William L. Anderson

hello@wlanderson.com | wlanderson.com | github.com/wlanderson0 | (612) 961-0497

Education

University of Wisconsin–Madison GPA: 4.0/4.0 Sept 2023 – May 2027 (Exp.)

- BS in Computer Sciences, Data Science, and Mathematics with Certificate in Public Policy
- Honors in Computer Sciences, Sophomore Research Fellowship

Research

AI Safety Research Fellow – **UChicago Existential Risk Lab** – Chicago, IL June 2025 – Present

- Leading a project on multi-agent AI safety and security advised by Lewis Hammond (Cooperative AI Foundation).
- Extending UKAISI Inspect to support complex multi-agent orchestration and monitoring protocols to study a novel threat model using diffuse, locally hard-to-detect attacks against multi-agent systems.
- Collaborating on projects which are: developing physical adversarial examples for military CV; analyzing feasibility of kinetic action for deterrence of superintelligence projects; studying impact of network topology on adversarial robustness in multi-agent systems. Papers expected for all projects.

Research Fellow – **Yiqiao Zhong Lab** – Madison, WI January 2025 – Present

- Studying on mechanistic effects of RL post-training on LLM robustness using evaluations and interpretability.
- Post-training LLMs with verifiable rewards using a Group Relative Policy Optimization training pipeline in VERL.
- Using a novel tree decomposition of CoT to study impact of problem topology on LLM reasoning robustness.

Student Researcher – **Junjie Hu Lab** – Madison, WI October 2024 – August 2025

- Developed and benchmarked a novel technique which used task vectors to speed up LLM safety fine-tuning.
- Applied vector decompositions and novel metrics to activation spaces of open-source LLM checkpoints to mechanistically study the emergence of generalization capabilities during pre-training.

Independent Study – **Alignment Research Engineer Accelerator (ARENA)** – Remote June 2024 – August 2024

- Built custom training pipelines for CNNs, VAEs, transformer language models, and RL agents.
- Re-implemented core algorithms including autograd, backprop, and PPO, as well as key mechanistic interpretability techniques such as linear probes, sparse autoencoders, and activation patching.

Leadership and Involvement

Wisconsin AI Safety Initiative, Ltd. 501(c)(3) – Madison, WI September 2023 – Present

Director April 2025 - Present

- Directing 30 person leadership team to coordinate large-scale education and upskilling programs, university relations, professional networking, and outreach for organization with 130+ members and $30,000 budget.
- Organizing an annual symposium with 320 experts and students to discuss the myriad impacts of AI on society.

Operations Lead April 2024 – April 2025

- Presented an interactive demo to 100+ U.S. Congressional staffers, policymakers, and journalists in Washington, D.C. on the capabilities and risks of multi-AI-agent systems at the Center for AI Policy's Advanced AI Exposition.
- Organized interdisciplinary panel on AI with professors from Computer Sciences, Philosophy, and Education.
- Awarded Open Philanthropy University Organizer Fellowship for facilitating advanced technical reading group, designing curricula, setting organizational strategy, scheduling 10+ weekly meetings, and leading projects.

Lifeguard Supervisor (Seasonal) – **City of Deephaven** – Deephaven, MN May 2024 - October 2024

- Conducted on-the-job training for new lifeguards, teaching emergency procedures, scanning, and rescues.

Lead Lifeguard (Seasonal) – **Minnetonka Aquatics** – Minnetonka, MN May 2021 – May 2024

- Directed beach teams of up to 6 lifeguards in emergency situations, developing leadership skills under pressure.

Presentations

- *Multi-Agent Security Risks*, International Conference on Information Systems, Nashville, TN, 2025 (Upcoming)
- *Replacement Models for Interpretability*, Yiqiao Zhong Reading Group, Madison, WI, 2025
- *A Case for Responsible AI*, IT Professionals Conference, Madison, WI, 2025
- *Security Risks in LLM Multi-Agent Systems*, Center for AI Policy Congressional Exposition, Washington, DC, 2025

Technical Knowledge

Languages: Python, R, C, Java, Lua, JavaScript, SQL, Bash

Frameworks: PyTorch, TF/Keras, Docker (Compose), Spark, HDFS, Arrow, Cassandra, BigQuery, Kafka, Inspect