Bharat Yalavarthi

bharatcy2@gmail.com | linkedin.com/in/bharat-yalavarthi | byalavar.github.io

RESEARCH INTEREST

My research focus is two-fold. First, I aim to advance the **interpretability** and understanding of deep learning models. Second, I leverage interpretability techniques to diagnose failure modes and improve the capabilities, **evaluation, safety, robustness, and trust.**

EDUCATION

- **University at Buffalo, The State University of New York** **Jan 25 - Present**
  Ph.D. in Computer Science - GPA: 3.8/4.0*, UB Presidential Fellowship*
  Advisors: Dr. Nalini Ratha, Dr. Venu Govindaraju
- **University at Buffalo, The State University of New York** **Aug 22 - May 24**
  M.S. in Computer Science - GPA: 3.9/4.0

SELECTED PUBLICATIONS

Visit my Google Scholar for complete list

- **Label-Free Mitigation of Spurious Correlations in VLMs Using Sparse Autoencoders** ICLR 26
  B.C Yalavarthi*, N. Ratha, V. Govindaraju*
- **Aligning Characteristic Descriptors with Images for Human-Expert-like Explainability** NeurIPS IAI Workshop 24
  B. C. Yalavarthi*, N. Ratha*
- **Enhancing Privacy in Face Analytics Using Fully Homomorphic Encryption** IEEE FG 24
  B. Yalavarthi*, A. R. Kaushik, A. Ross, V. Boddeti, N. Ratha*

RESEARCH EXPERIENCE

University at Buffalo (PhD Student) **Jan 2025 - Present**

- **(ICLR 2026)** Proposed Sparse Autoencoder (**SAE**) based VLM debiasing method that isolates and removes spurious correlations; improved WG accuracy on 5 standard benchmarks without labels, extra data, or retraining [Paper].
- Experimented with entropy regularization techniques on Concept Bottleneck Models (**CBMs**), cutting information leakage by 40%.
- Probed SAE concept formation across Gemma Scope scales: scaling yields fractal compositional expansion, not atomic convergence; narrow models collapse high-frequency concepts into coarse superpositions.

SUNY Research Foundation **Aug 2024 - Jan 2025**
Researcher - Full Time *Buffalo, NY*

- **Benchmark Creation:** Designed a Gestalt-grounded benchmark (1,200+ stimuli, 4 perceptual principles) revealing systematic visual reasoning failures in GPT-5, Claude-3.7, and Llama-3.2, absent from standard VQA benchmarks.
- Stress-tested SOTA MLLMs on OCR via targeted variations, isolating regimes where vision encoders (not language) break down.

SUNY Research Foundation **Jan 2023 – May 2024**
Research Assistant *Buffalo, NY*

- **Explainability:** Employed Mistral-7B and CLIP to generate human expert like justifications for face recognition and medical predictions; improved explanation faithfulness through characteristic descriptors **(NeurIPS-W 24)**. [GitHub]
- **Privacy & Secure ML:** Exposed attribute leakage in face-template protection schemes; FHE-based mitigation cut leakage **35%** over SOTA with no accuracy loss. **(IEEE FG 24, 25)**. Adapted FHE framework to a custom CNN achieving **99.3%** accuracy in secure sleep apnea detection from encrypted ECG signals **(ICPR 24)**.

INDUSTRY EXPERIENCE

Harman International (Samsung) **Oct 2020 – Jul 2022**
Software Engineer — ADAS & Navigation *Bangalore, India*

- **Delivered 20+** ADAS and route guidance features for VW Trucks/Coaches; modular redesign reduced SW maintenance costs by **25%**.
- **Led the product line expansion** of ADAS and Route Guidance modules for coaches, coordinating with multiple internal teams and clients. Integrated onboard traffic sign recognition module cutting incorrect sign events by **90%**.

MISCELLANEOUS

- **Skills:** PyTorch, Numpy, activation patching, causal tracing, probing, transcoders, red-teaming, evaluation, benchmark design, Hugging Face, Gemma Scope, SAELens, DDP, docker, Linux, LLMs, VLMs, diffusion models, Python, C++.
- **Reviewer:** CVPR, WACV, ICIP, IJDAR. **Presented** at **NSF CITeR** Program Review 2024.