Yan Zhou (Terry)

Cambridge, MA 02138 | +1 (857) 288-9283 | terryzhou1211@gmail.com
https://github.com/tz1211 | https://www.linkedin.com/in/yan-zhou-a9085b256/

Education

Harvard University Cambridge, United States
MS in Data Science, ***GPA: 4.0*** Sep 2025 – May 2027 (expected)

- **Selected Coursework**: AI Alignment and Safety, Deep Learning, Machine Learning, MLOps, Computing at Scale, Decision Theory, Statistical Inference

London School of Economics and Political Science (LSE) London, United Kingdom
BSc in Politics and Data Science, ***First Class Honors*** Sep 2022 – Jun 2025

- **Selected Coursework**: Artificial Intelligence, Machine Learning, Multivariate Calculus, Linear Algebra, Algorithm and Data Structure, Probability and Distribution Theory, Mathematical Proof and Analysis, Databases

Research Experience

ML Foundations Group, Harvard School of Engineering and Applied Sciences Cambridge, United States
Advisors: Prof. David Alvarez-Melis, Jonathan Geuter, Sara Kangaslahti Nov 2025 – Present

- Ongoing research: Investigating generalization behavior of boomerang-distillation on post-trained LLMs.
- Built model capability evaluation pipelines and optimized model distillation training workflow for multi-GPU cluster.
- Showed that post-trained teacher capabilities can be zero-shot transferred via layer patching onto base student models.

AI4LIFE Lab, Harvard School of Engineering and Applied Sciences Cambridge, United States
Advisors: Prof. Hima Lakkaraju, Prof. Yonatan Belinkov, Dr. Shichang Zhang Sep 2025 – Present

- Ongoing research: Activation steering for reasoning control in LLMs.
- Demonstrated that steering for specific model personas can lead to more concise model CoT and enhanced performance on reasoning benchmarks.

Data Science Institute, LSE London, United Kingdom
Advisor: Prof. Jonathan Cardoso-Silva Jun 2024 – Jun 2025

- Built an LLM chatbot for answering queries from student and faculty about LSE policy and regulations.
- Led backend development utilizing RAG framework and developed an automated pipeline to crawl, preprocess, and embed LSE regulations into a vector database.

Data Science Institute, LSE London, United Kingdom
Advisor: Prof. Jonathan Cardoso-Silva Jun 2023 – Oct 2024

- Built a machine learning framework to identify factors that drive MP rebellions in UK Parliament.
- Utilized speech embeddings to estimate political ideological differences, achieving a 20% increase in F1-score for predicting voting behavior.

Relevant Projects

Are Personas All You Need? Stress-Testing Persona Vectors as Alignment Tools Cambridge, United States
AI Safety Course Project (Advised by Prof. Boaz Barak, Roy Rinberg, Natalie Abreu) Nov 2025 – Dec 2025

- Evaluated the reliability of persona vectors as low-dimensional alignment tools for modeling behavioral traits.
- Analyzed geometric properties of persona vectors to quantify cross-trait interference during finetuning.
- Identified key robustness failures such that minor natural language variations produce persona vectors with cosine similarity as low as 0.6.

DataDetox – AI Model Lineage and Safety Tracking System Cambridge, United States
MLOps Course Project Sep 2025 – Dec 2025

- Developed an agentic AI system to trace model and training data lineage, allowing ML practitioners to identify hidden upstream risks and enhance model and data provenance.
- Engineered an asynchronous multi-agent backend that retrieves disparate metadata and queries from graph database to construct model and data lineage trees.
- Deployed a production-ready stack using Docker and Kubernetes for scalability.

Additional Experience

Harvard AI Safety Student Team Cambridge, United States
Technical AI Safety Fellowship Sep 2025 – Nov 2025

- Engaged in structured weekly reading group exploring advanced AI safety topics ranging across scheming, control, alignment, interpretability, and scalable oversight.

Technical Skills

Programing Languages:	Python (Advanced), R (Advanced), Typescript (Beginner), SQL, LaTeX
Machine Learning:	PyTorch, Transformers, Huggingface, Scikit-learn
ML Engineering:	vLLM, Finetuning, Model Distillation, Quantization, Agents, OpenAI API, RAG, Vector Databases, MongoDB, Neo4j, Docker, Kubernetes, Weights & Biases, Git, CI/CD, FastAPI, GCP, AWS, Slurm
Data & Visualization:	NumPy, Pandas, SciPy, Matplotlib, Plotly, Seaborn, Jupyter