David Crispell

714-931-8484 | davidglenrc@gmail.com | LinkedIn | GitHub | Website

Education

University of Colorado, Boulder — Boulder, Colorado

B.S. in Computer Science, B.A. in Mathematics — *Expected May 2027*

- Relevant Coursework: Machine Learning, Artificial Intelligence, Deep Reinforcement Learning, Probability & Statistics, Linear Algebra, Real Analysis, Data Structures & Algorithms, Computer Systems
- GPA: 3.874 (June 2026) | SAT: 1570 (October 2023)

Experience

Independent AI Safety Researcher — June 2026 – Present

Safety and Generalization — *Boulder, Colorado*

- Researching generalization and alignment failure modes in recursive-self-improvement (RSI) loops, especially whether hidden dispositions can pass through neutral model-generated data and strengthen in successor models
- Found that training-data order materially changes subliminal-transfer strength at fixed weight initialization, with evidence that number completion and wolf-related completions reuse some of the same circuitry
- Reproduced subliminal learning and rebuilt Neuronpedia's Jacobian-lens environment for controlled steering and custom experiments; secured $3,600 in BlueDot funding

AI Engineer Intern — May 2025 – August 2025

Gravity Foundation — *Boulder, Colorado*

- Led an independent internship project implementing a context-aware RAG scheme for agentic systems
- Developed and deployed techniques to classify and amend failure cases for large-scale database queries
- Minimized QA cloud spending by implementing a short circuit to detect and halt thrashing agents using a simple classifier
- Wrote hundreds of test cases, increasing coverage from 78% to 93% in an industry-scale codebase

Projects

April 2026 Mech Interp Puzzle Analysis — Apr. 2026

- Reverse-engineered two pretrained attention-only transformers from a mechanistic interpretability puzzle, recovering the learned max-of-list algorithm
- Causally validated the circuits with weight inspection, head ablations, and activation patching, explaining why the 2-layer digit-tokenized model required depth

President — Oct. 2024 – Present

Artificial Intelligence Student Initiative, CU Boulder — *Boulder, Colorado*

- Founded organization to foster student engagement with AI safety and interpretability
- Hosted reading groups, research challenges, discussions, and lectures; recruited students to tutor foundational AI safety and capabilities concepts

NeurIPS LLM Privacy Challenge | *Python, GitHub, Google Colab* — Aug. 2024 – Oct. 2024

- Collaborated with five researchers to red-team a fine-tuned Llama 3.1 8B model using contrastive activation addition
- Doubled the base rate of personally identifiable information regurgitation and documented the methodology, experiments, and findings in a technical report

Technical Skills

Languages: Python, C/C++, JavaScript, SQL, Swift
ML/Research: PyTorch, Hugging Face Transformers/PEFT, TransformerLens, SAELens, TensorFlow, OpenCV
Developer Tools: Git, Docker, RunPod, GCP, Prime Intellect, OpenRouter, Google Colab
Research Interests: AI safety, Reinforcement Learning, Subliminal Learning, Generalization, Mechanistic Interpretability, Model Organisms of Misalignment