Yiwen (Bruce) Zhang

New York, NY | yz8063@nyu.edu | linkedin.com/in/bruce-zhang-6a873421a/ | brucezhang.streamlit.app/

Education

New York University - Center for Data Science Sept. 2025 – May 2027
MS in Data Science (Exp.)
- Relevant Courses: Databases Systems; Machine Learning and Applications; Optimization; A/B Test Design.

University of California, Los Angeles (UCLA) Sept. 2021 – Jun. 2025
BS in Applied Mathematics, BS in Statistics and Data Science GPA: 3.74/4
- Dean's List for 7 quarters;
- Served as a Learning Assistant for R Programming and Chemical Structures, providing weekly office hours and review sessions to support 100+ students.

Skills

Programming and Analytics: Python (pandas, scikit-learn, PyTorch), R (tidyverse, quanteda, tm), SQL.

Big Data and Databases: SQL databases, erwin Data Modeler, MongoDB Atlas, Huggingface Dataset Hug.

Data Visualization: Tableau, PowerBI, kepler.gl, d3blocks, ggplot2, matplotlib, seaborn.

Other: Git, LaTeX, HTML/CSS, Streamlit.

Experience

Research Assistant, UCLA Mobility Lab – Los Angeles, CA June 2024 – Now
- Contributed to developing a machine learning pipeline leveraging annotated radar and trajectory data to predict vehicle and pedestrian motions, as part of the UCLA Mobility Lab team that won $750,000 in the U.S. Department of Transportation (USDOT)'s Intersection Safety Challenge.
- Implemented large-scale data pipelines integrating geospatial analytics and probabilistic modeling, improving POI-matching accuracy and enabling scalable urban mobility insights for city planning.
- Applied Hidden Markov Models with Viterbi decoding, combined with top-K POI probability weighting, to enhance the accuracy of activity-type inference by approximately 12% compared to baselines.
- Used Tableau, Kepler.io and d3blocks to build interactive dashboards to help visualize urban trajectory data and activity chain data.

Analyst Intern, Sunnet Systems Inc. – San Jose, CA Dec. 2022 – Jan. 2023
- Designed and executed SQL queries to extract, clean, and validate product inventory data, supporting accurate reporting across multiple business units.
- Built Excel-based dashboards with pivot tables, enabling the operations team to track 5,000+ products and reduce manual reporting time from 3 days on average to hours, improving delivery accuracy and reducing delays by 15%.

Projects

Visual Math Reasoning Benchmark for Multimodal LLMs (Ongoing) Sept. 2025
- Designed a full-stack benchmarking pipeline evaluating multimodal LLMs (Gemini 2.0, Mistral, Qwen, etc.) on generated math questions with diagrams, including automated validation, difficulty modeling, and topic-level performance analytics.
- Built data infrastructure with MongoDB Atlas + HuggingFace Datasets, authored unified MCQ/diagram schemas, and developed a Streamlit dashboard for run-level comparisons and statistical accuracy reporting (incl. CIs).
- Led methodology design (sampling, hybrid LLM validation, error profiling), enabling reliable comparison of student vs. master models across around 450 multimodal math problems.

Click-Through Rate (CTR) Prediction and Synthetic Data Usability Investigation Dec. 2024
- Evaluated the use of CTGAN synthetic data to handle highly imbalanced ad click-through datasets (98.5% one-sided), improving fairness and accuracy in real-world advertising systems.
- Boosted model performance with data transformations (e.g., down-sampling, synthetic sample diversification), achieving a 300% gain in F1-score and recall compared to baselines.
- Demonstrated that generating diverse minority samples enhances model generalization, with direct applications in ad revenue optimization.
- Analyzed how feature selection and imbalance adjustments impact synthetic data utility, providing guidelines for building more robust ML pipelines.

Sentiment & Usefulness Prediction of Patient Drug Reviews May 2025
- Built NLP pipeline on 215K+ drug reviews to classify sentiment (positive/neutral/negative) and predict review usefulness, applying TF-IDF, Random Forest, and XGBoost.
- Improved classification accuracy to 89% with Random Forest, outperforming baseline logistic regression by +19%, ensuring reliable sentiment detection.
- Predicted "usefulness" of reviews with F1-score of 0.87, enabling better prioritization of informative healthcare reviews for patients and clinicians.
- Applied BERTopic and LDA for theme extraction, revealing key patient concerns (e.g., side effects, treatment adherence), supporting data-driven healthcare insights.

Sales Call Optimization with Machine Learning (Salesmind.ai) Jun. 2025
- Analyzed 3,000+ customer call records and engineered features (duration, time-of-day, geography, recipient role), uncovering patterns behind low engagement (80%+ "Not Interested" responses).
- Built predictive models (Random Forest, Gradient Boosting) with SHAP explainability, improving call success classification accuracy by 20% over baseline; identified actionable insights: calls over 1 min increased appointment likelihood by 3×, executives (CEO/COO) were 40% more responsive than founders, and afternoon calls (1–5pm) yielded 2× higher engagement rates.