

Manan Wadhwa

manan-wadhwa.github.io wadhwamanan56@gmail.com github.com/Manan-Wadhwa Delhi, India

EDUCATION

Maharaja Agrasen Institute of Technology (GGSIPU), Delhi
B.Tech in Computer Science and Technology mait.ac.in

Expected Jul 2027

RESEARCH

* equal contribution

Known by Construction: Calibrating Model-State Instruments Against Models Whose State Was Installed, Hidden, or Never Present

Manan Wadhwa*, Shivam Dubey*

New in ML, NeurIPS 2026

State-reading tools agree with each other but track the prompt or training, not the state.

Dead Salmons Have a False-Positive Rate: A Calibrated Null-Model Test for Interpretability

Manan Wadhwa*, Shivam Dubey*

Interpretability as a Science, NeurIPS 2026

Interpretability methods find confident results in untrained models; we calibrate against them.

Steered Affect Moves a Partner Agent’s Activations, and the Logistic Probe That Reads It Returns a Tie-Break, Not an Estimate

Manan Wadhwa*, Shivam Dubey*, Debojeet Ghosh, Shreyansh Kumar, Ayushman Garg

Interpreting Agent Behavior, NeurIPS 2026

A steered agent’s emotion shows up in its partner; logistic probe directions fail to reproduce.

Toxin Circuits in ESM-2: Mechanistic Interpretability for Biosecurity Screening

Manan Wadhwa*, Shivam Dubey*

GenBio Workshop, ICML 2026

ESM-2 probes catch 93.9% of redesigned toxins that sequence screening misses entirely.

Distance to the Boundary Is Not Steering Resistance

Manan Wadhwa*, Shivam Dubey*

Under review, UniReps

Alignment training widens the decision margin $9.5\times$ while steering strength stays flat.

Doubt Is a Decision: A Sparse Circuit Commits Language Models to “I Don’t Know”

Sheryl Mehta*, Shagun Chadha*, Kanika Singhal*, Manan Wadhwa, Shivam Dubey *arXiv preprint, coming soon*

About 20 attention heads and 100 neurons decide when a model says it does not know.

EXPERIENCE

Co-founder , Redarc Labs	2026 – present
Special Projects Fellow , Secure AI Futures Lab	2026 – present
Research Fellow , MARS, Cambridge AI Safety Hub	Jul 2026 – present
Google Summer of Code Fellow , HumanAI	May – Aug 2026
Technical AI Safety Fellow , AI Safety Initiative, Georgia Tech	Jan – May 2026
AI and Web3 Intern , Lokachakra	Jun – Oct 2025

AWARDS

6th of 1800+, Haas F1 Challenge (video anomaly detection) 2025