Oliver Clive-Griffin

website | github | email | linkedin | most recent version of this document | (moving to) San Francisco / London

EXPERIENCE

Cambridge AI Safety Hub **Cambridge, UK / Remote**
Research Scholar - MARS (Mentorship for Alignment Research Students) *January 2025 - Present*
- **Mechanistic Interpretability** research under Bilal Chughtai investigating use-cases of **crosscoders** for model diffing.

Apart Research **Remote**
Apart Lab Studio Fellow - Part Time *January 2025 - Present*
- Invited to continue research using Sparse Autoencoder features to predict adversarial prompt success.

Sabbatical **Wellington, New Zealand**
Self-funded sabbatical focussed on upskilling in technical AI safety and ML engineering. *September - December 2024*
- Porting MLAgentBench to the UK AISI's **Inspect Evals** as part of Arcadia Impact's **AI Safety Engineering Taskforce**.
- **Won 2nd place** in Apart Research and Goodfire's "Reprogramming AI Models" hackathon. (link)
- Open source contribution to **SAELens**, a sparse autoencoder library.

Recurse Center **New York / Remote**
Participant - self-directed programming retreat *March - June 2024*
- Wrote "rax," a thousand lines of dependency-free Rust able to train neural networks.
- Worked through **ARENA (Alignment Research ENgineer Accelerator)**. Studied mechanistic interpretability and reinforcement learning. Isolated the mechanism for usage of for-loop variables in a 2 layer language model.
- Wrote a Lisp interpreter and bytecode compiler/VM in Rust while working through "Crafting Interpreters."
- Studied ML systems engineering: Sacha Rush's "GPU Puzzles," CUDA.

Halter **Auckland, New Zealand**
Halter is an "Operating System for Farming". It has raised over **$100M USD** from Bessemer Venture Partners, DCVC, Founders Fund, and others. In 2024 Halter was the **fastest growing company in New Zealand**, with **1500% 3-year growth**.

R&D Engineer *July 2023 - March 2024 (Full time), September - October 2024 (Part time Contract)*
- Developed multimodal **transformer-based models** for sparse spatiotemporal prediction.
- Led the design and implementation of a system for running ML jobs on over **100,000 hectares of satellite imagery daily**
- Created distributed pipelines for dataset generation with **Ray** and WebDataset.
- Built data visualisation tooling with ThreeJS, enabling rapid understanding of data and debugging.

Tech Lead *July 2022 - July 2023*
- Lead the design and development of Clover: a system used to model over **250,000 cattle** on **350,000 acres** of farmland.
- Inventor on a **patent** for a reversible, replayable geospatial modelling system: the core technical innovation behind Clover.

Junior Software Engineer *September 2021 - July 2022*
- Worked on a distributed ML inference pipeline that ran regular inference on the behaviour of ≈100,000 cattle.
- React Native mobile app development with a focus on data visualisation with D3.

SKILLS

Languages: Python, Rust, TypeScript, SQL, C, CUDA.
Libraries - Machine Learning: PyTorch, Jax, TransformerLens, SAELens Einops, Ray, WebDataset, Gym, Polars, Plotly.
Libraries - Software / Web: NestJS, TypeORM, lodash, Three.js D3.js, React (Native), React Query.
Platform: Docker, Terraform, Concourse CI, Postgres, PostGIS, AWS (ECS, SQS, SNS, RDS, S3, Batch, EC2).

SELECTED PROJECTS

Detecting Successful Adversarial Prompting From SAE Activations - Won **2nd place** in Apart Research hackathon.
Investigating for-loop Variable Usage in Toy Language Models - Mech-interp examination of a small Language Model.
Rax: A pure functional toy deep learning library written in Rust, loosely inspired by Jax.
Rusp: A lightweight Lisp interpreter and compiler + vm, written in Rust.