Zikai (Cyrus) Zhou

+1 (650) 512-5171 | Email: zikai@stanford.edu | https://zikai-zhou.github.io/

PhD student building AI systems whose decisions can be verified, challenged, and changed.

Education

Stanford University, PhD in Computer Science, Stanford NLP Group. 2024 - Present
University of Chicago, BS in Computer Science, Cum Laude, GPA: 3.9/4.0. 2020 - 2024

Selected Publications

1. **Zikai Zhou**, Yufei Jin, Yilin Xu, Chieh-Ju Chao, Monica S Lam, "Accountable AI with Grounded, Faithful, Consistent, Actionable Rationales: A Case Study in Clinical Trial Matching with VERDICT", *October 2026, The 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP).* **Paper Award Nominee.**
2. **Zikai Zhou**, Yufei Jin*, Yilin Xu*, Chieh-Ju Chao, Monica S Lam, "SatIR: Scalable High-Recall Constraint Satisfaction-Based Information Retrieval for Clinical Trials Matching", *October 2026, Third Conference on Language Modelling (COLM).* (* indicates equal contribution)
3. **Zikai Zhou**, Qizheng Zhang, Hermann Kumbong, Kunle Olukotun. "LowRA: Accurate and Efficient LoRA Fine-Tuning of LLMs under 2 Bits," *July 2025, The International Conference on Machine Learning (ICML).*
4. Bogdan Stoica, Utsav Sethi, Yiming Su, **Cyrus Zhou**, Suman Nath, Jonathan Mace, Madan Musuvathi, and Shan Lu, "If At First You Don't Succeed, Try, Try, Again...? Automatically Detecting Retry Bugs in Distributed Applications", *November 2024, The 30th Symposium on Operating Systems Principles (SOSP).*
5. Lefan Zhang, **Cyrus Zhou**, Michael L. Littman, Blase Ur, and Shan Lu, "Helping Users Debug Trigger-Action Programs", Jan 2023, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT/UbiComp).*

Research Experience

Stanford OVAL Clinical Trial Matcher [code] [project page] March 2025 - Present

- Project lead across research direction, system design, implementation, evaluation, and both papers; five-person Stanford–Mayo collaboration; designed the clinician evaluation with two collaborating physicians. Trial criteria and patient charts compile once into a shared formal representation, which then drives both stages: retrieval narrows a corpus of thousands to a handful of candidates, and each candidate gets an accountable eligibility decision – matching by what a trial actually requires rather than by textual similarity.
- **SatIR (retrieval):** built a database that splits each trial's criteria into separately checkable conditions with the time window each must hold over, mapped through SNOMED-CT so a chart fact matches a criterion written at a different level of detail. Search runs entirely in SQL with no model call, ruling a trial out only when a chart fact contradicts a criterion – so every rejection names it. Recovers 92-94% of the relevant-and-eligible trials any method finds, against 54-70% for TrialGPT, the strongest of six baselines that also include BM25 and four biomedical dense retrievers (SIGIR 2016); 0.80-0.86 recall against 0.25-0.34 on TREC 2022.
- **VERDICT** reaches F1 0.828 (TREC 2021) and reports what the chart left unresolved and which conditions would change the outcome, stating the requirement a trial imposes rather than any solver-invented value. A distilled Qwen2.5-7B matches GPT-5-mini at the same F1, beating its own end-to-end chain-of-thought by 16.5 points.
- Owned the ~200K-line Python system and released it open-source under Stanford OVAL, Apache-2.0 (github.com/stanford-oval/clinical-trial-matching), with reproduction recipes and a usable demos.

Efficient ML: Quantization, Kernels & DSL October 2022 - March 2025

- **LowRA** (Prof. Kunle Olukotun @ Stanford): proposed the idea, led a team of three. Quantization mapping/threshold search, precision assignment, and CUDA kernels pushed LoRA base weights to 1.15 bits, outperforming baselines at every precision. **ICML 2025**. Also prototyped a DSL for specifying quantized operations (Prof. Fredrik Kjolstad @ Stanford), and earlier UChicago work on low-precision quantization and SIMD dataflows.

Software Engineering, HCI, Security March 2022 - June 2024

- **Wasabi** (Prof. Shan Lu): static analysis plus LLMs plus exception injection to expose retry bugs in distributed systems such as Hadoop; built the per-unit-test coverage tool used to select bug-revealing tests. **SOSP 2024**. **TapDebug** (Profs. Lu, Ur): designed the user-study methodology and obstacle model. **ACM IMWUT 2023**.

Industry & Projects

Netpreme, Cambridge, MA - Member of Technical Staff Intern Summer 2025

- Built the LLM inference cost-modelling toolkit adopted as the internal baseline cost/throughput calculator: an LP optimizer with greedy and heuristic search over data placement (GPU/CPU/NVMe), offloading, and batch sizing on heterogeneous XPUs. Implemented a Mixture-of-Experts simulator for expert-layout and interconnect trade-offs.

Awards & Teaching

- Enlight Foundation Graduate Fellowship ($90K), Stanford; CRA Outstanding Undergraduate Researcher, Honorable Mention; Quad Undergraduate Research Scholar (3x); Jeff Metcalf Award.
- Head Teaching Assistant (Autumn 2026) and Project Mentor (Autumn 2025), CS224V - Conversational Virtual Assistants with Deep Learning, Stanford. Teaching Assistant for Introductory Courses at UChicago CS.