Clara Zeng

213-248-9142 | qiang_zeng@berkeley.edu | Berkeley, CA | **LinkedIn** | **GitHub**

EXPERIENCE

Cmind AI | *Associate Data Scientist* May 2024 – Sep 2025

- Architected and deployed modular machine-learning microservices for feature selection, class imbalance handling, and model training, enabling independent scaling and reducing production failure risk.
- Enhanced the Earnings Per Share Surprise Prediction system by implementing and automated retraining pipelines, training XGBoost, SVM, & RNN models on 15+ years of financial data to achieve 90% quarterly forecast accuracy.
- Led development of a customizable end-to-end sentiment-analysis pipeline using FinBERT (financial domain) and OpenAI APIs for tailored Q&A with explanations; designed Oracle Cloud ingestion for cost-efficient legacy-system compatibility and AWS S3 storage to support a smooth migration to modern client-facing analytics workflows.
- Leveraged MLflow for experiment tracking, model version control, and deployment monitoring, ensuring reproducibility, transparency, and compliance with internal MLOps standards.

Bluebono | *Data Analyst Intern* June – Aug 2025

- Designed and implemented a loan-to-value estimation and market-score evaluation pipeline for residential properties, covering the full workflow from project scoping and data acquisition to model deployment.
- Directed dataset acquisition efforts by evaluating vendors, negotiating contracts, and integrating a real-time data feed from the California Regional Multiple Listing Service (CRMLS) via the Trestle API.
- Engineered and processed 11.3M property records with over 1,000 market features; applied unsupervised feature clustering to group correlated variables, reducing dimensionality by 65% and improving model interpretability.

Guangfa Futures | *Quantitative Analyst Intern* June - Aug 2023

- Applied regression, time-series and K-Means clustering models to identify predictive patterns in futures markets.
- Performed sensitivity analyses to quantify volatility impacts, developed monthly and quarterly forecasting models, and delivered actionable reports to support trading strategy decisions.

EDUCATION

University of California, Berkeley Aug 2026 (expected)
Masters in Analytics

University of Southern California May 2025
B.S. in Computer Science & B.A. in Applied and Computational Mathematics GPA: 3.94/4.0

RESEARCH AND PROJECT EXPERIENCE

AI-Based Career Advisor Jan - May 2025

- Designed and developed an interactive AI-based tool to help users plan and explore career paths and provide realtime applicable link by leveraging user-input skills and interests. Deployed a Streamlit UI for streamlined user interaction.
- Utilized large language models (GPT APIs and Llama3) and 1.3 million entries datasets (O*NET, LinkedIn) for career path recommendations and real-time skill gap analysis, providing specific learning resources to bridge identified gaps.

Super Computing In Pocket (SCIP) Lab Jan 2024 – May 2025

- Conducted research on privacy-preserving knowledge distillation for large language models, aligning attention and hidden layers between GPT-2 and DistilGPT2 and integrating Differentially Private SGD to enhance data confidentiality without significant loss in performance, ensuring data protection
- Re-engineered and optimized a text dataset distillation framework using Hugging Face models; redesigned data-loading pipelines and initiated a full algorithm reimplementation to enable integration with differential privacy mechanisms.

Hate Speech Detection Project Aug - Dec 2023

- Implemented and optimized Naive Bayes and BERT models for hate speech detection on social media, applying k-fold cross-validation, early stopping, & extensive preprocessing to achieve 92% accuracy on a dynamically generated dataset.

TECHNICAL SKILLS

Programming & Tools: Python (TensorFlow, PyTorch), SQL, R, C++; Java, MLflow, Git/GitHub, Streamlit, RESTful API integration (OpenAI, Trestle), Web Scraping, MLflow, ETL/ELT pipelines

Machine Learning: SVM, Tree Ensembles (DT/RF/XGBoost), kNN, Naive Bayes, Unsupervised (k-Means, GMM, PCA), deep learning architectures (LSTM, Transformers like GPT, BERT/FinBERT, seq2seq+attn, word2vec/doc2vec), GAT, Q-Learning

Cloud & Databases: AWS (S3, Glue), Oracle Cloud, SQL Databases (MySQL, Redshift, Oracle DB), , NoSQL Databases (MongoDB)