Nicky Pochinkov

AI Safety Researcher

work@nicky.pro
www.nicky.pro
github.com/nickypro
in/nicky-pochinkov

EXPERIENCE

AI Safety Researcher *Oct 2022 - Present*
Independent / Long-Term Future Fund (LTFF) Grantee
- Formalised ideas and ran experiments attempting to understand representation of long-term planning in language models.
- Developed simple techniques for understanding & extracting paragraph-level information from language model token activations.
- Achieved state-of-the-art Machine Unlearning performance while studying Modularity & Separability in Language Models.
- Investigated & ran experiments on various topics, including Soft Optimisation with TRLx, Constrained RL, Active Inference, and Loss Landscapes.
- Built Taker, an interpretability library compatible with Huggingface, with support for ablation, steering, multi-gpu and quantized inference.

Independent Contractor *Nov 2024 - Present*
Anthropic, Automated Alignment Team

Teaching Assistant *Sep 2024 - Oct 2024*
Alignment Research Engineer Accelerator (ARENA 4.0)
- Presented short presentations on ML fundamentals to over 30 people.
- Provided hands-on support in understanding, implementing and debugging ML implementations, during an intensive 5-week program, including fundamentals, RL, mechanistic interpretability, and circuit discovery.

Research Lead *Jan 2024 - Aug 2024*
AI Safety Camp & Supervised Program for Alignment Research
- Mentored and Lead 14 people in teams to do interpretability research.
- Published work on Neuron Ablation techniques at XAI Conference.
- Wrote papers on Neural Task Specialisation in Transformer models.
- Studied how Language Models encode paragraph-scale information, forming frameworks for how to think about language model "goals".

ML Alignment Theory Scholar (MATS) *Jun 2022 - Sep 2022*
Stanford Existential Risk Initiative
- Investigated independent research agendas under John Wentworth.

Software Engineering Intern *Jun 2021 - Aug 2021*
Arista Networks
- Developed a Network Switch hardware feature start-to-end in C++.
- Understood ASIC hardware tables, performed hardware validation, wrote a design doc, made tests in python, and underwent code reviews.

Astrophysics & Computing Intern *Jun 2020 - Aug 2020*
Dublin Institute for Advanced Studies
- Created science website for viewing & rating meteor images (React, SQL).
- Contributed Python code for conversion between meteor data types.
- Used Bash & Docker to automate various tasks with systemd & cron jobs.

Patch Summer Accelerator Participant *Summer 2020*
Dogpatch Labs
- Worked in a team of 3 to build a Flutter iOS & Android App in 6 weeks.

PUBLICATIONS

Dissecting large language models 2023
NeurIPS SoLaR Workshop. *Pochinkov, N., & Schoots, N.*

Extracting Paragraphs from Token Activations 2024
NeurIPS MINT Workshop. *Pochinkov, N. et al.*

Modularity in Transformers: Investigating Neuron Separability... 2024
ArXiv. *Pochinkov, N., Jones, T., & Rahman, M.R.*

Investigating Neuron Ablation in Attention Heads... 2024
XAI World Conference. *Pochinkov, N., Pasero, B., & Shibayama, S.*

EDUCATION

Trinity College Dublin *Sep 2018 - May 2022*
Theoretical Physics
- First Class Honours (GPA 3.7+ Equivalent)
- Wrote a thesis on "Machine Learning of Many-Spin Quantum Systems" (2022)
- Hamilton Trust Research Internship on "accurate rigid-body simulation" (2019)

ACHIEVEMENTS

International Mathematical Olympiad
Represented Ireland in IMO 2018 (top 3 in Ireland)

Other International Olympiads
Represented Ireland in 2018 Chemistry Olympiad (IChO) and Benelux Math Olympiad (BxMO)

Naughton Scholarship
Awarded €20,000 for undergraduate excellence

Irish Leaving Certificate
Achieved maximum 625 points (top 0.3%)

Mozilla Builders Hackathon
Won 2nd Prize and $2000 (2021)

Townsend Prize
Achieved best exam results in my course (2019)

VOLUNTEERING

Chairperson, DU Vegan Society. *2019 - 2021*
- Organised and ran events, short-listed for "Best All-day Event" Award.

Technical Mentor, CoderDojo. *2017 - 2020*
- Mentored children, teaching HTML and CSS.

SKILLS

Programming Languages: python, javascript, C

Machine Learning: Transformers, Mechanistic Interpretability, Reinforcement Learning, PPO, PyTorch, TransformerLens, TRLX, TensorFlow

Backend: Linux, Nginx, Bash, Docker, MySQL, MongoDB, Cloud Firestore, Node, Express

Frontend: React, Redux, Flutter, Bloc, Figma

Languages: English (Native), Russian (Fluent), French (Basic), Irish (Basic)

Design: GIMP, Inkscape, DrawIO, Figma

Other: Leadership, Mathematics, Physics, Business, Finance, Accounting