Samuel Ratnam

samueljratnam@gmail.com • London, UK

Independent AI safety researcher on a gap year from Oxford, currently working on weight-space analysis of LLM personality. Cofounding Idealists Collective to empower people to imagine and fight for the futures they want. Interested in tools for human empowerment and LLM persona science.

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

Independent Research — BlueDot Impact Rapid Grants Jun 2026–present
Method for engineering trait entanglements in LLMs by training base models beneath frozen trait-eliciting LoRA adapters; used it to disentangle emergent misalignment (LessWrong). Currently: weight-space analysis of LLM personality, in collaboration with David Africa (Resolution).

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.

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
Empirical research on the coherence of LLM preferences, and theoretical work on parallels between human and LLM cognition — arguing that jailbreaks, persona modulation, and hallucination mirror hypnosis, dissociation, and confabulation as features of predictive architectures. link

Workshop Labs — Research Engineer Sep–Nov 2025
Infrastructure and research for personalized model training. Embeddings pipeline; 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.

Future Impact Group — Participant Nov 2024–Mar 2025
Built an interactive visualization of forecasted welfare of future digital minds under Bradford Saad.

Research

Engineering the Generalisation Landscape of LLMs. S. Ratnam. 2026. link

Why Jailbreaks Look Like Hypnosis: Parallels Between Human and LLM Cognition. S. Ratnam. 2026. link

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)*