Alex Chen

555-123-4567 alex.chen@email.com linkedin.com/in/alexchen github.com/alexchen

Technical Skills

Languages
Python, R, SQL, Scala, Java

ML/DL Frameworks
PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM

MLOps & Cloud
MLflow, Kubeflow, Docker, Kubernetes, AWS, GCP, Azure

Data Engineering
Spark, Kafka, Hadoop, Snowflake, Databricks

LLM & NLP
LangChain, Hugging Face, OpenAI API, BERT, GPT

DevOps
Jenkins, GitHub Actions, Terraform, Prometheus

Education

Stanford University Stanford, CA
M.S. in Computer Science *2020 – 2022*
- Focus: Artificial Intelligence

UC Berkeley Berkeley, CA
B.S. in EECS *2016 – 2020*
- GPA: 3.8/4.0, Magna Cum Laude

Certifications

AWS ML Specialty (2023)
GCP ML Engineer (2023)
TensorFlow Developer (2022)

Professional Summary

Experienced Machine Learning Engineer with 4+ years developing and deploying scalable ML systems in production. Expertise in end-to-end ML pipeline development, from data preprocessing to model deployment and monitoring. Proven track record of building recommendation systems, NLP applications, and computer vision models serving millions of users.

Experience

Senior Machine Learning Engineer Jan 2023 – Present
Meta *Menlo Park, CA*
- Led development of recommendation system serving 50M+ daily users, improving engagement by 15%
- Built real-time feature engineering pipeline using Kafka and Spark, reducing latency by 40%
- Implemented A/B testing framework for ML models, enabling data-driven deployment
- Mentored 3 junior engineers and established ML best practices across the team

Machine Learning Engineer Jun 2022 – Dec 2022
Uber *San Francisco, CA*
- Developed demand forecasting models using LSTM and Transformers, improving accuracy by 25%
- Built MLOps pipeline with MLflow and Kubernetes, automating training and deployment
- Optimized model inference latency from 200ms to 50ms using TensorRT and quantization
- Collaborated with product teams to integrate ML models into mobile applications

ML Engineering Intern Jun 2021 – Sep 2021
Google *Mountain View, CA*
- Implemented BERT-based text classification for content moderation, achieving 92% accuracy
- Developed data preprocessing pipeline handling 10TB+ of text data using Apache Beam
- Created model monitoring dashboard using TensorBoard and custom metrics

Projects

LLM-Powered Code Review Assistant | *Python, LangChain, OpenAI API* 2024
- Built intelligent code review system using GPT-4 and custom prompting strategies
- Integrated with GitHub API to analyze pull requests, processing 500+ reviews daily

Real-time Fraud Detection System | *Python, XGBoost, Kafka, Redis* 2023
- Designed ML pipeline for real-time fraud detection with sub-100ms latency
- Achieved 95% precision and 88% recall using ensemble methods

Medical Image Classification | *PyTorch, OpenCV, AWS SageMaker* 2022
- Developed CNN model achieving 94% accuracy using transfer learning
- Deployed with AWS SageMaker with automatic scaling and monitoring

Publications

"Scalable Real-time Recommendation Systems" | *KDD 2023* 2023
"Efficient Model Serving for Large-scale ML Applications" | *MLSys 2022* 2022