Jacob Davies

+44 (0) 7497 783212 • jacob_e_davies@mac.com • Jacob Davies
jacobdaviescam

Professional Profile

ML Research Engineer specializing in AI system evaluation, mechanistic interpretability, and empirical frameworks for assessing model capabilities. Combines deep technical expertise in transformer architectures and language models with proven ability to design robust evaluation pipelines and translate complex research findings into actionable insights.

Technical Skills

ML Research: **Transformer architectures, PyTorch, Reinforcement Learning, mechanistic interpretability, model capability assessment**

AI Evaluation: **Framework development, systematic benchmarking, automated evaluation systems, performance analysis**

Research Methods: **Experiment design, hypothesis testing, empirical evaluation, statistical analysis, large-scale model training**

Implementation: **Python, batch GPU processing, evaluation pipeline development**

Professional Experience

ARENA (Alignment Research Engineer Accelerator) **London, UK**
AI Safety Research Engineering *Sep 2025– Oct 2025*

- Completed intensive 4-week AI safety research program covering mechanistic interpretability and transformer architectures
- **Built GPT-2 from scratch** implementing full transformer architecture with custom attention mechanisms
- Developing expertise in mechanistic interpretability techniques for understanding model internal representations
- Completed Capstone project on identifying important steps in LLM reasoning traces using both black-box and white-box approaches.

LinkedIn **Dublin, IE**
AI Linguist *Nov 2024–Aug 2025*

- **Designed comprehensive evaluation framework** for AI agent systems with measurable performance criteria
- **Developed scalable automated evaluation pipelines** using Python, reducing evaluation time by 60%
- **Implemented systematic benchmarking protocols** for measuring AI capabilities across multiple dimensions
- **Built end-to-end human annotation systems** with LLM pre-labeling and human audit protocols
- Collaborated with engineering teams to implement performance monitoring protocols at production scale

University of Cambridge **Cambridge, UK**
AI and AI Policy Governance Researcher *Sep 2023–Jul 2024*

- **Conducted empirical research on AI system evaluation** with focus on safety assessment protocols
- **Developed evidence-based methodologies** for assessing AI system risks and alignment verification
- **Analyzed real-world AI deployment scenarios** in government contexts, identifying evaluation gaps
- Collaborated on cross-institutional research with Harvard University on AI transparency mechanisms
- Created technical recommendations for improving AI safety assessment frameworks

BAE Systems Digital Intelligence **London, UK**
Data Consultant Intern *Jun–Sep 2022*

- Performed large-scale data analysis for UK Health & Security Agency using SQL
- Contributed to AGILE development team building Azure platform for data engineers
- Designed analytical dashboards for computational efficiency analysis and performance insights

Education

University of Edinburgh **Edinburgh, UK**

MSc Speech and Language Processing, Distinction *2023–2024*

Research: Structural Generalisation in Dependency Parsers - Comparing Transition-Based and Graph-Based Parsers

- Designed empirical evaluation methodology comparing parser performance on structural generalization tasks
- Built complete evaluation pipeline from raw data to structured performance analysis
- Utilized supercomputing resources for large-scale model evaluation with batch GPU processing
- Implemented and compared multiple model architectures to identify failure modes and limitations

Relevant Coursework: Natural Language Processing, Machine Translation, Speech Recognition, Applied ML

St John's College, University of Cambridge **Cambridge, UK**

BA Computational Linguistics *2020–2023*

Research: An Exploration of Transformers for Metonymy Resolution under Time and Genre Shift

- Conducted systematic evaluation of BERT models across domains, identifying architectural limitations
- Analyzed cross-domain transfer capabilities revealing model weaknesses for emerging topics
- Implemented fine-tuned transformer models using rigorous experimental methodology

Selected Projects

GitHub Repository

SLP Dissertation Research *2024*

Comprehensive evaluation of transformer models for linguistic tasks with detailed performance analysis and statistical validation. Demonstrates proficiency in ML research methodology, experimental design, and systematic model comparison with reproducible research practices.

Github Repository

ARENA Capstone Project *2025*

A white-box AI safety research project extending the Thought Anchors methodology from reasoning analysis to practical intervention. This work demonstrates that targeted activation patching at critical reasoning layers can prevent harmful AI decisions while preserving general capabilities.

Awards & Recognition

McAuley Scholar, St John's College, University of Cambridge

Winner of the Harvard Book Award, 2019

Beyond Equality Facilitator