Zeyu Zhang

Email: zeyuzhang2028@u.northwestern.edu

Homepage: https://ZeyuZhang1901.github.io

Research Interests

- Large Language Models (LLMs) and Trustworthy AI
- Knowledge Leakage Detection and Mitigation in Complex Reasoning
- Reinforcement Learning-based Fine-Tuning (RLHF, DPO)
- LLM Interpretability and Black-Box Model Analysis
- Factual Memorization and Unlearning in LLMs

Education

- **Northwestern University** — **Evanston, U.S.**
  Statistics and Data Science, Doctor of Philosophy, GPA: 3.91/4.00 — *Sept 2023 - Present*
- **University of Science and Technology of China (USTC)** — **Hefei, P.R.China**
  Electronic Information Engineering, Bachelor of Engineer, GPA: 3.93/4.30 — *Sept 2019 - Jun 2023*
  - Wang Xiaomo Talent Program in Cyber Science and Technology
  - Talent Program in Information Science and Technology
  - ***China National Scholarship honored by Ministry of Education of the PRC, 2020-2021, top 1%***
  - ***Rank: 5/213 in School of Information Science and Technology***
- **University of Science and Technology of China (USTC)** — **Hefei, P.R.China**
  Artificial intelligence, Certificate, Minor, GPA: 3.93/4.30 — *Sept 2021 - Jun 2023*
  - Talent Program in Artificial Intelligence
  - ***Outstanding Undergraduate Honorary Rank, top 5%***

Research

- **Temporal Knowledge Leakage in LLM Prediction Tasks** — **Northwestern University**
  Advisor: Prof. Bradly C. Stadie — *Sep 2025 – Present*
  - Formalized **temporal knowledge leakage**—the inadvertent use of information not publicly available at a specified reference time—in LLM-based prediction tasks such as legal case outcomes, salary estimation, and stock ranking.
  - Introduced a four-phase detection pipeline with well-defined Shapley-weighted metric.
  - Proposed the **Temporal LLM Agent**, a multi-phase architecture that proactively prevents leakage during generation through iterative claim verification and regeneration.
  - Demonstrated substantial leakage reduction while maintaining prediction quality across legal case prediction, salary estimation, and stock ranking benchmarks.
- **LAMP: Linear Attribution Mapping Probe** — **Northwestern University**
  Advisor: Prof. Bradly C. Stadie — *Jan 2024 – Jan 2025*
  - Developed **Linear Attribution Mapping Probe** (**LAMP**), a framework for interpreting black-box language models by fitting locally linear surrogate models grounded in the model's own self-reported explanations.
  - Designed a perturbation-based method to probe model prediction sensitivity and extract decision surfaces without accessing gradients, logits, or internal states—enabling practical interpretability for proprietary LLMs.
  - Implemented extensive experiments across sentiment classification, controversial-topic detection, and harmful response auditing.
  - ***Accepted as Spotlight at AISTATS 2026.*** (`Link`: LAMP: Extracting Locally Linear Decision Surfaces from LLM World Models)
- **Factual Memorization and Learning in Large Language Models** — **Northwestern University**
  Advisor: Prof. Bradly C. Stadie — *Jan 2024 – Jan 2025*
  - Conducted a comprehensive analysis of LLM memorization behaviors under different fine-tuning paradigms, including SFT and DPO, building upon and extending insights from FinetuneBench.
  - Reproduced key experimental findings from Stanford's FinetuneBench, and systematically evaluated LLM robustness to input rephrasings and temporal shifts in question formulation.
  - Proposed a formal distinction between **passive memorization** (from exposure) and **positive memorization** (via direct QA supervision), demonstrating that the latter achieves superior memorization efficiency.
  - Designed temporal generalization experiments by injecting future-dated facts into the training set. Found no generalization beyond training, but demonstrated that a lightweight system prompt enforcing temporal consistency can mitigate this overfitting.
- **CUOLR: Unified Off-Policy Learning to Rank** — **Princeton University**
  Mentors: Prof. Mengdi Wang, Prof. Huazheng Wang — *May 2022 – Dec 2023*

- Formulated a unified framework that models various click models in off-policy Learning to Rank (LTR) as a *Markov Decision Process (MDP)*, enabling principled application of offline reinforcement learning.
- Proposed the **Click model-agnostic Unified Off-policy Learning to Rank** (**CUOLR**) algorithm that adapts to a wide range of click models *without requiring explicit debiasing or prior knowledge.*
- Demonstrated that offline RL methods (e.g., DQN, SAC, BCQ, CQL) can be applied effectively to LTR by leveraging the MDP formulation, maintaining consistency and robustness under diverse click model assumptions.
- Validated CUOLR on large-scale real-world datasets, achieving state-of-the-art performance and significantly improved robustness across heterogeneous click models.
- ***Published at NeurIPS 2023.*** (`Link`: Unified Off-Policy Learning to Rank: a Reinforcement Learning Perspective)

Projects

- **Signal Distortion Measurement Device Design** **USTC**
  Mentor: Dr. Wei Lu *Apr 2021 - Nov 2021*
  - Reduced distortion error to around **0.5%** with requirement of **3%** and extended measurement band width to **1k∼100k**.
  - Applied **window functions** to reduce Spectrum Leakage. Considering both effectiveness and feasibility, I chose Hanning window finally.
  - Designed an algorithm to **accurately detect the center spectrum** by adding energy from nearby spectrum lines.
  - Developed an LCD to visualize relevant data and input analog signals.

Publications

- **Unified Off-Policy Learning to Rank: a Reinforcement Learning Perspective**: Published at NeurIPS 2023.
  - **Zeyu Zhang**, Yi Su, Hui Yuan, Yiran Wu, Rishab Balasubramanian, Qingyun Wu, Huazheng Wang, Mengdi Wang
  - [Arxiv Link], [OpenReview], [Code]
- **LAMP: Extracting Locally Linear Decision Surfaces from LLM World Models**: Accepted as **Spotlight** at AISTATS 2026.
  - Ryan Chen, Youngmin Ko, **Zeyu Zhang**, Catherine Cho, Sunny Chung, Mauro Giuffrè, Dennis L. Shung, Bradly C. Stadie
  - [Arxiv Link], [OpenReview]

Honors and Awards

- ***National Scholarship honored by Ministry of Education of the PRC (top 1%)*** *Oct 2020*
- ***Outstanding Student Scholarship (Class A, Top 1%)*** *Sept 2020*
- The National Undergraduate Electronic Design Contest, 2nd Prize Nationally, 1st in Anhui Province *Nov 2021*
- Scholarship for Talent Program in Basic Disciplines (Class A, Top 3%) *Oct 2020*

Teaching Assistant

- **Data Science 2 with Python (STAT 303-2)** **Northwestern University**
  Assist with Prof. Emre Besler *Jan 2026 - Present*
  - This course introduces supervised machine learning in Python, with a focus on linear and logistic regression. It prepares students for learning advanced machine learning methods.
  - Responsibilities: Grade homework assignments and in-person exams.
- **Applied Multivariate Analysis (STAT 348)** **Northwestern University**
  Assist with Prof. Thomas Severini *Sept 2025 - Dec 2025*
  - Held weekly TA discussion sessions to answer questions related to course material.
  - Graded homework assignments.
- **Data Science 3 with Python (STAT 303-3-21)** **Northwestern University**
  Assist with Prof. Emre Besler *Apr 2025 - Jun 2025*
  - The course introduces non-linear statistical models such as splines, and tree-based methods such as random forests and boosting. It also introduces statistical concepts such as model bias and variance.
  - Graded homework assignments and exams.
- **Data Science 2 with Python (STAT 303-2-22)** **Northwestern University**
  Assist with Prof. Emre Besler *Jan 2025 - Mar 2025*
  - This course introduces supervised machine learning in Python, with a focus on linear and logistic regression. It prepares students for learning advanced machine learning methods.
  - Graded homework assignments and exams.
- **Introduction to Probability and Statistics (STAT 210-0-20)** **Northwestern University**
  Assist with Prof. Maxim Sinitsyn *Sept 2024 - Dec 2024*
  - This class covers descriptive statistics, probability, random variables, sampling distributions, confidence intervals, and significance tests.
  - Held discussion sessions twice a week to answer homework questions. Grade the midterms and final exam.