Kartikeya Chitranshi

kartikeya4524@gmail.com — +49 176 61668557 — Berlin, Germany
LinkedIn: https://www.linkedin.com/in/kartikeya-chitranshi/
GitHub: https://github.com/kartikeya1234

PROFILE

Machine Learning Researcher specializing in adversarial robustness and trustworthy AI for vision-language models. Two years at Zuse Institute Berlin implementing supervised fine-tuning pipelines for CLIP models using PyTorch and HuggingFace Transformers on multi-GPU clusters, with co-authored work on counterfactual training methods to reduce spurious correlations. Master's thesis investigated sparse adversarial attacks on multi-modal foundation models, systematically comparing adversarial versus counterfactual training paradigms across image classification and image-text retrieval benchmarks. Demonstrated expertise spanning both attack design and defense mechanisms, implementing adversarial perturbations to expose vulnerabilities while working to mitigate spurious correlations. Research interests focus on securing LLMs against adversarial manipulation while maintaining explainability and fairness, addressing critical challenges in deploying trustworthy AI systems.

RESEARCH INTERESTS

LLM Security & Robustness, Systematic Robustness Evaluation, Defense Mechanism Interactions, Attack Transferability, Trustworthy AI.

EDUCATION

Master: MSc. in Scientific Computing [Apr-2022] – [Jul-2025]
TU Berlin - Berlin, Germany.

Main focus: Adversarial Machine Learning, Trustworthy AI, Reinforcement Learning, Optimization Under Uncertainty.

Master Thesis, TU Berlin - Berlin, Germany. [Oct-2024] – [Apr-2025]
Topic: Robustness of Multi-Modal Foundation Models.

- Investigated **adversarial vulnerabilities in vision-language models**, implementing sparse attack algorithms that expose model weaknesses through imperceptible perturbations.
- **Systematic robustness evaluation** of CLIP models fine-tuned with adversarial versus counterfactual approaches across image classification and image-text retrieval tasks.
- **Implemented attack algorithms** and training pipelines using PyTorch and HuggingFace Transformers on multi-GPU infrastructure with **SLURM**.

Bachelor: BSc. in Physical Sciences [Jul-2018] – [Jun-2021]
University of Delhi - Delhi, India.

Main focus: Numerical Methods, Calculus and Matrices, Differential Equations, OOP.

PUBLICATIONS & PREPRINTS

- **Training on Plausible Counterfactuals Removes Spurious Correlations,**
Sadiku, S., Chitranshi, K., Kera, H., and Pokutta, S. (2025), *arXiv preprint, arXiv:2505.16583*.

WORK EXPERIENCE

Student Research Assistant [Sep-2023] – [Sep-2025]
Zuse Institute Berlin - Berlin, Germany.

- **Implemented supervised fine-tuning pipelines** for CLIP vision-language models using counterfactual and adversarial training to reduce spurious correlations, deploying training on multi-GPU clusters with SLURM using PyTorch and HuggingFace Transformers, culminating into co-authored work ”**Training on Plausible Counterfactuals Removes Spurious Correlations**”.
- **Worked on evaluation frameworks** comparing adversarially-trained versus counterfactual-trained foundation models on image classification and image-text retrieval tasks.
- **Implemented sparse adversarial attacks** on CNNs and vision-language models to systematically assess robustness vulnerabilities.
- **Investigated transferability of adversarial attacks**, especially Cross-model and Intra-model, across multiple machine learning models.

Machine Learning Intern [Jun-2021] – [Aug-2021]
TestAIng.com - Bangalore, India.

- **Implemented multiple adversarial attack algorithms** on deep neural networks using TensorFlow for robustness testing.
- **Created documentation** for reproducible experiments and systematic model evaluation.

KNOWLEDGE & SKILLS

Language skills:	German (A2, actively improving toward B1), English (C2), Hindi (Native).
Machine Learning:	Adversarial Attacks (FGSM, PGD, C&W, sparse perturbations), Defense Mechanisms (adversarial training, counterfactual training), Transferability Analysis, Robustness Evaluation & Benchmarking.
FM Expertise:	Vision-Language Models (CLIP, OpenFlamingo), Large Language Models, Vision Transformers, CNNs, Supervised Fine-Tuning.
Deep Learning:	PyTorch, HuggingFace Transformers, PyTorch Lightning, TensorFlow, Multi-GPU Distributed Training, SLURM, Multi-GPU cluster.
ML Development:	Python, Numpy, Pandas, Scikit-Learn, OpenCV, Matplotlib, Git, Docker, Kubernetes, Optuna.
Systems & Tools:	Linux, LaTeX, Vim.
Personal Skills:	Research Independence, Problem Solving & Analytical Thinking, Team Work, Communication & Presentation Skills.

ACADEMIC RECOGNITION

- Selected for Master’s programs at Imperial College London, University of Edinburgh, and Trinity College Dublin.

Kartikeya Chitranshi