New prompt-based technique to enhance AI security

June 24 2024

Algorithm 1 Word-level prompt-based attack (PAT)

Require: Input text (x, y), prompt function f_p, a pre-trained language model, target model f

Ensure: Adversarial example x_{adv} or FAILED

```
# Prompt Construction
x_p ← f_p(x, y)
# Candidate Generation
x_p ← Fill in blanks in x_p using the PLM
x_p ← Remove trigger text in x_p
y' ← f(x_p)
if y' ≠ y then
    x_adv ← x_p
    return x_adv
else
    return FAILED
end if
```

The diagram of the study's prompt-based attack approach (PAT). Credit: *Frontiers of Computer Science* (2023). DOI: 10.1007/s11704-023-2639-2

Researchers have developed a new approach to AI security that employs

text prompts to better protect AI systems from cyber threats. This method focuses on the creation of adversarial examples to prevent AI from being misled by inputs that are typically undetectable to humans.

The prompt-based technique streamlines the generation of these adversarial inputs, allowing for quicker response to potential threats without extensive computations. Preliminary testing has shown that this method can effectively safeguard AI responses with minimal direct interaction with the AI systems.

Dr. Feifei Ma, the lead researcher, outlines the process: "Our approach involved initially crafting malicious prompts to identify vulnerabilities in AI models. Following this identification, these prompts were utilized as training data, helping the AI to resist similar attacks in the future."

Subsequent experiments indicated that this training approach improved the robustness of AI systems. Models trained with adversarial prompts were less likely to succumb to similar attacks, demonstrating an enhancement in their defensive capabilities.

"This method allows us to expose and then mitigate vulnerabilities in AI models, which is especially critical in sectors like finance and health care," Dr. Ma noted.

The research, published in *Frontiers of Computer Science*, indicates that AI systems trained with these adversarial prompts are more capable of resisting similar manipulation tactics in the future, potentially improving their overall robustness against cyber threats.

It is a collaborative work between Chinese Academy of Sciences, University of Chinese Academy of Sciences, Stanford University, and National University of Singapore.

More information: Yuting Yang et al, A prompt-based approach to adversarial example generation and robustness enhancement, *Frontiers of Computer Science* (2023). DOI: 10.1007/s11704-023-2639-2

Provided by Higher Education Press