William L. Anderson

hello@wlanderson.com | wlanderson.com | github.com/wlanderson0

Education

University of Wisconsin–Madison GPA: 4.0/4.0 Sept. 2023 – Dec. 2025

- BS in Data Science with Certificate in Computer Sciences
- Sophomore Research Fellowship, OpenPhil University Organizer Fellowship

Research

Research Scholar – **ML Alignment and Theory Scholars** – Berkeley, CA / Oxford, UK Jan. 2026 – Present

- Developing robust multi-agent security benchmark and evaluating security-by-design frameworks.
- In the Multi-Agent Security stream under Dr. Christian Schroeder de Witt (Oxford).

Research Fellow – **UChicago Existential Risk Lab** – Chicago, IL June 2025 – Jan. 2026

- Leading a project on multi-agent AI security advised by Dr. 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 computer vision models; analyzing feasibility of kinetic action for deterrence of superintelligence projects.

Research Fellow – **Yiqiao Zhong Lab** – Madison, WI Jan. 2025 – Jan. 2026

- 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 TRL.
- Using a novel tree decomposition of CoT to study impact of problem topology on LLM reasoning robustness.

Student Researcher – **Junjie Hu Lab** – Madison, WI Oct. 2024 – Aug. 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.

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

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

Advisor Dec. 2025 - Present

- Continuing in an advisory role post-graduation.

Director April 2025 - Dec. 2025

- 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-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.

Pathfinder Mentor – **Kairos Project (Contract)** – Remote Aug. 2025 – Present

- Mentoring AI Safety group organizers, resulting in more than 2x group growth on average.

Presentations

- *Multi-Agent Security Risks*, International Conference on Information Systems, Nashville, TN, 2025
- *Panelist*, Wisconsin Technology Council Panel on AI Safety, Madison, WI, 2025
- *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