Ian McKenzie

(+1)341 777 9815 | ianmck98@gmail.com | irmckenzie.co.uk | naimenz | irmckenzie

Education

MSc Artificial Intelligence (Distinction)

University of Edinburgh (2020-2021)

- **MSc Project** on efficient inference with probabilistic first-order knowledge bases
- **Courses** covering deep learning, reinforcement learning, natural language processing, Bayesian inference

MPhys Theoretical Physics and Maths (1st)

University of St Andrews (2016-2020)

- **Brewster Prize** Awarded for best grades in final year across all physics undergraduate master's degrees — *2020*
- **Highly Commended Final Year Project** Finalist for best final year project in mathematics – *2020*
- **Fourth Year Class Medal (Theoretical Physics)** Awarded for best grades in year — *2019*
- **Dean's List** Excellent academic performance (average above first) — *2017, 2018, 2019, 2020*

Experience

Dangerous Capability Evaluations Research Engineer (contract)

Berkeley, CA

OpenAI (openai.com)

July 2023 - Present

- Building evaluations for measuring the dangerous capabilities of current state-of-the-art language models.
- Responsibilities include building and running evaluation environments, writing reports to inform policy and external collaborators, improving internal tooling.
- Current project: evaluating LM-based agents in a simulated online environment (similar to the work by ARC Evals on ARA).

Research Scientist

Berkeley, CA

FAR (far.ai)

February 2022 - June 2023

- Responsible for the Inverse Scaling Prize – an AI safety contest with up to $250,000 in prizes for participants that demonstrate inverse scaling in large language models.
- Supervised by Ethan Perez and Sam Bowman.
- Produced evidence for this phenomenon and developed a new framework for identifying examples of outer misalignment.
- Responsibilities included running weekly meetings of around 10 people, writing code to evaluate a variety of language models on few-shot tasks, maintaining all contest materials (including a codebase and public-facing documentation), and managing prize review.
- Lead author on *Inverse Scaling: When Bigger isn't Better*, accepted to TMLR and recommended for a Featured Certification, that outlines the theory behind inverse scaling and analyzes the prize-winning tasks.

REMIX Research Resident

Berkeley, CA

Redwood Research (redwoodresearch.org)

January 2023 - February 2023

- Research residency on model internals and mechanistic interpretability.
- Performed original research using TransformerLens on induction heads and context sensitivity in language models using Pythia and other open-source LMs.
- Replicated key results from Anthropic in-context learning paper.
- Developed a speculative new metric for context sensitivity that identifies induction heads and heads responsible for Indirect Object Identification.

Software Engineering Intern

San Francisco, CA

Ought (ought.org)

October 2021 - January 2022

- Job responsibilities include running machine learning experiments, working with other engineers and the product team to design and implement features into Elicit, an AI research assistant (`elicit.org`).
- Experiments include executing and evaluating prompt design and fine-tuning GPT3 and BERT-based models for complex classification tasks.
- Projects included collecting data and fine-tuning models for detecting whether an argument supported a claim, and detecting ill-formed or mismatched abstracts from Semantic Scholar, for which I created a sequence classification fine-tuning pipeline with HuggingFace Transformers and PyTorch to easily switch out the data and models used.
- Gained experience working on front-end (React) and back-end (Python), as well as working effectively as part of a small, fast-moving team.

Summer Research Associate

St Andrews, Scotland

University of St Andrews

July 2019 - August 2019

- Worked with a team of five other students on a problem in infinite group theory – to provide an elegant proof that Thompson's Group V has a finite presentation by transpositions.
- Wrote Python code to automate much of the required computation and visualise calculations.

Quantum Simulation Development Intern

St Andrews, Scotland

University of St Andrews

July 2018 - August 2018

- Designed and implemented interactive quantum mechanics visualisations for teaching undergraduate quantum physics modules.
- Used CSS for styling and JavaScript for animation and interactivity.

Publications

Inverse Scaling: When Bigger isn't Better

TMLR, Featured Certification, 2023

- https://openreview.net/pdf?id=DwgRm72GQF

Projects

Argument classification and non-abstract detection for Elicit

2021

- Part of my responsibilities as an intern at Ought.
- Collected data and fine-tuned models for detecting whether an argument supported a claim, and detecting ill-formed or mismatched abstracts from Semantic Scholar.
- Created a sequence classification fine-tuning pipeline with HuggingFace Transformers and PyTorch to easily switch out the data and models used.
- Launched the trained models to production with Docker and AWS.

Generalising first-order knowledge compilation to the hybrid setting

2021

- MSc project at Edinburgh, earned a distinction.
- Created a Python project to perform efficient (lifted) probabilistic inference in first-order logic knowledge bases.
- Rewrote and extended a Scala implementation.
- Wrote proofs for the validity of my generalisations to the earlier algorithms.

Pong with Deep Reinforcement Learning

2020

- Trained a Deep Q-Network (DQN) agent to play Pong.
- A self-directed project during the summer between degrees.
- Written in Python with NumPy, PyTorch, and OpenAI Gym.
- Developed from the paper without reference to other implementations.
- Full details written up at https://www.irmckenzie.co.uk/pong

Nearly Isostatic Networks and Allosteric Effects

2019-2020

- Highly Commended final year project, achieved 19 on St Andrews 20 point scale
- Involved implementing linear algebra- and graph-based optimisation techniques to understand and explain the computational results of a recent paper.
- Uses NumPy, SciPy, and networkx packages.

Predicting Insurance Claims Kaggle Competition

2020

- Project for module Knowledge Discovery and Data Mining, attained 92%.
- Written in Python with scikit-learn, XGBoost, and pandas
- Developed a model to predict whether a driver would make an insurance claim.
- Performed data exploration, cleaning, and preparation, and built, selected, tuned, and validated a variety of binary classification models.

Discrete Optimisation

2019

- Started as an assignment for a module (Advanced Symbolic Computation) to implement simulated annealing on Ising spin glasses, achieving 100%.
- I found it too interesting to stop after just doing one method on one problem, so I implemented both heuristic (simulated annealing) and exact (Branch-and-Bound, Held-Karp) optimisation methods on the Traveling Salesperson Problem and Ising spin glasses.
- Further details and code on https://www.irmckenzie.co.uk