Kaustubh Sharma

kaustubh_s@ee.iitr.ac.in | kaustubh202.github.io | github.com/kaustubh202/

Education

- **Indian Institute of Technology Roorkee** *2023-2027*
  B.Tech, Electrical Engineering | CGPA: 9.04/10 Roorkee, India

Areas of Interest

Mechanistic Interpretability, Foundation Models, AI Safety, Scientific ML

Publications and Pre-Prints

- **Dissecting Attention and MLP Roles: A Study of Domain Specialization in Large Language Models**
  Kaustubh Sharma, *Ojasva Nema, Abhiraj Bharangar, Manjot Singh, Srijan Tiwari*
  Under Review at ICLR, 2026 (First author; led ideation, method design, experiments, and writing)
- **Decoupled-Value Attention for Prior-Data Fitted Networks: GP Inference for Physical Equations**
  Kaustubh Sharma, *Simardeep Singh, Parikshit Pareek*
  Under Review at ICLR, 2026 (First author; developed DVA Attention and conducted experiments)
- **Explainable AI-Generated Image Forensics: A Low-Resolution Perspective with Novel Artifact Taxonomy**
  Kaustubh Sharma
  Proceedings of the International Conference on Computer Vision, APAI Workshop, 2025 (Solo author; full project ownership)
- **Image-Alchemy: Advancing Subject Fidelity in Personalized Text-to-Image Generation**
  Kaustubh Sharma, *Ojasva Nema, Amritanshu Tiwari, Cherish Puniani*
  DeLTa Workshop, ICLR, 2025 (First author; led method design, experiments and writing)

Research and Professional Experience

- **Undergraduate Research Assistant** *May 2025 - Ongoing*
  $\mathcal{P}^2$-Lab, Prof. Parikshit Pareek, Indian Institute of Technology Roorkee
  - ◦ Developed the DVA Attention mechanism for Gaussian Processes Inference in PFNs – Github
  - ◦ Working on building a foundational architecture for amortized kernel hyperparameter inference
- **Domain Circuit Discovery in LLMs for Safety - Mechanistic Interpretability** *April 2025 - Ongoing*
  Data Science Group, IIT Roorkee – Project Website (Ongoing)
  - ◦ Investigating domain-specific knowledge emergence in LLaMA 3-3B to locate specialized layers.
  - ◦ Evaluated Causal Effects, Probe Separability, Zero out tests, Hydra Effect and Fine Tuning shifts.
- **Sparsity-Aware Representation Learning for Jet Image Generation via Guided Latent Diffusion** *2025*
  Data Science Group, IIT Roorkee – Project Github
  - ◦ Developed a sparsity-aware latent diffusion framework to generate High Energy Particles Physics Data.
  - ◦ Crafted a custom variational autoencoder with a novel sparsity-focused reconstruction loss.
  - ◦ Introduced a mean-pulling mechanism during latent diffusion to minimize reconstruction artifacts.
- **Incoming Quantitative Research Intern** *Upcoming*
  Goldman Sachs
  - ◦ Selected for the quantitative strategist internship involving financial modeling and statistical analysis.
- **Educator at Edufabrica Pvt. Ltd.** *March 2025*
  - ◦ Delivered a workshop lecture for 2 days to 200+ students across India on Generative AI.

Achievements

- ◦ **Micron AI Hackathon** (2025, Micron Technology) — Second Runner Up in micron AI Hackathon.
- ◦ **JEE Advanced** (2023, Govt. of India) — AIR 1624 among 1 lakh+ applicants.
- ◦ **JEE Mains** (2023, Govt. of India) — AIR 1898 among 12 lakh+ applicants.
- ◦ **NTSE Scholar** (2021, Govt. of India) — National scholarship awarded to 2000 out of 9L+ candidates.
- ◦ **KVPY Fellow** (2022, IISc, Govt. of India) — AIR 509 among 2 lakh+ applicants.

Additional Information and Co-Curriculars

Technical Skills: Python (Pytorch, Transformers, NumPy, Diffusion), C++

ML Areas: Mechanistic Interpretability, Autoregressive Modelling, Diffusion Processes, Language Models

Math Interests: Probability Theory, Gaussian Processes, Optimization, Statistics

Languages: Hindi (Native), English (Advanced), French (Beginner)

Activities: Pianist — Music Section (IITR); Member — IITR Swimming Team