스마트 빌딩 냉난방공조 시스템에서 LLM 기반 네트워크 침입 탐지용 자동 프롬프트 생성 방법 설계

Design of an Automatic Prompt Generation Method for LLM-based Network Intrusion Detection in HVAC Systems of Smart buildings

요 약

폐쇄 환경에서만 운영되던 빌딩 자동화 시스템이 IT와 OT 기술의 융합으로 개방 환경으로 진화함에 따라 스마트 빌딩 시스템에 대한 사이버 보안을 고려하는 것이 필요하다. 본 논문에서는, 스마트 빌딩 냉난방공조(HVAC) 시스템을 위한 네트워크 기반 침입 탐지 기법을 대규모 언어 모델(LLM) 기반으로 구현할 때 프롬프트를 자동으로 생성하는 방법을 제안한다. 제안 방법은, 스마트 빌딩의 제어 네트워크에서 사용되는 Modbus-TCP 트래픽을 수집하고, 이를 기반으로 PLC(Programmable Logic Controller)와 HMI(Human Machine Interface)의 행위 문맥을 생성하여, MITRE ATT&CK(Adversarial Tactics, Technique, and Common Knowledge) 공격 기법 탐지 질의문과 결합하여 자연어 질의 형태의 프롬프트를 생성한다. 제안 기법을 검증하기 위해, 스마트 빌딩 HVAC 시스템의 정상 및 공격 행위를 포함하는 학습 데이터로 파인 튜닝된 라마 3.1을 LLM 기반 탐지 엔진으로 구현하여 실험하였다. 실험 결과, 네트워크 트래픽 기반 탐지가 가능한 6개의 MITRE ATT&CK 공격 기법을 탐지하고 사용된 공격 기법을 식별할 수 있었다.

1. 서 론

스마트 빌딩은 냉난방공조(HVAC), 조명, 차양 등의 하위 시스템으로 구성되며, 이를 제어하는 빌딩 자동화 시스템(Building Automation System, BAS)은 원래 폐쇄형 구조였으나 원격 통신과 클라우드 기반 제어 기술의 확산으로 개방형 구조로 전환되었다[1, 2]. 이에, BAS와 HVAC 시스템은 외부 네트워크와의 연결성이 증가하였고, 더불어 사이버 공격에 더 취약해졌다.

이러한 변화에 따라 BAS 대상 침입 탐지 시스템(Intrusion Detection System, IDS)에 대한 요구가 증가하고 있다. 기존 규칙 기반 침입 탐지는 알려진 공격에 대해 높은 탐지 정확도와 설명 가능성이 있지만, 시그니처를 지속적으로 업데이트해야 하고 유지보수가 복잡하다는 단점이 있다. 반면, 데이터 기반 탐지는 트래픽의 동적 특성을 반영한 대응이 가능하지만, 탐지 결과에 대한 설명이 어렵고 오탐 및 미탐의 위험이 존재한다.

침입 탐지 엔진으로 대규모 언어 모델(LLM)을 적용하면, 탐지 결과에 대한 설명이 가능하고, 네트워크 환경의 변화에 유연하게 대응할 수 있다. 그러나 침입 탐지 엔진을 LLM 기반으로 구현하기 위해서는 네트워크 기반의 프롬프트 생성이 핵심 요소로 작용하지만, 현재까지 이를 위한 효과적인 방법론은 존재하지 않는다. 본 논문에서는 스마트 빌딩의 제어 네트워크 환경에서 LLM 기반 NIDS(Network-based Intrusion Detection System)를 구현하는데 필요한 프롬프트 생성 기법을 제안한다. 구체적으로 스마트 빌딩 HVAC 네트워크로부터 수집한 트래픽 정보로부터 MITRE ATT&CK 공격 기법 탐지를 위한 프롬프트 생성 기법을 개발한다.

2. 관련 연구

Adjewa 등[3]은 네트워크 트래픽에서 새롭게 등장하는 공격을 탐지하고 분류하기 위해 LLM기반의 네트워크 침입 탐지 프레임워크를 제안하였다. 이 프레임워크는 악성 여부를 판단하는 탐지 모델과, 탐지된 트래픽의 공격 유형을 분류하는 식별 모델로 구성되며, 경량화된 BERT 모델을 파인튜닝하여 사용하였다.

Fu 등[4]은 차량 인터넷(IoV)을 위한 BERT 기반 침입 탐지 시스템을 제안하였다. 제안된 시스템은 기존 머신러닝 및 딥러닝 기반 방법과 달리, 차량 내부와 외부 네트워크 침입을 동시에 탐지할 수 있으며, 양방향 인코더 구조를 활용해 향상된 탐지 성능을 보였다.

Zhang 등[5]은 별도의 추가 학습 없이 프롬프트 기반의 in-context learning을 통해 무선 통신 영역에서 LLM이 자동으로 침입 탐지를 수행할 수 있음을 보여주었다. 이는 침입 탐지 시스템에 있어 프롬프트 생성 기법의 중요성과 확장 가능성을 시사한다.

Lin 등[6]은 클라우드 서비스의 수천만 개의 명령줄을 학습한 LLM을 활용하여 대규모 침입 탐지 시스템을 구현하고, 자기 지도 학습을 통해 시스템의 탐지 성능을 향상시켰다.

본 논문은 스마트빌딩 네트워크 패킷 데이터를 자연어 형태의 자산 행위 문맥으로 변환하고, 이를 바탕으로 MITRE ATT&CK 공격 기법 탐지용 질문과 결합한 프롬프트를 생성하여 파인튜닝된 모델에 질의하는 방식에서 기존 연구와 차별화된다.

3. LLM 기반 네트워크 침입 탐지 시스템

스마트 빌딩의 HVAC 제어 네트워크를 위한 LLM 기반 NIDS는, 네트워크 패킷 파서, 데이터 스토리지, 프롬프트 생성기, 파인튜닝 된 LLM으로 구성된다(그림 1 참조). 네트워크 트래픽 데이터를 수집한 후 전처리하여 스토리지에 저장하고, 이를 바탕으로 프롬프트를 생성하여 정상 및 공격 행위에 대해 학습한 파인튜닝 된 LLM에게 전달한다. LLM은 프롬프트를 입력받아 추론을

통해 탐지 결과를 응답한다.

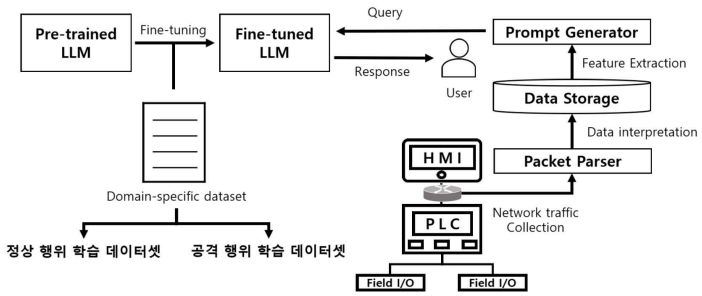

그림 1. LLM 기반 NIDS

4. 프롬프트 생성

프롬프트 생성 과정은 그림 2와 같다. 사전에 MITRE ATT&CK for ICS(Industrial Control System)에 포함한 공격 탐지에 사용할 질문을 생성한다. 실시간으로 수집한 Modbus-TCP 패킷을 분할하여 해석하고 데이터 스토리지에 저장한다. 이후, MITRE ATT&CK for ICS 공격 기법을 분석하여 지정된 패킷의 필드(특성 정보)를 데이터 스토리지에서 추출하고 이를 가공하여 자연어 질의 형태의 프롬프트를 생성한다.

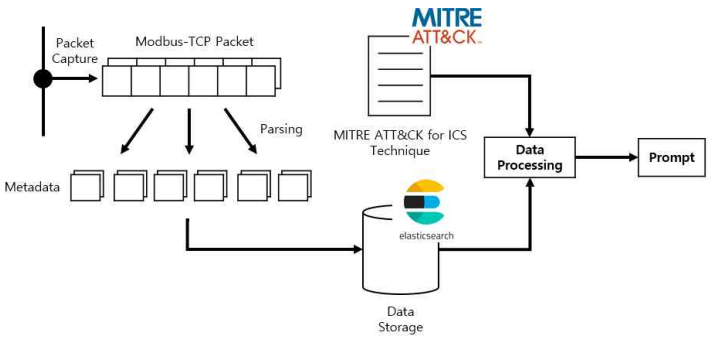

그림 2. 프롬프트 생성 과정

4.1 네트워크 탐지 가능한 공격 기법 식별 및 질의문 개발

네트워크 환경에서 탐지 가능한 MITRE ATT&CK for ICS 공격 기법을 식별하고, 해당 공격 탐지에 사용할 질문을 생성한다. HVAC 시스템에서 공격 대상 자산인 HMI와 PLC가 침해되었을 때, 네트워크 수준에서 탐지 가능한 공격 기법과 이들의 탐지에 사용할 질문을 표 1과 같이 작성하였다.

표 1. 탐지 가능한 공격 기법 및 탐지 질문

Technique name (Technique ID)	Question
Adversary-in-the-Middle (T0830)	Is there a value that falls outside the normal range, a response from a device other than expected, or more than one response to a single request?
Denial of Service (T0814)	Did a large number of packets occur in a short period of time?
Exploitation of Remote Services (T0866)	Check whether there is any packet using function code 43/13, and if found, determine whether it is normal for the source device of that packet to perform that function code operation.
Modify Parameter (T0836)	Identify any communications that involve write-related function codes (6 or 10), and determine whether the source device is unauthorized to use these function codes or if any of the written values fall outside the acceptable range.
Spoof Reporting Message (T0856)	Does a single device respond twice to a read request for the same variable within a very short time interval with different values? If so, would such redundant and inconsistent responses be considered unexpected behavior?
Unauthorized Command Message (T0855)	Identify any communications that involve write-related function codes (6 or 10), and verify whether the source device is authorized to use those function codes under normal operation.

4.2 프롬프트 생성을 위한 전처리 작업

수집한 Modbus-TCP 패킷은 전처리 과정을 거쳐 데이터 스토리지에 저장된다. 데이터 스토리지로는 elasticsearch를 사용하며, 패킷은 필드별로 분할하여 저장된다. 타임스탬프, IP, 데이터 필드 등의 필드 값은 사람이 이해하기 쉬운 자연어 형태로 표현하기 위해 각각 UTC 시간, 장치명, 완전히 해석된 데이터 값 등으로 전처리된 후 그림 3과 같이 저장된다.


```
"_index": "modbus250407",
"_id": "qVmnDZYBvUgt9JHKdx82",
"_score": 1,
"_source": {
  "timestamp": "2025-04-07T00:30:00.058099",
  "source_device": "plc1",
  "destination_device": "hmi1",
  "source_port": 502,
  "destination_port": 49868,
  "transaction_id": 3233,
  "protocol_id": 0,
  "length": 7,
  "unit_id": 1,
  "function_code": 3,
  "data": {
    "type": "response",
    "tag_name": "cooling_water_supply_value",
    "value": 28
  }
}
```


그림 3. Elasticsearch에 저장되는 패킷 예시

4.3 프롬프트 개발

네트워크 기반 침입 탐지를 위해 트래픽 관찰점에서 트래픽을 수집할 때마다 프롬프트를 생성하여 해당 트래픽의 이상 여부를 판단하는 단일 트래픽 기반의 탐지 방식은 네트워크 패킷의 데이터 필드의 비정상 유무 정도만 판단 가능하고, 기능 코드를 이용한 공격이나 서비스 거부 공격 같은 복수 트래픽을 통해 탐지해야 하는 공격을 탐지할 수 없다. 따라서 일정 주기마다 모여진 복수의 트래픽을 기반으로 프롬프트를 생성하여 비정상 여부를 판단한다.

4.3.1 주요 자산의 행위 문맥 생성

일정 시간 프레임 동안 HMI와 PLC의 네트워크 행위를 표현하는 행위 문맥(behavior context)을 생성한다. 이 행위 문맥을 LLM에 제공하여, 사전에 학습된 정상 및 공격 행위에 대한 지식을 바탕으로 이상 유무를 판단하도록 한다. 기능 코드의 정상 여부는 과거의 문맥과 비교해서 파악할 수 있고, 패킷의 전송 주기가 적절한지도 모델 질의를 통해 판단할 수 있다. 다음은 HVAC 가상 시뮬레이터에서 PLC와 HMI가 약 0.013초 동안 발생한 패킷 200개에 대한 행위 문맥 사례이다.

(1) At **00:31:00.026979**, HMI1 send a function code 3 request to PLC1 to read cooling_water_supply_value.

(중략)

(200) At **00:31:00.038917**, PLC2 respond to HMI1 with function code 3, value 3.0 for heating_valve_mode.

4.3.2 프롬프트 템플릿 설계

프롬프트에 포함될 내용이 그림 4에 나타나 있다. 프롬프트 템플릿(template)은 역할 부여, 질문, 행위 문맥의 3가지 항목으로 구성된다. 역할 부여 항목에는 '네트워크 기반 침입 탐지 시스템' 이라는 역할을, 질문 항목에는 PLC와 HMI1의 행위 문맥을 바탕으로 6가지 공격 기법 발생 여부를 확인하는 질문이 포함된다. 행위 문맥 항목은 공격 기법이 발생 여부를 판단한다.

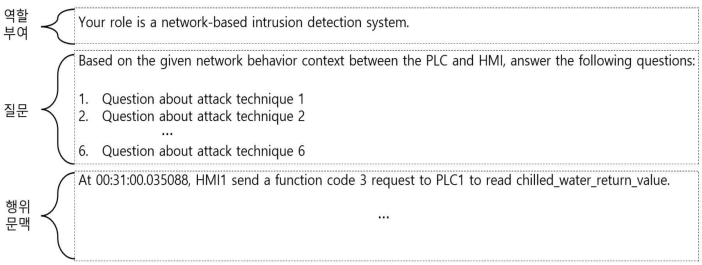

그림 4. 프롬프트 템플릿

5. 실험 및 결과

프롬프트의 유용성을 검증하기 위해, HVAC 시스템의 정상 및 공격 행위에 대한 학습 데이터로 라마 3.1을 파인튜닝하였다. 정상 행위 데이터는 ICSSIM[7]을 변형하여 개발한 HVAC 가상 시뮬레이터(IBEEMS SIM)의 제어 네트워크 동작을 기반으로 수집하였으며, 시스템 환경 구성과 네트워크 트래픽의 특성 정보(IP 주소, MAC 주소, 포트 번호, 기능 코드 등)를 포함한다. 공격 행위에 대한 학습 데이터는 MITRE ATT&CK for ICS의 공격 전술, 기법, 설명을 바탕으로 생성되었다.

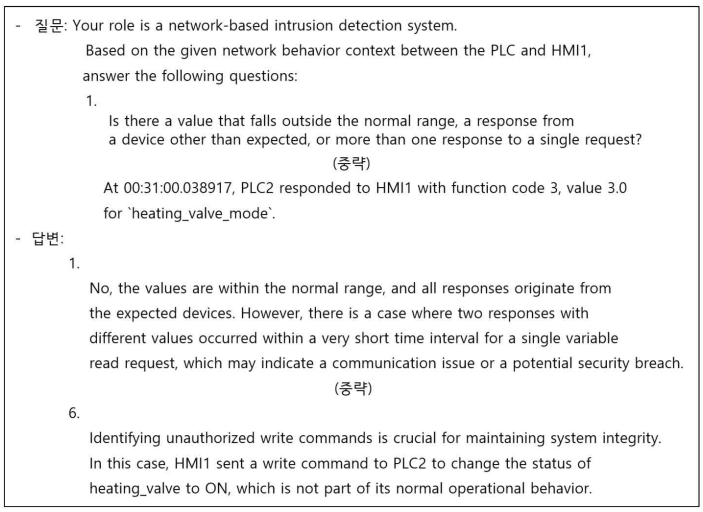

그림 5. 모델 질의응답 사례

그림 5는 파인튜닝된 LLM 모델에 질의하여 도출한 응답 사례이다. 제시된 2개의 사례에서 Adversary-in-the-Middle 공격과 Unauthorized Command Message 공격을 정확히 탐지하고, 탐지 근거를 충실히 설명하고 있음을 확인할 수 있다.

6. 결론 및 향후연구

본 연구는 스마트 빌딩 HVAC 네트워크 환경에서 MITRE ATT&CK 공격 기법 탐지를 위한 질의 프롬프트를 네트워크 트래픽으로부터 자동 생성하는 방법을 제안하였다. 생성된 프롬프트를 파인튜닝된 라마 3.1에 입력한 결과, 네트워크 트래픽 기반 탐지가 가능한 6개의 공격 기법을 정확히 탐지하고, 탐지된 공격 기법에 대한 설명을 제공함으로써 해석 가능한 탐지 기술 개발의 가능성을 보였다.

향후 연구에서는 PLC와 HMI 외에도 다양한 자산들을 탐지 대상에 포함시켜 복잡한 환경에서도 침입 탐지가 가능하도록 할 예정이다. 또한, Modbus-TCP 외의 산업용 이더넷 통신 프로토콜도 분석 대상으로 추가하여 실제 환경에 더 근접한 프롬프트 생성 프로세스를 구현할 계획이다.

참고 문헌

[1] Ciholas, Pierre, et al., "The security of smart buildings: a systematic literature review," arXiv preprint arXiv:1901.05837, 2019.

[2] Li, Guowen, et al., "A critical review of cyber-physical security for building automation systems," Annual Reviews in Control 55, pp. 237-254, 2023.

[3] Adjewa, Frederic, Moez Esseghir, and Leila Merghem-Boulahia., "LLM-based Continuous Intrusion Detection Framework for Next-Gen Networks," arXiv preprint arXiv:2411.03354, 2024.

[4] Fu, Mengyi, et al., "IoV-BERT-IDS: Hybrid network intrusion detection system in IoV using large language models," IEEE Transactions on Vehicular Technology , 2024.

[5] Zhang, Han, et al., "Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection," arXiv preprint arXiv:2405.11002, 2024.

[6] J. Lin, Y. Guo et al., "Intrusion Detection at Scale with the Assistance of a Command-line Language Model," 2024 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S), Brisbane, Australia, pp. 73-79, 2024. doi: 10.1109/DSN-S60304.2024.00031

[7] Dehlaghi-Ghadim, Alireza, et al., "ICSSIM—a framework for building industrial control systems security testbeds," Computers in Industry 148, 103906, 2023.