Barry (Zehui) Zhang

Machine Learning Engineer · Production ML systems, end to end

Irvine, CA barryzhangze@gmail.com linkedin.com/in/barry-zehui-zhang github.com/BarryZ807

Public edition. Employer-specific figures are left out; a detailed resume is available on request.

SUMMARY

Machine Learning Engineer who takes models to production and owns the systems end to end, from design to monitoring and measured impact. At CVS Health/Aetna, built and ran two production decision-automation systems, a real-time decision service and a daily review-reduction pipeline, leading delivery across business operations, Data Science and the automation team. Sole-author deep-learning research; 3 patent applications.

EXPERIENCE

CVS Health / Aetna
Machine Learning Engineer

Irvine, CA (Remote) · New York, NY
Feb 2025 – Nov 2026

Decision automation: decisions that used to need a human. Primary author of most of the team's engineering repositories.

Real-Time Decision System *end-to-end owner*

- Put a decision model into production as a real-time service that automation bots and an AI assistant call.
- Owned the whole system from design to production support: nightly data pipelines, a privacy-safe decision API on Kubernetes released behind automated and load tests, and a lookup service on a data product I built.
- Led the rollout across business operations, Data Science and the automation team: designed the API contract, ran the working sessions and staged a breaking change until every caller had switched.
- Built the monitoring that shows whether it works: daily metrics reconciled to source data, with missed decisions traced back to gaps in the model's coverage.

Manual-Review Reduction Pipeline *end-to-end owner*

- Built and ran the pipeline that finds false-positive review flags, from note matching to daily delivery to the business and automation teams, keeping those cases out of manual review.
- Implemented the matching engine (pattern extraction, fuzzy entity matching with edit-distance limits, three-way decisions) as an async Kubernetes service that Airflow starts and polls, with per-stage tracking, so one failure cannot take down the run.
- Cut runtime with async stages and tuned SQL, and shortened Data Science's path to production with an automated promotion workflow that became the team's reference pattern for new projects.
- Led the weekly sync with business operations, Data Science and the automation team, and turned each new Data Science rule into a tested production release.

Production Platform and Technical Leadership

- Owned the team's production services and Airflow pipelines across dev, test and prod, and moved them through platform migrations in orchestration, CI/CD and cloud infrastructure.
- Wrote the team's engineering handbook and served as the team's bridge to a partner organization.

CVS Health / Aetna
Machine Learning Engineer Intern

New York, NY
Jun 2024 – Aug 2024

- Built a full-stack GenAI cost-estimation tool (front end, back end and GCP pricing data) that prices architectures such as RAG in real time, plus a recommender that proposes options within budget.

Institute for Community Living (ICL)
Data Analyst Intern

New York, NY
Jun 2022 – Dec 2022

- Built auto-refreshing Power BI dashboards used across departments and quarterly KPI reports in Python, SQL and R.

RESEARCH

Independent Research: Risk-Aware Healthcare Cost Prediction *sole author* 2025 - Present

- Designed a multi-task FT-Transformer (PyTorch) that predicts expected cost, cost quantiles and high-cost risk from 72.8K person-years of US national survey data, with 0 quantile crossings vs. 1.54% for baselines.
- Doubled the high-cost spending captured vs. a calibrated baseline when flagging the top 0.5% (2.0x, p = 0.018), with significant gains across the top 0.5–2% and good calibration (ECE 0.006).
- Built to research-grade standards: 60 seeded, reproducible runs with 0 failed steps, survey-weighted bootstrap inference, and all 98 reported numbers machine-verified against run outputs.

INDEPENDENT PROJECTS

Wine Retail Commerce & Operations Platform Nov 2026

Portfolio build modeled on a family-owned wine retailer's public catalog · Next.js, React, TypeScript, PostgreSQL

- Designed and built the full platform: a bilingual storefront, wholesale ordering, and owner, advisor, warehouse and support workspaces with role-based access on a 41-table PostgreSQL schema.

Pawline: Custom Pet-Portrait Gift Store Nov 2026

Own brand concept, pre-launch · Next.js, React, TypeScript, PostgreSQL, Playwright

- Verified it like production: 117 automated tests with Chromium, Firefox and WebKit runs, a 25.9K-request load test (read p95 20–53 ms), exactly 120 charges for 120 concurrent orders and forced-crash recovery.

SKILLS

ML & Modeling	PyTorch, TensorFlow/Keras, scikit-learn, XGBoost/LightGBM, Optuna; tabular deep learning (FT-Transformer), multi-task heads, quantile regression, calibration, ablation, decision-curve analysis
Serving & MLOps	FastAPI, Pydantic, Docker, Kubernetes (GKE), REST API design, streaming, rate limiting, GitHub Actions, Jenkins, Snyk, SonarQube, pytest, HashiCorp Vault
Data & Cloud	GCP (BigQuery, Airflow on Cloud Composer 2 and 3, GKE, GCS, Cloud Monitoring/Logging, Secret Manager, IAM), AWS; SQL, pandas, NumPy
Observability	Grafana, Prometheus/PromQL, custom metrics, structured logging, OpenTelemetry, load testing
Full Stack	TypeScript, React, Next.js, Node.js, PostgreSQL, Prisma, Playwright
Languages & Domain	Python, TypeScript, SQL, R, Java, C++, Bash; Codex, Claude Code, Cursor, Git; healthcare data, RPA integration, HIPAA/PHI-safe design

EDUCATION

Weill Cornell Medicine, Cornell University New York, NY

M.S., Biostatistics and Data Science Dec 2024

Coursework: AI in Medicine, Computer Vision, Biostatistics I & II, Categorical & Censored Data Analysis, Statistical Learning

University of California, San Diego San Diego, CA

B.S., Joint Mathematics-Economics; Minor in Business; Provost Honors (4 quarters) Jun 2023

Coursework: Intro to Machine Learning, Data Structures & Algorithms (Java), Data Management, Probability, Mathematical Statistics I & II

MANUSCRIPTS & PATENT APPLICATIONS

Lead manuscript: Healthcare cost prediction when capacity is fixed: one model for expected cost, tail quantiles and high-cost risk (prepared for submission to JAMIA); 4 companion manuscripts in preparation.

Patent applications (China, CNIPA; filed 2026, pending)

- A method for predicting healthcare spending.
- A method for grouping a population by predicted risk.
- A method for allocating care resources from predicted risk.