# Dennis Akar

denizhanak@gmail.com | denden.dev | github.com/dendenakar | UK Citizen

## EDUCATION

**University of Cambridge** 2021 - 2022
*MPhil in Advanced Computer Science*
- Pass with distinction: 81.20%.
- Awarded the £5,000 ACS MPhil Scholarship for academic excellence.
- Researched geometric DL for molecular graphs (drug discovery) supervised by Prof Pietro Liò & Dr Cristian Bodnar.

**University of Manchester** 2018 - 2021
*BSc Computer Science and Mathematics*
- First Class Honours: 83.36%.
- Awarded Certificate of Excellence for top 10% graduating students.

## EXPERIENCE

**Long-Term Future Fund - AI Safety Researcher** Jan 2024 - Present
- Investigating machine unlearning, behaviour modelling, capability separability, and applying mechanistic interpretability methods such as SAEs for training and fine-tuning to improve safety in LLMs.

**ARENA - Teaching Assistant** Sep 2024 - Oct 2024
- Aided participants in equipping talented individuals with the skills, tools, and environment necessary for upskilling in ML engineering, for the purpose of contributing directly to AI alignment in technical roles.
- Provided hands-on support in understanding, implementing and debugging DL implementations, during an intensive 5-week program, including DL fundamentals, mechanistic interpretability, circuit discovery, and RL.

**MATS: Foundations of Mechanistic Interpretability (Lee Sharkey) - Research Fellow** May 2023 - Jan 2024
- Investigated "Attention Head Superposition" in language models with Chris Mathwin and Lee Sharkey.
- Proposed and implemented the gated attention block, resolving attention head superposition with the aim of making it easier for researchers to study individual attention heads.
- Facilitated Alignment 201 reading group for 5 MATS scholars.

**MATS: Mechanistic Interpretability (Neel Nanda) - Research Fellow** Nov 2022 - Jan 2023
- Applied the original and extended logit lens to the IOI task across a set of GPT-2 sized language models (extended DLA). Extended logit lens uses consecutive layers at the end of the model to map the residual stream to logit space.
- Found the tendency for certain models (e.g. GPT-Neo) to "flip" i.e. assign an *extremely low probability* throughout the model to the token that it will eventually output and used extended DLA to analyze how this tendency changes.

**CancerAI - Research Assistant** Jul 2022 - Oct 2022
- Researched explainable AI for use by clinical oncologists using **Tensorflow** and **PyTorch**.
- Developed front-end for VIIDA, an application for analyzing, modelling, explaining, and predicting cancer-related data with **Flask** and **React**.

**Cambridge Cancer Genomics - Software Engineer Intern** Jun 2019 - Sep 2019
- Integrated features and fixed bugs for the precision oncology platform OncOS backend using **Python** and **Flask**.
- Built a **full-stack** internal monitoring system for OncOS infrastructure to manage genomic data and processes.
- Researched variational autoencoder algorithms related to DNA sequence compression for SomaticNET, a neural network for evaluating tumor variants, using **Tensorflow (Python)**, **Bash**, **pysam** and **Annoy**.

## PUBLICATIONS

**Gated Attention Blocks: Preliminary Progress toward Removing Attention Head Superposition** Apr 2024
- In transformer language models, attention head superposition makes it difficult to study the function of individual attention heads in isolation. We study a particular kind of attention head superposition that involves constructive and destructive interference between the outputs of different attention heads. We propose a novel architecture - a "gated attention block" - which resolves this kind of attention head superposition in toy models. In future, we hope this architecture may be useful for studying more natural forms of attention head superposition in large language models.