

# LLM 기반 사이버 위협 분석 보고서 공격기술 자동 분석 기술<sup>☆</sup>

## LLM-Based Cyber Threat Analysis Report: Automated Attack Technique Analysis Technology

박 용 현<sup>1</sup>      최 창 희<sup>2\*</sup>  
Yonghyeon Park      Changhee Choi

### 요 약

최근 사이버 공격이 증가함에 따라, 이를 분석한 사이버 위협 보고서의 수도 급증하고 있다. 그러나 보고서 내 공격기술을 MITRE ATT&CK의 공격기술로 매핑하는 과정은 사이버 보안 전문가의 많은 시간과 노력을 요구한다. 이를 자동화하기 위해 본 논문에서는, 대형 언어 모델(LLM)을 활용하여 사이버 위협 분석 보고서로부터 MITRE ATT&CK의 공격기술을 자동으로 추출하는 방법을 제안한다. 이를 위해 MITRE ATT&CK의 공식 홈페이지를 크롤링하고, 사이버 위협 분석 보고서로부터 수동으로 공격기술을 라벨링하여 훈련 데이터셋을 구축하였다. 또한 LLaMA, Gemma, DeepSeek 등 최근 많이 사용되고 있는 LLM 모델들을 대상으로 프롬프트 엔지니어링 및 파인튜닝을 진행하였으며, 각 모델의 공격기술 추출 성능을 비교·분석하였다. 실험 결과, 제안하는 방법은 최대 61.1%의 정확도로 단일 문장으로부터 공격기술을 추출하였다. 또한, 공격기술 추출을 목적으로 한 파인튜닝 후에도 LLM 고유의 대화 성능이 유지됨을 확인하였다.

☞ 주제어 : LLM, 파인튜닝, MITRE ATT&CK, 사이버 공격 기술

### ABSTRACT

With the rise in cyber attacks, the number of cyber threat intelligence reports analyzing these incidents has increased significantly. However, mapping the attack techniques described in these reports to MITRE ATT&CK techniques requires significant time and effort from cybersecurity experts. To automate this process, we propose the method that leverages a Large Language Model (LLM) to automatically extract MITRE ATT&CK techniques from cyber threat intelligence reports. We crawled the official MITRE ATT&CK website and manually labeled attack techniques from cyber threat intelligence reports to construct a dataset. Additionally, we conducted prompt engineering and fine-tuning on widely used LLMs, including LLaMA, Gemma, and DeepSeek, and compared their performance in TTP extraction. Experimental results showed that the proposed method extracted attack techniques from individual sentences with an accuracy of up to 61.1%. Furthermore, even after fine-tuning the model for the specific task of technique extraction, it preserved the original conversational capabilities of the LLM.

☞ keyword : LLM, finetuning, MITRE ATT&CK, cyber attack technique

## 1. 서 론

최근 사이버 위협의 양과 다양성이 급격히 증가함에 따라, 이를 탐지하고 분석하는 과정에서 발생하는 데이터

및 보고서의 수가 폭발적으로 늘어나고 있다. 대부분의 사이버 보안 기관은 탐지된 공격을 분석한 후 사이버 위협 분석 보고서를 작성한다. 그러나 사이버 위협의 빈도가 증가함에 따라 이러한 보고서의 수도 급증하며 보고서의 활용도가 저하되고, 문제가 발생하고 있다. 더욱이, 보고서를 작성하는 기관마다 형식과 내용에 차이가 있어 통일성이 부족한 점이 문제로 지적된다.

이에 따라 사이버 위협을 표준화된 방식으로 표현하기 위한 여러 연구가 진행되어 왔다. 대표적으로 YARA rule, STIX와 같은 표준화된 형식들이 존재하며, 2018년 MITRE는 사이버 공격을 모델링한 ATT&CK v1.0을 발표한 이래로 v16.1까지 업데이트 하고 있다[1]. MITRE ATT&CK은 현재 사실상 사이버 보안 영역에서 표준으로

<sup>1</sup> 3rd, Agency for Defense Development, Seoul, 05661, Korea.

<sup>2</sup> Department of Cyber Defense, Sejong University, Seoul, 05006, Korea.

\* Corresponding author (choich@sejong.ac.kr)

[Received 15 March 2025, Reviewed 17 April 2025(R2 30 June 2025), Accepted 12 July 2025]

☆ This work was supported by the faculty research fund of Sejong University in 2025.

☆ This work was supported by Agency for Defense Development.

☆ 본 논문은 2024년 한국인터넷정보학회 추계학술대회 우수논문 추천에 따라 확장 및 수정된 논문임.

널리 사용되고 있으며, 최근 작성되는 사이버 위협 보고서에서도 공격 기술을 MITRE ATT&CK의 기술 목록에 맞추어 정리하는 방식이 점차 늘어나고 있다. 그러나 이를 위해서는 보고서 작성자가 400여 개에 달하는 공격 기술을 외우거나 일일이 검색해야 하는 번거로움이 있다. 반대로, 작성된 보고서를 읽고 MITRE ATT&CK 공격 기술로 라벨링하는 과정도 시간이 소요된다. 게다가 보고서 내의 키워드가 MITRE ATT&CK에서 정의된 공격 기술과 정확히 일치하지 않는다면, 독자는 이를 이해하기 위해 추가적인 노력을 기울여야 한다.

본 논문에서는 대형 언어 모델(LLM)을 활용하여 사이버 위협 분석 보고서로부터 공격기술을 자동으로 추출하는 새로운 방법을 제안한다. 이를 위해 MITRE ATT&CK 공식 홈페이지를 크롤링하고, 사이버 위협 분석 보고서에서 수집한 문단에 대응되는 공격기술을 수동으로 라벨링하여 문단-공격기술 데이터셋을 구축하였다. 이후, 해당 데이터셋을 활용하여 다양한 LLM 모델을 파인튜닝하고, 튜닝된 모델들의 공격기술 추출 성능을 비교·분석하였다. 실험 결과, 제안하는 방법이 공격기술 추출 성능을 효과적으로 향상시키는 것을 확인하였다. 특히, 파인튜닝된 LLAMA-8B 모델은 각 문장에 대응되는 공격기술을 약 61%의 정확도로 식별하였으며, 대화형 질의응답 기능 또한 유지하고 있음을 확인할 수 있었다.

## 2. 기존 연구

Husari 연구진은 보고서로부터 공격기술을 추출하는 방법인 TTPDrill을 제안하였다[2]. 이들은 자연어 처리 분야에서 사용되는 POS 파싱을 통해서 주어, 동사, 목적어를 추출하였다. 이를 BM25 및 TF-IDF 기반 방식으로 벡터화하고, 미리 구축한 사이버 위협 온톨로지와 매칭하여 공격기술을 매핑하였다. 그러나 이 방식은 온톨로지를 미리 정교하게 작성하고 주기적으로 업데이트 해야하는 유지 관리 부담이 있었다. 또한, 과거 버전의 MITRE를 사용하여 최근 사이버 공격을 완벽하게 설명하지 못하는 문제가 있다. Legoy 연구진은 이러한 단점을 극복하기 위해 rcATT라는 기술을 제안하였다. rcATT는 문서의 일부 문장만 추출하여 분석했던 TTPDrill과 달리, 보고서 전체를 입력으로 활용하였다. 또한 벡터화 방식으로 TF-IDF뿐 아니라 Word2Vec[4]을 추가로 사용하고, Decision Tree 및 SVC와 같은 분류 알고리즘을 도입하였다. 그 결과, TTPDrill에 비해 우수한 공격 기술 추출 성능을 보였다.

MITRE에서는 앞선 연구들의 단점을 보완한 TRAM을 발표하였다[5]. TRAM은 최신 자연어 모델 중 하나인 BERT[6]를 도입하여 문서 전체가 아닌 단락 단위로 분할하여 분석을 수행함으로써 정확도를 높였다. 하지만, TRAM은 공격기술 추출에 사용자의 개입이 필요한 점에서 앞선 기법과 구분된다.

Choi 연구진은 TRAM의 데이터 전처리가 충분히 정제되지 않아 긴 단락에서 성능이 저하되는 한계를 극복하기 위한 새로운 알고리즘을 제시하였다[7]. 이들은 PEGASUS 모델을 이용하여 각 단락을 간결하게 요약하고, 요약된 내용을 임베딩하여 공격 기술 추출 성능을 높이는 접근 방식을 제안하였다.

한편, Devlin 연구진은 공격 기술 간의 관계성을 분석하고 이를 지식 그래프 형태로 시각화하는 AttackKG를 개발하였다[8]. AttackKG는 개별 공격 행위를 템플릿 기반으로 공격기술 단위로 구조화한 뒤, 연관성을 표현한 그래프를 자동 생성하였다.

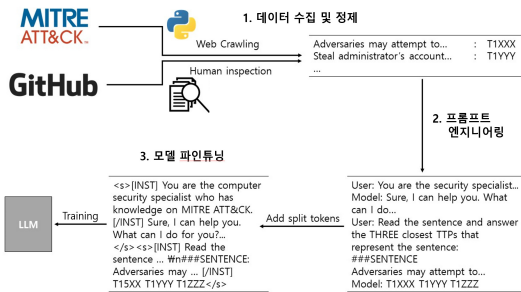
기존 연구들은 다양한 방법으로 공격 기술 추출 정확도를 높이고 연관성을 분석하였지만, 사용자가 사용하기에는 불편한 단점이 있었다. 본 연구에서는 사용자의 편의성을 개선하고, 대화형으로 질의응답이 가능하도록 LLM을 이용한 방법을 제안하였다.

## 3. 배경 지식

### 3.1 MITRE ATT&CK

MITRE ATT&CK은 미국의 비영리 단체인 MITRE가 공격자의 행동을 전술(Tactic), 기술(Technique), 절차(Procedure)로 나타낸 프레임워크이다. MITRE는 총 14개의 전술을 두며, 각 전술을 달성하기 위한 202개의 공격기술을 해당 전술에 매핑해 두었다. 또한, 각 공격기술은 세부기술로 나뉘기도 한다.

이러한 프레임워크는 실제 위협 그룹이 수행한 공격 행동을 분석하고, 이에 대응하는 방어전략을 수립하는데 큰 도움이 된다. 이러한 이점을 바탕으로 MITRE ATT&CK은 사이버 보안 분야의 다양한 연구에 폭넓게 활용되고 있다 [11-14].



(그림 1) 데이터 전처리 및 파인튜닝 과정  
(Figure 1) Data Pre-processing and Finetuning

### 3.2 LLM

LLM(Large Language Model, 대형 언어 모델)은 방대한 텍스트 데이터를 학습하여 자연어를 이해하고 생성하는 딥러닝 기반의 인공지능 모델이다. 주로 트랜스포머(Transformer) 아키텍처를 기반으로 하며, 번역, 요약, 질의응답 등 다양한 자연어처리(NLP) 작업에서 우수한 성능을 보인다. 대표적인 LLM으로는 LLaMA, Gemma, DeepSeek가 있으며, 학습 데이터의 종류나 매개변수 개수에 따라 다양한 변형이 존재한다. 본 논문에서는 널리 사용되는 LLaMA-3.1-8B-Instruct, gemma-2-9b-it, DeepSeek-R1-Distill-LLaMA-8B 모델을 실험에 활용하였다.

(표 1) TTP에서 전술과 기술의 관계  
(Table 1) The relationship between Tactic and Techniques in TTP

전술	권한 상승 (TA0004) Privilege Escalation
공격 기술	액세스 토큰 조작 (T1134) Access Token Manipulation 부팅/로그인 시 자동 시작 (T1547) Boot or Logon Autostart Execution

각 모델은 아키텍처, 학습 방식, 활용 목적에서 차이를 보인다. LLaMA는 메타(Meta)에서 개발한 모델로, 상대적으로 적은 수의 파라미터로도 강력한 성능을 내도록 최적화되었다. Gemma는 구글 딥마인드(Google DeepMind)에서 개발한 경량화된 구조로 높은 자연어처리 성능을 지향하는 모델이다. 클라우드 및 로컬 환경에서의 활용성을 고려하여 설계된 것이 특징이다. DeepSeek는 DeepSeek

AI에서 개발한 다국어 지원에 강점을 가진 모델이다. 아시아 언어 처리에 뛰어난 성능을 보이는 점이 영어 기반으로 개발된 여타 언어 모델과 차별화다.

### 4. 제안하는 방법

그림 1은 제안하는 방법의 전체적인 흐름을 보여준다. 우선 MITRE ATT&CK 홈페이지 및 GitHub CTI 보고서 저장소에서 공격 기술 및 설명을 수집하여 정제하였다. 이후 프롬프트 엔지니어링을 통해 학습 및 추론에 필요한 입력 프롬프트를 설계하고, 이를 활용하여 기본 모델을 파인튜닝하였다.

본 논문에서는 성능 향상을 위해 크게 두 가지 기법을 도입하였다. 먼저, 프롬프트 엔지니어링은 LLM이 정해진 형식의 응답을 생성하도록 입력 문장을 전략적으로 설계하는 기법이다. 이를 활용해 추출한 공격기술이 일관된 형식으로 출력되도록 유도하여, 활용 가능한 정보의 양을 증가시킨다. 다음으로, 파인튜닝은 LLM의 파라미터를 직접 조정하여 모델의 지식 수준 자체를 향상시키는 방식이다. 이는 모델이 사이버 공격 분야에 특화된 추론 능력을 갖추도록 돕는다.

이 두 기법은 공격기술 자동 추출을 달성하는 데 적절하게 기여한다. 데이터 수집 단계를 포함한 각 단계에 대한 자세한 설명을 아래에 제시한다.

#### 4.1 데이터 수집 및 정제

사이버 위협 분석 보고서로부터 공격기술을 추출하도록 파인튜닝하기 위해, 먼저 텍스트를 수집한 후 각 텍스트에 해당하는 공격기술을 식별하여 “텍스트-공격기술” 데이터셋을 구축하였다. 본 논문에서는 하나의 텍스트를 단일문장에서 최대 5~6문장 규모의 단일문단으로 구성하여, 짧은 문장에서 공격기술을 식별하는 능력뿐만 아니라 다수의 문장에서 사이버보안 관련 문장을 정확히 판별하는 능력도 함께 학습시켰다.

단일문단-공격기술 매핑의 정답 데이터(Ground truth) 확보를 위해, 가장 먼저 공격기술과 전술에 대한 자세한 설명을 제공하는 MITRE ATT&CK의 공식 홈페이지를 크롤링하였다. 이를 통해 모델이 다수의 문장에서 핵심 문장을 추출하고, 해당 내용을 공격기술과 매핑하는 능력을 향상시킬 수 있다. 또한, 공격기술들이 다양한 방식으로 패러프레이징(paraphrasing)되는 경우가 많아 여러 형태의 텍스트가 동일한 공격기술로 연결된 데이터가 존재한다.

이러한 데이터는 텍스트-공격기술 매핑을 학습하는 모델이 보다 일반화된 문장 이해 능력을 갖추도록 기여한다.

두 번째로, 단일문장-공격기술의 정답데이터를 확보하기 위해, 사이버보안 전문가가 사이버 위협 분석 보고서를 검토하고 1개 문장에 대해 1개의 공격기술을 수동으로 레이블링하였다[9]. 이 과정은 문장 해석의 다양성을 최소화하는 과정이다. 예를 들어, “The attacker created a new user account and added it to the administrators group”라는 문장은 직관적으로 T1136 Create Account와 매핑될 수 있지만, 앞뒤 문맥에 따라 거시적으로 해석될 경우 T1068 Privilege Escalation과도 매핑될 수 있다. 전문가는 해석의 다양성을 최소화하고, 이를 반영한 학습 데이터셋을 구축한다. 이렇게 구축된 데이터셋으로 파인튜닝된 모델은 단일 문장으로부터 공격기술을 정확하게 식별하는 능력을 확보한다.

## 4.2 프롬프트 엔지니어링

프롬프트 엔지니어링은 LLM이 원하는 출력을 생성하도록 입력 프롬프트를 설계하고 최적화하는 기술이다. 프롬프트를 적절히 구성하면 모델이 주어진 입력을 보다 효과적으로 해석하여 원하는 형태의 응답을 생성할 가능성을 높일 수 있다.

### 4.2.1 역할 부여 (Role Assignment)

역할 부여는 모델에게 특정한 역할을 지정하여 사고 방식과 응답의 스타일을 조정하는 기법이다. 본 연구에서는 사이버보안 연구원의 역할을 부여하기 위해, 아래와 같은 프롬프트를 사용하였다.

You are a computer security specialist who has rich knowledge in MITRE ATT&CK.

### 4.2.2 명확한 지시 (Clear Instruction)

명확한 지시는 모델이 수행해야 할 작업을 구체적으로 명시하는 방식이다. 본 연구에서는 아래와 같은 프롬프트를 사용하여 지시 내용을 명확히 전달하였다.

Read the sentence below and answer. What are the top 3 most relevant technique IDs?

### 4.2.3 프로그레시브 프롬프팅

(Progressive Prompting)

프로그레시브 프롬프팅은 대화를 반복적으로 이어가며 원하는 답변을 점진적으로 유도하는 방식이다. 이는 모델이 명령을 해석하는 방향을 단계적으로 조정할 수 있어, 보다 정교하고 정확한 응답을 이끌어내는 데 유용하다. 본 연구에서는 아래와 같이 연속적으로 질의하였다.

Q. Read the sentence below and answer. What are the top 3 most relevant technique IDs?  
SENTENCE: Adversaries may buy, lease, rent or obtain infrasturcture that can be used during targeting. A wide variety of...  
A. T1583 looks the closest answer. T1583 is ....  
Q. Without any explanation, please answer only with IDs. If there are several probable IDs, enumerate them splitting with a white space.  
A. T1583 T1588 T1584

### 4.2.4 퓨샷 프롬프팅 (Few-Shot Prompting)

퓨샷 프롬프팅은 원하는 출력 형식을 반복적으로 제공함으로써, 새로운 입력에 대해 같은 출력 형식으로 대응하도록 통제하는 강력한 프롬프팅 기법이다. 이 기법은 파인튜닝이 요구되지 않기 때문에 추가 학습을 위한 자원이나 시간이 필요하지 않으며, 동시에 모델의 파라미터를 바꾸지 않기 때문에 높은 범용성을 유지할 수 있다.

Refer to the examples below and follow their format.

Q. An attacker used Powershell for remote code execution.

A. T1059 T1203 T1569

Q. Scripts were obfuscated using Base 64 encoding.

A. T1027 T1059 T1140

Q. They dumped credentials using LSASS memory access.

A. T1003 T1055 T1082

**Algorithm 1: Finetuning an LLM with text-technique ID pairs**

Input: Pretrained Model, text-TTP pairs, hyperparameters  
Output: Fine-Tuned Model

# Step 1: Apply prompt template

```
1 dataset_with_prompt_template ← Preprocess(Dataset)
```

# Step 2: Set Quantization and LoRA Parameter

```
2 Quant_type ← 8-bit
3 Compute_type ← 16-bit
4 LoRA ← { layer_1, layer_2, ... }
```

# Step 3: Set Hyperparameters

```
5 lr ← 1e-5
6 epoch ← 1
7 batch_size ← 2
```

# Step 4: Initialize Model

```
8 Model ← from_pretrained(model_name)
```

# Step 5: Train the Model

```
9 for e ← 1 to epoch do
10   for batch in dataset_with_prompt_template do
11     train_model(Model, batch)
```

### 4.3 파인튜닝

#### 4.3.1 프롬프트 템플릿

프롬프트 템플릿은 LLM이 연속된 문장에서 각 문장의 발화 주체를 명확히 인식할 수 있도록 특수한 토큰이 추가된 문장 형태를 의미한다. 알고리즘1의 1번 줄은 프롬프트 템플릿을 적용하는 과정을 포함한다.

아래 표는, 4.2.3의 대화를 LLAMA2 토큰나이를 사용하여 프롬프트 템플릿을 적용한 예시를 보여준다. 특수토큰 <s>, </s>는 질문-대답의 한 라운드를 구분하며, 특수토큰 [INST], [/INST]는 질문 혹은 명령 부분을 구분한다. 다시 말해, <s> [INST] QUESTION [/INST] ANSWER </s>의 구조를 가지며, 긴 대화는 이러한 라운드의 반복으로 표현된다. 특수토큰의 종류는 모델의 종류마다 다르기에, 다른 토큰나이를 사용한 경우 템플릿 적용 결과는 달라질 수 있다.

```
<s>[INST] Read the sentence below and answer.
What are the top 3 most relevant technique IDs?
SENTENCE: Adversaries may buy, lease, rent or
obtain infrasturcture that can be used during
targeting. A wide variety of... [/INST] T1583 looks
the closest answer. T1583 is .... </s> <s>[INST]
Without any explanation, please answer only with
IDs. If there are several probable IDs, enumerate
them splitting with a white space. [/INST] T1583
T1588 T1584</s>
```

#### 4.3.2 매개변수 설정

알고리즘1의 2~4번 줄은 파라미터 저장 및 연산에 사용할 비트 수를 설정하는 단계이다. 학습 자원을 효율적으로 사용하기 위해, 본 연구에서는 정밀도를 8비트로 낮추는 양자화를 진행하였으며, 연산 시에 정밀도를 16비트로 상향하여 계산되도록 설정하였다. 또한, LoRA를 함께 적용하여 메모리 사용량을 줄였다. 5~7번 줄은 파인튜닝을 위한 주요 하이퍼파라미터를 설정하는 작업이다. 기존 성능을 저해하지 않는 범위 내에서 최적화로 수렴하도록  $lr=1e-5$ ,  $epoch=1$ 로 설정했으며 가용 VRAM을 고려하여  $batch\_size=2$ 로 설정하였다.

이후, 알고리즘1의 8번 줄에서는 사전학습된 모델을 불러와 파인튜닝을 위한 초기화 작업을 수행한다. 9~11번 줄에서는 전처리된 데이터셋과 앞서 설정한 하이퍼파라미터에 따라 학습을 진행한다. 메모리 초과 방지를 위해 전체 데이터를  $batch\_size$  크기로 나누어 학습하고, 이 과정을 전체 데이터에 대해  $epoch$ 번 만큼 반복하여 모델을 점진적으로 최적화하였다.

## 5. 실험 설정

본 논문에서는 MITRE ATT&CK 공식 홈페이지를 크롤링하고, 사이버 위협 분석 보고서를 수동으로 라벨링하여 총 24,348개의 텍스트-공격기술 쌍을 구축하였다. 학습 데이터와 테스트 데이터의 비율은 9:1로 설정하였으며, 텍스트의 길이는 균일하게 분포되도록 하였다. 파인튜닝을 위해 Nvidia GeForce RTX 3090 8개를 활용하였으며, 학습 데이터는 프롬프트 엔지니어링 과정을 통해 한 차례 가공되어 모델에게 전달되었다. 제안한 방식의 성능을 평가하기 위해 아래와 같은 평가 항목을 고려하였다.

1) 예측정확도 @Top-K: 본 논문에서는 문장 해석의 다양성을 반영하기 위해 Top-K 샘플링을 사용하여 예측정

확도를 측정한다. 구체적으로, 프롬프트 엔지니어링을 통해 모델이 정확히 K개의 공격기술 ID를 예측하도록 제어하고, 예측 결과 중 정답이 존재한다면 옳게 예측한 것으로 간주하며, 정답이 없으면 예측에 실패한 것으로 간주한다. 예를 들어, 문장 “The attacker created a new user account and added it to the administrators group”은 T1136 Create Account, T1098 Account Manipulation, T1069 Permission Groups Discovery로 총 3개의 공격기술로 해석될 수 있다.

위와 같이, 정답에 가까운 상위 K개의 예측을 생산하도록 유도하면, 다양한 공격 기법을 포괄적으로 식별하고 이에 대응하는 방어기법을 선제적으로 도출할 수 있다. 반대로, 과도하게 많은 가능성을 사전에 차단하는 효과도 기대할 수 있다. 모델이 수십 개에 가까운 공격기술을 예측하면 분석가가 의미 있는 기술과 의미 없는 기술을 구분하기 어려워진다. 따라서, 정답 예측 개수를 제한하여 실질적인 분석 효율성과 예측 신뢰도를 동시에 확보하는 데 유효하다.

결과적으로, Top-K 평가 방식은 복잡한 문장이 포함될 수 있는 다양한 해석 가능성을 수용하며, 보안 분야의 실사용 맥락에 더욱 부합하는 평가 지표를 제공한다.

2) 파인튜닝 적합도: 파인튜닝 전후 모델의 성능을 비교하여 파인튜닝이 적절히 수행되었는지를 확인한다.

3) Top-K 예측 개수 변화에 따른 예측정확도:

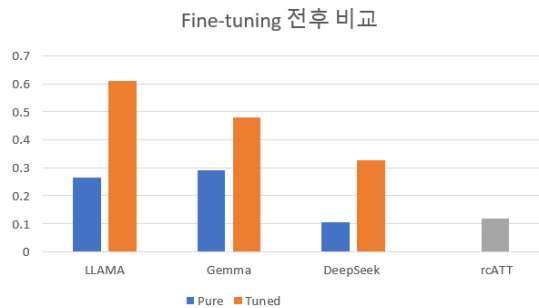
Top-K 구조에서는 K의 값에 따른 예측 정확도와 정밀도 간에 상관관계가 존재한다. 따라서 K를 1, 2, 3, 5 등 다양한 값으로 설정하여 실험을 수행하고, 각 경우에 정답이 예측 목록에 포함되었는지를 기준으로 예측정확도를 산출하였다. 이를 통해 예측 후보 수를 늘릴 때 모델 성능에 미치는 영향을 정량적으로 분석하고, 실제 활용 환경에서의 K값 선택 기준을 제시하고자 하였다.

4) 정밀도(Precision), 민감도(Recall) @Top-K: Top-K를 강제하면 정밀도는 0 또는  $1/K$ 의 값을 가진다. 또한, 문장 1개에 공격기술 ID 1개가 매핑된 데이터셋의 특성상 민감도는 0 또는 1의 값을 가진다. 따라서, 이들의 조화평균인 F1 스코어 역시 값이 제한된 이산값을 갖게 된다. 이러한 특성상 정밀도와 민감도는 모델의 성능을 정확하게 반영하지 못할 수 있지만, 비교기준으로서 의미를 가지므로 해당 값들을 정리하여 제시하였다.

5) LLM 범용성 평가: 파인튜닝 이후에도 범용적인 대화형 질의응답 기능이 보존되는지 확인한다. 이는 공격기술 추출 특화 성능과 일반적 언어 모델 성능 간의 Trade-Off를 평가하기 위한 항목이다.

## 5.1 예측정확도

예측정확도는 모델이 제시한 1개의 문장에서 추출될 수 있는 K개의 예측 공격기술 중 정답이 포함된 문장의 비율이다. 파인튜닝된 모델의 예측정확도를 평가하기 위해, 각 모델이 K=3 및 K=5 조건에서 예측한 공격기술 ID 목록에 정답이 포함되어 있는지를 확인하고, 그 결과를 그림 2 및 그림 3에 제시하였다.



(그림 2) 파인튜닝 전후 성능비교

(Figure 2) Performance comparison before and after finetuning

## 5.2 파인튜닝 이전의 예측정확도

그림 2에 따르면, 파인튜닝 전의 Top-3 예측 정확도는 모델의 종류에 의존함을 확인하였다. 기존 자연어처리 기법을 사용한 rcATT는 11.7%의 정확도를 보였다. rcATT는 LLM을 사용하지 않기 때문에 파인튜닝이 불가능하여 단일 측정값만 존재한다.

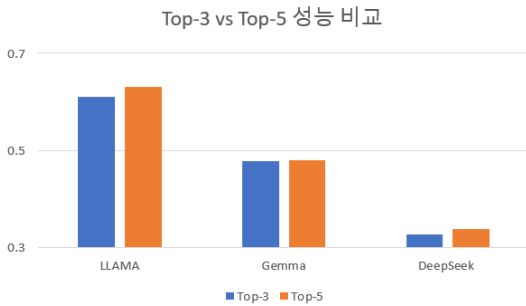
LLAMA-8B는 파인튜닝 전에 26.5%의 정확도를 보였으며, Gemma-9B는 29.19%의 정확도를 보이며 rcATT의 2.2~2.5배 성능을 달성하였다. 이에 반해 DeepSeek-distill-LLAMA-8B의 경우 파인튜닝 전 10.5%로 rcATT보다 낮은 정확도를 보였다.

## 5.3 파인튜닝 효과 분석

텍스트-공격기술 매핑을 활용하여 파인튜닝한 모델은 파인튜닝을 하지 않은 모델보다 정확도가 높음을 그림 2을 통해 알 수 있었다. 이는 테스트 데이터가 성공적으로 모델에 반영되었음을 시사한다.

다만, 파인튜닝을 통한 예측정확도 변화량은 아키텍처의 종류에 따라 다르게 나타났다. Top-3를 기준으로,

LLAMA-8B의 경우 파인튜닝 전에 26.5%의 정확도를 보였지만 파인튜닝 후 61.1% 수준으로 높아지며 가장 높은 상승폭을 보였다. Gemma-9B는 파인튜닝 전에 29.19%의 정확도를 보였지만, 파인튜닝 후 47.88%의 정확도를 보였다. DeepSeek의 경우 파인튜닝 전에 10.5%의 정확도를 보였지만, 파인튜닝 후 32.6%의 정확도를 보이며 가장 높은 비율의 개선폭을 보였다. 이는 rcATT의 2.8~5.2배에 달하는 성능으로, LLM 및 파인튜닝의 효용을 실증적으로 입증하는 결과이다.



(그림 3) Top-3대 Top-5 성능 비교  
(Figure 3) Top-3 vs Top-5 performance comparison

#### 5.4 Top-K 예측 개수 변화에 따른 예측정확도

K값에 따른 예측 정확도는 큰 차이를 보이지 않는 것으로 확인되었다. Finetuned 모델을 대상으로 실험한 결과, LLAMA의 경우 Top-3에 비해 Top-5가 약 2% 정확도 우위를 보였고, DeepSeek의 경우 약 1.5% 우위를 보였다. 하지만, Gemma는 0.05%에 불과한 매우 낮은 정확도 차이를 보였다. K값이 높아짐에 따라 precision이 낮아지는 것을 상쇄할 만큼의 정확도 향상이 없으므로, K=3 이하의 값을 사용하는 것이 적절하다는 결론을 내릴 수 있다.

#### 5.5 정밀도, 민감도, F1-Score

민감도(recall)과 정밀도(precision)는 오탐과 미탐을 판별하기 위해 사용하는 상호보완적인 성능 지표이다. 민감도를 높이기 위해 예측 개수를 과도하게 증가시키면, 정밀도가 매우 낮아진다. 정밀도를 높이기 위해 예측의 개수를 낮추면 민감도가 낮아질 우려가 있다.

민감도는 모든 정답 중 몇 개의 정답을 성공적으로 찾았는지 측정하는 지표이다. 예를 들어, 1개 문장에 10개의 TTP가 매핑되어 있고, LLM이 그 중 4개의 정답 TTP

를 찾아냈다면, 민감도는  $4/10 = 0.4$ 이다. 정밀도는 예측한 정답 중 실제 정답의 비율을 확인하는 지표이다. 예를 들어, LLM이 1개 문장에 대응되는 TTP 10개를 예측하고 그 중 8개의 정답 TTP를 찾아냈다면, 정밀도는  $8/10 = 0.8$ 이다. F1-Score는 정밀도와 민감도의 조화평균으로, 민감도와 정밀도 사이의 균형을 종합적으로 평가하기 위한 지표로 활용된다.

(표 2) 세 모델의 공격기술 추출 정밀도 및 민감도  
(Table 2) MITRE technique ID extraction precision and recall for the three models

튜닝여부		LLAMA	Gemma	DeepSeek	rcATT
전	민감도 (=정확도)	0.2648	0.2919	0.1054	-
	정밀도	0.0874	0.0955	0.0201	
	F1-score	0.1314	0.1439	0.0338	
후	민감도 (=정확도)	0.6111	0.4788	0.3261	0.1178
	정밀도	0.1582	0.3247	0.0762	0.0393
	F1-score	0.2513	0.3870	0.1235	0.0589

본 논문에서는, 1개 문장에 1개 공격기술이 매핑된 데이터에 대하여, 예측 개수를 3개로 고정하고 민감도와 정밀도를 측정하였다. 1개 문장의 공격기술을 예측하는 경우, 민감도의 분모는 항상 1이 된다. 또한, 예측의 성공 실패 여부에 따라 분자는 0 또는 1이 되므로, 민감도는 0 혹은 1의 값을 가진다. 정밀도의 경우, Top-3를 강제하기에 분모는 항상 3이 된다. 또한, 예측의 성공 실패 여부에 따라 분자는 0 또는 1이 되기에, 정밀도는 0 또는 1/3의 값을 가진다.

표 2는 모든 모델의 민감도와 정밀도를 측정하고, 평균을 구해 기록한 값이다. 실험 결과, 민감도는 LLAMA가 가장 높았지만 정밀도는 Gemma가 가장 높았는데, 이는 Gemma가 프롬프트에서 요구한 Top-3 예측을 일관되게 이행하지 않은 것과 관련이 있다. 구체적으로, Gemma는 3개의 예측값을 만들어달라는 프롬프트를 무시하고 1~2개의 예측만 제시하는 경우가 많았고, 이로 인해 정밀도가 더 높아진 것으로 해석하였다.

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP} \quad (1)$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

## 5.6 LLM 범용성

LLM은 특정 작업에 맞춰 파인튜닝될 경우, 해당 작업에 대한 성능은 향상되지만 기존의 범용적인 대화 능력이 저하될 가능성이 존재한다. 본 연구에서는 모델이 공격 기술을 정확히 추출하는 능력을 강화하면서도, 기존의 대화형 응답 성능이 얼마나 유지되는지를 확인하기 위해 실제 테스트를 통해 아래 네 가지 항목을 직접 평가하였다.

- 1) 정답의 정합성: 사용자의 질의에 의미적으로 적절한 응답을 생성하였는지 여부
- 2) 프롬프트 이행: 프롬프트의 요청을 반영한 적절한 응답 템플릿을 따르는지 여부
- 3) 사실적 정확성: LLM의 출력에 오류나 환각이 포함되지 않았는지 여부
- 4) 기타 작업 성능: 요약, 다음문장 예측, 마스크된 문장 복원 등 비지시적 작업을 잘 처리하는지 여부

Q. 안녕? 컴퓨터 보안과 관련된 질문을 하려고 하는데 받아줄 수 있을까?

A. 네! 무엇이든 물어봐주세요. 최선을 다해 답변드리겠습니다.

Q. 고마워. MITRE ATT&CK에 대해 알고있니?

A. 네, 알고 있습니다!

MITRE ATT&CK는 사이버 위협과 관련된 기술들을 기반으로 분류한 공통적인 취약점 데이터베이스입니다. 궁금하신 점이 있으시면 얼마든지 물어보세요!

Q. 조금 더 자세히 이야기해줄래?

A. 물론이죠! MITRE ATT&CK는 사이버 위협 행위자가 사용하는 기술들을 분류하고, 취약점과 공격 유형을 명확하게 설명하는 데 도움을 주려고 합니다. 이렇게 구성된 정보는 시스템 보안 강화에 필요한 사전 준비와 대처 방식 수립에 매우 효과적입니다.

ATT&CK는 다음과 같은 특징이 있습니다:

- 체계적인 분류: 기술은 공격 목표, 실행 메커니즘(예: 사용자 권리), 범위 등 기준으로 명확하게 분류되어 있어 이해하기 쉽습니다.
- 다양한 유형 지원: 시점별로 발행된 취약점 및 사례 연구를 통해 지속적인 업데이트가 이루어지며, 다양한 공격 유형에 대한 정보도 제공합니다.
- 공개 데이터베이스: 누구나 무료로 ATT&CK 데이터베이스에 접근하여 시스템 보안을 강화하는 데 활용할 수 있습니다. ATT&CK는 사이버 보안 분야에서 표준으로 사용되고 있으며, 다양한 보안 도구와 기술과 연동되어 적절한 방어 전략을 제시합니다.

Q. 방금 한 말을 요약해서 2~3문장으로 만들어줘

A. MITRE ATT&CK는 사이버 공격 기술들을 분류하고 기록한 공개 데이터베이스입니다. 실제 사례를 바탕으로 구체적인 정보를 제공하여, 시스템 보안 전략 수립 및 위협 예측에 큰 도움을 줍니다. 누구나 무료로 사용 가능하며 사이버 보안 분야에서 중요한 표준 역할을 합니다.

Q. 그 다음엔 무슨 문장을 쓰면 좋을까?

A. 다음은 몇 가지 아이디어입니다! 어떤 방향으로 이야기를 이어가고 싶으신지에 따라 선택하시면 됩니다.

사용자에게 질문:

- "ATT&CK 데이터베이스를 활용해 본 적이 있나요?"

- "어떤 분야에서 ATT&CK를 활용하고 싶은지 알려주세요. 자세히 설명해 드릴 수 있습니다."

ATT&CK의 활용 사례 소개:

- "ATT&CK는 실제로 어떻게 사용되는 걸까요? 몇 가지 사례를 알려드릴게요!"

- "기업들은 ATT&CK를 이용하여 보안 취약점을 파악하고, 효과적인 방어 전략을 수립하는데 활용합니다."

ATT&CK 관련 더 깊이 있는 정보 제공:

- "ATT&CK는 단순히 기술 분류를 넘어 다양한 분석 도구와 연동됩니다. 관심 있으시면 자세히 알려드릴 수 있습니다."

- "최근 ATT&CK 데이터베이스에 추가된 새로운 공격 기술들을 소개해 드릴까요?"

ATT&CK와 관련된 다른 주제로 전환

- "ATT&CK를 통해 공격자의 행동 패턴을 파악하면, 방어를 위한 팀도 얻을 수 있습니다."

- "사이버 보안은 매우 중요한 주제입니다. ATT&CK 외에도 알아야 할 다른 사항들이 많으니 함께 이야기해 볼까요?"

(그림 4) 파인튜닝 이후 모델의 범용성 평가

(Figure 4) Evaluation of model generalization after finetuning

Q. 사극 말투로 4~5줄 요약 부탁해

A. 참, 이 MITRE ATT&CK이라 하시는 것은 공격자들이 쓰는 무기를 분류한 책과도 같네. 실제 전투에서 사용된 기술들을 기록해서 보관하여, 미래에 어떤 위협을 받을지 예측하는데 활용하는 것이군요!

전략을 세우고 방어를 강화하려면 이 책을 참고하는 것이 절대적인 필수임이 드러납니다. 모든 사람이 무료로 볼 수 있으니, 이제부터 싸움에서 앞서가기 위해 본인의 지식을 키워 나가시길 바랍니다!

Q. 처음부터 끝까지, 사극 말투로 부탁해.

A. 여하 저, 묻는 것이 있으신가요? 컴퓨터 보안에 관하여라니, 이제 첨단 기술과도 엮인 수련의 길을 걸어간다면 그 속에서 자리 잡고 있는 MITRE ATT&CK이라는 말씀, 정말 알기 쉽게 하신네. 바로 공격자들이 사용하는 무기를 분류한 책과 같다고 할까요? 마치 전쟁 기술서를 만들었지만, 이는 오늘날 사이버 전쟁에 적용되는 것으로, 그 위험함은 아주 깊습니다! 실제 전투에서 사용된 공격 방식을 기록하고 보관하여 미래에 어떤 위협을 받을지 예측하는 데 쓰는 것이네. 이런 책이라면 아무리 첨단 기술로 변화하더라도, 전략을 세우고 방어를 강화하기 위해서 참고하는 것은 필수입니다! 저화와 함께 배우고, 깊이 이해하여 싸움에서 앞서가기 위한 지식을 키워 나가시길 바랍니다.

Q. 사극말투로 2~3줄 요약 부탁해. 처음부터 끝까지.

A. 군자, 이 MITRE ATT&CK이라는 것은 공격자가 쓰는 무기를 분류한 책과 같습니다. 실제 전쟁에서 사용된 기술들을 기록하여 미래에 어떤 위협을 받을지 예측하는 데 쓰이는 것이네!

(그림 5) 파인튜닝 이후 모델의 프롬프트 이행 여부 평가  
(Figure 5) Evaluation of model's prompt adherence after finetuning

실험 결과, 그림 4 및 그림 5에서 확인할 수 있듯이, 제안하는 방법은 사용자의 질의에 적절한 응답을 반환하고,



프롬프트의 요구 사항을 충실히 반영하며, 오류나 환각이 없는 출력을 안정적으로 생성하는 것으로 나타났다. 또한, 요약이나 예측과 같은 비지시적 작업에서도 모델의 성능이 일관되게 유지되어, 파인튜닝 후에도 범용적인 대화 능력이 크게 훼손되지 않았음을 확인하였다.

## 6. 결 론

본 논문에서는 LLM을 활용하여 사이버 위협 보고서로부터 공격기술을 자동으로 추출하는 방법을 제안하였다. 이를 위해 MITRE ATT&CK에서 제공하는 데이터 및 사이버 위협 보고서로부터 수동으로 추출한 텍스트-공격기술 데이터를 직접 구축하였다. 또한, 해당 데이터를 활용해 LLM을 파인튜닝 함으로써 공격기술 추출 성능을 강화하였다. 실험 결과 파인 튜닝이 공격기술 추출 성능을 강화함을 알 수 있었으며, 모델별로 차이가 있음을 알 수 있었다. 또한 파인 튜닝 후에도 LLM의 범용성을 유지하는 것을 알 수 있었다. 향후에는 정확도를 높이고 성능 지표를 보다 정밀하게 평가하기 위해 성능평가 지표를 개발할 예정이다.

## 참고문헌(Reference)

- [ 1 ] MITRE ATT&CK, <https://attack.mitre.org>
- [ 2 ] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, "TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources," in Proc. of the 33rd Annual Computer Security Applications Conference (ACSAC '17), pp. 103-115, 2017.  
<https://doi.org/10.1145/3134600.3134646>
- [ 3 ] V. Legoy, M. Caselli, C. seifert, and A. peter, "Automated retrieval of att&ck tactics and techniques for cyber threat reports," arXiv preprint arXiv:2004.14322, 2020.  
<https://doi.org/10.48550/arXiv.2004.14322>
- [ 4 ] C., Kenneth Ward, "Word2Vec," Natural Language Engineering, vol. 23, no. 1, pp. 155-162, 2017.  
<https://doi.org/10.1017/S1351324916000334>
- [ 5 ] J. Devlin, "BERT:Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019.  
<https://doi.org/10.48550/arXiv.1810.04805>
- [ 6 ] MITRE, Center for threat informed defense, TRAM:Threat Report ATT&CK Mapper, 2025.  
<https://github.com/center-for-threat-informed-defense/tram>
- [ 7 ] C. Choi, I. Lee, C. Shin, S. Lee, "Cyber threat campaign analyze based on PEGASUS and RoBERTa model," in proceeding KIMST 2023, 2023.
- [ 8 ] Z. Li, J. Zeng, Y. Chen, Z. Liang, "AttacKG: Constructing technique knowledge graph from cyber threat intelligence reports," in Proc. of the European Symposium on Research in Computer Security, pp. 589-609, 2022.  
[https://doi.org/10.1007/978-3-031-17140-6\\_29](https://doi.org/10.1007/978-3-031-17140-6_29)
- [ 9 ] APT & Cybercriminals Campaign Collection, [https://github.com/CyberMonitor/APT\\_CyberCriminal\\_Campaign\\_Collections](https://github.com/CyberMonitor/APT_CyberCriminal_Campaign_Collections), accessed September, 20, 2024.
- [ 10 ] CS Choi, Y. Jeong, S. Park, IH Won, HS Lim, SM Kim, Y Kang, C Yoon, J Park, Y Lee, HJ Lee, "Optimizing Language Augmentation for Multilingual Large Language Models:A Case Study on Korean," arXiv preprint arXiv:2403.10882, 2024.  
<https://doi.org/10.48550/arXiv.2403.10882>
- [ 11 ] Gee-hee Yun, Kyoung-jin Kim, "Enhancing the Cybersecurity Checklist for Mobile Applications in DTx based on MITRE ATT&CK for Ensuring Privacy," Journal of Internet Computing Services, vol. 24, no. 4, pp. 15-24, Aug. 2023.  
<https://doi.org/10.7472/jksii.2023.24.4.15>
- [ 12 ] Chang-hee Choi, Chan-ho Shin, Sung-uk Shin, "Cyber attack group classification based on MITRE ATT&CK model," Journal of Internet Computing Services, vol. 23, no. 6, pp. 1-13, Dec. 2022.  
<https://doi.org/10.7472/jksii.2022.23.6.1>
- [ 13 ] B. Al-Sada, A. Sadighian and G. Oligeri, "Analysis and Characterization of Cyber Threats Leveraging the MITRE ATT&CK Database," IEEE Access, vol. 12, pp. 1217-1234, 2024.  
<https://doi.org/10.1109/ACCESS.2023.3344680>
- [ 14 ] G. Ahn, J. Jang, S. Choi and D. Shin, "Research on Improving Cyber Resilience by Integrating the Zero Trust Security Model With the MITRE ATT&CK

Matrix,” IEEE Access, vol. 12, pp. 89291-89309, 2024.

<https://doi.org/10.1109/ACCESS.2024.3417182>

- [15] J. M. Kim, S. Y. Lee, W. S. Lee, “Discovering AI-enabled convergences based on BERT and topic

network,” KSII Transactions on Internet and Information Systems, vol. 17, no. 3, pp. 1022-1034, 2023.

<https://doi.org/10.3837/tiis.2023.03.018>

## ● 저 자 소 개 ●



### 박 용 현(Yonghyeon Park)

2019년 성균관대학교 컴퓨터공학과(공학사)

2022년 한국과학기술원 대학원 전산학부 정보보호대학원(공학석사)

2022년~2024년 국가보안기술연구소 연구원

2024년~현재 국방과학연구소 연구원

관심분야 : 정보보호, 머신러닝, LLM, 임베디드, 리버싱

E-mail : ryan8237@gmail.com



### 최 창 희(Changhee Choi)

2008년 연세대학교 컴퓨터과학과(공학사)

2010년 한국과학기술원 대학원 전산학과(공학석사)

2013년 한국과학기술원 대학원 전산학과(공학박사)

2013년~2025년 국방과학연구소 선임연구원

2025년~현재 세종대학교 사이버방학과 부교수

관심분야 : 정보보호, AI 기반 사이버보안, 디지털 포렌식

E-mail : choich@sejong.ac.kr