

Shivam Dubey

dubeyjiworks@gmail.com | punctualprocrastinator.github.io | linkedin.com/in/syntaxsavant | github.com/punctualprocrastinator

SUMMARY

AI safety researcher working on mechanistic interpretability and adversarial robustness: building white-box monitors for AI control, then attacking them with adaptive attackers. Ten accepted papers in 2025–26, including three at NeurIPS 2026 workshops with Manan Wadhwa and a first-author paper at ATTRIB (NeurIPS 2026). Now working on control and interpretability for non-model capability gains.

EXPERIENCE

MARS V, Meridian AI Research Summer • Fellow, White-Box Techniques for AI Control Jul 2026 – Present

- Compliance probes on four instruction-tuned models from three labs; a red/blue co-evolution loop across three attack families found jailbreak success and probe evasion mutually exclusive.
- Deployment lesson: every probe keeps a false positive rate, so monitoring needs a cascade, not one probe. Built Whetstone, the open-source harness. 8×B200 SLURM cluster.

Lexsi Labs • AI Researcher, Tabular Foundation Models Lab 2026 – Present

- Attributing tabular foundation model behavior to synthetic pretraining data. Three workshop papers (NeurIPS, ICML 2026).

Redarc Labs • Co-founder and Research Lead, Delhi 2026 – Present

- Lab behind all joint papers with Manan Wadhwa; now control for non-model capability gains.
- Run a six-person fellowship; mentor three first-author projects (two under review at NeurIPS 2026 workshops); wrote and teach an eight-session mechanistic interpretability curriculum.

Secure AI Futures Lab • Special Projects Fellow 2026 – Present

- Teaching the Redarc interpretability bootcamp in person at the Delhi hub; weekend reading groups.

Apart Research • Research Fellow Nov 2025 – Present

- Hessian-trace analysis of ViTs and ResNets under seven attacks: a five-order-of-magnitude gap between input- and parameter-space curvature. Lead author, TAIS 2026.

AI Institute of South Carolina • Research Intern Jan 2025 – Apr 2026

- Designed the fairness-aware speculative decoding in AMBEDKAR: religion bias down 77%, caste bias down 23%, at 6% latency.

SELECTED PUBLICATIONS

* equal contribution. Full list of 13 at punctualprocrastinator.github.io.

1. **Known by Construction: Calibrating Model-State Instruments Against Models Whose State Was Installed, Hidden, or Never Present.** Wadhwa*, Dubey*. *New in ML, NeurIPS 2026.*
2. **When Does Synthetic Relational Data Teach Models to Use Relations?** Dubey, Bouadi, Bouarour, Kulkarni, Tanna, Sankarapu. *ATTRIB, NeurIPS 2026.*
3. **Dead Salmons Have a False-Positive Rate: A Calibrated Null-Model Test for Interpretability.** Wadhwa*, Dubey*. *Interpretability as a Science, NeurIPS 2026.*
4. **Steered Affect Moves a Partner Agent’s Activations, and the Logistic Probe That Reads It Returns a Tie-Break.** Wadhwa*, Dubey*, et al. *Interpreting Agent Behavior, NeurIPS 2026.*
5. **Toxin Circuits in ESM-2: Mechanistic Interpretability for Biosecurity Screening.** Wadhwa*, Dubey*. *GenBio, ICML 2026.*
6. **Loss Landscape Response to Adversarial Perturbation Is Architecture-Dependent.** Dubey, Hoelscher-Obermaier. *TAIS 2026.*
7. **Spectral Regularization as a Safety-Critical Inductive Bias.** Dubey. *Reliable ML, NeurIPS 2025.*

OPEN SOURCE

- **whetstone:** Red×Blue harness for adversarial safety-monitor research. **dead-salmon-fpr:** a calibrated null for interpretability claims. **neuron_autolabel:** auto-interpretability scored by detection AUC.

EDUCATION

IIT Madras, BS in Data Science 2023 – 2027

Maharaja Agrasen Institute of Technology, B.Tech, Mechanical and Automation Engineering 2022 – 2026

SKILLS

Methods	QK/OV decomposition, direct logit attribution, activation patching, steering vectors, linear probing, SAE feature attribution, nanoGCG attacks, Hessian trace
Stack	Python, CUDA, PyTorch, JAX, TransformerLens, NNsight, Hugging Face, SLURM (multi-node)