Shadow: A Cross Domain, Mathematically Validated, Meta-Cognitive Reasoning Infrastructure for AI Models

Kommini, Kundana Chowdary (School: Wiregrass Ranch High School)

With the rise of Artificial Intelligence tools, society has seen a significant increase in their applications in high-stakes, sensitive industries. However, these tools often operate with opaque reasoning. Large Language Models (LLMs) are widely used for their conversational and human-like capabilities. Still, current techniques to uncover their reasoning are typically post-hoc, pattern-specific, computationally intensive, and costly, while also failing to capture the internal structure of models. My project introduces Shadow, a novel cross-domain, mathematically validated, meta-cognitive reasoning infrastructure for AI Models. Shadow analyzes LLM activations in situ, identifies abnormal reasoning pathways, and packages them into a meta-learning router to guide the model in recognizing and correcting its own deviations. Using a curated set of prompts, activations from Mistral 7B were analyzed and transformed into interpretable structures. A shared Universal Manifold allows this framework to generalize across multiple domains. The pipeline captures variance in activations with high accuracy ($p < 0.0001$) and identifies abnormal reasoning pathways that are statistically non-random. The meta-learning router achieved an Expected Calibration Error (ECE) of 0.802, indicating trustworthy alignment between confidence and accuracy, allowing the system to flag potentially inaccurate outputs in real time. The shared universal manifold, shows cosine similarity ($N=50$, $p<0.05$), facilitates generalization across transformer models. Shadow offers a scalable and cost-efficient approach to improving transparency, interpretability, and trustworthiness in LLM reasoning, providing experts in high-stakes industries with a tool for auditable and reliable AI outputs.

Awards Won:
Fourth Award of $600
Non-Trivial Ventures: Non-Trivial Fellowship Scholarship