

VEERARAJU ELLURU

92 Holly Rd, Iselin, NJ 08830 | +1 (945) 425-1534 | veerarajuel@gmail.com | [LinkedIn](#) | [GitHub](#) | [website](#)
B.Tech in Computer Science and Engineering, Indian Institute of Technology Jodhpur

Researcher with strong interests in AI safety, interpretability, and alignment, driven to make increasingly capable AI systems safe, secure, and reliable. Experienced in developing foundation models, interpreting and analysing anthropomorphic risks and LLM behaviors through rigorous research collaborations across academia and industry.

RESEARCH EXPERIENCE

External Collaborator, Jinesis Lab, University of Toronto, Canada (Remote) **Summer 2026**

Advisor: Dr. Zhijin Jin

- Ongoing work on influence functions as a novel black-box method to quantify [eval awareness](#) in LLMs.
- Ongoing work on generalizable mechanistic interpretability.

Undergraduate Research Assistant, [Image Analytics and Biometrics Lab](#), CSE Dept, IITJ **Nov. 2024 - May 2026**

Advisors: Dr. Mayank Vatsa and Dr. Shivang Agarwal

- Research on privacy-respecting, safe, and interpretable AI systems. Topics of research include Machine Unlearning, mechanistic interpretability, and adversarial attacks in Vision and Audio modalities.
- Representative [works](#) accepted to **ICCV-W** and under review at T-BIOM.

Research Intern, TAILS, Thoughtworks, Chicago, IL (Remote) **Summer 2025**

- Research on the **Fine-grained Incompleteness Evaluation** of Textual **Summarization** tasks LLMs.
- Co-development and evaluation of **Self-Attention via Lie-Algebraic Flows** for textual summarization.
- Representative [work](#) accepted at **PRICAI**.

Research Intern (REU), CDA, University of Illinois Urbana-Champaign, Champaign, IL **Summer 2024**

Advisor: Dr. Tiago Bresolin | [Animal Science Proceedings](#)

- **Developed Foundation Models** for Livestock Image Segmentation pipelines leveraging self-supervised, non-contrastive learning algorithms like Bootstrap Your Own Latent. Generated robust and precise cattle masks for large-scale **out-of-distribution** datasets via Supervised Fine-Tuning.
- Achieved **SoTA** on the data-specific segmentation - a **6%** improvement in Jaccard scores on large-scale cross-species datasets, at **only 2-5% labeled data** during SFT, hence debunking supervised algorithms.

PROFESSIONAL EXPERIENCE

Machine Learning and Data Analyst Intern, Fluxgen Technologies, Bengaluru, IN **Summer 2023**

- **Optimized** raw data from a live dashboard for analysis, enabling EDA and model training.
 - Improved **anomaly detection accuracy** using unsupervised algorithms - **Isolation Forest** and **OC-SVM**.
 - **Developed** data-driven water-level optimization insights, influencing **Tata Steel** amongst other clients.
 - Through anomaly detection, water usage was **optimized** by **14%** month-over-month at **3** steel plants.
-

EDUCATION

Indian Institute of Technology, Jodhpur

Bachelor of Technology in Computer Science and Engineering

2022 - 2026

CGPA: 9.33/10 (Rank 7/550)

RELEVANT COURSEWORK

Linear Algebra, Probability, Statistics & Stochastic Processes, Deep Learning, Natural Language Understanding, Computer Vision, Foundation Models & GenAI, Design & Analysis of Algorithms

SKILLS

Languages: Python, C, C++ | SQL | HTML, CSS, JavaScript

Frameworks and Tools: PyTorch | Django | AWS | Git, Bash

Deep Learning: HuggingFace, W&B, Ollama, Unsloth, Multi-gpu training

PUBLICATIONS

- **Veeraraju Elluru**, Arth Singh, Roberto Agüero, Hreetam Paul, Debojyoti Das, Ajay Agarwal. "Mechanistic Analysis of Model Compression in VLMs." **ICML Mech Interp Workshop 2026**. [URL](#).
 - Sai Siddhartha Chary Aylapuram, **Veeraraju Elluru**, and Shivang Agarwal. "Bias-aware machine unlearning: Towards fairer vision models via controllable forgetting." **ICCV U&ME Workshop, 2025**. [URL](#).
 - Manikandan Ravikiran, **Veeraraju Elluru**, Karthik Iyer, Prasanna Pendse, and Shayan Mohanty. "Lie algebra-based semantic flow for incompleteness detection in summarization." **PKAW, 2025**. [URL](#).
 - **Veeraraju Elluru**, Shreyansh Pathak, Shivang Agarwal, Mayank Vatsa, and Richa Singh. "Sparse Autoencoder-guided Audio Unlearning." **Transactions in Biometrics, Behavior, and Identity Science, 2025**. Under review.
 - Aarush Sinha, Ishan Garg, **Veeraraju Elluru**, Arth Singh, Kushal Garg. "Mechanistic Analysis of Alignment Algorithms in Language Models." **NeurIPS ED Track 2026**. Under review.
 - **Veeraraju Elluru**, Ahmad Sallam, Anderson Alves, and Tiago Bresolin. "Towards unsupervised latent representations for cattle image segmentation." **Animal Science proceedings, 16:560–562, 2025**.
-

PROJECTS

"Fantastic Biases and Where to Find Them" | [GitHub](#)

Nov. 2025

- Applied **mechanistic interpretability** (linear probing, activation and attribution patching, automated circuit discovery) to locate **gender and racial bias** in OSS models (GPT2-medium, large, GPT-Neo) to identify *sparse* circuits responsible for biased representations.
- Utilized the **TransformerLens** framework to trace information flow across Multihead Attention Heads and MLP layers to output logits on StereoSet and WinoGender datasets.
- Performed activation engineering-based **steering**, including additive, projective, and SAE-based steering to surgically **mitigate bias** by up to **40%** in the residual streams of models at **inference-time**, without degrading MMLU scores (<2%).

Multi-view 3D scene reconstruction, Course: Computer Vision | [GitHub](#) / [HF](#)

Apr. 2025

- **Evaluated** multiple scene-reconstruction methods such as NeRFs, GANs, Gaussian Splatting, and classical Structure-from-Motion (SfM) and Multi-view Stereo (MVS) techniques.
 - **Benchmarked** the above models across a plethora of 2D-3D datasets to illustrate the best performance.
 - Deployed a **front-end** application using Streamlit to support easy tinkering.
-

ACHIEVEMENTS

Outstanding Reviewer, ICML Mechanistic Interpretability Workshop

Jun. 2026

FINALIST, Anthropic Fellowship Program

Apr. 2026

Selected to attend the annual **Microsoft Research India Academic Summit**

Jun. 2025

Ranked 7/500 throughout undergraduate

2022-26

LEADERSHIP EXPERIENCE

Co-founder (Technical Research), Safety And Alignment Research ([SAAR-India](#))

March 2026-Present

- Founded SAAR-India, a community with close to 500 members actively working on technical AI Safety and Alignment research.
- Fostering more than 10 ongoing research projects that resulted in papers/technical blogs.