Journal of KIIT. Vol. 24, No. 8, pp. 45-55, Aug. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.8.45

CTI 문서 TTP 자동 추출을 위한 데이터 정제 기반 하이브리드 LLM-BERT 프레임워크

김현석*, 정유철**

Data Preprocessing-based Hybrid LLM-BERT Framework for Automatic TTP Extraction from CTI Documents

Hyeonseok Kim*, Yuchul Jung**

본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성(IITP-2026-RS-2022-00156394) 및 (2026)년도 교육부 및 경상북도의 재원으로 경북RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)-(아이디어 창업밸리)의 결과입니다(2026-rise-15-105).

요 약

사이버 위협 인텔리전스(CTI, Cyber Threat Intelligence) 보고서는 비정형 장문 텍스트로 작성되는 경우가 많아, 분석가가 수작업으로 MITRE ATT&CK 기반 TTP(Tactics, Techniques, and Procedures)를 매핑하는 데 상당한 시간과 비용이 소요된다. 문서 단위 다중 라벨 환경에서는 장문 문맥 처리, 희소 라벨 분포, 데이터 노이즈로 자동 추출 성능이 저하된다. 본 연구는 TRAM2와 AnnoCTR 데이터셋에서 URL, IOC 등 노이즈를 체계적으로 제거하고, Kanana-1.5 2.1B와 CySecBERT를 결합한 하이브리드 TTP 추출 프레임워크를 제안한다. LLM으로 텍스트를 입력받아 TTP 후보를 생성하고 BERT로 정밀 검증하는 구조이다. 실험 결과 TRAM2에서는 규칙 기반 노이즈 제거가, AnnoCTR에서는 URL 제거가 최적 성능을 보였다. 하이브리드 구조는 TRAM2, AnnoCTR 모두에서 F1 점수 향상을 보였으며, 특히 Precision 중심의 성능 개선이 두드러졌다.

Abstract

Cyber Threat Intelligence (CTI) reports are often written as unstructured long-form text, requiring substantial time and cost for analysts to manually map them to MITRE ATT&CK-based Tactics, Techniques, and Procedures (TTPs). In document-level multi-label settings, automatic extraction performance is degraded by long-context processing, sparse label distributions, and data noise. This study proposes a hybrid TTP extraction framework that systematically removes noise such as URLs and IOCs from the TRAM2 and AnnoCTR datasets and combines Kanana-1.5 2.1B with CySecBERT. The framework first uses an LLM to generate TTP candidates from input text and then applies BERT for precise verification. Experimental results show that rule-based noise removal achieved the best performance on TRAM2, while URL removal was most effective on AnnoCTR. The hybrid structure improved F1 scores on both TRAM2 and AnnoCTR, with particularly notable improvements in precision-oriented performance.

Keywords

cyber threat intelligence, TTP extraction, data denoising, multi-label classification, framework

* 국립금오공과대학교 컴퓨터공학과 학사과정
- ORCID: https://orcid.org/0009-0009-7851-3977
** 국립금오공과대학교 컴퓨터공학부 부교수(교신저자)
- ORCID: https://orcid.org/0000-0002-8871-1979

· Received: May 08, 2026, Revised: Jun. 26, 2026, Accepted: Jun. 29, 2026
· Corresponding Author: Yuchul Jung
Dept. of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea
Tel.: +82-54-478-7536, Email: jyc@kumoh.ac.kr

Ⅰ. 서 론

사이버 공격의 고도화와 빈발화에 따라 CTI(Cyber Threat Intelligence)는 조직의 보안 방어 체계에서 핵심적인 역할을 담당하고 있다[1]. 그러나 CTI 보고서는 일반적으로 비정형 자연어 텍스트로 제공되며, 문서 하나에 다수의 공격 행위와 절차가 복합적으로 서술되는 경우가 많다. 이러한 특성으로 인해 분석가가 수작업으로 MITRE ATT&CK 기반 TTP(Tactics, Techniques, and Procedures)를 식별하고 매핑하는 과정에는 많은 시간과 비용이 요구된다. 특히 대규모 위협 인텔리전스가 지속적으로 수집·갱신되는 환경에서는 수작업 기반 분석만으로는 확장성과 일관성을 확보하기 어렵기 때문에, 자동화된 TTP 추출 기술의 필요성이 점차 커지고 있다[2][3].

최근 CTI 보고서로부터 ATT&CK 기반 TTP를 자동 추출하기 위한 다양한 연구가 수행되고 있다[3]-[6]. 그러나 기존 방법들은 문서 수준(Document-level) 다중 라벨 환경에서 여전히 구조적 한계를 가진다. 첫째, 하나의 문서에 다수의 TTP가 동시에 존재하는 경우가 많아 장문 문맥을 안정적으로 처리해야 한다. 둘째, TTP 라벨은 long-tail 분포를 보이는 경우가 많아 상위 빈도 라벨에 대한 편향이 쉽게 발생한다. 셋째, CTI 문서에는 URL, IOC, 메타데이터, 해시 문자열과 같이 분석에 직접 기여하지 않거나 오히려 모델을 교란할 수 있는 텍스트 노이즈가 포함되어 있어 성능 저하를 유발할 수 있다[4]. 기존 연구들은 분류 모델, 생성 모델, 또는 개체 인식 기반 방법을 제안해 왔으나, 이러한 노이즈와 희소 라벨 문제를 문서 수준에서 동시에 다룬 연구는 상대적으로 제한적이다.

본 연구는 이러한 문제를 해결하기 위해, 문서 단위 다중 라벨 TTP 추출에서 데이터 정제 전략과 하이브리드 추론 구조를 함께 고려하는 접근을 제안한다. 구체적으로, TRAM2와 AnnoCTR 데이터셋을 대상으로 URL, IOC, 메타데이터 및 규칙 기반 문장 제거 전략을 단계적으로 적용하여 데이터셋별 최적 정제 방식을 분석하였다. 또한 생성 기반 대형 언어모델인 Kanana-1.5 2.1B[7]과 분류 기반 언어모델인 CySecBERT를 결합하여, LLM이 입력을 기반으로 후보를 제안하고 BERT가 이를 정밀 검증하는 후보 생성 - 검증형(Candidate generation - verification) 하이브리드 파이프라인을 설계하였다.

본 연구의 기여는 다음과 같다.

1. CTI 문서에 포함된 주요 노이즈 유형을 단계적으로 제거하고, 각 정제 방식이 문서 단위 TTP 추출 성능에 미치는 영향을 데이터셋별로 비교 분석하였다

2. 생성 기반 후보 생성 단계와 분류 기반 정밀 검증 단계를 결합한 하이브리드 TTP 추출 파이프라인을 제안하였다.

3. 두 공개 데이터셋에서 제안 방식의 성능을 평가하여, 단일 분류 모델 대비 일관된 개선 경향과 Precision 중심 향상 가능성을 확인하였다.

4. 데이터셋 특성에 따라 상이한 정제 전략이 필요함을 보임으로써, CTI 자동 분석 시스템 설계 시 데이터 전처리의 중요성을 실험적으로 제시하였다.

Ⅱ. 관련 연구

2.1 자동 TTP 추출 연구

자동 TTP 추출 연구는 크게 개체 인식 기반 접근, 분류 기반 접근, 생성 기반 접근으로 구분할 수 있다[3]. 초기 연구에서는 NER(Named Entity Recognition) 기반 접근을 통해 CTI 텍스트 내 ATT&CK 관련 개체나 기술 표현을 탐지하려는 시도가 이루어졌다[3][8]. 이러한 방법은 명시적으로 드러나는 기술 표현을 포착하는 데는 유효하였으나, 문맥적 추론이 필요한 암시적 TTP 표현에 대해서는 한계를 보였다[4].

이후에는 BERT와 같은 사전학습 언어모델(SciBERT, SecureBERT 등) 계열을 활용하여 문장이나 문서 단위로 여러 개의 TTP를 동시에 예측하는 multi-label 분류 방식이 활발히 연구되었다[3][5]. CySecBERT와 같은 도메인 특화 모델은 기존 NER 기반 접근보다 전반적으로 높은 예측 성능을 보였으나, 문서 길이가 길고 라벨 분포가 불균형한 환경에서는 여전히 희소 TTP 예측 성능이 저하되는 문제가 보고되었다[3][9][10]. 특히 문서 수준 TTP 추출에서는 하나의 입력에 다수의 TTP가 동시에 존재하므로, 상위 빈도 라벨에 대한 편향이 심화되기 쉽다.

또한 국내 사이버보안 분야에서도 네트워크 침입 탐지와 공격 유형 분류를 위해 딥러닝 및 전이학습 기반 모델을 적용하려는 연구가 수행되고 있다[11]. 이는 보안 도메인 데이터의 특성을 반영한 모델 설계와 비교 분석의 필요성을 뒷받침한다.

최근에는 대형 언어모델(LLM)을 활용한 접근도 제안되고 있다[3][6]. Few-shot prompting, retrieval-augmented generation, supervised fine-tuning 등의 전략이 적용되었으나, realistic open-set 환경에서는 프롬프트 설계, 예제 순서, 검색 품질 등에 따라 성능 변동성이 크게 나타나는 문제가 보고되었다[3]. 즉, 생성 모델은 폭넓은 후보 탐색 측면에서는 장점을 가지지만, 데이터 부족으로 인한 출력 안정성과 라벨 정밀성 측면에서는 여전히 한계를 가진다.

본 연구는 생성 모델의 후보 탐색 능력과 분류 모델의 판별 능력을 결합한 후보 생성 - 검증 구조를 통해, 생성 기반 접근에서 발생할 수 있는 불필요한 후보를 줄이고 문서 단위 TTP 추출 성능을 개선하고자 한다. 그림 1은 CTI 문서에 포함된 URL, 해시값, IP 주소 등 분석 목적과 직접 관련되지 않는 노이즈가 정제 전후에 어떻게 변화하는지를 개념적으로 보여준다. 본 연구에서는 이러한 정제 과정을 통해 모델이 공격 행위와 절차를 설명하는 핵심 문맥에 더 집중하도록 하였다.

2.2 데이터셋 노이즈 제거 연구

CTI 문서는 위협 행위 기술 외에도 URL, IP 주소, 해시값, 도메인, 파일 경로, 메타데이터 등 다양한 부가 정보를 포함한다. 이러한 정보는 일부 상황에서는 유용한 단서가 될 수 있으나, 문서 수준 TTP 추출 관점에서는 핵심 의미와 직접 관련되지 않거나 오히려 모델의 주의를 분산시키는 노이즈로 작용할 수 있다[4]. 기존 연구에서는 CTI 데이터가 비구조적이며 노이즈가 많다는 점이 반복적으로 지적되어 왔지만[3][4], 노이즈 유형을 단계적으로 분리하여 각 정제 전략의 성능 영향을 체계적으로 비교한 연구는 상대적으로 부족하다.

본 연구는 URL 제거, 메타데이터 제거, IOC 제거, 규칙 기반 문장 제거를 각각 독립적으로 적용하여 성능 변화를 측정하고, 데이터셋별로 어떤 정제 전략이 효과적인지를 비교 분석한다. 이를 통해 데이터 정제가 단순한 전처리 단계가 아니라, 문서 단위 다중 라벨 TTP 추출 성능에 직접적인 영향을 주는 핵심 설계 요소임을 보이고자 한다.

III. 연구 접근 및 개선

3.1 데이터셋 설정

본 연구에서는 문서 단위 다중 라벨 TTP 추출 성능을 평가하기 위해 TRAM2와 AnnoCTR 데이터셋을 사용하였다[3][12][13]. 본 연구에서는 별도의 라벨 재정의나 수동 매핑을 수행하지 않고, TRAM2와 AnnoCTR 데이터셋에서 제공하는 MITRE ATT&CK 기반 TTP 라벨 체계를 그대로 사용하였다. 따라서 학습, 후보 검증, 최종 평가는 각 데이터셋의 기존 라벨 공간을 기준으로 독립적으로 수행하였다.

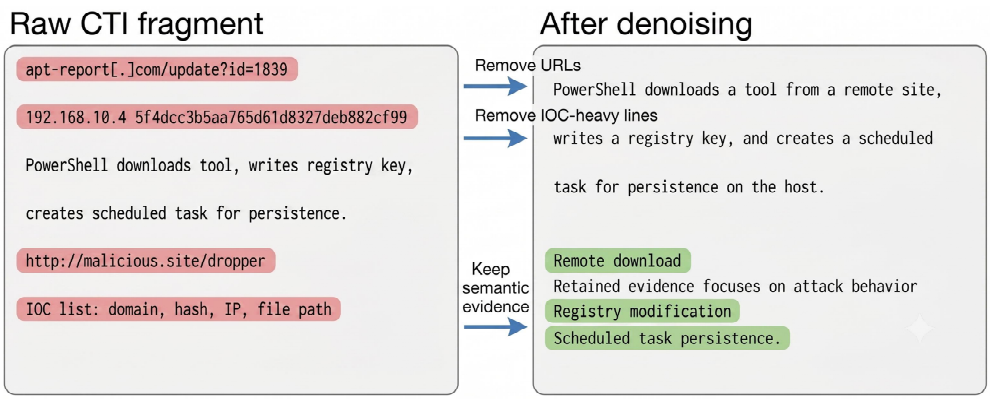

그림 1. 데이터 정제 전후 개념 예시
Fig. 1. Examples of concepts before and after data cleaning

두 데이터셋은 모두 CTI 문서를 입력으로 하고, 각 문서에 하나 이상의 ATT&CK 기반 TTP 라벨이 부여된 구조를 가진다. 표 1은 본 연구에서 사용한 TRAM2와 AnnoCTR 데이터셋의 문서 수, 라벨 유형 수, 문장 수 및 후처리 이후의 변화를 요약한 것이다.

표 1. 사용 데이터셋
Table 1. Used dataset

	TRAM2	AnnoCTR
Documents	120	70
Label types	50	125
Sentences	30716	6679
Post-processed documents	119	62
Post-processed label types	50	61
Post-processed sentences	3257	1242

두 데이터셋은 문서 수와 라벨 수, 후처리 이후 문장 수에서 차이를 보이므로 데이터셋별 정제 전략과 성능 변화를 별도로 분석하였다. 실험은 각 데이터셋이 제공하는 학습 및 평가 분할을 기준으로 수행하였으며, 전처리와 모델 학습은 각 분할의 역할을 유지하는 방식으로 진행하였다. 라벨 공간의 정의는 각 데이터셋의 공식 분할을 기준으로 설정하였다. 다만 학습 분할에 전혀 나타나지 않는 라벨에 대해서는 모델이 충분한 학습 신호를 획득하기 어렵기 때문에, 본 연구에서는 학습 가능한 라벨 집합을 기준으로 평가를 수행하였다. 이때 사용된 라벨 집합과 제외 기준은 실험 절차 내에서 고정하여 비교 실험 전반에 동일하게 적용하였다.

또한 데이터 노이즈가 성능에 미치는 영향을 분석하기 위해, 기본 데이터셋(RAW)을 기준으로 총 4가지 정제 방식을 개별적으로 적용하였다. 각 정제 단계는 독립적으로 설계되었으며, 이를 통해 어떤 유형의 노이즈 제거가 성능 향상에 기여하는지를 체계적으로 분석하였다.

3.2 하이브리드 모델 구조 및 학습 전략

본 연구는 문서 단위 다중 라벨 TTP 추출을 위해 생성 기반 후보 확장과 분류 기반 정밀 검증을 결합한 하이브리드 파이프라인을 제안한다.

그림 2는 제안한 하이브리드 파이프라인의 전체 흐름을 나타낸다.

입력 CTI 문서는 정제 과정을 거친 뒤 LLM을 통해 후보 TTP가 생성되고, 이후 CySecBERT 기반 검증 단계를 통해 최종 TTP 라벨이 결정된다.

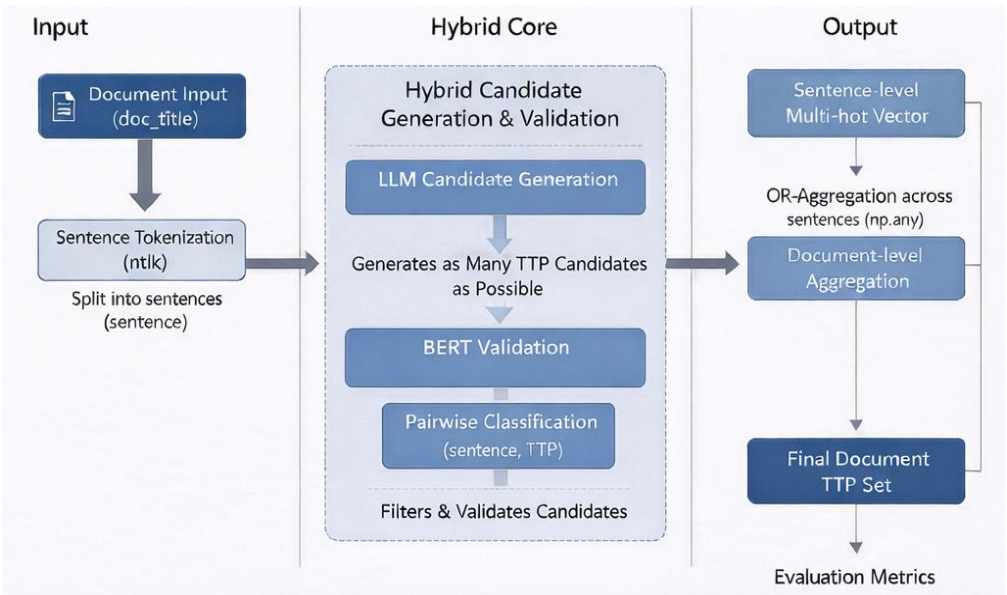

그림 2. 하이브리드 파이프라인 구조
Fig. 2. Hybrid pipeline structure

후보 생성 단계에서는 Kanana-1.5 2.1B를 사용하여 입력 문서 d로부터 TTP 후보 집합 C_LLM(d)를 생성한다. 이 단계의 목적은 입력에 포함된 텍스트 정보를 기반으로 잠재적으로 관련 가능한 TTP 후보를 생성하는 것이다. 생성된 후보는 TRAM2와 AnnoCTR 데이터셋에서 제공하는 MITRE ATT&CK 기반 TTP 라벨 체계를 기준으로 검증된다.

후보 검증 단계에서는 CySecBERT 기반 분류 모델을 사용하여 각 후보의 타당성을 평가한다. 구체적으로, BERT는 입력 문서 d에 대해 라벨별 확률 p_BERT(y|d)를 산출하며, 생성된 후보 $y \in C_LLM(d)$에 대해서는 해당 확률이 임계값 T를 초과하는 경우 최종 예측 라벨로 채택한다. 즉, 최종 예측 집합 $\hat{Y}(d)$는 '생성 후보에 포함되고, 동시에 BERT 확률이 임계값 이상인 라벨의 집합'으로 정의된다. 이 구조는 LLM이 문맥적으로 가능한 후보를 제안하고, BERT가 데이터셋 학습을 통해 획득한 판별 능력으로 허위 후보를 제거하도록 설계되었다. 결과적으로 생성 기반 접근의 후보 탐색력과 분류 기반 접근의 안정적 결정 능력을 상호 보완적으로 활용할 수 있다.

기존의 LLM-Classifier 결합 연구가 주로 LLM이 생성한 라벨을 학습 데이터로 사용하는 teacher-student 형태의 지식 증류에 초점을 두었다면 [14][15], 본 연구는 추론 단계에서의 직접 결합에 초점을 둔다. 이는 추가적인 재학습 없이도 LLM 기반 후보 생성과 분류 기반 판별을 결합할 수 있으며, 모듈 단위 교체 및 확장이 용이하다는 장점이 있다.

표 2은 실험에 사용한 모델별 주요 하이퍼파라미터와 재현성 설정을 정리한 것이다. CySecBERT는 문장 단위 입력에 대해 다중 라벨 분류를 수행하도록 학습하였다. 학습률은 5×10-5, 최대 에폭은 100으로 설정하고 검증 성능 기준의 early stopping을 적용하였다. Kanana-1.5 2.1B는 문서 수준 후보 생성을 위해 3에폭 학습하였으며, 과적합을 방지하기 위해 학습 반복 수를 제한하였다. BERT는 문장 단위 표현을 기반으로 후보 라벨의 타당성을 판별하는 역할을 수행하며, LLM은 입력을 기반으로 후보 라벨 집합을 생성하는 역할을 담당한다.

Ⅳ. 실험 및 분석

4.1 실험 설정

본 실험은 문서 단위 다중 라벨 TTP 추출 환경에서 수행되었다. 평가 지표로는 문서 단위 다중 라벨 분류에 적합한 Micro-F1, Precision, Recall 지표[16]을 사용하였다. 모든 모델은 문장 단위로 학습하였으며, 추론 단계에서는 문장별 라벨 확률을 문서 단위 확률로 집계한 뒤 최종 예측을 수행하였다. 문장 단위 예측 결과를 문서 단위 예측으로 변환하기 위해 Sum pooling, Mean pooling, Max pooling, Noisy-OR를 비교하였으며, 집계 함수별 성능 비교 결과는 표 3에 제시하였다. 표 3에서 Max pooling은 TRAM2에서 가장 높은 F1을 보였고 AnnoCTR에서도 비교적 안정적인 성능을 보여, 이후 실험에서는 Max pooling을 기본 문서 단위 집계 함수로 사용하였다.

CySecBERT는 학습 파라미터 수가 상대적으로 적고 입력 시퀀스 길이가 짧아 안정적인 학습이 가능하므로 학습률 5×10-5, 최대 100 epoch로 설정하고 검증 성능 기반 early stopping을 적용하였다.

표 2. 모델별 주요 하이퍼파라미터 및 재현성 설정
Table 2. Key hyperparameters and reproducibility settings by model

Category	Model	Input unit	Key settings	Max length	Training/inference conditions
Baseline	CySecBERT	Sentence	lr=5e-5, epoch=100, early stopping	512	Binary multi-label classification
Generator	Kanana-1.5 2.1B	Sentence	epoch=3, fine-tuned for candidate generation	3072	Candidate generation
Hybrid	LLM + CySecBERT	Sentence + candidates	threshold τ = 0.5	N/A	LLM generates candidates, then BERT validates
Reproducibility	Common settings	-	seeds = {42, 43, 44}	-	Pytorch 2.6.0+cu124, RTX 6000 ada 48GB

반면 Kanana-1.5 2.1B 모델은 상대적으로 파라미터 수가 많고 정제된 데이터셋에서 미세조정을 수행하므로, 후보 생성 능력을 유지하면서 과적합 위험을 줄이기 위해 3 epoch만 학습하였다.

4.2 데이터 정제

본 연구에서는 노이즈가 문서 단위 다중 라벨 TTP 추출 성능에 미치는 영향에 대한 각 정제 종류별 영향을 독립적으로 평가하기 위해, 총 4가지의 데이터 정제 방식을 개별적으로 적용하고 성능 변화를 분석하였다.

각 정제 전략은 다음과 같다

(1): URL 제거: 정규표현식을 이용하여 URL 문자열을 제거

(2): 메타데이터 제거: 문서 헤더, 작성자 정보, 시간 정보 등 본문 의미와 직접 관련되지 않은 메타데이터 제거

(3): IOC 필터링: IP 주소, 도메인, 해시값, 파일 경로 등 주요 IOC 표현 추상화

(4): 규칙 기반 문장 제거: 매우 짧은 문장, 독립적인 IP/URL/Hex 패턴 위주의 문장을 제거.

특히 규칙 기반 문장 제거는 길이 3 이하 문장, 8자리 이상의 16진수 문자열 중심 문장, 단독 IP 형식 문장, 단독 URL 형식 문장을 우선 식별하고, 이 중 1개 이상의 조건을 만족하는 경우 제거하도록 설계하였다. 이를 통해 문맥적 의미보다 패턴 정보에 치우친 문장을 선택적으로 제거하고자 하였다.

표 4는 데이터 정제 단계별 제거 문장 수와 문서 길이 변화를 나타내며, 표 5와 표 6는 각각 TRAM2와 AnnoCTR에서 데이터 정제 방식별 Precision, Recall, F1 변화를 비교한 결과이다. 정제 과정에서 제거되는 문장의 양은 전체 문서 길이에 비해 크지 않지만, URL, IOC, 패턴성 문장과 같은 노이즈를 선별적으로 제거함으로써 모델 입력의 의미적 밀도를 높이는 효과를 기대할 수 있다.

실험 결과, URL 제거는 두 데이터셋에서 공통적으로 성능 향상에 기여하였다. 반면 메타데이터 제거는 두 데이터셋 모두에서 성능 저하를 보였는데, 이는 일부 메타데이터가 문서 출처나 공격 맥락을 간접적으로 반영할 가능성을 시사한다. 또한 규칙 기반 문장 제거는 TRAM2에서는 성능 향상에 기여하였으나, AnnoCTR에서는 오히려 성능이 감소하였다. 이는 두 데이터셋이 포함하는 문서 구조와 노이즈 분포가 상이하기 때문으로 해석된다.

표 3. 문장-문서 집계 함수 비교
Table 3. Comparison of sentence-to-document aggregation functions

Aggregation method	Description	TRAM2 F1	AnnoCTR F1	Advantages	Considerations
Sum pooling	Sum of sentence-level scores	46.96	63.03	Sensitive for long documents	May introduce document length bias
Mean pooling	Average of sentence-level scores	0	1	Reduces length bias	May lead to conservative predictions
Max pooling	Uses the maximum sentence-level score	71.27	61.23	Effective for detecting strong signals	Sensitive to noisy sentences
Noisy-OR	Probabilistic combination	51.08	62.47	Easy theoretical interpretation	Requires implementation/tuning

표 4. 데이터 정제 단계별 제거량 및 문서 길이 변화
Table 4. Removal volume and document length changes by data cleaning stage

Dataset	Preprocessing step	Removed sentences	Removed token ratio (%)	Avg. document length (Before)	Avg. document length (After)
TRAM2	RAW → 1 (URL)	49	0.08%	24.23	24.21
TRAM2	RAW → 4 (Rule-based)	4	0.03%	24.23	24.25
AnnoCTR	RAW → 1 (URL)	94	0.45%	19.63	19.61
AnnoCTR	RAW → 3 (IOC)	20	0.08%	19.63	19.63

표 5. TRAM2 데이터 정제 종류별 변화
Table 5. TRAM2 data denoising changes

Stage	Precision	Recall	F1
RAW	70.34	68.55	69.44
1	64.57	75.92	69.79
2	72.72	63.45	67.77
3	73.17	62.6	67.48
4	63.1	81.86	71.27

표 6. AnnoCTR 데이터 정제 종류별 변화
Table 6. AnnoCTR data denoising changes

Stage	Precision	Recall	F1
RAW	69.69	55.02	61.49
1	65.77	58.85	62.12
2	68.71	53.58	60.21
3	68.20	56.45	61.78
4	69.13	53.58	60.37

4.3 하이브리드 후보 생성 실험

하이브리드 파이프라인의 후보 생성 단계에서는 4.2절에서 확인된 데이터셋별 최적 정제 전략을 반영하여 학습 데이터를 구성하였다. 구체적으로 TRAM2에는 규칙 기반 문장 제거 정제 결과를, AnnoCTR에는 URL 제거 정제 결과를 적용하였다. 이후 TRAM2 및 AnnoCTR에서 각각 학습된 Kanana-1.5 2.1B 중 가장 높은 점수인 모델을 사용해 문서 수준 TTP 후보를 생성하고, TRAM2 및 AnnoCTR에서 각각 학습된 CySecBERT 모델을 통해 후보를 검증하였다.

그림 3과 4는 각각 TRAM2와 AnnoCTR에서 모델별 성능을 나타내며, 표 7은 제안한 하이브리드 방식과 단일 CySecBERT 모델, 기존 연구의 비교 모델 간 성능을 나타낸다.

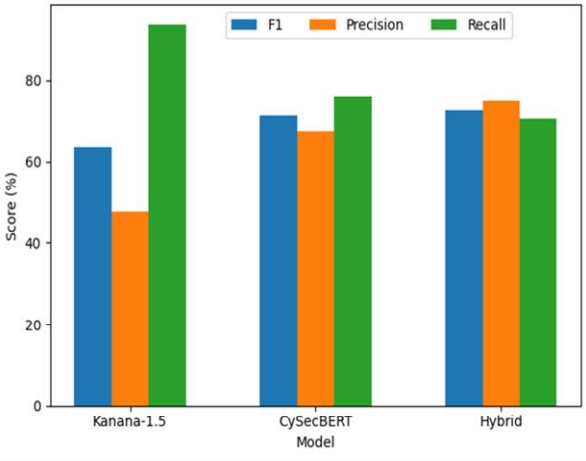

그림 3. 모델 성능 비교 (TRAM2)
Fig. 3. Model performance comparison (TRAM2)

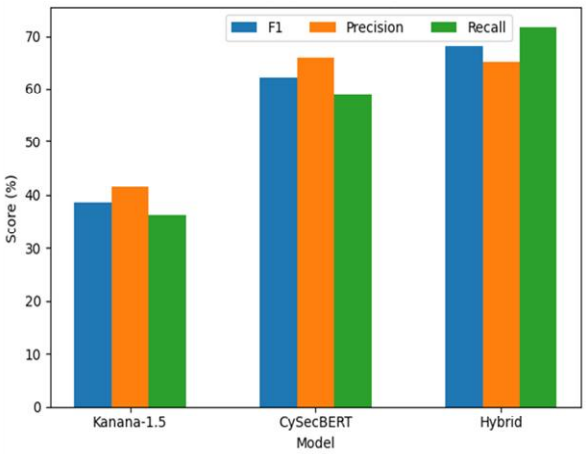

그림 4. 모델 성능 비교 (AnnoCTR)
Fig. 4. Model performance comparison (AnnoCTR)

실험 결과, 제안한 하이브리드 방식은 두 데이터셋 모두에서 단일 분류 모델 대비 성능 향상을 보였다. TRAM2에서는 Micro-F1이 70.11에서 74.62로 향상되었고, AnnoCTR에서는 62.12에서 63.64로 향상되었다.

표 7. 하이브리드 후보 생성 성능 비교
Table 7. Hybrid candidate generation performance comparison

Evaluation dataset	Model	Precision	Recall	F1
TRAM2	Hybrid	72.34	74.78	74.62
	CySecBERT	66	74.78	70.11
	RoBERTaLarge[3]	77.59	68.39	70.55
AnnoCTR	Hybrid	82.64	51.74	63.64
	CySecBERT	65.77	58.85	62.12
	CySecBERT[3]	60.75	67.86	62.75

특히 AnnoCTR에서는 Precision이 크게 개선되었는데, 이는 생성 단계에서 구성된 후보 집합을 분류 단계의 임계값 기반 필터링으로 재검증함으로써 최종 예측 집합이 보다 보수적으로 형성된 결과로 해석된다. 기존 연구와의 비교에서도 제안 방식의 특징을 확인할 수 있다. TRAM2에서 기존 RoBERTaLarge 기반 모델은 Precision이 77.59로 높았으나 Recall이 68.39에 머물러 F1은 70.55를 기록하였다. 반면 본 연구의 하이브리드 방식은 Precision은 72.34로 다소 낮지만 Recall을 74.78까지 확보하여 F1 74.62를 달성하였다. 이는 제안 방식이 특정 라벨에 대해 지나치게 보수적으로 예측하기보다, 후보 생성과 검증 과정을 통해 Precision과 Recall의 균형을 개선했음을 의미한다. AnnoCTR에서는 기존 CySecBERT 기반 결과와 비교하여 F1 향상 폭은 제한적이었으나, Precision이 크게 향상되었다. 따라서 본 연구의 개선 효과는 단순한 Recall 증가보다는 허위 양성 후보를 줄이는 Precision 중심 개선에 더 뚜렷하게 나타난다. 본 연구에서는 하이브리드 구조에서 임계값 tau가 최종 성능에 미치는 영향을 확인하기 위해 threshold 변화에 따른 성능 변화를 분석하였다. 그림 5와 그림 6은 각각 AnnoCTR과 TRAM2에서 tau 변화에 따른 F1 점수 변화를 나타낸다.

전반적으로 임계값이 낮을수록 더 많은 후보가 최종 라벨로 채택되고, 임계값이 높아질수록 예측이 보수적으로 수행되는 경향을 보였다. 본 연구에서는 이러한 변화 양상을 고려하여 Precision과 Recall의 균형이 비교적 안정적인 tau = 0.5를 기본 임계값으로 사용하였다. 추가적으로 표 8은 하이브리드 구조의 구성 요소별 기여도를 분석한 어블레이션 실험 결과이다.

LLM 단독 방식은 Recall은 높지만 Precision이 낮아 불필요한 후보를 다수 생성하는 경향을 보였다. 반면 Full Hybrid 방식은 LLM의 후보 생성 결과를 CySecBERT가 재검증함으로써 Precision을 크게 개선하였고, 결과적으로 두 데이터셋 모두에서 가장 안정적인 F1 성능을 보였다.

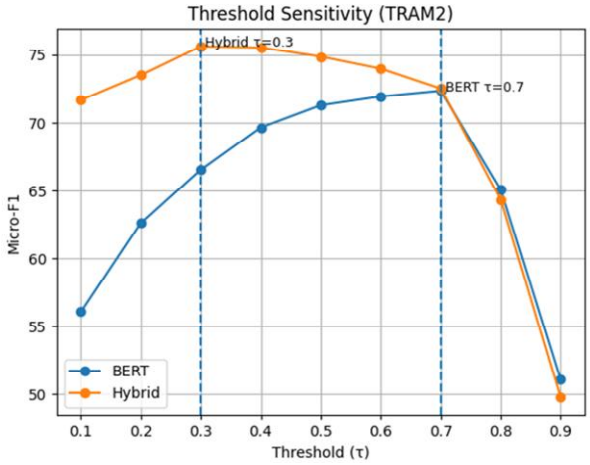

그림 5. Tau에 따른 점수 변화(TRAM2)
Fig. 5. Score Variation with Respect to τ (TRAM2)

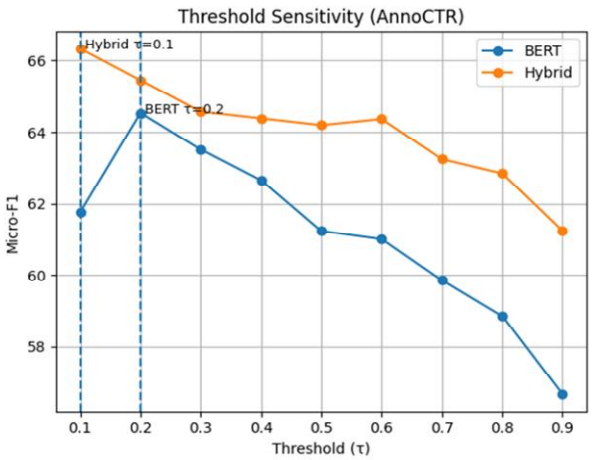

그림 6. Tau에 따른 점수 변화(AnnoCTR)
Fig. 6. Score Variation with Respect to τ (AnnoCTR)

표 8. 하이브리드 구조 어블레이션 실험
Table 8. Ablation analysis of the hybrid framework

Setting	TRAM2 precision	TRAM2 recall	TRAM2 F1	AnnoCTR F1
CySecBERT only	63.1	81.86	71.27	61.23
LLM only	47.33	93.2	62.78	50.19
LLM candidates + rule filtering	44.46	96.6	60.89	49
Full hybrid	72.3	77.62	74.86	63.83

4.4 모델 성능 비교 및 개선 전략

본 연구의 성능 향상은 크게 두 가지 요인에서 기인한다. 첫째, 데이터셋별 최적 정제 전략을 적용함으로써 문서 내 불필요한 패턴성 노이즈를 줄이고, 모델이 핵심 의미 정보에 더 집중할 수 있도록 하였다. 둘째, 생성 기반 후보 확장과 분류 기반 정밀 검증을 결합한 하이브리드 구조를 통해 단일 모델이 갖는 구조적 한계를 보완하였다.

TRAM2에서는 규칙 기반 문장 제거가 가장 효과적이었으며, 이는 해당 데이터셋에 상대적으로 패턴성 노이즈가 많이 포함되어 있음을 시사한다. 반면 AnnoCTR에서는 URL 제거만으로도 가장 좋은 성능을 보였는데, 과도한 정제가 오히려 의미 단서를 제거할 가능성이 있음을 보여준다. 이러한 결과는 데이터 정제 전략이 모델 외부의 보조적 요소가 아니라, 데이터셋 특성을 반영하여 조정되어야 하는 핵심 설계 변수임을 의미한다.

또한 하이브리드 구조는 입력을 바탕으로 가능한 TTP 후보를 생성한 뒤 판별적으로 재검증하는 절차를 통해, 문서 수준 다중 라벨 추출에서 Precision과 Recall 간의 trade-off 양상을 보이며 성능 향상을 달성할 수 있음을 보여준다. 특히 단일 분류기가 상대적으로 보수적인 예측을 수행하는 상황에서, 생성 기반 후보 생성 단계는 문맥적으로 관련 가능한 라벨 공간을 구성하며, 분류 단계에서의 판별을 위한 후보 집합을 제공하는 역할을 수행하였다.

4.5 분석

표 9는 반복 실험에서 얻은 F1 점수의 평균과 표준편차를 통해 성능 향상 폭의 안정성을 확인한 결과이다.

표 9. 향상폭 안정성 검증표
Table 9. Stability analysis of performance improvements

Dataset	Model	Mean	Std
TRAM2	CySecBERT	69.71	1.15
	Hybrid	73.98	0.97
AnnoCTR	CySecBERT	60.9	1.17
	Hybrid	63.24	0.42

제안 방식의 성능 향상이 특정 난수 초기화에 의존하는지를 확인하기 위해, 동일한 데이터 분할과 전처리 조건에서 seed를 달리한 반복 실험을 수행하였다. 각 실험에서는 모델 학습 초기화와 데이터 로딩 순서를 동일한 seed로 고정하였으며, 최종 성능은 Micro-F1의 평균과 표준편차로 보고하였다.

TRAM2에서는 Hybrid 모델의 평균 F1이 73.98로 CySecBERT의 69.71보다 높았고, AnnoCTR에서도 Hybrid 모델의 평균 F1이 63.24로 CySecBERT의 60.90보다 높았다. 또한 Hybrid 모델의 표준편차가 두 데이터셋 모두에서 더 낮게 나타나, 제안 방식이 단일 분류 모델보다 비교적 안정적인 성능을 보였음을 확인하였다.

표 9의 반복 실험 결과는 제안한 하이브리드 구조가 단일 CySecBERT 모델 대비 평균적으로 높은 F1을 보이며, 일부 성능 향상이 특정 seed에만 의존하지 않음을 보여준다. 다만 이러한 분석 결과에도 불구하고, 본 연구는 몇 가지 한계도 함께 가진다. 첫째, 본 실험은 TRAM2와 AnnoCTR이라는 두 공개 데이터셋에 기반하고 있으므로, 실제 산업 현장의 CTI 보고서가 가지는 문체 다양성과 작성 관행을 충분히 반영하지 못할 수 있다. 따라서 향후에는 실제 조직 환경에서 수집된 CTI 문서를 포함한 외부 검증이 필요하다. 둘째, 제안한 하이브리드 구조는 후보 생성 단계에서 입력 정보를 활용하여 TTP 후보 집합을 생성한다는 점에서 구조적 특징을 가지지만, 저빈도 TTP에 대한 예측 성능이 충분히 개선되었다고 보기는 어렵다. 특히 희소 라벨에 대해서는 후보 생성 부분의 개선과 이후의 정밀 검증 기준 및 threshold 설정 전략이 추가적으로 정교화될 필요가 있다. 셋째, LLM을 포함하는 추론 구조는 단일 분류 모델에 비해 계산 비용이 증가한다. 본 연구에서는 Kanana-1.5 2.1B를 사용하여 비교적 안정적으로 실험을 수행하였으나, 더 큰 모델을 사용할 경우 메모리 사용량과 추론 시간이 크게 증가할 수 있다. 따라서 대규모 실시간 CTI 처리 환경에 적용하기 위해서는 경량화, 모델 캐싱, 배치 추론 최적화 등의 후속 연구가 필요하다.

V. 결 론

본 연구에서는 문서 단위 다중 라벨 환경에서

CTI 보고서로부터 MITRE ATT&CK 기반 TTP를 자동 추출하기 위해, 데이터 정제 전략과 하이브리드 LLM-BERT 프레임워크를 함께 고려한 접근을 제안하였다. TRAM2와 AnnoCTR 데이터셋을 대상으로 URL, IOC, 메타데이터, 규칙 기반 문장 제거 등 다양한 정제 전략을 비교한 결과, 데이터셋마다 최적의 정제 방식이 상이함을 확인하였다. 이는 CTI 데이터 전처리가 단순한 보조 절차가 아니라, 문서 수준 TTP 추출 성능에 직접적인 영향을 주는 핵심 설계 요소임을 보여준다.

또한 Kanana-1.5 2.1B를 이용한 후보 생성과 CySecBERT를 이용한 정밀 검증을 결합한 하이브리드 구조를 통해, 단일 분류 모델 대비 두 데이터셋 모두에서 성능 개선 경향을 확인하였다. 특히 LLM 기반 후보 생성과 분류 기반 검증의 결합은 문서 수준 다중 라벨 예측에서 성능을 보다 안정적으로 확보하는 데 유효한 방향임을 보여주었다.

향후 연구에서는 희소 TTP에 대한 예측 성능을 보다 정밀하게 분석하고, 실제 CTI 운영 환경을 반영한 외부 데이터셋 검증, 경량화된 추론 구조 설계, 문서 길이 및 라벨 빈도에 따른 세분화 성능 평가를 수행할 필요가 있다. 이를 통해 제안한 접근을 실무형 CTI 자동 분석 시스템으로 확장할 수 있을 것으로 기대한다.

References

[1] P. Santos, R. Abreu, M. J. C. S. Reis, C. Serôdio, and F. Branco, "A Systematic Review of Cyber Threat Intelligence: The Effectiveness of Technologies, Strategies, and Collaborations in Combating Modern Threats", Sensors, Vol. 25, No. 14, Article No. 4272, Jul. 2025. https://doi.org/10.3390/s25144272.

[2] C. Meng, Z. Jiang, Q. Wang, X. Li, C. Ma, F. Dong, F. Ren, and B. Liu, "Instantiating Standards: Enabling Standard-Driven Text TTP Extraction with Evolvable Memory", arXiv preprint, arXiv:2505.09261, May 2025. https://doi.org/10.48550/arXiv.2505.09261.

[3] M. Büchel, T. Paladini, S. Longari, and M. Carminati, "SoK: Automated TTP Extraction from CTI Reports-Are We There Yet?", Proc. USENIX Security Symposium, Seattle, WA, USA, pp. 4621-4641, Aug. 2025.

[4] N. Rani, B. Saha, V. Maurya, and S. K. Shukla, "TTPXHunter: Actionable Threat Intelligence Extraction as TTPs from Finished Cyber Threat Reports", arXiv preprint, arXiv.2403.03267, Mar. 2024. https://doi.org/10.48550/arXiv.2403.03267.

[5] Y. Park and W. You, "A Pretrained Language Model for Cyber Threat Intelligence", Proc. 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Singapore, pp. 113-122, Dec. 2023. https://doi.org/10.18653/v1/2023.emnlp-industry.12.

[6] Y. Zhang, X. Zhou, H. Wen, W. Niu, J. Liu, and H. Wang, "Tactics, Techniques, and Procedures in Interpreted Malware: A Zero-Shot Generation with Large Language Models", arXiv preprint, arXiv:2407.08532, Jul. 2024. https://doi.org/10.48550/arXiv.2407.08532.

[7] S. Baeg, J. Cho, T. Eo, S. Jung, J. Kang, E. Kim, and K. K. Team, "Kanana: Compute-Efficient Bilingual Language Models", arXiv preprint, arXiv:2502.18934, Feb. 2025. https://doi.org/10.48550/arXiv.2502.18934.

[8] I. Keraghel, S. Morbieu, and M. Nadif, "Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study", arXiv preprint, arXiv:2401.10825, Jan. 2024. https://doi.org/10.48550/arXiv.2401.10825.

[9] L. Yuan, X. Xu, S. Sun, H. Yu, Y. Wei, and J. Zhou, "Research of Multi-Label Text Classification Based on Label Attention and Correlation Networks", PLoS ONE, Vol. 19, No. 10, Art. no. e0311305, Oct. 2024. https://doi.org/10.1371/journal.pone.0311305.

[10] M. Bayer, P. Kuehn, R. Shanehsaz, and C. Reuter, "CySecBERT: A Domain-Adapted

Language Model for the Cybersecurity Domain", arXiv preprint, arXiv:2212.02974, Dec. 2022. https://doi.org/10.48550/arXiv.2212.02974.

[11] S. S. Joo, S. H. Lee, J. H. Yu, D. S. Moon, and J. H. Bae, "Study on a Convolutional Neural Network Model for Attack Type Classification in the Field of Network Intrusion Detection", Journal of Korean Institute of Information Technology, Vol. 22, No. 7, pp. 45-54, Jul. 2024. https://doi.org/10.14801/jkiit.2024.22.7.45.

[12] L. Lange, M. Müller, G. Haratinezhad Torbati, D. Milchevski, P. Grau, S. Pujari, and A. Friedrich, "AnnoCTR: A Dataset for Detecting and Linking Entities, Tactics, and Techniques in Cyber Threat Reports", arXiv preprint, arXiv:2404.07765, Apr. 2024. https://doi.org/10.48550/arXiv.2404.07765.

[13] Center for Threat-Informed Defense, "Threat Report ATT&CK Mapper", GitHub, 2020. https://github.com/center-for-threat-informed-defense/tram. [accessed: Apr. 20, 2026].

[14] N. Pangakis and S. Wolken, "Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels", arXiv preprint, arXiv.2406.17633, Jun. 2024. https://doi.org/10.48550/arXiv.2406.17633.

[15] Y. Lu and K. Smith, "Feeding LLM Annotations to BERT Classifiers at Your Own Risk", arXiv preprint, arXiv.2504.15432, Apr. 2025. https://doi.org/10.48550/arXiv.2504.15432.

[16] M. L. Zhang and Z. H. Zhou, "A Review on Multi-Label Learning Algorithms", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 8, pp. 1819-1837, Aug. 2014. https://doi.org/10.1109/TKDE.2013.39.

저자소개

김 현 석 (Hyeonseok Kim)

2022년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터공학과 학사과정

관심분야 : LLM 최적화, 자동분류

정 유 철 (Yuchul Jung)

2011년 2월 : 한국과학기술원 (KAIST) 전산학과(공학박사)

2009년 1월 ~ 2013년 7월 : 한국전자통신연구원 (ETRI) 선임연구원

2013년 8월 ~ 2017년 8월 : 한국과학기술정보연구원(KISTI) 선임연구원

2017년 8월 ~ 2022년 2월 : 국립금오공과대학교 컴퓨터공학과 조교수

2022년 2월 ~ 현재 : 국립금오공과대학교 인공지능공학과 부교수

관심분야 : 거대언어모델, 자연어처리, 지식그래프, 한국어 음성 인식/합성, AI응용, AI기반 CTI, 배터리 SOH예측