Vishnu K. Chhabra

Columbus, OH | 612-500-6474 | vishnu.kabirc@gmail.com
linkedin.com/in/vishnu-chhabra1999 | github.com/vishnuKC

Summary

PhD candidate (expected May 2027; major in Artificial Intelligence, minors in Statistical Learning and High Performance Computing) working on AI safety through mechanistic interpretability. I study how safety mechanisms in language models, such as refusal behavior and steering directions, change under compression, fine-tuning, and corruption, and I build lightweight methods that make models safer. 10 papers (6 first-author) at NAACL, MobiCom, CSCW, and ICML workshops. Seeking Research Scientist roles in AI safety and interpretability.

Education

The Ohio State University **May 2022 – Expected May 2027**
PhD in Computer Science and Engineering (GPA: 3.9) *Columbus, Ohio*
Major: Artificial Intelligence. **Minors**: Statistical Learning, High Performance Computing.
Advisors: Xueru Zhang, Mohammad Mahdi Khalili. **Coursework**: Advanced Machine Learning, Trustworthy AI, Natural Language Processing, Computer Vision, Artificial Intelligence, Algorithms, Probability Theory, Real Analysis

University of Minnesota – Twin Cities **September 2018 – May 2022**
Bachelor's in Computer Science and Mathematics *Minneapolis, Minnesota*
Advisors: Joseph Konstan, Ju Sun

Publications

AI Safety and Interpretability

- **V.K. Chhabra**, D. Zhu, M.M. Khalili. *"Neuroplasticity and Corruption in Model Mechanisms: A Case Study of Indirect Object Identification"*. **NAACL 2025**
- **V.K. Chhabra**, D. Zhu, Z. Zhu, X. Zhang, M.M. Khalili. *"Towards Understanding and Improving Refusal in Compressed Models via Mechanistic Interpretability"*. **ICML 2025 Workshop**
- **V.K. Chhabra**, X. Zhang, M.M. Khalili. *"Speaking the Steering Vector: Steering Vectors as Universally Transferable System Prompts"*. *Under Review*
- **V.K. Chhabra**, Z. Zhu, X. Zhang, M.M. Khalili. *"Scoping the Emergence of Steering Vectors in Large Language Models"*. *Under Review*
- **V.K. Chhabra**, X. Zhang, M.M. Khalili. *"Alignment Gradients are Pseudo-Refusal Directions"*. *Under Review*
- **V.K. Chhabra**, D. Zhu, Z. Zhu, X. Zhang, M.M. Khalili. *"Fixing What Fine-Tuning Breaks: A Simple and Efficient Approach"*. *Under Review*
- S. Gupte, **V.K. Chhabra**, M.M. Khalili. *"On the Transferability of Sparse Autoencoders for Interpreting Compressed Models"*. *Under Review*

Machine Learning, Systems, and HCI

- A. Banerjee, X. Zhao, **V.K. Chhabra**, K. Srinivasan, S. Parthasarathy. *"HORCRUX: Accurate Cross Band Channel Prediction"*. **MobiCom 2024**
- D. Zhu, **V.K. Chhabra**, M.M. Khalili. *"ECG Signal Denoising Using Multi-scale Patch Embedding and Transformers"*. **ICML 2024 Workshop (NGSM)**
- R. Kong, C. Zhang, R. Sun, **V.K. Chhabra**, T. Nadimpalli, J. Konstan. *"Multi-Objective Personalization in Multi-Stakeholder Organizational Bulk E-mail: A Field Experiment"*. **CSCW 2022**

Research Experience

The Ohio State University **May 2022 – Present**
Graduate Associate, AI Safety and Mechanistic Interpretability *Columbus, Ohio*

- **Safety of compressed models.** Showed how pruning and quantization change the mechanisms behind refusal in language models, and introduced **AIRD**, a lightweight algorithm that improves the safety of compressed models by up to **41%** (ICML 2025 Workshop). Studying whether Sparse Autoencoders transfer to compressed models.
- **Fine-tuning and corruption.** Discovered *neuroplasticity* in language model mechanisms, giving a mechanistic explanation of how task-specific and harmful fine-tuning change a model (NAACL 2025). Developed a simple, efficient approach to repair model behavior broken by fine-tuning (under review).
- **Steering vectors and refusal directions.** Studying how steering vectors emerge and evolve in LLMs and showing that they can act as universally transferable system prompts; showed that alignment gradients behave as pseudo-refusal directions (3 papers under review).
- **Ongoing work.** Aligning model safety by pruning jailbreak-related features with Sparse Autoencoders; steering text embeddings in multimodal vision-language models; interpretability of agentic safety in long-horizon settings.
- **Mentoring.** Mentored undergraduate students in the AIEdge REU (2023) on resource allocation and deepfake detection projects.

Additional Experience

Scriptulate **May 2021 – August 2021**
Software Engineering Intern *Remote*

- Built databases and anomaly detection on healthcare data using K-NN, SMOTE, spectral clustering, K-Means, and SVM; built knowledge graphs to analyze doctor–patient relationships.

University of Minnesota – Twin Cities **September 2020 – May 2022**
Teaching Assistant *Minneapolis, Minnesota*

- Taught Java and OCaml to undergraduates remotely during the pandemic; designed grading schemes, project assignments, and debugging solutions.

Selected Research Projects

Cross Band Channel Prediction | *Python, PyTorch, Diffusion Models* **May 2022**

- Designed an end-to-end framework based on a conditional diffusion model that outperforms the prior state of the art by **2 dB** SNR for downlink channel prediction; collaborated on the physics-based HORCRUX approach (MobiCom 2024).

Facial Recognition for Dental-Assisting Robots | *Python, PyTorch* **June 2023**

- Generated synthetic benchmark data with Stable Diffusion and evaluated AdaFace and MogFace, reaching **90%** accuracy on a new benchmark for dual-factor authentication in dental-assisting robots.

Food Calorie Estimation using WiFi | *MATLAB, Python, Spectral Analysis* **March 2023**

- Estimated calories and nutrients of liquid and solid foods from WiFi signals using spectral analysis, with up to **10%** error and robustness to signal interference and food containers.

Deep Reinforcement Learning for Automated Trading | *Undergraduate Thesis* **January 2022**

- Benchmarked DQN, DDPG, PPO, SAC, A2C, TD3, and multi-agent DDPG for cryptocurrency trading, modeling the market as a POMDP and evaluating with **Monte Carlo simulations**. Awarded a **Distinguished** Poster.

Technical Skills

Research: Mechanistic interpretability, Sparse Autoencoders, steering vectors, circuit analysis, LLM safety and refusal evaluation, model compression (pruning, quantization), fine-tuning, statistical learning, diffusion models, reinforcement learning
Languages: Python, C/C++, Java, OCaml, SQL, CUDA
ML and Interpretability Libraries: PyTorch, Hugging Face Transformers, TransformerLens, SAELens, TensorFlow, scikit-learn, NumPy, pandas
Training and Inference: PyTorch Distributed, DeepSpeed/FSDP, Horovod, vLLM, Weights & Biases
HPC and Infrastructure: Slurm, MPI, Linux, Docker, Kubernetes, AWS SageMaker
Applications and Tools: FastAPI, LangChain, OpenAI API, Git/GitHub, MATLAB