Samuel Ratnam

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

AI safety researcher on a gap year from Oxford. Researching alignment pretraining with Geodesic. Interested in LLM psychology, multi-agent scaffolding, 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

Geodesic Research — Research Collaborator Oct 2025–present

First controlled study of how AI discourse in pretraining shapes alignment. Trained 6.9B LLMs; upsampling aligned discourse reduced misalignment 45%→9%. Led misalignment evals. arXiv:2601.10160

London Impact Research Groups — Mentor 2025–present

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.

Encode Oxford — Chapter Lead Oct 2024–Jul 2025

Led Oxford AI safety student group; events and community.

Extrian — Software Engineer (Contract) Mar–Apr 2025

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

Future Impact Group — Participant Nov 2024–Mar 2025

Interactive visualization of welfare of future digital minds (Bradford Saad).

ARBOx (Alignment Research Bootcamp Oxford) Jan 2025

2-week ML safety bootcamp (David Quarrel): GPT-2 from scratch, mechanistic interp, RL.

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.