LLM 기반 사이버 공격-방어 시뮬레이션 프레임워크 개념 설계

이성호, 이정식, 한인성
국방과학연구소
sungho_lee@add.re.kr, godsider@add.re.kr, insung-han@add.re.kr

A Conceptual Design of an LLM-Based Cyber Attack-Defense Simulation Framework

Lee Sungho, Lee Jung Sik, Han Insung
Agency for Defense Development

요 약

최근 사이버 공격의 정교화와 다변화에 따라, 기존의 정적 보안 시스템은 실시간 대응과 지속적인 위협 탐지에 한계를 보이고 있다. 본 논문에서는 이러한 문제를 해결하기 위한 새로운 접근 방식으로, 두 개의 사전 학습된 대형 언어 모델(LLM)이 공격자와 방어자 역할을 수행하며 상호작용을 통해 학습하는 대화형 시뮬레이션 프레임워크를 제안한다. 공격형 모델은 네트워크 환경의 취약점을 분석하고 공격 시나리오를 생성하며, 방어형 모델은 이에 대응하는 방어 전략을 수립한다. 각 시나리오의 결과는 강화학습 기반 피드백 과정을 통해 양측 모델에 전략적 역량을 점진적으로 향상시킨다. 본 연구는 개념 설계에 초점을 맞추어 프레임워크의 구조와 학습 메커니즘을 구체화하였으며, 향후 실제 구현과 시뮬레이션 실험을 통해 성능을 정량적으로 검증할 계획이다.

Ⅰ. 서 론

현대의 사이버 보안 환경은 점점 더 지능화되고 있으며, 기존의 정적인 룰 기반 보안 시스템만으로는 실시간으로 진화하는 다양한 위협에 대응하는 데 한계가 있다. 이에 따라 인공지능, 특히 대형 언어 모델(LLM; Large Language Model)을 활용한 지능형 보안 대응에 대한 관심이 높아지고 있다. 본 논문에서는 두 개의 사전 학습된 LLM이 각각 사이버 공격자와 방어자의 역할을 맡아 서로 대화하고 대응하는 과정을 통해 상호 학습하는 새로운 개념의 시뮬레이션 프레임워크를 제안한다. 이 프레임워크에서 Red 모델은(공격자)는 가상의 네트워크 환경에서 가능한 취약점을 탐색하고 공격 시나리오를 생성하며, 이에 대해 Blue 모델은(방어자) 실시간으로 대응 전략을 제시하고 방어 기법을 발전시켜 나간다. 이러한 양방향 상호작용 기반 학습 구조는 사이버 보안 분야에서의 LLM 활용 가능성을 넓히고, 더욱 정교하고 적응적인 자동화 보안 시스템의 구현에 기여할 수 있을 것으로 기대된다.

Ⅱ. 연구 배경

최근 사이버 보안 위협은 정교하고 다양한 형태로 진화하고 있으며, 이러한 위협에 효과적으로 대응하기 위해서는 고정된 룰 기반 보안 시스템을 넘어서는 지능적이고 적응 가능한 방어 체계가 요구된다. 특히, 대형 언어 모델(LLM)을 기반으로 하는 보안 응용 기술은 자연어를 이해하고 생성할 수 있는 특성을 바탕으로, 위협 탐지·분석·대응 전 과정에서의 자동화 가능성을 제시하며 많은 주목을 받고 있다. 이러한 흐름 속에서 LLM 간의 상호작용을 통해 지속적 학습과 시뮬레이션이 가능한 보안 프레임워크를 구축하려는 시도들이 나타나고 있으며, 이는 기존의 정적인 학습 구조를 넘어서는 새로운 방향성을 제시하고 있다.

Ferrag et al.[1] 연구진은 LLM을 활용한 사이버 보안의 다양한 응용 사례를 분석하며, 특히 LLM이 DDoS, 악성코드 생성 등의 악의적 명령어를 공격자의 입장에서 학습 및 생성할 수 있는지에 대한 가능성과 이에 대응하기 위한 프레임워크 필요성을 강조하였다. 이 논문은 총 42개의 LLM 모델을 비교 분석하며, 보안 분야에서 LLM의 공격/방어 양면성에 대한 체계적인 분류를 시도하였다. 해당 연구는 LLM 기반 보안 연구의 포괄적 이해를 제공하며, 향후 상호작용형 LLM 프레임워크의 필요성을 뒷받침하는 기반이 된다.

Shah & Deshpande[2]은 LLM을 활용하여 웹 보안 환경에서 발생 가능한 데이터 오염(data poisoning) 및 모델 조작(model manipulation) 공격을 분석하고, 이를 사전 예측하고 방어할 수 있는 프레임워크를 제안하였다. 이들은 LLM 기반 공격 탐지 시스템이 단순 탐지기를 넘어서, 공격 시뮬레이션 및 방어 전략 설계에 있어 대화형 상호작용 구조가 유용함을 시사하였다. 이러한 관점은 본 논문이 제안하는 LLM 상호 학습 기반 프레임워크의 필요성과 연관된다.

Ⅲ. 제안하는 프레임워크 개요 및 구조

본 연구에서 제안하는 프레임워크는 두 개의 사전 학습된 대형 언어 모델(Large Language Model, 이하 LLM)이 각각 사이버 공격자 및 방어자의 역할을 수행하며 상호작용하는 대화형 시뮬레이션 구조로 설계되었다. Red 모델은 공격자의 역할로서 네트워크 환경에서 존재할 수 있는 취약점을 분석하고, 그에 기반한 공격 시나리오를 생성한다. 반면 Blue 모델은 방어자의 입장에서 해당 시나리오에 대응하기 위한 탐지 및 방어 전략을 설계한다. 이와 같은 상호작용은 반복적으로 수행되며, 각 에피소드의 결과는 양측 모델의 성능 개선을 위한 학습 피드백으로 활용된다. 프레임워크의 핵심은 단순한 응답 생성이 아닌, 지속적인 상호 학습 루프를 통해 LLM이 실제 환경에 근접한 형태의 사이버 보안 역량을

점진적으로 획득할 수 있도록 설계된다는 점에 있다.

프레임워크는 그림 1과 같이 네트워크 시뮬레이션 환경, Red 모델, Blue 모델 세 가지로 구성된다. 네트워크 환경은 Mininet, NS-3 등의 시뮬레이터를 기반으로 구축되며, MITRE ATT&CK 모델에 기반한 실제 위협 데이터를 반영한다. Red 모델은 사전에 APT 공격 보고서 데이터셋으로 파인 튜닝된 모델로 취약점 탐색 및 공격 벡터 생성 기능을 수행하고 MITRE ATT&CK Evaluations[3] 에서 정의하는 APT 공격 시나리오의 흐름을 따라 시뮬레이션 환경에 맞게 공격 방식을 제안한다. Blue 모델은 Red 모델에서 제안한 공격 시나리오에 대해 분석 후, 이에 대한 방어 전략을 생성하며, 전략은 탐지 로직 개선, 시그니처 업데이트, 룰 기반 방어 설정의 형태로 구성된다. 이러한 공격 및 방어 전략은 시뮬레이션 환경에 실제 공격 모사, 보안 정책 생성 및 방화벽 룰 변환 등 실행 가능한 출력으로 반응된다.

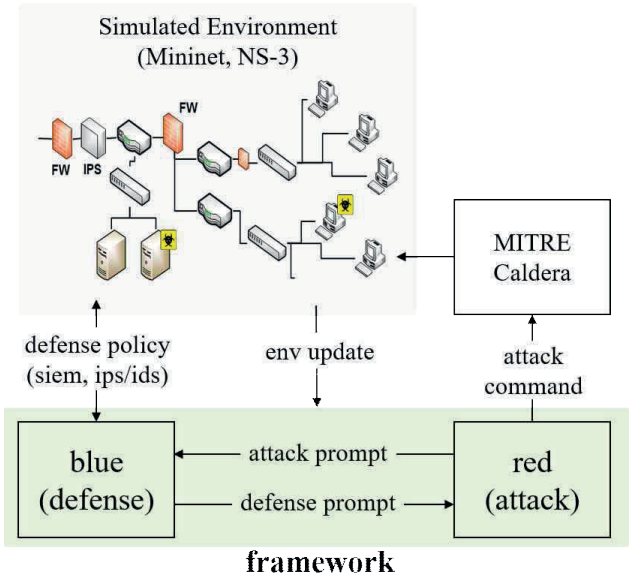

그림 1. 사이버 공격-방어 프레임워크 설계

LLM 간의 상호작용은 지속적인 학습 루프를 통해 각 모델의 역량을 강화한다. Red 모델은 프롬프트와 환경 데이터를 바탕으로 공격 전략을 설계하고 실행한다. Blue 모델은 해당 공격에서 나타나는 취약점을 예측하고 이를 통해 대응책을 제시한다. 이후 시나리오 실행 결과에 따라 공격 성공 여부 및 탐지 정확도를 평가하고, 평가 결과를 기반으로 각각의 모델에 대해 강화가 이루어진다. 공격 실패 시 Red 모델에서는 실패한 공격 벡터와 유사한 취약점 기반으로 재탐색을 진행한다. Blue 모델은 탐지 실패 시 방어 로직을 재구성하거나 시그니처를 추가적으로 학습한다. 이러한 과정은 강화학습 기반으로 구성되고, 보상 함수로 Red 모델은 공격 성공률 또는 탐지 회피에 성공했을 경우 점수를 얻는다. Blue 모델은 탐지에 성공했을 때 탐지의 정확도에 따라 점수를 얻는 방식으로 구성된다. 또한, 모델 간 상호작용 기록은 로그로 저장되며, 이를 기반으로 후속 학습 혹은 Human Feedback 방식의 보강 학습도 가능하다.

그림 2는 본문에서 제안한 프레임워크에서 Red 모델과 Blue 모델이 반복적인 학습을 통해 어떻게 전략을 발전시키는지를 시각적으로 나타낸 것이다. 위에서부터 아래로 Red 모델이 초기 공격 전략을 생성하고, Blue 모델은 이를 탐지하거나 방어하는 초기 전략을 시도한다. 이후 공격/방어 성공 여부 및 환경 상태를 기반으로 평가를 진행하고 각 모델에 보상을 준다. 각 모델은 보상과 현재 환경의 상태를 참고해 각자의 모델을 최적화한다.

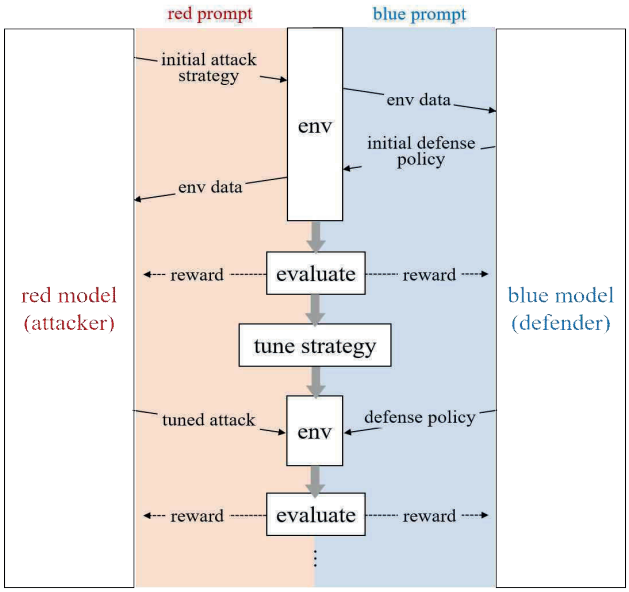

그림 2. 모델 간 학습 흐름

Red와 Blue 모델은 각각 강화학습 보상 함수를 기반으로 학습하며, 각자의 성공 또는 실패가 상대 모델의 학습을 위한 자극으로 작용한다. 공격자가 탐지 회피에 성공할 경우, 방어자는 그 시나리오를 기반으로 탐지 모델을 보완하고 반대로 방어에 성공한 경우 공격자는 보다 정교한 공격 벡터를 학습하게 된다. 이러한 상호 관계를 기반으로한 학습 구조는 자기지도 학습(self-supervised learning)의 대화형 확장 형태로 최종적으로는 Blue 모델의 전략적 지능 강화를 목표로한다.

Ⅳ. 결론 및 향후 연구

본 논문에서는 사이버 보안 분야에서의 대형 언어 모델(LLM)의 활용 가능성을 탐색하기 위해, 공격자와 방어자 역할을 수행하는 두 개의 LLM이 상호작용하며 반복적으로 학습하는 개념적 시뮬레이션 프레임워크를 제안하였다. 제안된 구조는 동적인 위협 탐지 및 대응 능력을 갖춘 자동화 보안 시스템의 실현 가능성을 제시하며, 대화형 상호학습이라는 새로운 접근 방식을 통해 LLM의 적용 범위를 확장할 수 있음을 보여준다.

본 논문에서는 개념적인 설계에 중점을 두었으며, 향후 연구에서는 실제 LLM 모델을 기반으로 한 프로토타입 구현과 시뮬레이션 실험을 통해 성능을 검증하고, 다양한 공격 시나리오와 방어 전략에 대한 효과성을 정량적으로 평가할 예정이다.

참 고 문 헌

[1] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, “Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities,” Elsevier, Amsterdam, Netherlands, pp. 1–25, 2025.

[2] S. P. Shah, A. V. Deshpande, “Addressing Data Poisoning and Model Manipulation Risks using LLM Models in Web Security,” IEEE Press, Piscataway, NJ, USA, pp. 100–110, 2024.

[3] MITRE Corporation, “MITRE ATT&CK Evaluations”, MITRE, https://evals.mitre.org/, accessed May 6, 2025.