Towards Trustworthy Multimodal AI Systems

Chirag Agarwal

University of Virginia
chiragagarwal@virgina.edu

In the proposed talk, I will present an overview of my research under the theme "*Towards Trustworthy Multimodal AI Systems*." My talk will consist of two main parts. The first section will survey my work on multimodal interpretability, including benchmarks for quantifying the limitations of existing multimodal explanation algorithms and challenges of scaling current techniques (especially, mechanistic interpretability tools) to multimodal models. The second section will focus on my research in addressing safety and reasoning problems of frontier medical models.

Rethinking Explainable Artificial Intelligence in the Era of Multimodal AI

With the increasing development of multimodal AI models, it becomes imperative that their predictions are readily explainable to relevant stakeholders and practitioners. However, the field of explainable artificial intelligence (XAI) has not kept pace with the multimodal surge, and most existing techniques remain unimodal, designed for models with a single input type. To accurately explain a multimodal model, we must develop modality-aware techniques that reveal how multiple modalities interact to produce their predictions and satisfy certain multimodal properties.

Lack of Multimodal XAI Techniques. In my talk, I will show why post-hoc heatmaps, token attributions, and mechanistic interpretability (a sub-field of XAI that aims to reverse-engineer models for understanding their inner workings) tools are insufficient for explaining multimodal models (Agarwal 2025; Petkar et al. 2025). Next, I will present a multimodal XAI benchmark for vision–language models that scores explanations on faithfulness (*does the explanation match the model's internal usage of evidence?*) and plausibility (*do humans make better decisions with it?*) (Agarwal et al. 2024). We will also discuss a framework for counterfactual and compositional explanations that reflect how a model learns evidence across modalities. The aim is to transition from pretty saliency maps to decision-relevant, testable explanations.

Mechanistic Interpretability (MechInterp). While there is a rise in the use of mechanistic interpretability tools to reverse engineer large language and multimodal models, there is little to no work on understanding their robustness and safety implications. Here, I will discuss our work on showing the sensitivity of state-of-the-art unsupervised probing techniques, a widely used MechInterp method (Sadiekh et al. 2026), and share some preliminary results from our ongoing work on multimodal interpretability tools for understanding visual-language models.

Safety and Reasoning for Medical AI Models

The safety and reasoning performance of frontier models are paramount in deploying them in critical applications like healthcare and clinical decision-making. This part will focus on two complementary threads: i) domain-specific safety and related trustworthy evaluation that surfaces risks missed by generic red-teaming, focusing on multilingual and distribution-shifted settings; and ii) methods that explicitly train and assess reasoning in medical LLMs.

Domain-specific Benchmarks. Here, I will summarize lessons from MEDSAFETY-style (Han et al. 2024) evaluation bench that probe clinical assistants with adversarial prompts, rare adverse-event scenarios, and distribution shifts. Considering the global impact of frontier models in healthcare, a key theme of my discussion will be on multilingual safety. I will highlight the CLINIC (Ghosh et al. 2025) line of work that systematizes multilingual clinical safety evaluation and shows how safety gaps compound under translation, code-switching, and noisy OCR. Specifically, as health models expand beyond English, toxicity, misinformation, and instruction-following failures worsen in low-resource languages.

Domain-Specific Reasoning. Clinical decision-making relies on structured reasoning rather than point predictions. Here, I will target the reasoning process of medical language models. Concretely, we elicit and train multilingual reasoning chains using long chain-of-thought (CoT) that follows clinician-familiar scaffolds (Onyame et al. 2025) and we ensure evidence grounding to guidelines or cited sources when available. Finally, I will discuss our training paradigm, where we pair a curriculum-informed reinforcement learning framework with cold-start initialization, code-switching–aware supervised fine-tuning (SFT), and Group Relative Policy Optimization (GRPO) on stepwise rationales over multilingual reasoning qualities.

References

Agarwal, C. 2025. Rethinking Explainability in the Era of Multimodal AI. *arXiv*.

Agarwal, C.; et al. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from LLMs. *arXiv*.

Ghosh, A.; et al. 2025. CLINIC : Evaluating Multilingual Trustworthiness in LLMs for Healthcare. *Under review*.

Han, T.; et al. 2024. Medsafetybench: Evaluating and improving the medical safety of LLMs. *NeurIPS*.

Onyame, E.; et al. 2025. CURE-Med: Curriculum-Informed Reinforcement Learning for Multilingual Medical Reasoning. *Under review*.

Petkar, S.; et al. 2025. A Graph Talks, But Who's Listening? Rethinking Evaluations for Graph-Language Models. *arXiv*.

Sadiekh, S.; et al. 2026. Polarity-Aware Probing for Quantifying Latent Alignment in Language Models. *AAAI*.