

대규모 언어 모델과 검색 증강 생성 기법을 사용한 스마트 빌딩용 침입 탐지 시스템*

김도익⁰¹ 안석현² 박수현³ 조성제³ 김홍근⁴

단국대학교 모바일시스템공학과¹

단국대학교 인공지능융합학과²

단국대학교 소프트웨어학과³

동국대학교 국제정보보호대학원 정보보호학과⁴

{doik0920, seokhyun, parksh, sjcho}@dankook.ac.kr, hgkim4044@gmail.com

An Intrusion Detection System for Smart Buildings Using Large Language Models and Retrieval-Augmented Generation

DoIk Kim⁰¹ SeokHyun Ann² SuHyeon Park³ SeongJe Cho³ HongGeun Kim⁴

¹Dept. of Mobile Systems Engineering, Dankook University

²Dept. of AI-based Convergence, Dankook University

³Dept. of Software Science, Dankook University

⁴Graduate School of International Affairs & Information Security, Dongguk University

요약

스마트 빌딩 자동화 시스템은 편리성과 에너지 효율성을 제공하지만, 네트워크 기반 제어 특성상 다양한 사이버 위협에 노출될 수 있다. 본 논문에서는, 대규모 언어 모델(Large Language Models, LLMs)과 검색 증강 생성(Retrieval-Augmented Generation, RAG) 기반으로, 스마트 빌딩의 냉난방공조(HVAC) 시스템에 대한 사이버 공격을 탐지하는 기법을 제안한다. 제안하는 침입탐지는 도메인 지식 기반 질문-답변 데이터셋을 활용하여 정상 및 비정상 행위를 탐지한다. 정상 행위 데이터는 HVAC 가상 시뮬레이터인 IBEEMS SIM을 기반으로 수집하였으며, 공격 행위는 MITRE ATT&CK for ICS 프레임워크를 바탕으로 구축하였다. 또한, 다양한 벡터 검색 엔진과 임베딩 모델, 온프레미스로 구동되는 LLM들을 비교·평가하여 최적의 침입탐지 기법을 개발하고자 노력하였다. 최종적으로 제안한 RAG 기반 스마트 빌딩 침입 탐지 시스템은, 도메인 지식이 반영된 검색 및 응답을 통해 정상 여부를 판별하며, 설명 가능한 침입탐지 기법 개발의 가능성을 보여 준다.

1. 서론

스마트 빌딩은 다양한 센서와 제어 장치를 포함한 빌딩 자동화 시스템(Building Automation System, BAS)을 기반으로, 에너지 효율성과 사용자 편의성을 극대화한다. 특히 냉난방공조(HVAC) 시스템은 건물 내 환경 제어의 핵심 요소로, PLC(Programmable Logic Controller)와 HMI(Human Machine Interface) 간의 네트워크 통신을 통해 제어된다. 그러나 이러한 시스템은 편의성과 효율성에 중점을 두고 설계되었기 때문에, 보안 측면에서는 상대적으로 취약해 사이버 공격의 주요 표적이 될 수 있다[1]. 이에 따라 스마트 빌딩 환경에 적합한 침입 탐지 시스템(IDS, Intrusion Detection System)의 필요성이 대두되고 있다.

* 본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제임 (No. RS-2021-KP002461). 또한 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학사연구사업 ICT 핵심인재양성사업의 연구결과로 수행되었음 (IITP-2023-00259867). 또한 본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음 (No. 2024-0-00035).

기존의 침입 탐지 시스템은 오용 탐지 또는 이상 탐지 기반 방식으로 동작한다. 오용 탐지 방식은 알려진 공격 패턴을 식별하는 데는 효과적이거나, 새로운 유형의 공격에는 취약하며[2], 이상 탐지 기반 방법은 높은 오탐률로 인해 실질적인 보안 대응에 한계가 있다[3]. 특히, 스마트 빌딩의 HVAC와 같은 특수 도메인에서는 해당 도메인에 대한 깊은 이해와, 탐지 결과에 대한 설명력을 갖춘 접근이 필수적이다.

본 논문에서는 기존 연구의 한계를 극복하기 위해, RAG(Retrieval-Augmented Generation)와 대규모 언어 모델(Large Language Model, LLM)을 결합한 침입탐지 기법을 스마트 빌딩 HVAC 제어 환경에 구현한다. LLM은 대량의 데이터를 학습하여 언어의 문맥을 심층적으로 이해하고, 텍스트를 생성하는 데 특화된 모델이다[4]. 그러나 이러한 모델은 훈련 데이터에 포함되지 않은 정보에 대해서는 정확한 응답을 제공하는 데 한계가 있다. 이를 보완하기 위해 제안된 RAG 구조는 외부 지식 기반에서 관련 정보를 검색한 후, 해당 문서를 바탕으로 LLM이 최종 응답을 생성하는 방식으로 동작한다. 이로써 문맥에 적합하면서 보다 정확한 응답 생성이 가능해진다[5].

본 연구는 이러한 RAG의 강점을 스마트 빌딩 HVAC 제어 시스템

의 보안 강화에 적용하여, 네트워크 기반 제어 행위의 정상 여부를 도메인 지식에 기반해 판단하고, 설명 가능한 침입 탐지 시스템의 실현 가능성을 제시한다.

2. 제안 방법

RAG 기반 침입 탐지 시스템은 그림 1과 같이 문서 검색기, 지식 베이스, 사전 학습된 LLM으로 구성된다. HVAC 시스템 내에서 발생하는 네트워크 행위가 질의 형태로 입력되면, 문서 검색기가 지식 베이스에서 유사한 문서를 검색하여 해당 질의와 함께 LLM에 입력한다.

LLM은 검색된 문서를 바탕으로 도메인 지식을 반영해 해당 행위의 정상 여부를 판단하며, 그 판단의 근거를 자연어로 설명함으로써 사용자가 결과를 직관적으로 이해할 수 있도록 지원한다.

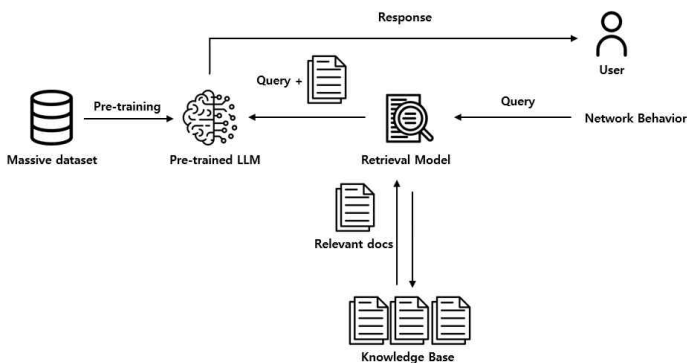


그림 1. RAG 기반 침입탐지 기법의 구성

3. RAG 기반 침입탐지 시스템 개발

3.1 데이터셋

실험 데이터셋은 정상 행위와 공격 행위로 구성된 질문-답변 형식으로 구축되었다. ICSSIM[6]을 변형하여 본 연구진이 직접 개발한 HVAC 가상 시뮬레이터(IBEEMS SIM)의 제어 네트워크 동작을 기반으로 정상 데이터를 수집하였고, 공격 행위 데이터는 MITRE ATT&CK for ICS (Adversarial Tactics, Techniques and Common Knowledge for Industrial Control System) 프레임워크의 네트워크 기반 공격 시나리오를 바탕으로 구성했다.

3.2 데이터 전처리

본 연구에서는 질문-답변 형식의 데이터를 효과적으로 벡터화하기 위해, 문서를 적재한 후 줄 단위로 내용을 분할하고, 질문이 시작되는 부분을 기준으로 하나의 질문-답변 단위를 구분하였다. 각 질문과 그에 대응하는 답변은 하나의 완결된 의미 단위로 간주되며, 이들을 청크 단위로 구성하여 Document 객체로 저장하였다.

이렇게 전처리된 문서는 벡터 임베딩 및 검색의 기반이 되며, 관련성 높은 질문-답변 쌍의 의미적 일관성을 유지하고, 대규모 언어 모델이 보다 정확하고 문맥에 맞는 응답을 생성할 수 있도록 한다.

3.3 임베딩 모델

본 연구에서는 정확도와 연산 효율성의 균형을 고려하여 e5-base-v2 모델을 채택하였다. 해당 모델은 적은 연산으로도 우수한 정보 검색(Retrieval) 성능을 발휘하여[7], RAG 기반 침입 탐지 시스템에 적합하다.

스마트 빌딩 HVAC 환경 기반의 질문-답변 데이터셋을 활용하여

코사인 유사도 기준 상위 5개의 문서를 검색한 결과, e5-base-v2 모델은 평균 코사인 유사도 0.8504를 기록하며 우수한 성능을 보였다.

3.4 벡터 검색 엔진

문서 검색기는 벡터 검색 엔진으로 구현되어, 임베딩된 질문 벡터와 데이터베이스 내 문서 벡터 간의 유사도 검색을 수행한다. 실험에서는 벡터 검색 엔진의 성능을 비교하기 위해, HVAC 제어 환경에서 발생할 수 있는 네트워크 행위를 기반으로 생성된 100개의 정상/공격 프롬프트를 쿼리로 사용하였다. 각 쿼리에 대해 상위 5개의 유사 문서를 검색하였으며, 모든 벡터는 코사인 유사도 기반의 HNSW 인덱스로 구축하였다.

검색 성능은 단일 CPU 환경에서 순차적으로 검색 요청을 처리하며, 각 벡터 검색 엔진의 평균 지연 시간(ms)을 측정하여 비교하였다. 실험 결과, 표 1과 같이 FAISS는 지연 시간 면에서 가장 좋은 성능을 보였으며 또한, 전문가 평가에서도 높은 관련성을 보여 최종 벡터 검색 엔진으로 선정하였다.

표 1. 벡터 검색 엔진별 평균 검색 지연 시간

유형	검색 엔진	평균 지연 시간(ms)
벡터 검색 라이브러리	FAISS	0.1166
벡터 데이터베이스	Chroma	4.2811
	Elasticsearch	9.3435
	Redis	0.8245
	Weaviate	6.1109
	Milvus	3.9275
	Qdrant	2.6764

4. RAG 기반 침입 탐지 시스템의 언어 모델 성능 평가

사전 학습된 대규모 언어 모델은 네트워크 행위 기반 침입 탐지 시스템에서 보안상의 이유로 온프레미스 환경에서 동작해야 하며, 응답의 정확도 또한 중요하게 고려되어야 한다.

본 연구에서는 온프레미스 환경에서 활용 가능한 대규모 언어 모델인 LLaMA 3.1 8B, Mistral 7B, Gemma 7B, Falcon 7B를 대상으로 성능을 비교 평가하였다.

실험은 Ollama 환경에서 수행되었으며, HVAC 제어 환경에서 발생할 수 있는 네트워크 행위를 기반으로 생성된 100개의 정상/공격 프롬프트를 사용하여 평가하였다.

표 2와 표 3의 비교 결과, RAG 기반 침입 탐지 시스템은 RAG를 적용하지 않은 언어 모델 대비 최대 73%포인트(%p)의 precision 향상, 그리고 최대 78%포인트(%p)의 recall 향상(Mistral 7B 기준)을 보였다. 이를 통해, RAG 구조의 적용이 분류 성능 개선에 미치는 영향이 크다는 것을 알 수 있다.

또한, RAG 기반 침입 탐지 시스템은 문맥 정보를 바탕으로 판단의 근거를 명확히 제시하여, 보다 정교하고 일관된 응답을 생성하였다. 반면, RAG를 적용하지 않은 언어 모델은 문맥 없이 일반적인 답변을 하거나 “판단할 수 없다”고 응답하는 경우가 많았다.

표 2. RAG 기반 시스템 평가 결과

Model	Precision	Recall
LLaMA 3.1	0.98	0.97
Mistral	0.94	0.96
Gemma	1.0	0.98
Falcon	0.84	0.86

표 3. RAG 미적용 언어 모델 평가 결과

Model	Precision	Recall
LLaMA 3.1	0.48	0.64
Mistral	0.21	0.18
Gemma	0.67	0.80
Falcon	0.52	0.54

다음은 RAG 기반 침입 탐지 시스템이 도메인 특화된 문맥을 바탕으로 도출한 판단 예시다.

(1) **질문:** HMI1 sent function code 14 (Read File Record) to PLC2. Is this normal behavior?

- RAG 모델 **답변:** Based on the information provided in the Context, I would say that the behavior described is not normal.

According to Question Q and A in the Context, HMI-1 in IBEEMS SIM sends Function code 3 (read) to PLC-1 and PLC-2 to retrieve data. Additionally, Question Q states that this is considered abnormal behavior and may indicate malicious actions by an attacker if any Function code other than 3 is sent.

Function code 14 (Read File Record) is not mentioned as a normal or expected function code for HMI-1 in the Context. In fact, it appears to be unrelated to the standard read operations (Function code 3). Therefore, based on the information provided, sending Function code 14 to PLC2 does not match the normal behavior of HMI-1 and may indicate an intrusion or malicious activity.

이러한 결과는, 도메인 지식 기반 문서 검색을 통해 문맥 정보를 제공받는 RAG 기반 침입 탐지 시스템이 네트워크 행위 분석에서의 판단 정확도뿐만 아니라, 근거가 명확한 설명 제공에도 크게 기여할 수 있음을 시사한다.

5. 결론 및 향후 연구

본 논문에서는 스마트 빌딩 HVAC 제어 환경에 특화된 RAG 기반 침입 탐지 시스템을 제안하고, 임베딩 모델, 벡터 검색 엔진, 사전 학습된 언어 모델의 조합에 따른 성능을 종합적으로 비교·평가하였다.

실험 결과, 제안한 RAG 기반 침입 탐지 시스템은 RAG를 적용하지 않은 언어 모델을 사용한 방식에 비해 더 높은 탐지 성능을

기록하였을 뿐 아니라, 문맥 기반 설명 제공을 통해 시스템 신뢰성과 해석 가능성 또한 향상시켰다.

다만, 본 연구에서 사용된 데이터셋은 시뮬레이터 기반 네트워크 행위에 한정되어 있어, 실제 스마트 빌딩 환경에서 발생할 수 있는 다양한 예외 상황이나 복합 공격 시나리오를 충분히 반영하지 못한 한계가 있다.

따라서, 향후 연구에서는 실제 건물 네트워크 환경에서 데이터를 수집하고, 실환경 기반 시나리오를 반영한 고도화된 데이터셋을 구축하여, 제안 시스템의 현실 적용 가능성과 확장성을 검증할 예정이다.

참고 문헌

- [1] D. Chen et al., "Defending against cyber-attacks in building HVAC systems through energy performance evaluation using a physics-informed dynamic Bayesian network (PIDBN)," *Energy*, vol. 322, p. 135369, 2025.
- [2] Z. Chiba et al., "Intelligent and Improved Self-Adaptive Anomaly based Intrusion Detection System for Networks," *International Journal of Communication Networks and Information Security*, vol. 11, no. 2, pp. 312-330, 2019.
- [3] "K. Al Jallad et al., "Anomaly detection optimization using big data and deep learning to reduce false-positive," *Journal of Big Data*, vol. 7, no. 1, article 68, 2020.
- [4] Y. Yao et al., "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.
- [5] M. Arslan et al., "A Survey on RAG with LLMs," *Procedia Computer Science*, vol. 246, pp. 3781-3790, 2024.
- [6] Alireza Dehlaghi-Ghadim et al., "ICSSIM – A framework for building industrial control systems security testbeds", *Computers in Industry*, vol. 148, 103906, 2023.
- [7] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *arXiv preprint*, arXiv:2212.03533, 2022.