Measuring the Faithfulness and Safety of Large Reasoning Models

Dr. Jan Philip Wahle

Abstract: Large language models increasingly solve complex tasks by generating step-by-step reasoning before producing an answer. But does this verbalized reasoning actually reflect how a model reaches its decision? This project investigates the faithfulness of reasoning in large language models and its implications for AI safety. The project develops new benchmarks and intervention methods to test whether changes in a model's reasoning causally affect its decisions, particularly in realistic safety-critical scenarios, such as deciding how to respond to a warning about a potentially compromised food cold chain. It then connects these observable reasoning traces to the model's internal representations to identify when and how decisions are formed inside the model. Building on these insights, the project develops monitoring and intervention methods that can detect misleading, unreliable, or unsafe reasoning. Ultimately, the goal is to better understand when model explanations can be trusted and to develop methods that make the reasoning and decisions of advanced language models more transparent and reliable.