

Shivam Dubey

dubeyjiworks@gmail.com | punctualprocrastinator.github.io | linkedin.com/in/syntaxsavant | github.com/punctualprocrastinator
AI Safety, Mechanistic Interpretability, Adversarial Robustness, White-Box AI Control

RESEARCH FOCUS

AI safety researcher working on mechanistic interpretability and adversarial robustness. Current focus is white-box monitoring for AI control: building activation-level monitors, then attacking them with adaptive attackers. Core methods are QK/OV circuit decomposition, direct logit attribution, activation patching, steering vectors, SAE feature attribution and spectral dynamics. Now extending this to non-model capability gains: what control and interpretability look like for the harness a model runs in, not only the model. The approach throughout is understand, exploit, then defend.

EDUCATION

Indian Institute of Technology Madras 2023 – 2027
BS in Data Science. Focus: AI safety, interpretability, fairness, governance

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

RESEARCH EXPERIENCE

MARS V, Meridian AI Research Summer Jul 2026 – Present
Fellow, Lindley Lentati Track: White-Box Techniques for AI Control. Cambridge

- Trained linear compliance probes on residual-stream activations across four instruction-tuned models from three labs (Qwen2.5-7B, Llama-3.1-8B, Gemma-2-9B, ablated Llama-3). Compliance is linearly decodable and tracks refusal itself rather than surface phrasing.
- Ran a red-team/blue-team co-evolution loop against the probes across three attack families (suffix-gradient, discrete search, activation steering). **Jailbreak success and probe evasion came out mutually exclusive**: perturbations that evade the monitor also stop jailbreaking.
- The deployment lesson is the false positive rate, not detection: probes do find the jailbreak direction, but every probe keeps a false positive rate, so deployment needs a monitor cascade rather than one probe or a handful.
- Extending into trained attack policies (Redwood open problem: stress-testing white-box monitors) behind a six-stage de-risking protocol, with an ablation that separates behavior from instruction-following. Built Whetstone, the open-source Red×Blue harness behind it. Multi-GPU SLURM cluster (8×B200).

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

- Capability research on tabular foundation models: attributing model behavior to synthetic pretraining data. Two papers at the ATTRIB workshop at NeurIPS 2026 (first author on one) and one at FMSD at ICML 2026.

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

- AI safety research lab co-founded with Manan Wadhwa; all our joint papers come from it. Started with interpretability for existential risk; now focused on control and interpretability for non-model capability gains (what an agent can do because of its harness rather than its weights).
- Run a six-person research fellowship. Cohort 01 is built around one thesis: expressed confidence in LLMs is separable from internal knowledge, implemented partly through entropy neurons that modulate output entropy without changing token predictions.
- Mentoring three first-author workshop projects; two papers under review at NeurIPS 2026 workshops.
- Wrote and teach *Redarc Fellows*, an eight-session mechanistic interpretability curriculum taught with PyTorch hooks and Hugging Face directly, with no interpretability library in the stack.

Secure AI Futures Lab 2026 – Present
Special Projects Fellow. Delhi hub

- Teaching the Redarc Labs mechanistic interpretability bootcamp in person at SAFL's Delhi hub, and running reading groups on the weekends.

Apart Research Nov 2025 – Present
Research Fellow. Mentor: Jason Hoelscher-Obermaier

- Measured geometric divergence of ViTs and ResNets under seven adversarial attack algorithms using Hutchinson's stochastic Hessian trace estimator. Found a five-order-of-magnitude gap between input-space and parameter-space curvature: standard robustness metrics misclassify internal model instability. Lead author, TAIS 2026.

Artificial Intelligence Institute of South Carolina Jan 2025 – Apr 2026
Research Intern. Mentor: Dr. Amitava Das

- Co-author of AMBEDKAR (arXiv:2509.02133, cited by MIT Technology Review). Designed the fairness-aware speculative decoding algorithm, which reduces religion bias by 77% and caste bias by 23% at 6% latency overhead.

PUBLICATIONS

* equal contribution.

ACCEPTED

1. **Known by Construction: Calibrating Model-State Instruments Against Models Whose State Was Installed, Hidden, or Never Present.** M. Wadhwa*, S. Dubey*. *New in ML, NeurIPS 2026*.
2. **Dead Salmons Have a False-Positive Rate: A Calibrated Null-Model Test for Interpretability.** M. Wadhwa*, S. Dubey*. *Interpretability as a Science, NeurIPS 2026*.
3. **Steered Affect Moves a Partner Agent’s Activations, and the Logistic Probe That Reads It Returns a Tie-Break, Not an Estimate.** M. Wadhwa*, S. Dubey*, D. Ghosh, S. Kumar, A. Garg. *Interpreting Agent Behavior, NeurIPS 2026*.
4. **When Does Synthetic Relational Data Teach Models to Use Relations? Tracing Predictive Structure from Pretraining Data to Model Behavior.** S. Dubey, M. Bouadi, N. Bouarour, V. Kulkarni, A. Tanna, V. K. Sankarapu. *ATTRIB Workshop, NeurIPS 2026*.
5. **From Behavior to Provenance: Attributing Tabular Foundation Models to Synthetic Pretraining Data.** M. Bouadi, N. Bouarour, S. Dubey, A. Tanna, V. K. Sankarapu. *ATTRIB Workshop, NeurIPS 2026*.
6. **Toxin Circuits in ESM-2: Mechanistic Interpretability for Biosecurity Screening.** M. Wadhwa*, S. Dubey*. *GenBio Workshop, ICML 2026*.
7. **What You Pretrain On Matters: Synthetic Task Distributions Determine Tabular Foundation Model Quality.** M. Bouadi, N. Bouarour, V. Kulkarni, S. Dubey, A. Tanna, V. K. Sankarapu. *FMSD Workshop, ICML 2026*. arXiv:2605.18971, as “Shaping the Prior”.
8. **Loss Landscape Response to Adversarial Perturbation Is Architecture-Dependent: Comparing ViT and ResNet.** S. Dubey, J. Hoelscher-Obermaier. *TAIS 2026, Oxford*.
9. **Physics-Informed Neural Operators with Multi-Method Explainability for Ocean Temperature Prediction.** S. Dubey. *XAI4Science Workshop, AAAI 2026*.
10. **Spectral Regularization as a Safety-Critical Inductive Bias.** S. Dubey. *Reliable ML Workshop, NeurIPS 2025*.

UNDER REVIEW AND PREPRINTS

11. **AMBEDKAR: A Multi-level Bias Elimination through a Decoding Approach with Knowledge Augmentation for Robust Constitutional Alignment of Language Models.** S. Mukhopadhyay, A. Kasat, S. Dubey, R. Karthikeyan, D. Sood, V. Jain, A. Chadha, A. Das. *arXiv:2509.02133, 2025*. Cited by MIT Technology Review.
12. **Distance to the Boundary Is Not Steering Resistance.** M. Wadhwa*, S. Dubey*. *Under review, UniReps*.
13. **Doubt Is a Decision: A Sparse Circuit Commits Language Models to “I Don’t Know”.** S. Mehta*, S. Chadha*, K. Singhal*, M. Wadhwa, S. Dubey. *arXiv preprint, forthcoming*.

SELECTED PROJECTS

- **Isolating Refusal and Compliance Heads** (Qwen2.5-7B): a zero-shot harm classifier at AUC 0.943 with zero false positives; unconditional bipolar assertion holds nanoGCG attack success to 33%, against 66% undefended.
- **Pre-commitment Probing** (OLMo-3-7B-Think): a residual-stream probe predicts answer correctness before any reasoning token (AUC 0.780); a Layer 15 direction transfers across six languages ($p = 0.0017$).
- **The Geometry of Control: Spectral Attractors** (Llama-3.2-3B, Qwen-2.5-3B): refusals, bias and reasoning each live in their own spectrally rigid low-dimensional subspace; projecting into the safety subspace yields glitch tokens, not refusals.
- **Context Poisoning in Tabular ICL**: the first logit lens for tabular in-context learning; 1.2% poisoned context hijacks it completely, and larger models are more vulnerable, not less.

OPEN-SOURCE TOOLING

- **whetstone**: Red×Blue harness for adversarial safety-monitor research, built for MARS.
- **neuron_autolabel**: auto-interpretability for LLaMA-style models, scored by detection AUC rather than simulation.
- **dead-salmon-fpr**: calibrated null-model test; corrects interpretability false positives on untrained models.
- **mechanistic-tamper-verification**: detects tampered model updates by where weight change concentrates.

TECHNICAL SKILLS

Languages	Python, CUDA, MATLAB, Bash, JavaScript
Frameworks	PyTorch, JAX, TransformerLens, NNSight, Hugging Face, DevInterp, LangChain, Docker, SLURM (multi-node, 8×B200)
Methods	QK/OV circuit decomposition, direct logit attribution, activation patching, steering vectors (CAA), logit lens, linear probing, SAE feature attribution, nanoGCG and suffix-gradient attacks, Hutchinson Hessian trace, SVD and spectral dynamics