MERIT: Mechanistic Explainability of Reasoning Integrity and Transparency

Shan, Richard (School: North Carolina School of Science and Mathematics)

Large language models (LLMs) are increasingly trusted in domains such as education, research, and decision making. Despite their widespread adoption, fundamental questions remain about alignment and logic in LLM reasoning. This assurance gap poses significant risks in high-stakes domains. This study develops a mechanistic framework for reasoning verification by examining internal model computations rather than relying solely on behavioral outputs. I analyzed DeepSeek-R1 Distill Llama-8B latent feature activations across 2,000 mathematical problems of varying difficulty and domain. Employing Sparse Autoencoders for feature extraction, I identify specific internal features directly tied to reasoning behavior. I develop six novel metrics to quantify reasoning quality. Through causal intervention analysis, I demonstrate that general features fracture into domain-specific specialists under increased task complexity. Experimental manipulation of identified features significantly boosts reasoning behavior, establishing that internal features functionally control aspects of reasoning. At higher mathematical difficulties, domain-experts emerge for geometry, number theory, and other subdomains. Additionally, I novelly demonstrate the existence of distinct feature-governed reasoning modalities: a concise calculation-oriented mode and a verbose explanation-oriented mode. Finally, I create an LLM optimization harness with my mechanistic findings, and successfully increase LLM reasoning accuracy by 22% on a test set of 1,000 advanced mathematical problems. This framework enables assurance and optimization of LLMs with implications for all reasoning-enabled AI applications.

Awards Won:
Third Award of $1,200
Midwest Microelectronics Consortium: Four cash awards of $3,000 each.