Felix Hofstätter, MEng

FelixAHofstaetter@gmail.com

+436642475686

felixhofstaetter

@Felhof1

Felhof

felhof.com

Education

Alignment Research Engineering Bootcamp, *ARENA*
05/2023 – 06/2023 | London, UK

MEng Mathematics & Computer Science (First Class Hons),
Imperial College London
2017 – 2021 | London, UK

Professional Experience

MATS Scholar, *MATS*
01/2024 – 09/2024 | Berkeley, USA
- Evaluate sandbagging capabilities in LLMs (see publications).

AI Safety Researcher, *AI Safety Hub Labs*
07/2023 – 09/2023 | Vienna, Austria
- Evaluate deceptive behavior in LLMs. Write up results on the Alignment Forum .

Software Consultant, *TNG Technology Consulting*
08/2021 – 12/2022 | Vienna, Austria
- Develop an E-Commerce platform.
- Hold workshops on Python programming.

Data Engineer (Internship), *Swoop Finance*
06/2020 – 09/2020 | London, UK
- Build a database of financial information on UK companies.

Undergraduate Research Project (Software Verification),
Imperial College London
07/2019 – 09/2019 | London, UK

Projects

Contribute to Open Source Interpretability Libraries
- TransformerLens , e.g. implementing Grouped Query Attention .
- SteeringVectors , e.g. improving steering .

Investigating the information flow inside Large Language Models, *ARENA Capstone Project*
- Investigate how interventions on the attention mechanism influence LLM capabilities.

Deep Reinforcement Learning Algorithms in PyTorch, *GitHub*
- Implement and evaluate various standard RL algorithms, mostly based on OpenAI's Spinning Up.

Writing about AI Alignment, *Medium and LessWrong*
- On LessWrong and Medium , summarize AI safety research.

Publications

AI Sandbagging: Language Models can Strategically Underperform on Evaluations 2024
Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, Francis Rhys Ward
- Conduct experiments to show how LLM-capabilities can be hidden during evaluations, e.g. by implanting a backdoor using synthetic data.

Skills

Programming
Python, Typescript, Java

Large Language Models
PEFT Fine-Tuning, Synthetic Data Generation, Prompt Engineering

Machine Learning
PyTorch, HF transformers, NumPy, pandas, Implementing Reinforcement Learning algorithms

Software Development
Test Driven Development, Agile Methods, Git