Kola Ayonrinde

koayon@gmail.com GitHub +44 7531 251209

Research Experience

Research Scholar, MATS 2024

- Extending Free-Energy Equilibria results from ACS to practical application showing the benefits of Active Inference-style agents in learning effective world models.
- Conducted research into efficient and performant SAEs for the Mechanistic Interpretability Community, utilising a more principled activation function and allowing for zero dead latents.

Research Scholar, Cavendish Labs 2023 - 2024

- Built the first open-source implementation of the AtP* algorithm, an efficient method for localising LLM behaviour to components for Mechanistic Interpretability.
- Extended Towards Monosemanticity Paper to analyse modular models with SAEs.
- Wrote Python library for Interpretability of Mixture of Expert (MoE) models.

Research Engineer, UnlikelyAI 2023 - 2024

- Used Contrastive Decoding approaches to improve faithful reasoning performance by 20% on internal benchmarks whilst retaining latency.
- Built internal Copilot product to accelerate engineer and data labeller productivity by 50%.

***ML Researcher,* University of Cambridge** 2019

- Developed predictability measures as a proxy for musical interestingness; correlated with field research to provide evidence for “medium complexity” theory for artistic appreciation.
- Optimised model performance by developing method for representing musical notation.
- Research published in advisor’s book from Cambridge University Press.

Publications

K Ayonrinde. ***Feature Choice Top-m Sparse Autoencoders***. Aim to submit at ICLR 2025. [WIP]
Ayonrinde et al. ***Mechanistic Explanation of MoE Transformers.*** Forthcoming, w/ EleutherAI

Industry Experience

***Senior Software Engineer (Data & ML),* Beam** 2021 - 2023
***Product Data Scientist,* what3words** 2020 - 2021

Education

University Of Cambridge
MA, BA Mathematics
Final Year Project: Natural Language Processing
Relevant Modules: Logic, Optimisation, Programming, Statistics

CaMLAB - Technical Alignment Research Course

Skills & Interests:

Programming Languages: Python, SQL, Ruby on Rails, TypeScript

Tools: PyTorch, AWS, DSPy, NNsight, Pandas, WandB, Airflow, FastAPI, 🤗

Creator and Maintainer of StateCraft - Package for using, saving and sharing SSM states. We show that combinations of states can be effective and are building a community resource of reusable states
Writer of Technical ML blog - 10k monthly readers
Creator of PromptHeart - early site for sharing, discovering and remixing LLM prompts
Musical Director for live bands; former Touring & Session Musician
Host of Weekly Paper Discussions for a group of ~15 researchers
ML Writing published in The Gradient magazine - 70k+ readers, 400 points on HackerNews etc.