Gaurav Hadavale

hadavalegaurav@gmail.com | Portfolio | LinkedIn | GitHub

RESEARCH INTERESTS

Explainable AI (XAI), Mechanistic Interpretability, Causal Auditing of LLMs, Trustworthy Medical Imaging, and AI Safety.

EDUCATION

Sardar Patel Institute of Technology (University of Mumbai) Mumbai, India
B.Tech in Electronics & Telecommunication Engineering Aug 2023 – June 2027 (Expected)

- **Minor in AI & Machine Learning:** CGPA 9.0/10
- **Relevant Coursework:** Deep Learning, Medical Image Analysis, Computer Vision, Linear Algebra, Probability & Statistics.

PREPRINTS & PUBLICATIONS

- **G. Hadavale**. "Beyond Accuracy: An Interpretability-Driven Audit of Deep Learning Models for Pneumonia Detection from Chest X-Rays." *TechRxiv*. DOI: 10.36227/techrxiv.176739845.56601492/v1. (Abstract Submitted to CARS/ECR 2026).
- **G. Hadavale**. "Causal Auditing of Latent Affect in Language Models via Activation Steering." *TechRxiv*. DOI: 10.36227/techrxiv.176823105.52829404/v1.

RESEARCH EXPERIENCE

Beyond Accuracy: An Interpretability-Driven Audit of Pneumonia Detection Models *PyTorch, RISE*

- **Problem:** SOTA DenseNet-121 models often exhibit "Clever Hans" behavior, relying on non-pathological shortcuts such as text markers rather than clinical pathology.
- **Objective:** To design a causal auditing framework that identifies and mitigates spurious correlations in deep learning-based medical diagnostics.
- **Approach:** Developed a statistical pipeline using RISE and Grad-CAM++ to isolate High-Confidence False Positives (HCFP) and conducted marker injection/removal experiments for causal validation.
- **Results:** Improved specificity from 0.49 to 0.8675 via targeted data cleaning and demonstrated zero-shot robustness on the adult NIH Chest X-ray 14 dataset (AUROC 0.65).

Causal Auditing of Latent Affect in Language Models via Activation Steering *TransformerLens, GPT-2*

- **Problem:** Input-level explanations fail to isolate internal causal mechanisms governing complex affective behaviors in Large Language Models (LLMs).
- **Objective:** To identify and causally modulate latent affective concepts (specifically "anger") in a transformer-based model via activation-level interventions.
- **Approach:** Used Contrastive Activation Analysis to isolate an "anger" direction in the residual stream (Layer 8) and applied inference-time activation steering ($h' = h + \alpha\theta$).
- **Results:** Proved that steering modulates discourse structure (e.g., negation rate, sentence length) and established specificity using norm-matched random control vectors.

Manifold-Constrained Counterfactual Explanations for Medical Diagnostics *PyTorch, VAE*

- **Problem:** Standard gradient-based explanations frequently produce "off-manifold" adversarial noise that lacks clinical actionability and plausibility.
- **Objective:** To generate counterfactual explanations that respect the underlying data distribution to provide meaningful medical interventions.

- **Approach:** Formulated a manifold-constrained objective using Variational Autoencoders (VAEs) and validated interpretability through PCA latent space visualization.
- **Results:** Achieved an ROC AUC of 0.9132 and reduced critical False Negatives by 66% through Youden's J Statistic-based threshold calibration.

TECHNICAL SKILLS

- **Explainable AI (XAI):** Mechanistic Interpretability, Counterfactual Generation, Manifold Learning, Causal Interventions, Grad-CAM/++, SHAP, Concept Bottlenecks.
- **Data Analysis & Statistics:** Statistical Modeling, Hypothesis Testing, Youden's J Statistic, ROC/AUC Analysis, Spectral Analysis (FFT), Experimental Design.
- **Computer Vision:** Vision Transformers (ViT), CNNs (DenseNet, ResNet, ConvNeXt), Medical Image Analysis, Image Forensics, Saliency Mapping.
- **Deep Learning:** Variational Autoencoders (VAE), Self-Supervised Learning, Random Forests, XGBoost, PCA, Clustering, Transfer Learning, Optimization Algorithms.
- **Languages/Tools:** Python, Julia, SQL, C++, PyTorch, TransformerLens, Scipy, NumPy, Pandas, Scikit-learn.