Machine Learning and Knowledge Extraction, vol.7, no.3, pp.77-1-77-24, September, 2025.

Evaluating Prompt Injection Attacks with LSTM-Based Generative Adversarial Networks: A Lightweight Alternative to Large Language Models

S. Rashid; E. Bollis; L. Pellicer; D. Rabbani; R. Palacios Hielscher; A. Gupta; A. Gupta

Abstract-

Generative Adversarial Networks (GANs) using Long Short-Term Memory (LSTM) provide a computationally cheaper approach for text generation compared to large language models (LLMs). The low hardware barrier of training GANs poses a threat because it means more bad actors may use them to mass-produce prompt attack messages against LLM systems. Thus, to better understand the threat of GANs being used for prompt attack generation, we train two well-known GAN architectures, SeqGAN and RelGAN, on prompt attack messages. For each architecture, we evaluate generated prompt attack messages, comparing results with each other, with generated attacks from another computationally cheap approach, a 1-billion-parameter Llama 3.2 small language model (SLM), and with messages from the original dataset. This evaluation suggests that GAN architectures like SeqGAN and RelGAN have the potential to be used in conjunction with SLMs to readily generate malicious prompts that impose new threats against LLM-based systems such as chatbots. Analyzing the effectiveness of state-of-the-art defenses against prompt attacks, we also find that GAN-generated attacks can deceive most of these defenses with varying levels of success with the exception of Meta’s PromptGuard. Further, we suggest an improvement of prompt attack defenses based on the analysis of the language quality of the prompts, which we found to be the weakest point of GAN-generated messages.

** **
** **

Index Terms- AI Cybersecurity; adversarial prompts; large language models; Generative Adversarial Network

Citation:
Rashid, S.; Bollis, E.; Pellicer, L.; Rabbani, D.; Palacios, R.; Gupta, A.; Gupta, A. "Evaluating Prompt Injection Attacks with LSTM-Based Generative Adversarial Networks: A Lightweight Alternative to Large Language Models", Machine Learning and Knowledge Extraction, vol.7, no.3, pp.77-1-77-24, September, 2025.