Deven Choudhary

+91-8791810013 | devensodyssey@gmail.com | linkedin.com/in/choudharydeven | devenchoudhary.com

Summary

Independent mechanistic interpretability researcher, working on where and how models represent things internally – circuits, features, and where their explanations break down. Final-year B.Tech CS (AI/ML) at Bennett University, graduating 2027.

Publications

When Sentiment Circuits Don't Cross Borders – sole authorPreprint; targeting **BlackboxNLP @ EMNLP 2026**
- Used hook-based logit-diff activation patching across all 144 attention heads (native PyTorch) to localize the English sentiment circuit (dominant head L7H2, max IE +1.21).
- Found partial transfer to Devanagari Hindi (+0.85), near-zero transfer to Romanized Hindi ($n = 922$, max IE +0.32), and full transfer in regionally pre-trained MuRIL (+1.42).
- Identified token fragmentation (53.9% pair filtration) as a mechanical barrier; the MuRIL contrast gives converging evidence that pretraining exposure, not architecture alone, drives cross-lingual circuit formation.

Loyal Lies – Auditing Secret Loyalties Under Attack – co-author Apart Research Hackathon, July 2026
- Showed activation probes are unreliable loyalty detectors: a probe trained on two clean seeds of the same recipe, with no loyalty signal present, hit AUROC 1.000 – a representation-entanglement confound.
- Built a grey-box logprob audit that correctly ranked the true principal 1 of 120 ($z = 4.57$) while keeping clean and decoy controls in the base band; an obfuscation fine-tune attack still left one detection channel standing.
- Project | GitHub

Research & Projects

Crosswise – Fabrication Detection for NLA Interpretability | *PyTorch, SAEs* GitHub
- Cross-checks natural-language activation explanations against independent sparse-autoencoder feature evidence to flag unsupported claims.
- Catches 83% of fabricated claims at 67% precision on a 212-position benchmark.

AuditAI – Causal Bias Audit of Gemma-2-2B | *PyTorch, LoRA* GitHub
- Measured directional caste-name bias in model approval decisions using counterfactual surname pairs and sequence-level log-probability scoring; largest observed approval gap +1.94 (Malhotra vs. Khatik).
- Built a Measure/Flag/Fix pipeline with LoRA fine-tuning (r=32, bf16, all projections) to test mitigation, served behind a lightweight API. Submitted to Google Solution Challenge 2026 (Top 100).

Circuit Surgeon – Automated Circuit Discovery | *TransformerLens, PyTorch* GitHub
- Implements ACDC-style edge attribution patching with greedy pruning and hierarchical node-to-edge refinement, validated by causal scrubbing (complement, circuit, and random baselines) with faithfulness scoring.

Experience

Authentic Memory (GAMI) Munich, Germany (Remote)
Software Engineering Intern, Cryptographic Provenance Infrastructure *Aug 2026 – Present*
- Backend and integration engineering on open cryptographic infrastructure for long-term archival integrity and provenance verification.

Technical Skills

Interpretability: Activation Patching, Edge Attribution (EAP), Circuit Tracing (ACDC, Attribution Graphs), Causal Scrubbing, SAEs, Transcoders/CLTs, Crosscoders, NLAs (Natural Language Autoencoders), Feature Steering
ML/DL: PyTorch, Hugging Face, TransformerLens, SAE Training, LoRA
Engineering: Python, FastAPI, Docker, PostgreSQL, Redis, Qdrant, Git, Linux, Bash

Programs & Talks

HAAISS 2026 – Human-Aligned AI Summer School, Prague July 2026

BlueDot Impact – Technical AI Safety Course & Project Sprint *(participant, both)* Completed 2026

Education

Bennett University Greater Noida, India
B.Tech in Computer Science Engineering (AI/ML Specialization) *Expected May 2027*