

Alex Chen

alex.chen@demo.example.com | linkedin.com/in/alex-chen-demo | alexchen.demo.example.com | Austin, TX (US, remote)

PROFESSIONAL SUMMARY

Senior AI engineer with 6 years shipping production ML and LLM systems. Led an ML platform at a Series B fintech (2020-2024): 2 models to 15+ in production, real-time fraud detection at 99.7% precision, MLOps for 4 teams. Now focused on LLM infrastructure, evaluation, and observability - the part where great models actually meet production traffic.

CORE COMPETENCIES

LLM infrastructure & model serving

Model evaluation & observability

Real-time ML pipelines (Kafka, Feature Store)

MLOps: registry, CI/CD, drift monitoring

Python + distributed systems

WORK EXPERIENCE

TechFin Corp

2020 - 2024

Senior ML Engineer / ML Platform Lead

Austin, TX

- Led ML platform team (3 engineers). Built internal LLM-serving stack: model registry, A/B routing, feature store, per-tenant rate limiting - serving 4 product teams end-to-end.
- Designed real-time fraud detection pipeline: Kafka to feature computation to model inference to decision engine. Sustained 99.7% precision at 50 ms p99, saving ~\$2M/year in chargebacks.
- Shipped LLM reranking on top of collaborative filtering for lending products; 18% conversion uplift measured across 12 A/B weeks.
- Cut model deploy time from 2 weeks to 4 hours by wiring CI/CD (GitHub Actions + SageMaker) with automated eval gates against a held-out set.
- Stood up model monitoring: drift detection dashboards (Grafana), performance SLIs, automated retraining triggers; on-call load dropped ~35%.

DataStartup Inc

2018 - 2020

ML Engineer

Remote

- Built NLP pipeline for legal-document classification (BERT fine-tuning), 94% accuracy on held-out corpus of 12k contracts.
- Implemented learning-to-rank models for internal search; NDCG@10 up 22% vs the BM25 baseline.
- Set up experiment tracking with MLflow and model versioning; established the shared eval harness the team still uses.

PROJECTS

LLM Eval Toolkit [Open Source](#)

Evaluation framework for LLM applications: custom metrics, regression testing, and CI integration. Directly relevant to model observability and rollout gates.

Python, pytest, OpenAI/Anthropic SDKs

FraudShield [Open Source](#)

Real-time fraud detection reference stack (~500 GitHub stars). Distilled the production system into a runnable, testable blueprint.

Kafka Streams, feature store, model serving

EDUCATION

MS Computer Science UT Austin	2018
Focus: machine learning systems, distributed systems.	
BS Computer Science UC Berkeley	2016

CERTIFICATIONS

SKILLS

ML / AI: PyTorch, Hugging Face, LangChain, LLM eval, prompt engineering

MLOps: SageMaker, MLflow, Kubeflow, Airflow, feature store, drift monitoring **Infra:** Kubernetes, Kafka, Redis, PostgreSQL, AWS

Languages: Python, Go, TypeScript, SQL