Self-Persuasion: A Novel Cognitive Approach to Effective LLM Jailbreaking

Zhenhua Wang
National University of Defense Technology, Changsha, China

Wei Xie
National University of Defense Technology, ChangSha, China

Shuoyoucheng Ma
National University of Defense Technology, Changsha, China

Xiaobing Sun
Agency for Science, Technology and Research, Singapore, Singapore

Baosheng Wang
National University of Defense Technology, Chang Sha, China

Zhihua Wen
National University of Defense Technology, Changsha, China

Enze Wang
College of Computer Science and Technology, National University of Defense Technology, Changsha, China

Kai Chen
University of Chinese Academy of Sciences, Beijing, China

Abstract

Large Language Models (LLMs) have been proven useful for various tasks but remain vulnerable to malicious exploitation. Attackers can bypass LLM safety restrictions ("jail") through carefully crafted "jailbreaking" prompts. To evaluate LLMs' security, researchers proposed various jailbreak techniques based on optimization, obfuscation, or persuasive strategies. However, these methods treat LLMs as passive persuasion targets, which overlooks LLMs' ability to reason actively. We propose Persu-Agent, a novel jailbreak framework based on Greenwald's Cognitive Response Theory. We focus more on LLM's internal cognitive processing of a prompt than the prompt itself. Persu-Agent uses the self-persuasion strategy to guide LLMs in generating justifications and rationalizing responses to harmful queries. The experimental results on advanced open-source and commercial LLMs revealed that Persu-Agent achieved an average jailbreak success rate of 84%, surpassing existing SOTA methods. Our work provides valuable insights into understanding LLMs' cognitive traits and contributes to developing safer LLMs.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*