

Abstract

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 where reasoning integrity is critical. 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 by logical structure, self-corrections, and coherence. Through correlation and causal intervention analysis across difficulty levels, I demonstrate that general reasoning features fracture into domain-specific specialists under increased task complexity. Experimental manipulation of identified features produces significant causal effects on reasoning behavior, establishing that internal features functionally control specific aspects of reasoning quality. At higher mathematical difficulties, domain-experts emerge for geometry, number theory, and other mathematical subdomains. Additionally, I demonstrate the existence of distinct feature-governed reasoning modalities: a concise calculation-oriented mode and a verbose explanation-oriented mode. I establish that LLM reasoning is predictable by its internal feature structure, which is difficulty-dependent and domain-specialized. This framework enables assurance of reasoning models based on legitimate internal monitoring, rather than output correctness alone, with implications for all reasoning-enabled AI applications.