Adam Karvonen

adam.karvonen@gmail.com • github.com/adamkarvonen/ • linkedin.com/in/adam-karvonen/
Google Scholar

Summary: Machine learning researcher and engineer specializing in AI safety and interpretability. First author of three ML main conference papers (NeurIPS, ICML, COLM), including an oral presentation at the ICML Mech Interp workshop. Demonstrated success in applying LLM research to real-world problems. Published blog posts on ML projects with significant attention (3x Hacker News front page, 700k Twitter views).

Education: B.S., Computer Science, Southern New Hampshire University, graduated 03/2023, **GPA: 4.0**

Skills

- Python, Pytorch, Pandas, Scala, FastAPI, Pydantic, Docker, LLMs, Machine Learning, Git, SQL
- Clear communication of complex technical concepts, project management

Work Experience

Machine Learning Alignment and Theory Scholars, ML Researcher, Berkeley, CA, 04/24-present

- Conducted research on interpretability on LLMs under Google Deepmind researcher Neel Nanda
- Obtained 1 year independent research grant, now working with Owain Evans in MATS 8.0
- Infrastructure lead and research co-lead for the SAE Bench project, an open-source suite of Sparse Autoencoder (SAE) evaluations, involving 10+ collaborators
- Published three first-author main conference papers on interpretability and SAEs
- Co-developed and maintained dictionary_learning, a popular open source SAE training repo.
- Authored an "Introduction to Sparse Autoencoders" guide, now ranked ~#1 on Google

ML Consultant, Independent, 09/24-01/25

- Used SAE steering to prevent unwanted codegen behavior for a YC startup, convincingly beating baseline methods, with a codebase that is "awesome / very readable" - Dan Mossing, OpenAI

Galois, Research Engineer, Minneapolis, MN, 06/22-02/24

- Lead development of tools using LLMs to safely refactor thousands of lines of code and formally verify software
- Authored 2 Galois website blog posts about these tools: Using GPT-4 to Assist in C to Rust Translation, Applying GPT-4 to SAW Formal Verification
- Developed an LLM knowledge graph query and construction tool that was successfully integrated into client-delivered software.
- Developed backend features of human subjects experiment application using FastAPI, Docker, SQL, and Pytorch

Rapid Design Solutions, Robot and CNC programmer, Seattle, WA, 09/20 - 09/21

- Programmed, set up, and optimized industrial robots to automate tasks
- Had to wear many hats and solve unfamiliar problems at a small, busy company

Thompson Precision, CNC Programmer, Kalispell, MT, 05/19-09/20

- Went from green trainee to machining $20,000 5 axis rocket engine prototypes in 1.5 years

Projects

Chess-GPT's Internal World Model

- Trained a GPT to play chess at 1500 Elo. Used interpretability based intervention on the model's internal activations to increase its skill level by 260%.
- Blog posts: Chess-GPT's Internal World Model, Manipulating Chess-GPT's World Model

FIRST Robotics

- Started the team as a freshman and was captain and head programmer for all 4 years
- Handled business of the team - recruiting, fundraising, and finances

Publications

2025

- SAEBench: A Comprehensive Benchmark for Sparse Autoencoders
  - **Adam Karvonen**, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Callum Stuart McDougall, Yeu-Tong Lau, Eoin Farrell, Arthur Conmy, Kola Ayonrinde, Demian Till, Matthew Wearden, Samuel Marks, Neel Nanda
  - *ICML main track*
- Learning Multi-Level Features with Matryoshka Sparse Autoencoders
  - Bart Bussmann, Noa Nabeshima, **Adam Karvonen**, Neel Nanda
  - *ICML main track*
- Revisiting End-To-End Sparse Autoencoder Training: A Short Finetune Is All You Need
  - **Adam Karvonen**
- Steering Fine-Tuning Generalization with Targeted Concept Ablation
  - Helena Casademunt, Caden Juang, **Adam Karvonen**, Samuel Marks, Senthooran Rajamanoharan, Neel Nanda
  - *NeurIPS submission*
- Robustly Improving LLM Fairness in Realistic Settings via Interpretability
  - **Adam Karvonen**, Samuel Marks
  - *ICLR submission*

2024

- Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models
  - **Adam Karvonen**
  - *1st Annual Conference on Language Modeling (COLM)*
- Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models
  - **Adam Karvonen,** Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Verdun, David Bau, Samuel Marks
  - *NeurIPS main track, Oral Presentation at the 2024 ICML Mech Interp Workshop*
- Evaluating Sparse Autoencoders on Targeted Concept Removal Tasks
  - **Adam Karvonen**, Can Rager, Samuel Marks, Neel Nanda
  - *NeurIPS ATTRIB workshop*

2023

- Leveraging Manifold Learning and Relationship Equity Management for Symbiotic Explainable Artificial Intelligence
  - Eric Davis, Sourya Dey, **Adam Karvonen,** Ethan Lew, Donya Quick, Panchapakesan Shyamshankar, Ted Hille, Matt LeBeau
  - *9th International Conference on Human Factors in Robots, Drones and Unmanned Systems*

Blog Posts

- Sieve: SAEs Beat Baselines on a Real-World Task (A Code Generation Case Study)
  - **Adam Karvonen***, Dhruv Pai*, Mason Wang, Ben Keigwin
- Applying GPT-4 to SAW Formal Verification (Official Galois Company Blog Post)
  - **Adam Karvonen**
- Using GPT-4 to Assist in C to Rust Translation (Official Galois Company Blog Post)
  - **Adam Karvonen**