Thesis Project Portfolio

CaseEdit: Enhancing Localized Commonsense Reasoning via Null-Space Constrained Knowledge Editing in Small Parameter Language Models.

(Technical Report)

Cognitive-Inspired Neural Architectures: Enhancing Interpretability and Ethical Alignment for Safe Decision-Making in High-Impact Sectors.

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Varun Reddy

Spring, 2025

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Introduction: My Capstone project and my STS research explore how AI can be made safer, more trustworthy, and better aligned with human needs in real-world settings. My technical Capstone *focused* on enhancing small parameter language models' commonsense reasoning through CASEEDIT, a localized knowledge editing framework. My STS project, on the other hand, *examined* how cognitive science-inspired interpretability methods could help AI models behave more ethically and transparently, especially in high-stakes sectors like healthcare, justice, and education. Together, these projects grapple with the technical and societal challenges of ensuring that AI systems function effectively and behave in ways that align with human values.

Capstone Project: My Capstone *addressed* the problem of how small, edge-deployed language models struggle with personalized commonsense reasoning. In environments like smart homes, users often need AI to adapt to unconventional but intuitive uses of everyday objects—something large models handle better, but at a cost of impractical computational resources. To address this, we *developed* CASEEDIT, a knowledge editing framework that *enabled* small models to integrate personalized commonsense adaptations while minimizing ripple effects on unrelated knowledge. By enabling AI systems to better reflect human behavior patterns, CASEEDIT *aimed* to make AI assistants more practical, ethical, and useful. Considering the societal impacts of this technology *is* essential because commonsense reasoning *is* deeply cultural, context-dependent, and often invisible to model designers. Poorly adapted AI *can* perpetuate misunderstandings or even create risks in household environments if models act on rigid or inappropriate assumptions.

STS Research Project: In my STS research, I *explored* how cognitive science and neuroscience-inspired methods, such as mechanistic interpretability and chain-of-thought reasoning, could be used to decode AI systems' internal processes. I *critically analyzed* how interpretability techniques reveal how AI models function technically and how they reflect broader social norms, biases, and ethical standards. Using STS frameworks, I *argued* that AI interpretability must be culturally sensitive and sector-specific, especially in healthcare, law, and education, where the risks of AI errors *are* highest. My research *showed* that simply improving technical transparency *is* insufficient; AI systems must also be regulated and designed with an understanding of how different societies perceive fairness, trust, and accountability.

Conclusion: When considered together, my Capstone and STS research *reveal* that improving AI’s technical adaptability and transparency must go hand-in-hand with broader societal and ethical considerations. CASEEDIT *demonstrated* how models could be engineered to reason more flexibly within human-specific contexts. However, without STS-driven insights, such technical improvements could still fail by embedding hidden biases or ignoring ethical risks. Bridging the gap between technical and societal understanding *ensures* that future AI systems are not only functionally capable but also socially responsible, trustworthy, and truly aligned with the diverse ways humans live and reason.