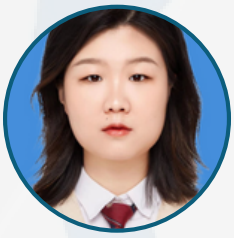

*Wenjie Wang, Yukai Zhou

ShanghaiTech University, Shanghai, China

Jailbreaking LLMs by Suppressing Rejection

Abstract: Ensuring the safety alignment of Large Language Models (LLMs) is crucial to generating responses consistent with human values. Despite their ability to recognize and avoid harmful queries, LLMs are vulnerable to "jailbreaking" attacks, where carefully crafted prompts elicit them to produce toxic content. One category of jailbreak attacks is reformulating the task as adversarial attacks by eliciting the LLM to generate an affirmative response. However, the typical attack in this category GCG has very limited attack success rate. In this study, to better study the jailbreak attack, we introduce the DSN (Don't Say No) attack, which prompts LLMs to not only generate affirmative responses but also novelly enhance the objective to suppress refusals. In addition, another challenge lies in jailbreak attacks is the evaluation, as it is difficult to directly and accurately assess the harmfulness of the attack. The existing evaluation such as refusal keyword matching has its own limitation as it reveals numerous false positive and false negative instances. To overcome this challenge, we propose an ensemble evaluation pipeline incorporating Natural Language Inference (NLI) contradiction assessment and two external LLM evaluators. Extensive experiments demonstrate the potency of the DSN and the effectiveness of ensemble evaluation compared to baseline methods.

Keywords: Large Language Models, Jailbreak attacks, Evaluation

Biography: Wenjie Wang is an assistant Professor (tenure-track) at School of Information Science and Technology in the ShanghaiTech University. Wenjie completed her Ph.D. in 2023 at Emory University (@Atlanta, US) under the supervision of Dr.Li Xiong. Before that, she received the Bachelor degree at Huazhong University of Science and Technology(@Wuhan, China) in 2017. Her major research interest include safety alignment and privacy preserving of large scale models, and LLM with social science such as LLM debias methods, especially gender debias.

ISBN: 978-1-0685851-7-3