

Keltin Grimes

kgrimes@andrew.cmu.edu | 984-242-6692 | Website: keltin13.github.io

EDUCATION

Carnegie Mellon University, graduated 2023, **4.0 GPA**, University Honors

Bachelor of Science in **Statistics and Machine Learning**, Additional Major in **Computer Science**

Relevant Coursework: Advanced Intro to ML (PhD, current), Intro to Deep Learning (Masters), Probability Theory, Statistical Inference, Linear Algebra, Statistical Learning, Software of Security and Privacy

Skills: Advanced Python and PyTorch • Intermediate C, C++, Java, R, SQL, various Web Dev.

WORK/RESEARCH EXPERIENCE

Associate Researcher – Software Engineering Institute, Carnegie Mellon, June 2023 – Present

Promoted from Assistant Researcher during first promotion cycle for exceeding expectations of position.

- Developed new method for trojan attacks on LLMs using model editing: faster and more effective than fine-tuning, and enabling poisoning of high-level concepts, a novel class of trojans. Submitted paper to top ML conference (under review).
- Led data-science project to develop ML models for evaluating the fuel savings of aircraft modifications based on in-flight sensor data. Delivered prototype which is actively being used by US Air Force. Presented work at DATAWorks 2024 workshop. Featuring in SEI's 2024 Year in Review.
- Performed literature review on Machine Unlearning for Computer Vision, published extended abstract at IEEE S&P DLSP Workshop proposing improvements to unlearning evaluations.
- Developed advanced distributed model trainer in PyTorch for an Adversarial ML project.
- Visiting Expert at the Laboratory for Analytic Sciences 2024 SCADS program.
- Various confidential research projects and presentations.

Research Assistant – Materials Science Department, Carnegie Mellon, January 2021 – May 2023

- Explored deep learning models to predict properties of molecules as part of Dr. Noa Marom's lab, developing active learning techniques using Bayesian neural networks. Set up a distributed hyperparameter optimization scheme and trained large models using PyTorch and Tensorflow.
- Presented poster on work at Materials Science & Technology 2022

Machine Learning Intern – ObjectSecurity, November 2022 – May 2023

- Performed literature review of adversarial attacks on NLP models and replicated results from various papers. Built API for interfacing with proprietary data science platform.

Software Development Intern – Amazon, May-August 2021 and 2022

- **2022** - Designed and implemented a suite of APIs using AWS's serverless compute platform. Automated the Alexa Shopping division's use of a system integral to Alexa's speech recognition performance.
- **2021** - Developed a tool to automate the process of adapting Amazon Alexa's Automatic Speech Recognition models to new use-cases. Used audio generated from neural networks to save 3-4 weeks of calendar effort compared to the previous manual annotation system.

Research Intern – The Laboratory for Analytic Sciences, May-August 2020

- Created an application allowing users to rapidly triage a large collection of documents, with efficiency gains of over 50% compared with standard techniques. Built with PyTorch and ReactJS.
- Utilized state-of-the-art techniques in Natural Language Processing and active learning to enable the training of an accurate text-classifier on extremely small amounts of data.

AWARDS

- Ultimate Jailbreaking Championship First-break prize on 5th most difficult model, Gray Swan AI, 2024
- Leadership Challenge Winner, American Statistical Association, 2019
- Short Story Competition Honorable Mention for *Iluula*, Raleigh Fine Arts, 2019

PUBLICATIONS

Grimes, Keltin, et al. "Concept-ROT: Poisoning Concepts in Large Language Models with Model Editing." *Under review (2024)*.

Casper, Stephen, et al. "The SaTML'24 CNN Interpretability Competition: New Innovations for Concept-Level Interpretability" *arXiv preprint arXiv:2404.02949 (2024)*.

Grimes, Keltin, et al. "Gone but Not Forgotten: Improved Benchmarks for Machine Unlearning." *Deep Learning Security and Privacy Workshop (2024)*.

PRESENTATIONS

Grimes, Keltin. "Rank-One Trojaning: Fast Insertion of Concept-Level Trojans Using Model Editing." Intelligence Community conference (name withheld), November 2024, Washington, DC.

Ratchford, Jasmine and Grimes, Keltin. "AI Engineering and LLMs." Summer Conference on Applied Data Science, Laboratory for Analytic Sciences, 17 June 2024, Raleigh, NC.

Grimes, Keltin. "Gone but Not Forgotten: Improved Benchmarks for Machine Unlearning." IEEE Security and Privacy, Deep Learning Security and Privacy Workshop, 24 May 2024, San Francisco, CA.

Grimes, Keltin. "Statistical Validation of Fuel Savings from In-Flight Data Recordings." Defense and Aerospace Test and Analysis Workshop, 18 April 2024, Washington, DC.

PROJECTS

- Replicated every plot from Toy Models of Superposition by Elhage et. al., 2024 [\[link\]](#)
- Ode to Transformer – fine-tuned a (at the time) state-of-the-art language model to write poems. Created an interactive web interface in ReactJS and backend in Python, 2020
- Salvation – developed a 3D first-person shooter game from scratch in Python as a class project. Included detailed player mechanics, path-finding enemies, random level generation, and a level editor. One of 12 students selected to present their project to the rest of the 450-person class, 2019. [\[link\]](#)

EXTRACURRICULARS

- CMU Data Science Club, member, 2019 End-of-Semester Data Challenge Winner, 2019-2023
- CMU Running Club, Prime Minister and Student Pandemic Safety Ambassador, 2019-2023
- Ironman 70.3 Triathlon Finisher, 2021 Ohio Ironman 70.3
- Long-time violinist, self-taught pianist.
- Hobbies: soccer, skiing, chess, crossword puzzles, and reading.