

Samuel Ratnam

samueljratnam@gmail.com • 07453 250785 • samuelratnam.xyz • London, UK

AI safety researcher on a gap year from Oxford. Cofounding Idealists Collective to empower people to imagine and fight for the futures they want. Interested in tools for human empowerment, LLM psychology, and model welfare.

Education

University of Oxford — MCompSciPhil (Computer Science & Philosophy) 2024–present
Currently on gap year

Wallington County Grammar School — 5 A*s (Maths, Further Maths, Physics, Philosophy, EPQ) 2017–2024

Experience

Idealists Collective — Cofounder Jan 2026–present

Community of philosophers, artists, and technologists pursuing utopian futures.

AFFINE — Seminar Participant May 2026

Month-long seminar on superintelligence alignment and agent foundations, with mentorship from ex-MIRI, DeepMind and Astera researchers.

Extrian — Software Engineer (Contract) Apr 2026–May 2026

Led a complex refactor of the production codebase for an audio data collection platform.

Paradigm Machines — Machine Learning Researcher (Contract) Apr 2026

Research into diversity-augmented reinforcement learning and GFlowNets for scientific creativity in ML models.

Workshop Labs — AI Researcher (Contract) Feb 2026–April 2026

Research evaluating LLMs for personalisation.

Geodesic Research — Research Collaborator Oct 2025–Feb 2026

First controlled study of how AI discourse in pretraining shapes alignment, training 6.9B parameter LLMs end-to-end under varying conditions. arXiv:2601.10160

London Impact Research Groups — Mentor Oct 2025–Jan 2026

Mentored research on how fine-tuning objectives shape safety, robustness, and persona drift in LLMs. arXiv:2601.12639

Supervised Program for Alignment Research (SPAR) — AI Safety Researcher Sep–Dec 2025

Research on coherence of LLM preferences; interpretive frames for understanding LLMs.

Workshop Labs — Research Engineer Sep–Nov 2025

Infrastructure and research for personalized model training. Embeddings pipeline 350k tok/s (DB→TurboPuffer); fine-tuning and synthetic data pipelines.

ERA Cambridge — Fellow Jun–Aug 2025

DELTA: framework for discovering behavioral differences between LLMs via adversarial prompts and interpretable hypotheses. Applications: reverse-engineering system prompts, LLM personality testing. Supervised by Usman Anwar.

Extrian — Software Engineer (Contract) Mar–Apr 2025

Web apps and LLM pipelines for phishing simulation and cybersecurity training.

Publications

Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment. C. Tice, P. Radmard, **S. Ratnam**, et al. 2026. arXiv:2601.10160

Objective Matters: Fine-Tuning Objectives Shape Safety, Robustness, and Persona Drift. D. Vennemeyer et al., **S. Ratnam**. 2026. arXiv:2601.12639 (*Mentored*)

Selected Projects

AI Alignment Research Graph — Apart Research Hackathon, 1st Place (\$1k). Graph-based tool for navigating AI safety research (team of 4).

AI Poem Copilot — EPQ. RoBERTa fine-tuned on 30k+ poems; vector DB for retrieval.