and transformed with Oracle SQL and Pandas, reducing forecast error by 22% over baseline

Projects

Cerberus: Real-Time Fraud Detection ML Pipeline | *Kafka, Feast, Redis, FastAPI, Docker* — *Jan – Mar 2026*

- Developed a feature engineering pipeline using **Kafka** and **Feast**, synchronizing real-time data to **Redis** and historical records to PostgreSQL to achieve **sub-10ms** feature retrieval latency
- Integrated an XGBoost model to identify high-risk transactions and containerized the entire lifecycle (producer, stream processor, and FastAPI inference service) using **Docker**

MacroForge: Multi-Agent Nutrition System | *Google ADK, Gemini, GCP, Vertex AI* — *Oct – Nov 2025*

- Architected a multi-agent system using Google **Agent Development Kit** (ADK) to autonomously coordinate **specialized sub-agents** for recipe generation, food substitution, and shopping optimization
- Orchestrated a resilient, grounded agent pipeline by integrating **custom tools**, MCP servers, and Google Search APIs, with dynamic routing and **Vertex AI** for scalable production deployment

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Information Systems Management (GPA: 3.91, Highest Distinction) — *Aug 2019 – Dec 2020*

Fudan University — Shanghai, China
Bachelor of Information Security (School of Computer Science) — *Sep 2015 – Jun 2019*

Technical Skills

ML Frameworks: Pytorch, TensorFlow, Transformers, Hugging Face, LightGBM, XGBoost, Scikit-learn, spaCy
MLOps: AWS (Kubeflow, SageMaker), GCP (Vertex AI), Docker, FastAPI, GitLab CI/CD, JFrog, Poetry, UV
Data & Streaming: Kafka, Redis, PostgreSQL, MongoDB, Snowflake, Feast, Spark
Large Language Models: OpenAI SDK, LangChain, Google ADK, Pydantic, DSPy, Qdrant