Otto Mättas
245867IAXD

A Novel Framework for Verifying Safety Properties in Large Language Models Using Formal Methods

SIGNIFICANCE STATEMENT

For Nature Machine Intelligence

Large Language Models (LLMs) are increasingly deployed in critical applications, from healthcare to financial systems, raising concerns about their safety and reliability. Despite their widespread use, there has been no systematic way to verify their safety properties formally. Our research introduces a groundbreaking framework that combines abstract interpretation with probabilistic model checking to verify safety properties in LLMs. This approach can detect potential failure modes and safety violations before deployment, reducing risks in real-world applications. We demonstrate our framework's effectiveness by successfully identifying previously unknown safety vulnerabilities in several widely-used LLMs. This work represents a significant advance in AI safety, providing developers and organisations with a practical tool for ensuring their AI systems meet critical safety requirements.