

Zelong Xu

📍 501 North Henry Street, Apt 406, Madison, WI 53703, USA

✉ zxu684@wisc.edu ☎ +1 608-609-4547

EDUCATIONAL BACKGROUND

University of Wisconsin-Madison

Sep 2024 – Jan 2027

Major: Statistics

(Expected)

Certificate: Computer Science, Mathematics

- GPA: 3.90/4.00
- Core Courses: Interpretable Machine Learning, Matrix Methods in Machine Learning, Deep Learning, Numerical Analysis

Shandong University

Sep 2022 – Jun 2024

Major: Finance

- GPA: 88.1/100
- Core Courses: Linear Algebra, Calculus, Statistics, Probability, Econometrics, Big Data Application with Machine Learning

RESEARCH PROJECTS

Confidence-Calibrated Circuit Discovery for Large Language Models

May 2026 – present

- Developing a statistical framework that brings finite-sample confidence guarantees to mechanistic circuit discovery in LLMs, pairing randomization-based causal effect tests (per-edge p-values) with Model-X knockoff false discovery rate (FDR) control to separate genuine circuit components from spurious selections.

DOG-DPO: Dynamic Optimization in Geometry for Multi-Dataset DPO

Jan 2026 – May 2026

- Proposed DOG-DPO, the first framework to model multi-dataset preference data as a structured geometric object, recasting DPO data selection from per-sample scoring into a geometric coverage problem over anchor and residual subspaces, with a provable spanning-then-filling phase transition.
- Built a training-free pipeline (no teacher LLM or reference DPO run), 15–35× faster than DEITA and Selective-DPO; across six safety and jailbreak benchmarks on LLaMA-3.2-3B and Gemma-3-1B, it matches full-data DPO using only 11.6% of preference pairs while reducing inter-dataset selection bias.

EmoPair: Counterfactual Paired Listwise Alignment for Emotion-Aware Speech LLMs

Nov 2025 – May 2026

- Proposed EmoPair, a counterfactual paired listwise alignment framework for emotion-aware speech LLMs that synthesizes same-text-multiple-emotion speech variants (7 Ekman emotions) via Qwen3-TTS and trains the model with listwise preference optimization plus a paired counterfactual constraint anchoring emotional responses against a neutral baseline.
- Outperformed binary-GRPO (PALLM) and Likert-GRPO (ParaS2S) baselines on the same Qwen3-Omni-30B backbone, improving response appropriateness from 72.4 to 77.2 on IEMOCAP and 47.9 to 55.8 on held-out RAVDESS (OOD), and raising counterfactual swap consistency from 0.62 to 0.81.

BR-LOCAL: A Solver-Verified Benchmark for Ripple Effects under In-Context Knowledge Editing

Oct 2025 – April 2026

- Designed BR-LOCAL, the first solver-verified benchmark for ripple-effect evaluation under in-context knowledge editing (ICE), addressing the 33–76% annotation error in Wikidata-grounded predecessors (MQuAKE, RippleEdits): minimal patches on synthetic Horn knowledge bases yield episodes with deterministic gold change-sets and continuous per-query distance-to-patch $d(q)$, including an unreachable ∞ -stratum; paired COLD-START and ONLINE-FULL tracks isolate static reasoning from in-context edit propagation.
- Built a procedural Horn-KB synthesizer and end-to-end evaluator harness producing 5K episodes; benchmarked SOTA LLMs like GPT-5.2, DeepSeek-R1, Claude-4.6, and Qwen3-235B, activating the reasoning chain raises Change-F1 from 0.12 to 0.90, establishing that ripple-effect capability is gated by inference-time reasoning rather than parametric knowledge.

Multi-Emotional Probing for Controllable Language Generation

Mar 2025 – May 2025

- Proposed and implemented a novel probe-and-inject framework to achieve fine-grained emotional control in LLMs, successfully extracting emotion vectors from BERT’s latent space and steering Llama-2 model’s generative output.
- Discovered that emotion representations are localized in BERT’s middle layers (5-17) and exhibit a near-orthogonal geometric structure, enabling precise control with high F1-scores for targeted emotions like ‘joy’ (0.93) and ‘anger’ (0.90) via vector injection.

Enhancing Fake News Detection with Machine Learning

Apr 2024 – Jun 2024

- Developed a fake news detection system by systematically benchmarking multiple classifiers (Logistic Regression, SVM, Naive Bayes) on a 49,910-article dataset, using a TF-IDF based NLP pipeline for feature extraction.
- Optimized the top-performing model via hyperparameter tuning, achieving 96.15% accuracy and 93% recall, effectively distinguishing between real and fake news articles.

INTERNSHIP EXPERIENCE

Icarus Fund, Remote

Jul 2024 – Aug 2024

Quantitative Analyst

- Developed and backtested Python-based quantitative models for S&P 500 companies, optimizing portfolio strategies with Sharpe Ratios up to 1.42.

SKILLS & LANGUAGES

Technical Skills: Python, Java, JavaScript, HTML, R, SQL, PyTorch, TensorFlow, Scikit-learn, SPSS, Stata

Languages: Chinese (Native), English (Fluent, TOEFL: 109/120)