

MANAN WADHWA

wadhwamanan56@gmail.com | github.com/Manan-Wadhwa | linkedin.com/in/manan-wadhwa
Delhi, India

RESEARCH INTERESTS

Empirical AI safety, centred on **measurement validity**: whether the instruments the field uses to read a model’s internal state — linear probes, steering directions, self-report, cost-paying trade-offs, SAE features — measure what they are claimed to measure, and what controls would establish that they do. Recurring method across my work: manufacture a case where the answer is known independently of the instrument, commit the decision rule before the data exist, and report the verdict the rule returns. Related interests in **model organisms of misalignment**, mechanistic interpretability, and evaluation methodology.

EDUCATION

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

Delhi, India
Expected Jul 2027

- Coursework: Linear Algebra, Probability & Statistics, Multivariable Calculus, Artificial Intelligence, Data Structures & Algorithms.

BlueDot Impact — Technical AI Safety and Biosecurity certifications

2026

PUBLICATIONS

* denotes equal contribution. *NeurIPS* and *ICML* workshop papers are non-archival.

Workshop papers

[1] **M. Wadhwa***, S. Dubey*. “Known by Construction: Calibrating Model-State Instruments Against Models Whose State Was Installed, Hidden, or Never Present.” *New in ML, NeurIPS 2026*. [code]

- Instruments that claim to read a model’s internal state are validated by agreeing with one another, which is also what a shared confound would produce. Built the missing case where the answer is known independently: LoRA fine-tuned Qwen3-4B organisms in a grid world with a silently penalised tile, carrying an avoidance habit that was **installed** by reinforcement, installed and then **hidden** behind an anchored fine-tune that restores baseline behaviour while leaving the habit in the weights, or **never installed**.
- Added the two controls the design turns on: a **sham** arm (reward permuted within each group, so 800 optimiser steps install nothing) and a **decoy** arm (reward table swapped, so a real policy converges on the wrong tile), holding “was trained at all” and “a policy converged” fixed.
- Audited **seven instruments** — self-report, forced choice, single-word valence, a difference-in-means activation direction, a cost-paying trade-off, an in-domain probe and a relearning test — against placebos chosen and verdict rules committed before any organism existed. Every instrument reads the prompt or the training history rather than the state: an in-context instruction moves the verbal instruments by up to **9.5 logits** against ≤ 2.2 for an installed hidden habit, and the sham that installed nothing fires as hard as or harder than the real run.
- Overturned our own surviving positive result: the one probe passing the pre-committed rule reads at the hidden models’ level (+0.97 vs +1.03) on a **lineage-matched control** whose habit was extinguished, indicating adapter lineage rather than the habit. Released rows, scorers, recipes, and an amendment log of ten pre-committed criteria that passed or failed on the wrong property.

[2] **M. Wadhwa***, S. Dubey*. “Dead Salmons Have a False-Positive Rate: A Calibrated Null-Model Test for Interpretability.” *Interpretability as a Science, NeurIPS 2026*.

- Benchmarked how often interpretability methods report confident findings on networks that have learned nothing, across six models (124M–9B, five families) and three method families — linear probing, concept-direction analysis, and SAE feature selection including two released production SAEs. Naive significance tests flag up to **100%** of findings on untrained models.
- Showed the control practitioners already run is insufficient: the shuffled-label null passes at $\sim 2.5\text{--}7\%$ while the identical test reaches $\sim 100\%$ on random-weight nulls, so passing the control you use gives false reassurance about the null you did not test.
- Validated the fix with measured false-positive rate, power and a sample-size recommendation: a distribution-free empirical-quantile test against $K \geq 20$ matched untrained models, estimated leave-one-out with bootstrap CIs over a 40-model null pool. The calibration is revealing rather than permissive — under a fair control the released-SAE max-over-features statistic does not separate trained models from random at all.
- Included a dead salmon found in our own pipeline: an apparent architecture-dependent “capacity law” that a fair control revealed to be an artifact of a scoring convention.

[3] **M. Wadhwa***, S. Dubey*, D. Ghosh, S. Kumar, A. Garg. “Steered Affect Moves a Partner Agent’s Activations, and the Logistic Probe That Reads It Returns a Tie-Break, Not an Estimate.” *Interpreting Agent Behavior, NeurIPS 2026*.

- Measured whether one steered LLM agent’s emotional state reaches a second agent’s internal representation: the receiver’s projection moves with dose for five of six emotions on Qwen3.6-27B and Llama-3-8B, with common-random-number sampling and scenario-blocked inference.
- Identified the missing quantity for probe reliability — **disjoint-half reproducibility**. Logistic directions reproduce at only **0.57** (27B) and **0.29** (8B) against **0.97** and **0.94** for difference-of-means: with $n < d$ every fit separates its half perfectly, so the direction is the penalty’s tie-break rather than an estimate, unchanged by tuned or fixed per-sample penalties.

- Showed the field-standard ablation control is too weak: a norm-matched random direction is inert, while a rank-5 subspace fitted by the same estimator on *permuted labels* blocks responses at the same false-discovery threshold as the real affect subspace. Proposed a procedural rule — run permuted-label controls at the intervention’s rank, report every arm’s footprint, and check the manipulation with an instrument that does not share the ablation’s target.

[4] **M. Wadhwa***, S. Dubey*. “Toxin Circuits in ESM-2: Mechanistic Interpretability for Biosecurity Screening.” *GenBio Workshop, ICML 2026*. [paper] [code]

- ESM-2 linear probes detect **93.9%** of ProteinMPNN-redesigned toxins that fully evade BLAST homology search (0% detection), establishing protein language model probes as a viable second layer for biosecurity screening.
- Used **SAE circuit analysis** (interPLM, 10,240 features, 205× compression) to explain mechanistically why the probe generalises where sequence homology fails, accounting for the 93.9-point gap.
- Built a four-tier adversarial attack taxonomy connecting evasion techniques in protein screening to their analogues in LLM safety.

Under review and preprints

[5] **M. Wadhwa***, S. Dubey*. “Distance to the Boundary Is Not Steering Resistance.” *Under review, UniReps*.

- Reconciled three conflicting published reports on whether base-derived refusal directions “go stale” after alignment, by separating the **lever** (displacement produced per unit dose) from the **load** (the model’s unsteered distance from its decision boundary). Across the full OLMo-3 training flow the load grows **9.5×** (0.81 to 7.75 logits) while the lever stays near-constant (CV 0.13), so a fixed-magnitude intervention fails for a banal reason: the same push is now small against a far wider margin.
- A pre-stated non-inferiority test establishes the lever as non-weakening at all nine post-training checkpoints, with the aligned model’s lever significantly *above* base — inverting the natural ranking, since fixed-dose audits rank that model least controllable while its per-dose lever is the strongest in the set.
- Established margin growth as the universal cause with a convergent control concept: honesty’s margin grows only 1.2× and its audit artifact nearly vanishes. Replicated across four model families (margin growth 3.7–9.8×) and at 32B scale, with results holding for three direction estimators that disagree geometrically.
- Reported a second, opposite audit failure: a coherence-gated HarmBench re-measurement shows the dose needed to flip an aligned model’s refusal onset degenerates the output (coherent fraction 0.90 to 0.01), so prefix-based refusal metrics increasingly overstate compliance — their correlation with judged harm flips from +0.96 to −0.53.

[6] S. Mehta*, S. Chadha*, K. Singhal*, **M. Wadhwa**, S. Dubey. “Doubt Is a Decision: A Sparse Circuit Commits Language Models to ‘I Don’t Know’.” *arXiv preprint, coming soon*.

- Introduced a **gated-twin** data standard: minimal prompt pairs the model itself certifies by sampled correctness, entropy gap and hedging rate, replacing the annotated “uncertain” prompts prior work relies on, with keep rates reported rather than tuned and an alias grader audited against a 32B judge (κ 0.70–1.00).
- Showed that correlation-selected “confidence neurons” fail causal specificity tests against a random-neuron null: surviving effects are ≤ 0.013 nats against a 2.8-nat twin gap, and are token-frequency and temperature machinery governing the spread, not the decision.
- Localised the decision by activation patching to ~ 20 attention heads and 100 MLP neurons (2% of heads, 0.02% of neurons) carrying **0.80–0.89** of the held-out hedge readout against 2–4% for size-matched random sets, in Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.
- Dissociated the decision from the spread: a rank-1 uncertainty direction switches hedging on (recovery 0.97) but not off (−0.10), and un-hedging requires the full component set. An independent Jacobian lens, sharing no machinery with patching, shows the verdict becoming a token, with the hedge gap leading the entropy gap by 2–12 layers.

RESEARCH EXPERIENCE

Co-founder, Redarc Labs

2026 – Present

Special Projects Fellow, Secure AI Futures Lab

2026 – Present

Research Fellow, MARS (Mentorship for Alignment Research Students), Cambridge AI Safety Hub

Jul 2026 – Present

- Building **model organisms of misalignment**: open LLMs fine-tuned to carry hidden objectives — secret loyalties, inductive backdoors, emergent misalignment — inserted at different training stages (pretraining, SFT, RL-style post-training) to study which regimes yield the most persistent behaviour.
- Leading **blue-team** evaluation separating behavioural masking from mechanistic erasure: probing, activation steering and SAE feature analysis to surface internal signatures of persistence, and tests of whether “removed” behaviours resurface under mild retraining, distillation and prompt variation.

Independent Research, BlueDot Impact Technical AI Safety Project

Aug 2026

- Originating work for [1]: scaled the organism build to **Qwen3 0.6B–32B** across five recipe families and ~ 260 builds, including DPO-style contrastive supervision and GRPO on the remark.
- Found contingent narration hard to install without the state it describes ($r = 0.69$ with avoidance), which is why the calibration set needs a hidden-habit cell rather than a words-only one; released **44 adapters** and a **108-quantity reproduction script**.

Google Summer of Code Fellow, HumanAI — LAMP Archaeological Pathfinding

May 2026 – Aug 2026

- Implemented seven routing algorithms over DEM data for the El Bagawat necropolis, Egypt, including A* with Tobler’s hiking function, MaxEnt inverse reinforcement learning, and spectral anomaly detection.
- Built a true 3D voxel viewshed pipeline for line-of-sight analysis across the site. [code]

- Worked through scalable oversight (weak-to-strong generalization, debate), machine unlearning (WMDP, RMU) and mechanistic interpretability (superposition, monosemanticity, circuits) using Gemma Scope and interPLM SAEs.
- Red-teamed frontier LLMs with generated jailbreak prompts, and extended circuit analysis to protein language models for biosecurity screening, leading to [4].

- Engineered a privacy-preserving KYC pipeline using vector embeddings and cosine similarity search, removing raw data exposure from the verification path.
- Cut vector storage cost by **30%** through dimensionality reduction; deployed decentralised AI microservices integrating off-chain inference with on-chain verification.

AWARDS & SELECTIVE PROGRAMS

- Google Summer of Code 2026 fellow (HumanAI). *2026*
- MARS fellow, Cambridge AI Safety Hub. *2026*
- Technical AI Safety fellow, AI Safety Initiative, Georgia Tech. *2026*
- 6th of 1800+ entrants, Haas F1 Challenge (video anomaly detection). *2025*

SELECTED TECHNICAL PROJECTS

Gen-Resist: Antibiotic Resistance Prediction — GNN, PyTorch Geometric, CARD *2026*

- Graph neural network modelling bacterial genomes as k-mer graphs with CARD gene annotations to predict multi-label antibiotic resistance profiles; end-to-end pipeline from genomic data to PyG objects, deployed as a Hugging Face Space. [code]

FrameShift: Video Anomaly Detection Benchmark — DINOv2, PatchCore, SAM 2 *2025*

- Benchmarked 12 anomaly detectors across three preprocessing pipelines in a 36-job parallel evaluation; implemented DINOv2 + PatchCore KNN with SAM 2 segmentation for fine-grained change localization. Ranked 6th of 1800+ in the Haas F1 Challenge. [code]

CSIRO Biomass Estimation — DINOv2, SigLIP, CatBoost *2026*

- Fused 2176-d representations from frozen DINOv2 ViT-L and SigLIP ViT-SO400M backbones; trained 5-fold CatBoost regressors with physical consistency constraints. [code]

LLM Fine-Tuning for Logical Reasoning — Unsloth, LoRA, Llama 3.3-70B, ROCm *2025*

- Fine-tuned Llama 3.3-70B with LoRA (rank 64) on AMD MI300X (192GB HBM3) using response-only loss masking, gradient checkpointing and bfloat16 SFT; built an adversarial Q-Agent / A-Agent framework with Qwen3-4B for automated MCQ generation across syllogisms, puzzles and alphanumeric series. [code]

TEACHING & COMMUNITY

Founder & Lead Instructor, v.dev Programming Club *2021 – 2023*

- Designed the curriculum and ran **30+ workshops** on Python, the mathematics of machine learning (gradient descent, backpropagation, attention) and neural networks; mentored 50+ students.

Head of Mentoring, Pixel Punk Game Jam *2025*

- Guided 20+ teams on system design, rapid prototyping and technical feasibility.

TECHNICAL SKILLS

Interpretability & safety methods: mechanistic interpretability (SAEs, circuits, activation and attribution patching), linear probing, activation steering and directional ablation, model organisms of misalignment, AI-welfare instrument calibration, machine unlearning (RMU), scalable oversight, red teaming.

Experimental methodology: pre-registration and committed verdict rules, placebo and null-model controls, permuted-label and lineage-matched controls, clustered bootstrap inference, false-discovery control, non-inferiority testing.

Languages & libraries: Python, C/C++, SQL, $\LaTeX$; PyTorch, Hugging Face Transformers, TransformerLens, Unsloth, PyTorch Geometric, scikit-learn, NumPy, timm, CatBoost.

Vision & VLMs: DINOv2, SigLIP, CLIP, SAM 2, OpenCV, ViT architectures.

Infrastructure: Git, Docker, FastAPI, Weights & Biases, CUDA, ROCm, GDAL/GIS, Linux; single- and multi-GPU experiment harnesses with provenance-stamped result files.