NAVRAJ SINGH

ms22110@iisermohali.ac.in | +91 7988241385 | Website | Github

EDUCATION & PROFESSIONAL TRAINING

Indian Institute of Science Education and Research (IISER) Mohali Punjab, India
BS-MS Dual Degree Student (Fifth Year) Oct 2022 – Present
CPI: 9/10

ARENA 9.0 (Alignment Research Engineer Accelerator) London, UK
Attended at LISA Offices, Shoreditch, London October 2026 - November 2026

Note: Detailed reports and code for the projects available at Github.

AI SAFETY & ALIGNMENT RESEARCH

Evaluation Awareness in Large Language Models *Pre-print on arXiv*
Lead Researcher | Supervisor: Maheep Chaudhary Feb – June 2026

- Mapped the representational scaling of evaluation awareness across compact architectures (Gemma 3, Phi-3, Llama-3 8B). Utilized Chain-of-Thought analysis, representation probing, and Integrated Gradients attribution for the investigation.
- Engineered a dual-pathway mitigation pipeline combining automated prompt sanitization with activation counter-steering, to neutralize evaluation artifacts across 200 high-confidence prompts, achieving an average behavioral flip rate of 70.58% across all evaluated architectures.

Interpretability Frameworks for Diffusion Models *Ongoing MS Thesis*
Research Student | Supervisor: Prof. Ashutosh Modi May 2026 – Present

- Building mechanistic evaluation tools to analyze feature representation and concept formulation during the reverse denoising process.
- Applying the Schrödinger bridging framework to track multi-step trajectories from Gaussian noise to structured outputs, identifying precise layers where early generation choices stabilize.

Narrow Secret Loyalty Extension Using CAI
Research Fellow | BlueDot Impact Technical AIS Project | USD1,600 research grant June - August2026

- Extended Lamerton & Roger (2026) via SL-CAI and RL-CAI pipelines on a 7B Qwen-2.5-Instruct model to probe for hidden loyalty behaviors under adversarial affordances.
- Found action-breadth entropy increased monotonically across baseline → SL-CAI → RL-CAI, and intent selectivity improved substantially after SL-CAI.Black-box detection rates stayed statistically indistinguishable from baseline across most affordance levels which is evidence that standard black-box audits can miss loyalty shifts that are present internally.
- Submitted to Apart Research's Secret Loyalties Sprint. LW write-up in progress.

Bluedot Technical AI Safety Course
Scholar – 6-Day Intensive Cohort June 2026

- Selected for a technical training cohort covering alignment theory, mechanistic interpretability foundations, and frontier model vulnerability profiles.

VEIL: Mechanistic Upstream Guardrails for Generative Biology *Apart Research Hackathon*
Core Developer | AIxBio Track

- Trained an L_1-regularized linear probe on ProteinMPNN decoder latents to identify structural markers of pathogenic risk, isolating causal features independent of token length.
- Analyzed feature polysemanticity and exposed generalization boundaries on unseen biological targets (e.g., Ricin) where safety vectors conflated structural threat with structural rigidity.

Mechanistic Analysis of Simplicity Bias and Feature Steering *PRECOG Lab*
Independent Researcher

- Trained Sparse Autoencoders (SAEs) on penultimate activations to disentangle compressed neural features.
- Proved a mechanistic variance between simple and robust features via targeted feature steering interventions, applying token-level gradients to successfully manipulate downstream classifications.

Replication Studies

- Completed replication studies of various foundation papers in the field of Foundational DL and AI Safety.

OTHER TECHNICAL & QUANTITATIVE INTERNSHIPS

Fully Homomorphic Encryption in AI (CKKS) *Supervisor: Prof. Mridula Verma*
Jul 2025 – Nov 2025

- Implemented encrypted neural networks running evaluations directly on ciphertexts to ensure strict data privacy.
- Managed CKKS noise budgets and scaling factors across deep architectures to optimize the precision-compute tradeoff.

Valid Circuit Detection in Hexa-board PCB *Supervisor: Prof. Satyajit Jena*
May 2025 – Jul 2025

- Engineered an end-to-end computer vision object-detection pipeline leveraging a hybrid SSD architecture (ResNet-50 backbone paired with an auxiliary ResNet-18 classifier) for multi-scale anomaly detection.

Computational Methods for Option Pricing *Supervisor: Prof. Anindya Goswami*
Jun 2024 – Aug 2024

- Developed high-performance C++ implementations of continuous-time option pricing and stochastic algorithms.

Grain Size Distribution Analysis Using Mask-RCNN *Supervisor: Prof. Yunus Ali Pulpadan*
Jun 2024 – Aug 2024

- Fine-tuned a custom Mask-RCNN framework on drone telemetry to segment and calculate granular distribution frequencies.

TECHNICAL SKILLS

- **Languages & Libraries:** Python, C++, PyTorch, TensorFlow, CUDA, R, NumPy, Scikit-learn, Pydantic, aiosqlite.
- **Core Competencies:** Mechanistic Interpretability, Sparse Autoencoders (SAEs), Circuit Discovery, Linear Probing & Steering, Causal Interventions (Activation Patching), Fully Homomorphic Encryption (CKKS).
- **Mathematics:** Linear Algebra, Measure Theoretic Probability, Cryptography & Number Theory, Group Theory, Stochastic PDEs.

ACHIEVEMENTS & RELEVANT COURSEWORK

- **Achievements:** JEE Main (Top 1 Percentile), IISER Aptitude Test (AIR-1345), KVPY Written Test (Qualified), SOF Science Olympiad (State Rank 51), SOF Mathematics Olympiad (State Rank 102).
- **Key Courses:** Machine Learning (Stanford CS229), Deep Learning, Quantum Information & Computation, Theory of Computation, Measure Theoretic Probability, Functional Analysis, Non-Linear Dynamics.