ANSHUL KAUSHAL

AI Researcher
M.Tech. Artificial Intelligence
Indian Institute Of Technology, Patna
Focus: LLM Safety, Mechanistic Interpretability, Multimodal Learning, Computer Vision

kaushal1anshul@gmail.com
github.com/Enthusiast101
linkedin.com/in/kaushal1anshul
enthusiast101.github.io
Google Scholar

Education

Degree/Certificate	Institute/Board	CGPA/Percentage	Year
M.Tech. (AI)	Indian Institute of Technology, Patna	–	2026 – Present
B.E. (CSE)	Panjab University	9.07 / 10	2022 – 2026

Summary

AI researcher with 5 research internships and 5 publications spanning LLM safety, mechanistic interpretability, multimodal learning, computer vision, medical image segmentation, sparse signal reconstruction, and explainable AI. Hands-on experience with PyTorch, TensorFlow, Hugging Face, DINOv2, MetaCLIP, Mamba, Conditional GANs, Retrieval-Augmented Generation (RAG), model auditing, and representation learning.

Research Experience

- **Lexsi Labs (Aurionpro)** *Remote, Maharashtra, India*
  Research Intern – Mechanistic Interpretability and Safety in LLMs Feb 2026 – Aug 2026
  – Built **SAFETUNE**, a unified AI safety framework integrating **118+ methods** for auditing, repairing, and preserving safety in fine-tuned language models.
  – Applied **mechanistic interpretability**, safety circuit discovery, and steering vectors to analyze multilingual LLM safety and support model alignment.

- **Indian Institute of Technology Jodhpur** *Jodhpur, Rajasthan, India*
  ACM IKDD Scholar – Self-Supervised Vision and Multimodal Learning May 2025 – Jan 2026
  – Engineered a CLIP-like image-text representation model by integrating the **DINOv2** self-supervised vision encoder into **MetaCLIP** for downstream multimodal large language model (MLLM) applications.
  – Aligned the vision backbone using **99K CC and LAION image-caption pairs**, lowering data and compute requirements for multimodal representation learning.
  – Raised **zero-shot classification by approximately 2%** across ImageNet-1K, ObjectNet, and iNAT2021 benchmarks.

- **National Astronomical Observatory Japan** *Remote, Mitaka, Japan*
  Research Intern – Sparse Signal Reconstruction Nov 2024 – Nov 2025
  – Reconstructed sparse **Gamma-Ray Burst (GRB) light-curve signals** across four categories: Good, Break, Flares, and Break with Flares.
  – Benchmarked Conditional GANs, **Mamba state-space models (SSMs)**, and polynomial fitting for sparse signal reconstruction.
  – Lowered uncertainty relative to prior studies to **37% for** $\log T_a$**, 38% for** $\log F_a$**, and 41% for** α **parameters**.

- **IISER Bhopal** *Bhopal, Madhya Pradesh, India*
  Research Intern – Medical Image Segmentation Jun 2024 – Jul 2024
  – Studied **Vision-Mamba state-space models** for multi-domain medical imaging and semantic segmentation.
  – Augmented CNN-Transformer segmentation models with an adaptive learnable **fuzzy logic module** across four public datasets.
  – Raised accuracy by **2%** while reducing FLOPs, memory consumption, and parameter count by **30%**.

- **IIIT Allahabad** *Remote, Allahabad, India*
  Research Intern – Emotion Classification using EEG Mar 2024 – Jun 2024
  – Classified **24 emotions** from high-density EEG data for affective computing.
  – Optimized hyperparameters with a genetic algorithm, increasing classification accuracy and F1-score by **4%**.
  – Interpreted model outputs with **explainable AI (XAI)** to identify EEG frequency bands associated with specific emotional states.

Publications

- P. Seth*, **Anshul Kaushal***, S. Sadhu*, V. K. Sankrapu, "Drift, Then Repair: A Safety Audit of Fine-Tuned Language Models," **NeurIPS 2026 FLLMPT Workshop**. Library.
- P. Seth*, S. Sadhu*, **Anshul Kaushal***, V. K. Sankrapu, "SAFETUNE: A Unified, Faithful Library for Auditing and Repairing Safety Drift in Fine-Tuned LLMs," **EMNLP 2026 Demo Track**. Library.
- **Anshul Kaushal**[1], K. Jangid[2*], V. K. Kurmi[3], "CLFSeg: A Fuzzy-Logic based Solution for Boundary Clarity and Uncertainty Reduction in Medical Image Segmentation," **BMVC 2025, Core-A**. Paper.
- **Anshul Kaushal**[1], A. Manchanda[2], M. G. Dainotti[3*], et al., "Multi-Model Framework for Reconstructing Gamma-Ray Burst Light Curves," **Journal of High Energy Astrophysics**. Paper.

- A. Manchanda[1], **Anshul Kaushal**[2], M. G. Dainotti[3*], et al., "Gamma-Ray Burst Light Curve Reconstruction: A Comparative Machine and Deep Learning Analysis," **The Astrophysical Journal**. Paper.
- N. Ali[1], M. Asif[2*], **Anshul Kaushal**[3], U. Singh, U. S. Tiwari, "Optimizing Emotion Recognition in EEG Data: A Genetic Algorithm Approach with XAI Insights," **ICCCNT 2024**. Paper.

Projects

- **COVID-19 Lung Infection Segmentation**
  Semantic segmentation, U-Net variants, medical image analysis GitHub
  – Implemented Trans, Residual, Squared, and Recurrent **U-Net** variants for COVID-19 lung infection segmentation, achieving a **Jaccard Index** $>$ **0.87**.
  – Deployed model inference with **FastAPI** and **Streamlit** for real-time predictions.
- **Academic Research RAG-bot**
  Retrieval-Augmented Generation for research-paper question answering GitHub
  – Built a **Retrieval-Augmented Generation (RAG)** pipeline with OpenAI GPT-3.5 Turbo and text embeddings for research-paper question answering.
  – Implemented **Pinecone vector database** retrieval with dot-product similarity for semantic search over embedded research data.
- **Rock-Paper-Scissors Classifier**
  Transfer learning, computer vision, model interpretability GitHub
  – Fine-tuned a pre-trained **ResNet-18** with data augmentation for image classification, reaching **85% training accuracy**.
  – Applied **Grad-CAM** and t-SNE visualizations with OpenCV to interpret learned image features and model behavior.

Key Courses Taken

- Machine Learning, Deep Learning, Representation Learning, Optimization, Generative AI, Unsupervised Learning, Recommender Systems, Reinforcement Learning, Mathematics for Machine Learning, Linear Algebra, Data Science Math Skills

Technical Skills

- **Programming:** Python, C, C++, SQL, MATLAB, LaTeX, JavaScript, HTML, CSS
- **Machine Learning:** Machine Learning, Deep Learning, Representation Learning, Self-Supervised Learning, Transfer Learning, Hyperparameter Optimization, Generative AI, Computer Vision, Semantic Segmentation, Model Evaluation, Explainable AI (XAI)
- **LLM / Multimodal AI:** Large Language Models (LLMs), LLM Safety, Mechanistic Interpretability, Model Alignment, Model Auditing, Safety Circuit Discovery, Steering Vectors, Multimodal Large Language Models (MLLMs), Retrieval-Augmented Generation (RAG), Text Embeddings, Vector Databases
- **Models / Methods:** DINOv2, MetaCLIP, CLIP, Mamba SSMs, Conditional GANs, CNN-Transformers, U-Net, ResNet-18, Genetic Algorithms, Fuzzy Logic, Grad-CAM, t-SNE
- **Frameworks / Libraries:** PyTorch, TensorFlow, Hugging Face, Keras, Scikit-learn, NumPy, Pandas, OpenCV, Matplotlib, Seaborn, FastAPI, Streamlit
- **Tools / Platforms:** Linux, Git/GitHub, Docker, CUDA, AWS, GCP, Jupyter, Google Colab, Anaconda, Power BI, MS Office

Positions of Responsibility

- **Core Team,** Training and Placement Cell *Oct. 2023 – Present*
  – Collaborated in a **15+ member team** to organize **10+ workshops and events** on machine learning and AI, impacting **200+ students**.

Achievements

- MEXT 2027 Scholarship – top 186 candidates from 7,300 applicants (top 2.5%)
- ACM IKDD Uplink Research Internship 2025 – 2% acceptance rate
- Oxford Machine Learning Summer School (OxML) 2025 – shortlisted
- IELTS Academic – 8.0 Overall Band (L: 9.0, R: 9.0, W: 7.5, S: 6.5)

Certifications

- MLx Fundamentals (OxML): Machine Learning, Deep Learning, Representation Learning, Optimization, Generative AI
- DeepLearning.AI (Coursera): Unsupervised Learning, Recommender Systems, Reinforcement Learning
- Imperial College London (Coursera): Mathematics for Machine Learning – Linear Algebra
- Amazon Web Services: AWS Elements of Data Science
- Duke University (Coursera): Data Science Math Skills