

JACK HOPKINS

@ jack.hopkins@me.com +44 7549 949362 51 Durham Road, London E12 5AY JackHopkins

Technical lead specializing in machine learning systems with experience managing R&D teams. 10 years of experience in language model research, published in top venues and journals (ACL, ICML, Neurips, New Scientist). Interested in safety research and LLM environment development.

WORK

Fellow - AI Safety June 2025 – Feb 2026

Anthropic Berkeley / London

Determined that fine-tuned LLM lie detectors don't generalise. Identified a novel technique to get LLMs to leak secrets through targeted weight amplification. Collaborated with Fabien Roger and Rowan Wang. MATS 8.0 Megastream.

Safety Researcher June 2025 –

MATS 8.0 - 8.2 London

Research / Cofounder 2021 – 2024

Paperplane (YC W23) New York / London

- Led technical development of multi-modal conversational intelligence platform
- Raised venture capital and attended Y Combinator (W23)
- Deployed and sold software to several UK/US technology customers.

CTO / Cofounder 2016 – 2020

Spherical Defence London

- Developed proprietary VAE variant achieving state-of-the-art accuracy on API security benchmarks
- Built and led engineering team serving enterprise customers including Meta, LSEG and Twilio
- Secured venture funding and established technical innovation partnership with AWS

Consulting 2017 – 2026

Developed commercial AI systems for UK and US companies including Skillbrium, Ditchley Foundation, Turnbull, Goldcast and GoodStart Labs.

SKILLS

ML: PyTorch, Transformers, VLMs, (V)AEs, LSTMs, SFT, RL
Research interests: Mechanistic interpretability, formal verification of models, language modelling, model merging, information retrieval, AI safety, API security.

AWARDS

- Winner** [OpenAI 20b Red-Teaming Competition] (2025)
5th Atlantic Rowing Race (2018-19)
1st ECM Cambridge Tech Challenge (2015)

REFERENCES

- Dr Akbir Khan - Anthropic
@ akbir@anthropic.com
- Dr Minqi Jiang - Meta
@ mnqjng@gmail.com
- Mike Seltenreich - Anthropic
@ michael@anthropic.com
- Jack Kleeman - Anthropic
@ jackk@anthropic.com

EDUCATION

MPhil Advanced Computer Science 2015 - 2016

Darwin College, University of Cambridge Distinction

BSc Computer Science 2011-2014

Royal Holloway College, University of London 1st Class

OPEN-SOURCE

Factorio Learning Environment

An open-ended agent environment for evaluating LLMs in unbounded scenarios in Factorio, a game centered on automation and exponential resource production. 1k+ stars on [Github].

Tanuki - Prompt engineering for developers

Easily build LLM-powered apps that get cheaper and faster over time. Python and Typescript implementations. 700 stars on [Github].

PAPERS

Self-Attribution Bias

When AI Monitors Go Easy on Themselves
Jack Hopkins et al. 2026. NeurIPS Preprint. [Paper].

Overthinking

Amplifying Reasoning Weights to Extract Learned Secrets
Jack Hopkins et al. 2026. ICML. [Paper].

Factorio Learning Environment

Jack Hopkins et al. 2025. NeurIPS. [Paper].

Automatically Generating Rhythmic Verse with Neural Networks

Jack Hopkins et al. 2017. ACL Oral (Volume 1: Long Papers, Pages 168-178.) [Paper].

Open Game Tournaments in STARLITE

Jack Hopkins et al. 2015. AAMAS 2015.

PIRASA

Strategic Protocol Selection for E-Commerce Agents
Jack Hopkins et al. 2019. EM '19: Electronic Markets, Vol. 29, No. 2, 06.2019, p. 239-252.

OTHER

-  **I rowed across the Atlantic Ocean in 2019**
5000 kilometres over 40 days, unsupported, taking Columbus' route from La Gomera to Antigua, raising £22,000 for Multiple Sclerosis charities. Team Atlantic Discovery.
-  **I launched a drinks company during COVID.**
Created a brand of gin to help raise funds for an expedition. Raised £50,000 for the expedition and won a 'best taste' award. *Northwest Passage Expedition Gin*