Richard Zhuang

richardzhuang0412.github.io richardzhuang0412@berkeley.edu richardzhuang0412

EDUCATION

University of California, Berkeley Expected Graduation: June 2025

Double Major in Applied Mathematics and Computer Science *GPA:* ***4.0/4.0***

Relevant Coursework: Machine Learning, Foundations of Large Language Models, Generative Models, Statistical Learning Theory, Probability Theory and Stochastic Processes, Principles and Techniques of Data Science, Design and Analysis of Computer Algorithms, Linear Algebra, Multivariable Calculus.

ACADEMIC PROJECTS

LLM2Vec: Learning Representations of Large Language Models Jan 2024 - Present

Advised by Ph.D. Tianhao Wu

- Designed and implemented a **matrix factorization model** and an **encoder-decoder based autoregressive model** to predict correctness of LLM responses on unseen queries.
- Conducted probing experiments to understand what information is contained in the learned embeddings.

PokerBench: Training LLMs to Become Better Poker Players Jan 2024 - Sep 2024

Advised by Ph.D. candidate Akshat Gupta

- Established a benchmarking dataset with 20K test cases and 560K training cases consisting of game scenarios and optimal decision labels verified by poker strategy solvers.
- Performed **supervised fine-tuning** on Llama-3-8B to improve test set accuracy from 26.0% to 78.3%.
- First-authored work under review in AAAI 2025.

LLM Behavior Analysis as AI Collectives Sep 2023 - Feb 2024

Advised by Ph.D. Yujin Kwon

- Assisted prompt engineering and performed simulations of Prisoner's Dilemma game and Story Relay game to evaluate LLM interactions as a group.
- Work accepted as a position paper in ICML 2024 (https://arxiv.org/pdf/2402.12590).

EXPERIENCES

CMU Delphi Group May 2024 - Sep 2024

Software Development Engineer Intern

- Assisted in developing R packages for real-time epidemiology forecasting through dogfooding.
- Enhanced package documentations by establishing tutorial notebooks on time-series modeling using **ARIMA**, **ETS**, and **Fourier Decomposition methods**.

UCSF Mindscape Lab Sep 2023 - Present

Machine Learning Researcher Intern

- Developed an agent-based **monte-carlo simulation** for disease transmission within hospitalized settings.
- Implemented a **hierarchical likelihood model** that learns the distribution of patient movement patterns.
- Currently experimenting with image reconstruction algorithms like **Masked Autoencoder** and **Vision Transformer** to recover omniscent transmission pattern from partial observations.

Sports Analytics Group at Berkeley Aug 2022 - May 2024

Data Analyst

- Fall 2022: Created a **decision tree regression model** to assess the fair market value of free agent players for the Los Angeles Sparks in WNBA.
- Fall 2023: Utilized **similarity scoring** and **clustering algorithms** to produce a scouting algorithm that identify cost-effective players for the Minnesota Lynx in WNBA.

ADDITIONAL

- Languages: Python, Java, R, SQL, MATLAB, JavaScript
- Tools & Packages: PyTorch, HuggingFace, vLLM, LitGPT, NumPy, Pandas, Matplotlib, AWS