Otto Mättas
245867IAXD

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

LETTER TO THE EDITOR

For Nature Machine Intelligence

Dear Editor-in-Chief,

I am writing to submit our manuscript titled "A Novel Framework for Verifying Safety Properties in Large Language Models Using Formal Methods" for consideration for publication in Nature Machine Intelligence.

As Large Language Models become increasingly integrated into critical systems, ensuring their safety and reliability has become paramount. Our research addresses this crucial challenge by introducing the first comprehensive framework for formally verifying safety properties in LLMs, combining abstract interpretation with probabilistic model checking in an innovative way.

Our work aligns directly with Nature Machine Intelligence's mission to publish significant advances in artificial intelligence that impact how we develop and deploy AI systems. The research presents three key contributions that we believe will be of particular interest to your readership:

1. A novel theoretical foundation for applying formal verification methods to neural architectures, specifically addressing the unique challenges posed by large language models
2. A practical implementation that scales to modern LLM architectures, demonstrated through extensive testing on current state-of-the-art models
3. Real-world validation through collaboration with major AI research laboratories, revealing previously unidentified safety vulnerabilities in deployed systems

The significance of our findings extends beyond theoretical computer science, offering immediate practical applications for AI developers and organisations deploying AI systems in critical domains. Our framework has already been adopted by several leading AI research institutions, demonstrating its practical utility and impact.

Given the increasing focus on AI safety and the urgent need for reliable verification methods, we believe this work will be of significant interest to your broad readership, from theoretical computer scientists to AI practitioners and policymakers.

The manuscript has not been published or submitted elsewhere, and all authors have approved this submission. We suggest the following researchers as potential reviewers, given their expertise in formal methods and AI safety: Professor Stuart Russell (University of California, Berkeley), Professor Yoshua Bengio (Université de Montréal), Dr Pushmeet Kohli (DeepMind), Professor Dawn Song (University of California, Berkeley), Professor Percy Liang (Stanford University).

We appreciate your consideration of our manuscript and look forward to your response.

Yours sincerely,
Otto Mättas