미래사회
Journal of Future Society
2026. Vol. 17, No. 1, 176~193.
https://doi.org/10.22987/jifso.2026.17.1.176

LLM 기반 CTI 자동 분석 평가 프레임워크 제안*

유 진 호** **조 효 진*****

연세대학교

최근 개인정보 유출과 랜섬웨어 등 사이버 위협이 증가함에 따라 Cyber Threat Intelligence (CTI) 분석의 중요성이 커지고 있다. 그러나 CTI 보고서는 비정형 자연어로 작성되어 수동 분석에 한계가 있으며, 이를 해결하기 위해 대규모 언어 모델(LLM)을 활용한 자동 분석 연구가 활발히 진행되고 있다. 본 연구는 2023~2025년 발표된 주요 연구를 분석하여, 기술 흐름이 단순 객체 추출에서 지식 구조화 및 운영 산출물 생성으로 확장되고 있음을 확인하였다. 하지만 기존 연구들은 서로 다른 데이터셋과 평가 지표를 사용해 정량적 비교가 어렵고, 환각 문제와 근거 추적성 부족으로 실무 적용에 제약이 있다. 이에 본 논문은 LLM 기반 CTI 분석 기술을 객관적으로 평가하기 위한 CTI 자동 분석 평가 프레임워크를 제안한다. 제안 프레임워크는 단계별 태스크 번들, 산출물 표준화, 근거 추적성 평가, 운영 효율성 지표, 그리고 다양성을 고려한 데이터셋 구축 원칙을 포함하며, 이를 통해 신뢰할 수 있는 자동 위협 분석 기술의 발전에 기여하고자 한다.

주요어: 사이버 위협 인텔리전스, 대규모 언어 모델, 위협 지식 구조화, 탐지 정책 자동 생성, 벤치마크 프레임워크

---

* 본 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발).
** 주저자: 유진호/연세대학교 정보대학원 정보보호 대학원생/서울 서대문구 연세로 50 새천년관 B114호/Tel: 02-2123-4526/E-mail: yujinho000@yonsei.ac.kr
*** 교신저자: 조효진/연세대학교 정보대학원 정보보호 부교수/서울 서대문구 연세로 50 새천년관 403호/Tel: 02-2123-4526/E-mail: hyojin.jo@yonsei.ac.kr

Ⅰ. 서론

ICT(Information & Communication Technology) 기술의 발전과 디지털 전환이 빠르게 가속화됨에 따라, 조직들이 보호하고 관리해야 할 자산이 증가하고 있다. 이에 따라, 다양한 공격 표면(Attack Surface)이 확장되고 사이버 보안 위협이 증가하고 있다. 또한, 생성형 AI(e.g., ChatGPT)의 확산은 공격 도구 제작 및 변형의 진입 장벽을 낮추어 공격자들은 손쉽게 공격을 생성하고 고도화·변형시켜서 확산할 수 있게 되었다. AhnLab에서 발표한 '2025년 사이버 위협 동향 및 2026년 전망 보고서'에서도 이러한 위협에 대해서 경고한 바 있다(AhnLab, 2025).

이처럼 보호해야 할 자산이 증가하고, 고도화되는 공격에 대응하기 위해서 조직은 수동적 대응 방식을 넘어 사전에 위협을 식별·차단하는 능동적 대응 전략(e.g., Threat Hunting)을 채택하고 있다.

수동적 대응 전략은 조직이 외부 환경 변화나 위협이 발생한 이후에 대응을 수행하는 전략으로 사후 대응 중심의 전략이다. 이에 반해, 능동적 대응 전략은 조직이 환경 변화와 발생할 수 있는 위협을 예측하고, 선제적으로 행동하여 사전에 위협을 예방하고 위협을 최소화하려는 전략이다. 이는 사전 대응 중심의 전략이다.

능동적 대응 전략을 위해서는 위협을 사전에 인지·예측하고 선제적으로 대응하기 위해서 공격자의 의도, 능력, 공격 방식, 위협의 발생 가능성 등에 대한 정보가 필수적이다. 이러한 정보를 구조화하여 제공하는 것이 Cyber Threat Intelligence(CTI) 기술이다.

CTI는 단순한 보안 이벤트, 로그 정보가 아니라 사이버 위협과 관련된 데이터를 수집·분석·가공하여 의사결정에 활용할 수 있는 지식으로 구조화하여 제공하는 정보를 의미한다. 이를 통해 조직을 대상으로 발생할 수 있는 위협을 식별하고, 대응 방향과 우선순위를 결정하는 기준 자원으로 활용한다.

1.1. 연구 배경

능동적 대응 전략에서 CTI는 위협을 사전에 인지하고 대응 방향을 결정하는 핵심 자원으로 활용되고 있으나, 실제 조직에서 CTI를 활용하는 것은 여러 구조적 한계에 직면해 있다.

현재 CTI의 상당수는 보안 벤더, 기관, 커뮤니티 등에 의해 생성된 비정형 자연어 기반 문서 형태로 제공된다. 이러한 CTI 보고서는 공격 캠페인, 공격자의 전술·기술·절차(Tactics, Techniques, and Procedures; TTPs), 취약점 정보 등이 자연어로 서술되어 있어 조직이 대응해야 할 위협인지 바로 판단하기 어렵고, 시스템에 즉시 활용하기 어렵다.

CTI 분석과 활용 과정은 여전히 보안 전문가의 수동 분석에 크게 의존하고 있다. 보안 분석가는 다수의 CTI 보고서를 검토하고, 그 중 조직에 관련된 위협을 선별하여 탐지 정책, 대응 방안, 위협 헌팅 시나리오로 정제해야 한다. 그러나 이러한 방식은 분석가의 지식과 경험, 숙련도에 의해 편향이 생길 수 있다. 또한, 한정된 보안 인력과 자원으로 새롭게 보고되는 방대한 양의 CTI 보고서를 모두 검토하고 반영하기 어렵기 때문에, 이는 결과적으로 최신성 저하와 대응 공백으로 이어질 수 있다.

이러한 한계는 양질의 CTI 보고서가 공개되어 있어도, 조직의 능동적 대응 전략에서 CTI가 충분히 활용되지 못하는 결과로 이어진다. 따라서 비정형 CTI를 효율적으로 해석하고 구조화하여, 제한된 인력과 자원에서도 실질적으로 활용할 수 있는 방안에 관한 연구가 요구된다.

이러한 문제의식하에, CTI를 자동으로 분석하여 핵심 Entity를 추출하고 이를 운영자에게 제공하고자 하는 연구가 꾸준히 수행되어 왔다. 특히 최근에는 Large Language Model (LLM)의 자연어 이해·추론 능력을 활용하여 CTI를 자동으로 분석하고 구

조화하려는 연구가 활발히 진행되고 있다. 이러한 LLM 기반 접근은 AI 기술이 단순한 자동화 도구를 넘어 조직의 운영 방식과 전략적 의사결정을 재구성하는 핵심 요소로 작용한다는 논의와도 맞닿아 있다(이완형, 2024).

선행 연구들은 LLM이 비정형 CTI 보고서로부터 공격 행위나 공격자의 TTP 등의 정보를 추출하거나, 이를 MITRE ATT&CK(MITRE, 2020)과 같은 표준 프레임워크에 매핑하는 데 활용될 수 있음을 보여준다. 이러한 연구들은 비정형 CTI를 자동으로 구조화할 수 있다는 가능성을 제시한다.

1.2. 연구 목적 및 기여

본 연구는 조직의 능동적 대응 전략에서 핵심 자원으로 활용되는 CTI가 비정형성 때문에 보안 운영 환경에 즉시 활용되기 어렵다는 점에 주목한다. 이를 해소하기 위해 제안된 LLM 기반 CTI 자동 분석·구조화 기법에 관한 선행 연구를 수집하고 체계적으로 정리하고자 한다. 또한, 대표적인 활용 시나리오를 기반으로 요구사항 분석을 수행함으로써, 조직 환경에서 CTI 자동 분석 파이프라인이 지속 가능하게 운영 가능한지 탐구한다.

이에 본 연구는 선행 연구들에 대한 체계적인 비교·분석을 통해 현재 기술의 한계를 식별하고, 이를 극복하기 위한 CTI 자동 분석 평가 프레임워크를 제안하는 것을 주된 목적으로 한다. 이를 위해 본 연구는 2023~2025년 발표된 대표 연구들을 대상으로 (i) LLM 활용 방식, (ii) 산출물 유형, (iii) 검증 데이터셋 및 평가 방식, (iv) 적용 가능성 관점에서 심층적인 비교를 수행한다.

분석 결과, 현재 기술은 표준화된 평가 기준 부재로 인해 객관적 검증이 어렵다는 한계를 식별하였다. 이에 본 연구는 향후 연구들이 동일한 기준에서 성능을 검증하고 발전할 수 있도록 단계별 태스크 번들(Task Bundle), 산출물 표준화, 근거 추적성 및 운영 효율성 평가를 포함하는 평가 프레임워크를 핵심 과제로 제시한다. 이는 궁극적으로 조직이 신뢰할 수 있는 자동화된 위협 대응 체계를 구축하는 데 기여할 수 있을 것으로 기대된다.

1.3. 연구 방법 및 범위

본 연구는 LLM을 활용하여 비정형 CTI 보고서를 자동으로 해석·구조화하는 선행 연구들을 분석하고, 이를 바탕으로 객관적 성능 검증을 위한 CTI 자동 분석 평가 프레임워크를 도출하는 것을 범위로 한다. 문헌 수집은 arXiv 및 주요 해외 학술/학회 프로시딩에서 (“Cyber Threat Intelligence” OR “CTI”) AND (“Large Language Model” OR “LLM”) 키워드를 기반으로 1차 후보군을 수집하고, 제목·초록을 기준으로 CTI 보고서를 LLM을 통해 자동으로 분석하여 구조화된 지식과 산출물을 추출하는 파이프라인을 제안한 연구만을 포함하도록 선별하여 10편의 논문을 심층 분석을 수행하였다. 본 연구는 이러한 선행 연구 분석을 통해 현재 기술의 평가 방식의 문제점을 식별하고, 이를 해결하기 위한 평가 프레임워크와 데이터셋 구축 원칙을 제안한다.

1.4. 논문 구성

본 논문은 다음과 같이 구성된다. 1장에서는 연구 배경과 필요성을 제시하고 연구 목적 및 범위를 정의한다. 2장에서는 관련된 연구를 이해하기 위한 배경 지식인 CTI와 CTI 표현 모델, LLM 기반 문서 해석 및 분석 품질을 향상시키는 다양한 기법을 정리한다. 3장에서는 CTI 자동 분석의 적용 시나리오를 제시한다. 4장에서는 선별한 10편의 연구를 분류 체계에 따라 비교·분석하며, 데이터셋·평가·운영 관점에서 차이점을 정리한다. 5장에서는 기술 동향 변화와 선행 연구들의 한계점을 논의하고 이를 극복하기 위해, 단계별 태스크 번들과 평가 지표

를 포함하는 CTI 자동 분석 평가 프레임워크를 제안하고 구체적인 구성 요소를 설명한다. 마지막으로 6장에서는 결론을 통해 연구 결과를 요약한다.

Ⅱ. 이론적 배경

본 장에서는 LLM 기반 CTI 자동 분석 기술 동향을 이해하는 데 필요한 이론적 배경을 정리한다. 먼저 CTI의 정의와 활용 목적에 관해 설명하고 이를 구조화·공유하기 위한 대표적인 산출물의 특성과 차이를 기술한다. 이후 비정형 문서 기반 CTI를 자동 분석하는 과정에서 활용되는 LLM 기법의 개념을 정리한다.

2.1. CTI(Cyber Threat Intelligence)

CTI는 조직이 사이버 위협을 식별(Identify), 평가(Assess), 감시(Monitor), 대응(Respond)하는 데 활용할 수 있는 정보 전반을 의미한다. CTI는 침해사고 및 공격 캠페인 분석 결과, 침해 지표(Indicator of Compromise; IoC), 공격자의 TTP, 탐지·차단·완화에 필요한 권고 조치 등 다양한 정보를 포함하고 있다.

CTI는 단순한 정보의 나열 및 조합이 아니라, 수집된 데이터가 의사결정에 활용될 수 있을 만큼 해석·정리된 지식이다. 또한, CTI는 여러 출처에서 생성되며, 공유를 통해 상호 보완될 때 더욱 풍부한 지식이 된다.

2.2. CTI 표현 모델 및 산출물

CTI는 다양한 출처(기관 보고서, 벤더 블로그 등)에서 비정형 텍스트로 생성되는 경우가 많아 교환·연계·재사용을 위해서 표준화된 표현 모델 또는 구조화 산출물이 필요하다. 대표적으로 Structured Threat Information Expression (STIX), 지식 그래프, MITRE ATT&CK 기반 TTP 표현이 널리 활용되고 있다.

STIX는 사이버 위협 및 관측 정보를 표현하기 위한 표준으로, 위협 행위자·캠페인·악성코드·인프라 등을 객체(Entity)로 정의하고, 이들 사이의 연관을 관계(Relationship)로 기술함으로써 일관된 스키마를 제공한다. 이러한 표준화는 기관/제품 간 상호운용성과 재사용성을 높여, CTI 공유와 활용을 가능하게 한다.

지식 그래프는 객체와 관계를 노드(Node)/엣지(Edge)로 모델링하여 CTI 보고서에 산개된 지식을 질의·추론 가능한 형태로 구성한다. STIX가 상호운용을 위한 표준 스키마 중심인 데 반해, 지식 그래프는 분석 목적에 따라 온톨로지, 스키마를 유연하게 설계할 수 있고 관계 예측, 경로 질의 등의 그래프 분석 기법과 결합하기 용이하다. 다만, 지식 그래프는 스키마 설계와 객체의 정규화, 관계의 일관성 등 품질 관리가 중요하며, 자동 추출 단계의 오류는 그래프 전체 품질로 전이될 수 있다.

MITRE ATT&CK 프레임워크의 TTP는 실제 관측 기반 공격자의 행위를 전술(Tactic)–기술(Technique)–절차(Procedures)로 정리한 지식 베이스로, 보안 운영에서 행위 기반 분석과 탐지 설계의 공통언어로 활용된다. 전술은 공격의 목표(why), 기술/세부기술은 그 목표를 달성하기 위한 방법(how)에 해당하며, CTI 보고서를 TTP로 매핑하면 서로 다른 CTI 간 공격자의 행위 및 위협 요소들을 비교·재사용하기 쉬워진다.

2.3. LLM 기반 문서 해석

비정형 CTI 보고서는 문서의 길이와 구조가 일관되지 않고, 동일 개념이 다양한 표현으로 기술되며, 기술적 세부 용어가 혼재한다. 따라서 문서 해석–정보 추출–정규화/구조화 과정을 자동화하기 위해 최근 LLM이 활용되고 있다. 위와 같은 파이프

라인의 성능을 향상시키기 위해 대표적으로 다음과 같은 LLM 기법들이 적용된다.

In-Context Learning(ICL)은 모델의 파라미터를 업데이트하지 않고, 프롬프트에 작업 목표와 예시를 포함해 모델이 그 문맥 안에서 해당 과제를 수행하도록 유도하는 방식이다(Brown, 2020). 예시를 제공하지 않고 지시문만으로 수행하는 경우를 Zero-shot이라고 하며, 예시를 1개 제공하면 One-shot, 여러 개 제공하면 Few-shot이라 한다. Zero-shot은 구현이 단순하지만, 출력 형식이 불안정해 작업 목표가 모호한 경우 성능이 저하될 수 있다. One-shot은 단일 예시를 통해 출력 스타일과 형식을 고정할 수 있지만, 출력 편향이 발생할 수 있고 Few-shot은 다양한 예시를 통해 안정성을 높일 수 있지만, 프롬프트 길이가 증가하면서 비용과 지연이 커질 수 있다. ICL의 성능은 예시 개수 자체보다는 예시의 대표성과 작업 목표의 명확성에 의해 크게 좌우될 수 있다.

프롬프트 기반 기법에서 또 다른 중요한 요소는 프롬프트 설계와 제약 조건 부여이다. 프롬프트는 단순한 질문이 아니라 수행해야 할 작업과 출력 요구조건을 규정하는 인터페이스로 기능할 수 있으며, 지시문(Instruction), 역할(Role), 출력 형식(Output Format), 제약(Constraint) 등으로 구성되는 경우가 많다. 특히 Prompt Constraint는 모델의 자유도를 의도적으로 줄여 출력의 안정성과 정확성을 높이려는 방법으로 자주 활용되고 있다.

Chain-of-Thought(CoT)는 결론만 제공하도록 요구하는 대신, 문제 해결 과정을 단계적으로 유도하여 복잡한 작업의 정확도를 높이기 위한 기법이다(Wei, 2022). 특히, 문장 간 근거 관계를 따라가며 판단해야 하는 경우 관련 문장 식별 - 근거 요약 - 결론 도출과 같은 중간 단계를 명시하여 오류의 가능성을 낮출 수 있다.

Retrieval-Augmented Generation(RAG)은 모델이 답변을 생성하기 이전에 외부 지식 저장소에서 관련 정보를 검색하여 컨텍스트로 주입하고, 그 범위 내에서 생성하도록 하는 방식이다. RAG는 일반적으로 문서를 적절한 단위로 분할(Chunking)하고 인덱싱한 뒤, 질의에 대한 관련 문서를 검색하고 필요시 재정렬한다. 해당 검색 및 정렬 결과를 프롬프트에 포함시켜 생성하는 파이프라인으로 구성된다. 모델의 사전 학습 과정에서 충분히 반영되지 못한 최신 정보나 도메인 지식을 보완할 수 있고, 답변이 어떤 문서에 기반해 생성되었는지 연결할 수 있다는 장점이 존재한다(Lewis, 2020).

Fine-tuning은 특정 작업·도메인 데이터로 모델의 파라미터를 업데이트하여 관련 작업에 대한 성능과 출력 형식의 일관성을 높이는 기법이다. 프롬프트 기반 기법만으로는 출력이 불안정하거나, 동일 입력에 대해 결과 변동성이 큰 경우 주로 활용된다. Fine-tuning은 모델이 해당 작업에서 요구하는 표현과 데이터, 판단 기준을 강하게 반영할 수 있어 분류·추출·변환과 같이 정답 형식이 비교적 명확한 작업에서 안정적인 성능 개선을 기대할 수 있다.

Ⅲ. 활용 분야 및 시나리오 분석

본 장에서는 LLM 기반 CTI 자동 분석 기술이 실제 보안 운영에서 적용될 수 있는 대표 시나리오를 두 가지로 구분하고, 각 시나리오에서 요구되는 품질 및 운영 요구사항을 분석한다.

3.1. CTI 표준 기반 지식 공유

비정형 CTI 보고서가 운영 환경에서 반복적으로 재사용되고 공유되기 위해서는, 자연어 기반 텍스트에서 공유 및 연계할 수 있는 표준 형태로 변환되어야 한다. 대표적으로 STIX는 사이버 위협 및 관측 정보를 표현하기 위한 구조화된 언어로, 위협 요소를 객체로 정의하고 객체 간 관계를 통해 지식을 구조화하도록 설계되어 있다. 또한, Trusted

Automated Exchange of Intelligence Information (TAXII)는 조직 간 CTI를 교환하기 위한 RESTful API 기반의 전송 표준을 제공한다.

LLM 기반 자동 분석은 보고서의 텍스트에서 핵심 객체를 추출하고 객체 간 관계를 식별해 STIX로 재구성할 수 있다. 이를 배포·동기화 가능한 형태(e.g., TAXII)로 변환하면, 결과적으로 CTI는 표준 객체 단위로 검색·공유가 가능한 지식으로 축적될 수 있다. 결과적으로 조직 내·외로 구조화된 위협 지식을 자동으로 공유하는 파이프라인으로 확장할 수 있다.

3.2. 운영 산출물 자동 생성 및 배포

최근 CTI 자동 분석 연구는 지식 구조화를 넘어, 실제 탐지 엔진에서 실행할 수 있는 운영 산출물(탐지 정책/쿼리) 생성으로 확장되고 있다. CTI 보고서를 기반으로 탐지 정책을 자동 생성하면, 탐지 정책 반영 시간을 단축하고 운영자의 반복 작업을 완화할 수 있다. 생성되는 산출물은 조직의 운영 스택에 따라 로그 기반 탐지 정책(e.g., Sigma)과 네트워크/호스트 기반 탐지 정책(e.g., Snort/YARA)으로 구분하여 적용할 수 있다.

LLM은 CTI에서 탐지에 활용할 수 있는 정보(e.g., 프로토콜, 포트, 식별할 수 있는 문자열 및 패턴)를 추출 및 정규화한 뒤 이를 각 정책 포맷의 구성 요소에 맞게 생성할 수 있다. 예를 들어 Snort 정책은 Rule Header와 Rule Option으로 구성되어 있다. 따라서 CTI에서 추출된 통신 특성을 Header와 Option의 적절한 필드에 배치함으로써 정책 초안을 자동으로 생성하는 파이프라인을 구축할 수 있다. 또한, 탐지 정책 포맷과 작성 규칙은 공식 문서와 저장소가 폭넓게 제공되므로, 이러한 문서/예시를 외부 지식으로 참조하여 생성 과정에서 문법 오류를 줄이고 포맷 정합성을 높일 수 있다.

Ⅳ. LLM 기반 CTI 분석 기술 비교

본 장에서는 LLM을 활용하여 비정형 CTI 보고서를 자동 분석하는 대표 연구들을 비교·분석한다. <표 1>은 각 연구의 입력, LLM 활용 기법, 환각 완화 전략 등 해당 연구가 제안한 방법론의 주요 특징에 대해서 정리하였다. <표 1>의 분류 내용을 보면, LLM 기반 CTI 자동 분석 연구들은 최종 산출물 유형에 따라서 (i) 지식 그래프(Knowledge Graph; KG)/표준 구조화 산출물 생성, (ii) TTP 추출 및 Intelligence 구조화, (iii) 운영 산출물(탐지 정책·쿼리) 자동 생성의 세 가지 유형으로 분류할 수 있다.

4.1. 지식 그래프 및 표준 구조화 산출물 생성

첫 번째 유형은 CTI 보고서를 입력으로 받아 객체(Entity)·관계(Relationship)를 추출하여 지식 그래프 또는 STIX와 같은 표준 구조로 변환하는 연구들이다. 대부분의 연구는 추출된 정보를 <주체, 관계, 객체> 형태의 Triplet으로 표현하며, 이를 연결하여 그래프 구조를 형성하거나 STIX 객체로 매핑한다. Triplet은 연구에서 정의한 스키마/온톨로지에 따라 달라질 수 있으며, 일부 연구는 텍스트의 문법적 구조를 그대로 따르는 주어–관계–목적어(Subject–Predicate–Object; SPO) 형태의 Triplet을 추출하는 경우도 있다. 예를 들어, “Attacker used PowerShell to download a payload”라는 문장에서 SPO Triplet은 <Attacker, used_ to_download, payload>와 같이 문장 의미를 반영하는 반면, 일반적인 Triplet은 <Attacker, uses, PowerShell>, <PowerShell, downloads, payload>와 같이 표준화된 관계로 분해되어 표현된다. 이를 통해 각 요소 및 객체 간 관계를 체계적으로 추출하여, 이후 검색·추론·이해 등 운영자의 후속 분석이 가능한 형태로 구조화하는 것을 목표로 한다.

<표 1> 제안된 LLM 기반 CTI 자동 분석 프레임워크 비교

관련 연구	LLM	LLM 활용	환각 완화 전략	도메인	구조화 산출물	운영 산출물
aCTIon	GPT-3.5 Turbo	Zero-shot	Constraint Prompting (Self-Validation)	Enterprise	STIX (KG)	-
(Liu, 2023)	ChatGPT	Zero-shot	Constraint Prompting (On-Topology)	Enterprise	KG	-
LLM-TIKG	Llama 2	Few-shot, Fine-tuning	Constraint Prompting (Limit Label-space)	Enterprise	TTP, KG	-
(Fieblinger, 2024)	Zephyr-7B	Few-shot, Fine-tuning	Constraint Prompting (On-Topology)	Enterprise	KG	-
IntelEX	GPT-4o-mini	Few-shot, RAG, GraphRAG	LLM-as-a-judge	Enterprise	TTPs	Sigma (Splunk)
TTPFshot	GPT-4o, Llama 3	Dynamic Few-shot	Constraint Prompting (Limit Label-space)	Enterprise	TTPs	-
FALCON	GPT-4o	Agentic (RAG)	Semantic Validator	Enterprise	-	Snort, YARA
(Dong, 2025)	Qwen 2.5	Fine-tuning	Constraint Prompting (Limit Label-space)	Enterprise	TTPs	-
CTINEXUS	Qwen 2.5	Dynamic Few-shot	Constraint Prompting (On-Topology)	Enterprise	KG	-
LLMCloudHunter	GPT-4o	Few-shot, Multimodal	Majority Voting	Cloud	API, Entity	Sigma (Splunk)

LLM은 실제 추론에 사용된 모델을 의미하며, 여러 모델을 평가한 연구의 경우 보고된 최고 성능 모델을 기재하였다.

Siracusano(2023)는 비정형 CTI 보고서가 표준화된 구조로 즉시 변환되지 않아 공유 및 재사용이 어렵다는 한계를 해결하기 위해, CTI 보고서를 운영자가 활용할 수 있는 STIX 번들 형태의 구조화된 CTI로 자동 변환하는 LLM 기반 파이프라인 aCTIon를 제안하였다. 보고서의 텍스트에서 Entity와 Relationship을 추출한 뒤 이를 STIX 객체(Domain/Relationship Object)로 매핑하고, Schema 제약을 만족하도록 정규화·검증 단계를 결합해, STIX 기반의 상호운용 가능한 구조화 산출물로 변환하는 데 초점을 둔다. 데이터셋은 실제 공개 CTI 보고서 204개와 이에 대응하는 STIX 번들을 3개의 독립적인 분석가 그룹이 수동으로 큐레이션한 결과를 사용하였다. 평가는 동일 벤치마크에서 선행 솔루션과 비교하였을 때 주요 Entity 추출에서 F1-score가 개선되는 결과를 보고하였다.

Liu(2023)는 CTI 보고서를 분석하여 공격자나 악성코드 같은 주요 Entity와 이들 간의 주어-관계-목적어 형태로 구조화한 SPO Triplet을 추출하여 지식 그래프로 변환하는 방법론을 제안하였다. STIX를 바탕으로 구성한 CTI 온톨로지에 맞춰 ChatGPT를 활용하여 Entity를 추출하고, 후보 SPO Triplet을 생성한 뒤, 후보 중 적절한 Triplet을 선별하는 파이프라인을 구축하였다. 평가는 13개의 CTI 보고서에 대해 전문가가 라벨링 한 데이터셋을 기준으로 수행되었으며, 그 결과 Entity 추출에서

F1-score 0.78, Triplet 추출에서 0.56을 기록하여 기존 모델 대비 성능이 향상되었음을 입증하였다.

Hu(2024)는 LLM이 직접 지식 그래프를 추출하는 경우 비용과 지연시간이 크고, 소형 모델을 학습하기엔 라벨 데이터가 부족하다는 한계를 지적하였다. 이를 해결하기 위해 GPT에 Few-shot 기법을 활용하여 라벨을 생성·증강해 학습 데이터를 구축한 뒤, Llama2-7B 모델을 Fine-tuning하여 지식 그래프를 효율적으로 구축하는 프레임워크인 LLM-TIKG을 제안하였다. 평가 데이터는 위협 인텔리전스 공유 플랫폼에서 크롤링한 CTI 콘텐츠 12,545건(블로그, 뉴스, 보안 분석 보고서 포함)을 기반으로 하며, Entity 추출 F1-score 0.85 수준을, TTP 분류는 약 0.98 수준을 보였다고 보고한다.

Fieblinger(2024)는 오픈 소스 LLM을 활용해 CTI 텍스트에서 Triplet을 추출하고 이를 기반으로 지식 그래프를 구축한 뒤, 운영자가 활용할 수 있는 링크 예측까지 확장해 검토하였다. Llama2, Mistral, Zephyr 등 오픈 소스 LLM을 대상으로 프롬프트 기반, 가이던스 기반, Fine-tuning 전략을 비교하여 최적의 조합을 선택하여 대규모 Corpus에서 지식 그래프를 생성한 뒤, 생성된 데이터가 사전에 정의된 스키마 규격에 부합하도록 자동으로 보정하는 후처리 프로세스를 적용하였다. 평가는 2015~2022년 기간의 36개의 악성코드 패밀리와 관련된 CTI 보고서 120개에서 선별한 768개의 문단과 공개된 CTI 보고서 약 12,000개를 활용해 평가한 결과 가이던스 기법에서 Triplet 추출 성능이 개선되었음을 보고하였다.

Cheng(2024)은 규칙/서명 기반 접근과 대규모 Fine-tuning 기반 연구 사이 간극을 해결하기 위해 최적화된 ICL 만으로 효율적인 CTI 분석과 사이버 보안 지식 그래프(Cyber Security Knowledge Graph)를 구축하는 프레임워크인 CTINEXUS를 제안하였다. 자동 프롬프트 구성 및 k-NN(k-Nearest Neighbors) 기반 예시 검색을 통해 추출 성능을 최적화하고 Entity 정렬 및 장거리 관계 예측을 도입하여 그래프의 연결성을 강화하였다. 10개 플랫폼에서 수집한 150개의 CTI 보고서를 통해 평가한 결과, Triplet 추출 단계에서 F1-score 0.87로 우수한 성능을 입증하였다.

4.2. TTP 추출 및 Threat Intelligence 구조화

두 번째 유형은 CTI 보고서에서 공격 행위를 의미하는 문장과 문단을 식별한 뒤, 이를 MITRE ATT&CK의 TTP로 매핑하여 운영자가 해석 가능한 형태로 위협 지식을 구조화하는 연구들이다. CTI 보고서에 기술된 공격 활동을 TTP 기반으로 구조화하여, 공격의 핵심 행위와 절차를 정리하여 Threat Hunting, 정책 설계 등 후속 보안 운영에서 재사용할 수 있는 지식으로 구조화하는 것이 목표이다.

Hamzić(2025)는 라벨 데이터가 부족한 환경에서 CTI의 문단을 MITRE ATT&CK TTP로 정확하게 분류하기 어렵다는 문제를 해결하기 위해, 라벨이 부족한 환경에서도 동작하는 Retrieval 기반 Few-Shot TTP 분류 프레임워크인 TTPFShot을 제안하였다. ATT&CK의 문서를 문장 단위로 분할해 임베딩하여 외부 지식 저장소에 저장한 후, 입력 문장과 의미상으로 유사한 N개의 예시를 검색하여 Few-shot 프롬프트를 자동으로 구성하고, LLM이 이를 근거로 해당 문장의 TTP 라벨을 예측하도록 설계하였다. 평가는 문장 단위 데이터셋과 문서 단위 데이터셋을 활용해 수행하였다. 문장 단위 데이터셋은 기존 TTPHunter(Rani, 2023) 기반 데이터셋을 활용하였고, 문서 단위 데이터셋은 CISA 기반 CTI 보고서 77건을 활용하였다. 문장 단위 평가에서는 F1-score 0.96 수준으로 높은 성능을 보고하였으며, 문서 단위 평가는 성능이 제한되었지만, 기존 방법 대비 개선된 평가 결과를 보고하였다.

Dong(2025)은 기존 TTP 추출이 CTI 보고서의

TTP를 포함하지 않는 불필요한 문장을 충분히 걸러내지 못하고, 매핑 결과만을 제공하여 해석 가능성이 낮다는 점을 문제로 제기하였다. 이를 위해 CTI 보고서를 문장 단위로 분해한 뒤, (i) Technique 포함 여부 판단 (ii) Technique 분류 (iii) Sub-technique 식별로 구성된3단계 파이프라인을 구축하고, 단계마다 추론 근거(Rationale)를 함께 생성하여 해석 가능성을 높였다. 라벨 부족 문제는 GPT-4o로 합성 데이터를 생성하였으며, 이를 바탕으로 Qwen 2.5-7B 모델을 Fine-tuning 하여 최적화된 TTP 추출 모델을 구현하였다. MITRE ATT&CK와 TRAM(MITRE, 2023) 데이터를 통합·정제한 데이터셋으로 평가한 결과 F1-score 0.97을 보고하며 기존 SOTA(State-of-the-art) 및 상용 LLM에 대비 우수한 성능과 일반화 능력을 보고하였다.

4.3. 운영 산출물 자동 생성

세 번째 유형은 CTI 보고서로부터 실제 보안 운영에 즉시 사용 가능한 운영 산출물(탐지 정책, 쿼리 등)을 생성하는 연구들이다. 관련된 연구들은 공통적으로 CTI에서 탐지에 필요한 정보를 구조화하고, 정책/쿼리 문법에 맞는 산출물을 생성하는 구조로 되어있다. 특히, 문법 오류와 잘못된 정책 생성을 줄이기 위한 검증 루프가 결합된 것이 특징이다.

Xu(2024)가 제안한 IntelEX는 CTI 보고서로부터 단순 Technique 식별을 넘어 공격 수준(Attack-level) TTP와 절차를 포함하는 위협 지식을 자동으로 추출하고, 이를 보안 운영에 활용할 수 있는 탐지 정책으로 변환하는 것을 목표로 한다. 이를 위해 MITRE ATT&CK 지식을 중심으로 ICL을 적용하되, ATT&CK 문서를 임베딩한 외부 지식 저장소를 활용해 RAG와 GraphRAG를 결합하여 추출 성능을 보강하고 별도 LLM-as-a-judge 모듈을 결합하여 환각 및 과잉 추출을 억제하도록 설계하였다. 또한, 추출된 TTP를 기반으로 Sigma 정책을 생성하고, 쿼리 변환 과정에서 발생하는 실행 및 문법 오류를 LLM에 피드백하여 오류가 해소될 때까지 반복 수정하는 파이프라인을 구축하였다. 평가는 1,769건의 Cisco Talos CTI 보고서와 16건의 수동 라벨 보고서를 활용해 수행되었고, Atomic Red Team 기반 61개의 Splunk 테스트를 통해 생성된 탐지 정책의 F1-score 0.929를 보고하여 운영 적용성을 평가하였다.

Schwartz(2025)가 제안한 LLMCloudHunter는 텍스트와 이미지가 혼재된 클라우드 CTI 보고서에서 클라우드 보안에 핵심 요소로 활용되는 API Call을 추출하고 이를 통해 탐지 정책을 자동 생성하는 것을 목표로 한다. GPT-4o 기반 파이프라인으로, 보고서 내 이미지를 정보성과 비정보성 이미지로 분류한 뒤 정보성 이미지에 대해 텍스트를 추출해 원문에 통합한다. 이후 문단 단위로 API Call, IoC, TTP를 다수결(Majority Voting) 방식으로 반복 추출하여 초기 정책 집합을 구성하고 정책 최적화 과정을 거쳐 최종 탐지 정책을 도출하는 방식으로 설계되었다. 평가에서는 20개의 CTI 보고서를 기반으로 Entity 추출 성능의 정밀도는 0.83을 보였고, 총 260개의 정책 중 99.18%가 문법적으로 컴파일에 성공하였다.

Mitra(2025)는 비정형 CTI로부터 실제 배포 가능한 침입 탐지 시스템의 정책을 자동 생성하고, 생성된 결과가 검토한 CTI와 정합한 지 검증하여 최종 정책을 생성하는 것을 목표로 한다. 해당 연구는 단순 정책 생성을 넘어 기존 정책과의 정합성을 고려하여 업데이트하거나, 신규 정책을 생성하며 배포 가능 여부를 자율 검증하는 에이전트(Agentic) 기반 파이프라인인 FALCON을 제안하였다. 비정형 CTI에서 IoC, 프로토콜, 행위 서술과 같은 서명(Signature) 생성 단서를 활용해 관련 정책을 검색한 뒤, 검색된 정책과 CTI를 결합해 생성 프롬프트를 구성하고, Rule Generator LLM Agent가 Snort 또는 YARA 초기 정책을 생성한다. 이후 생성된

<표 2> 제안된 LLM 기반 CTI 자동 분석 연구의 평가 데이터셋과 지표 비교

관련 연구	평가 데이터셋	데이터셋 공개	라벨링	구조화 산출물 평가 방법	운영 산출물 평가 방법
aCTIon	공개된 CTI 보고서	○	수동	STIX 식별	–
(Liu, 2023)	공개된 CTI 보고서	×	수동	Triplet 식별	–
LLM-TIKG	공개된 CTI 보고서	△	수동	Triplet 식별	–
(Fieblinger, 2024)	공개된 CTI 보고서	×	수동	ROUGE	–
IntelEX	공개된 CTI 보고서	△	수동	TTP 식별	문법 정합성 탐지 성능
TTPFshot	공개된 CTI 보고서	○	수동	TTP 식별	–
FALCON	공개된 탐지 정책에 대한 CTI 보고서 구축	○	수동	–	문법 정합성, 전문가 정성 평가
(Dong, 2025)	공개된 CTI 보고서에서 문장 데이터셋 구축	○	수동	TTP 식별	–
CTINEXUS	공개된 CTI 보고서	○	수동	Triplet 식별	–
LLMCloudHunter	공개된 클라우드 CTI 보고서	△	수동	Entity 식별	컴파일 성공률 문법 정합성

O: 데이터셋 공개, △: 데이터셋 출처 공개, X:데이터셋 미공개

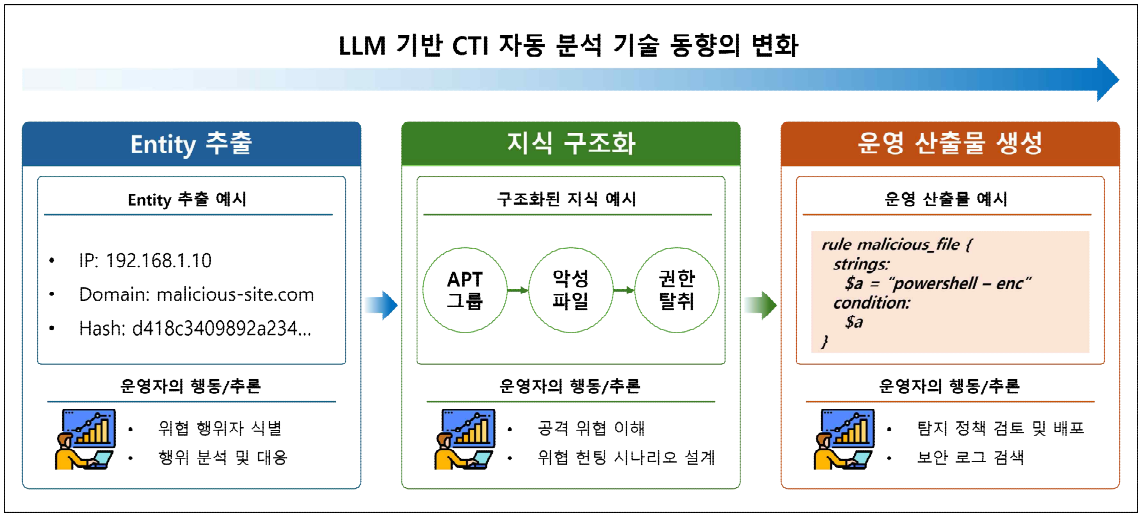
정책과 CTI의 문맥과 일치하는지 평가하기 위해 Bi-encoder 기반 Semantic Scorer를 통해 후보 정책을 검증 및 선택하는 구조로 동작한다. 데이터셋은 공개된 Snort 4,017개와 YARA 4,587개의 정책을 기반으로 각 정책에 대응하는 CTI를 구축하여 평가하였다. 생성된 탐지 정책은 전문가 평가를 통해 평균 약 95% 수준의 정합도를 보여 실제 배포 가능성을 확인하였다.

4.4. 비교 분석 및 데이터셋

LLM을 활용해 CTI 보고서를 자동으로 분석하는 방법론을 제시한 선행 연구에서 사용된 각 데이터셋과 평가 지표는 <표 2>와 같다. 다수의 연구는 연구자들이 각자 수집한 출처가 다른 CTI(블로그, 기관·기업 보고서 등)를 수집·정제하여 평가 데이터셋으로 활용하였다. 이 경우 보고서의 작성 주체, 문체, 포함 정보, 문서 길이 및 분석 난이도가 서로 달라질 수 있으며, 이는 LLM의 추출·구조화·추론 성능에 영향을 미친다. 또한, CTI는 비정형 자연어 문서로서 구조가 불균일하고 동일한 개념이 다양한 표현으로 기술되는 특성이 있으므로, 데이터 출처의 차이는 결과적으로 문서 난이도 분포의 차이로 이어진다.

또한, 대부분의 연구가 라벨링을 '수동'에 의존한다. 일부 연구는 다양한 전문가 그룹을 활용하여 라벨링을 수행하였으나, 라벨의 정의와 기준이 연구마다 달라질 수 있고, 라벨러의 전문성과 일관성 수준에 따라 결과의 신뢰도가 달라질 수 있다. 이러한 요인은 평가의 이질성을 심화시키는 요인으로 작용할 수 있다.

LLM 기반 CTI 자동 분석 연구들은 공통적으로 비정형 CTI를 활용 가능한 산출물로 구조화하는 것을 목표로 하지만, 평가 대상과 측정 방법이 연구마



[그림 1] LLM 기반 CTI 자동 분석 기술 동향의 변화 도식도

다 상이하다. 구조화 산출물 평가는 크게 STIX 식별, Triplet 식별, TTP 식별 등 서로 다른 평가 목표로 분기되며, 일부 연구는 생성된 텍스트의 유사도로 평가하기도 한다. 운영 산출물 평가도 컴파일 성공률, 전문가 정성 평가, 탐지 성능 등 다양한 지표를 기반으로 평가된다.

이처럼 서로 다른 평가 지표는 연구 간 성능을 동일 기준에서 비교하기 어렵다는 문제로 이어질 수 있다. 예를 들어 컴파일 성공률은 산출물이 실행 가능한 형식으로 제공되었는지 확인하는 데에는 유효하지만, CTI의 문맥을 충분히 반영했는지까지 보장하지 않는다. 반대로 전문가 정성 평가는 의미적 타당성과 효용성은 판단할 수 있으나, 평가 기준과 평가자에 따라 결과가 달라질 수 있다. 결국, 현재 연구들은 CTI 자동 분석이라는 같은 목표를 공유하지만, 각 연구가 검증한 품질이 서로 달라, 하나의 공통 지표로 성능을 표준화하기 어렵다.

V. 논의 및 제안

5.1. 기술 동향의 변화

LLM 기반 CTI 자동 분석 연구는 비정형 CTI를 운영에 활용할 수 있는 형태로 구조화한다는 공통 목표를 유지하면서, 연구가 다루는 핵심 산출물이 점진적으로 확장되는 양상을 보인다. [그림 1]은 이러한 변화 과정을 개념적으로 정리한 것이다. 초기에는 CTI 보고서로부터 식별 가능한 IoC 및 Entity를 추출하여, 운영자가 후속 분석을 수행할 수 있도록 정보의 단위를 정리해 주는 보조 프레임워크였다. 이 수준에서는 텍스트 내 문자열 패턴(IP/도메인/해시), 단순 Entity 추출과 같이 추출 대상이 비교적 명확한 정보를 대상으로 하여, 문장 수준의 의미 추론보다는 정형 필드 추출에 가까웠다.

이후 연구는 단순 Entity 추출을 넘어, CTI의 핵심 가치인 공격자가 무엇을 어떻게 수행하였는지를 표현하는 TTP 수준의 공격 문맥을 구조화하는 것으로 확장되었다. 즉, IoC/Entity 중심의 단편적인 정보를 넘어 MITRE ATT&CK 기반 TTP 매핑,

관계(Triplet) 추출, 지식 그래프 구성으로 공격 문맥을 구조화하는 방향으로 확장되었다. 특히, LLM은 문서 구조가 불균일한 CTI에서 공격 문맥을 자연어로부터 복원하는 데 활용되어 추출–정규화–매핑 다단계 파이프라인을 통해 안정적으로 구조화된 위협 지식을 생성하는 방향으로 발전하였다.

최근에는 한 단계 더 나아가, 구조화된 지식을 보안 운영 환경에서 직접 실행 가능한 운영 산출물로 변환하는 연구 주제가 부상하고 있다. 대표적으로 탐지 정책(e.g., Snort/YARA, Sigma)이나 보안 로그 검색 쿼리와 같은 산출물은 단순 정보 요약을 넘어 직접적인 대응과 통제로 이어질 수 있는 산출물이다. 이러한 연구 흐름은 CTI 보고서 이해보다는 이해 결과를 보안 제품과 운영 프로세스에 반영하는 것이 실제 현업에서 더 큰 효용을 갖는다는 문제의식과 밀접하다. 결과적으로 기술 동향은 (i) IoC/Entity 추출, (ii) TTP 매핑 및 위협 지식 구조화 (iii) 운영 산출물 생성으로 확장되며, LLM은 각 단계에서 CTI 보고서의 공격 문맥을 이해하고 정형화하는 핵심 구성 요소로 자리 잡고 있다.

5.2. 선행 연구의 한계

이러한 발전에도 불구하고, 기존 연구에는 다음과 같은 한계가 존재한다.

첫째, 환각과 근거 추적성 부족은 CTI 자동 분석의 신뢰도를 구조적으로 제한할 수 있다. CTI는 기술 용어·취약성 문맥 등 정밀성이 요구되는 영역인데, LLM은 도메인 특화 문맥에서 오해나 환각으로 거짓 위협 지식을 생성할 위험이 있다. 이 한계점은 단순히 정확도 저하에 그치지 않고 운영 의사결정을 왜곡할 수 있다는 점이다. 따라서 결과가 원문 내 어떠한 근거에 기반하여 도출되었는지를 추적할 수 있도록 설계하는 방향이 필요하다.

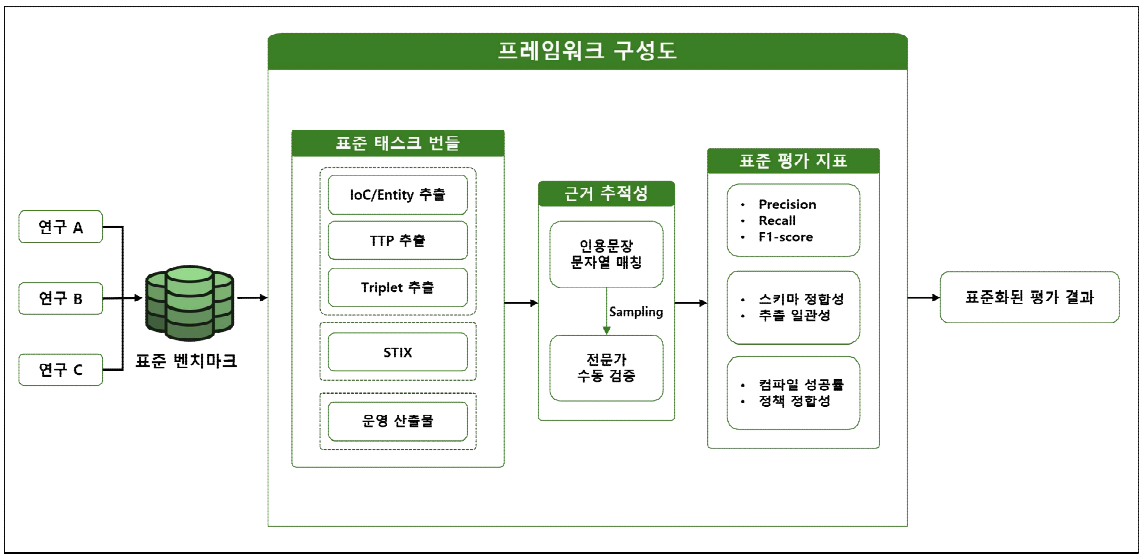
둘째, 서로 다른 데이터셋과 수동 라벨 의존으로 인해 연구 간 성능을 정량적으로 비교하기 어렵다. 다수 연구가 각자 수집·정제한 CTI를 평가 데이터셋으로 활용하면서, 문체·길이·표현·난이도 분포가 달라지고, 이는 LLM의 성능에 직접적인 영향을 준다. 또한, 라벨링이 주로 수동에 의존하며, 라벨의 정의, 라벨러의 전문성에 따라 평가 이질성이 심화될 수 있다.

셋째, 서로 다른 평가 지표 역시 비교 가능성을 저해하는 요소이다. 구조화 산출물 평가는 STIX/Triplet/TTP 식별 등 평가 목표가 다르고, 일부 연구는 ROUGE 같은 텍스트 유사도로 평가를 진행하였다. 운영 산출물도 컴파일 성공률, 전문가 정성 평가, 탐지 성능 등 평가 지표가 공통된 연구 목표를 가지더라도, 각 연구가 검증한 품질의 속성이 달라 하나의 공통 척도로 성능을 표준화하기 어렵다는 한계점이 존재한다.

5.3. 기존 벤치마크 연구 동향 및 한계

이와 같은 비교 가능성의 한계를 완화하기 위해, 최근에는 CTI 영역에서 LLM을 공통 기준으로 평가하기 위한 벤치마크 데이터셋을 제안하는 연구가 등장하고 있다. CTIBench(Alam, 2024)는 CTI 특화 과제에서 LLM을 객관적으로 평가할 수 있는 표준 평가 데이터셋의 부재가 문제임을 명시하고 이를 해결하기 위한 벤치마크를 제안한다. SEvenLLM(Ji, 2024)은 CTI 평가를 위해 객관식과 질의응답 기반의 SEVENLLM–Bench를 구성하였으며, 전문가 검증 과정을 통해 테스트셋의 정합성과 타당성을 확보하였다.

이러한 초기 연구들이 표준으로 수렴하기에는 포괄해야 하는 범위가 광범위하다. 언어, 지역, 출처 편향을 포함한 현실적인 제약이 존재하기 때문에 향후 CTI 분석 연구가 동일한 조건에서 비교할 수 있는 표준 데이터셋 구축을 정립해야 한다.



[그림 2] 제안하는 CTI 자동 분석 평가 프레임워크 도식도

5.4. 표준 평가 프레임워크 제안

이에 본 연구는 환각, 근거 추적성 부족, 평가 이질성 문제를 해결하고 LLM 기반 CTI 자동 분석 연구를 동일한 조건에서 비교·검증하기 위한 CTI 자동 분석 평가 프레임워크를 제안한다. 제안하는 프레임워크의 도식도는 [그림 2]에 정리하였다. 해당 프레임워크는 단순 데이터셋을 넘어 평가 체계 및 데이터셋 구축 원칙을 포함하는 통합 평가 체계로 구성한다.

첫째, 태스크 번들은 단일 과제가 아니라 CTI 자동 분석의 단계적 목표를 반영한 핵심 태스크 묶음으로 구성한다. (i) IoC/Entity 추출, (ii) ATT&CK 기반 TTP 매핑 (iii) Triplet 추출 (iv) 구조화 산출물 생성 (v) 탐지 정책 및 쿼리 등 운영 산출물 생성을 단계화하여, 연구가 어느 수준의 자동화를 달성하였는지 분리할 수 있도록 설계한다.

둘째, 산출물 표현은 결과를 자유 서술로만 평가하지 않고, 가능한 범위 내에서 STIX 2.1과 ATT&CK 분류 체계를 기준으로 정규화된 출력 형식을 요구한다. 이를 통해 연구별 출력 포맷 차이로 인한 비교 불가능성을 완화하고, 실무 도입 시 상호 운용성을 확보한다.

셋째, 근거 추적 기반 평가를 필수 항목으로 포함한다. LLM은 추출한 각 항목(e.g., Entity, TTP)에 대해 해당 정보가 도출된 원문의 인덱스를 함께 출력하거나 로깅하도록 한다. CTI 도메인의 추출 대상(e.g., IoC, ATT&CK ID)은 정형화된 식별자가 중심이므로 문자열 매칭 기반 자동 검증을 기본으로 한다(Siracusano, 2023; Cheng, 2024). 근거 추적성은 전체 추출 항목 중 인용된 원문 문장에 대해 해당 항목이 실제로 존재하는 비율로 정의하며, 본 연구에서 분석한 기존 연구의 성능과 높은 정밀성 요구를 고려하여 0.9 이상을 권장 기준으로 제시한다. 자동 검증 결과의 신뢰성 확보를 위해 전체 결과 중 일부를 샘플링하여 경력 3년 이상인 전문가 3인 이상이 수동 검토하는 방식을 병행한다.

넷째, 공통 평가 지표는 <표 3>에 정리한다. 본 지표는 추출·구조화·운영 산출물의 세 단계로 구분되며, 각 단계에 적합한 평가 기준을 제시한다.

IoC/Entity 추출, TTP 매핑, Triplet 추출에는 정밀도·재현율·F1-score를 적용하고, 구조화 산출물에는 스키마 정합성과 관계 일관성을 평가한다. 또한, 탐지 정책 및 쿼리와 같은 운영 산출물은 문법 정합성과 적용 가능성을 기준으로 평가하여 실제 활용 가능성을 고려한다. 이를 통해 단순 성능 비교를 넘어 CTI 자동 분석 결과의 품질과 실용성을 함께 평가할 수 있다.

<표 3> 제안하는 공통된 평가 지표

평가 범주	평가 대상	성능 지표
정보 추출	IoC, Entity 추출	P/R/F1
	TTP 매핑	P/R/F1
	Triplet 추출	P/R/F1
산출물 품질	산출물 생성 (STIX)	스키마 정합성
		추출 일관성
운영 산출물	탐지 정책, 쿼리	컴파일 성공률
		정책 정합성
P: 정밀도, R: 재현율, F1: F1-score		

다섯째, 운영성 평가를 포함한다. CTI 자동 분석의 최종 목표는 보안 운영 환경에서의 활용이므로, 정확도뿐 아니라 생성 비용·처리시간, 운영자 검증 등을 함께 보고하도록 설계해야 한다. 특히, 정책·쿼리 생성은 최소한의 자동 검증(정합성 검사)을 포함하도록 한다.

마지막으로 벤치마크 데이터셋 구축 원칙을 명시한다. 프레임워크의 대표성과 일반성을 위해 (i) 다양한 출처의 공개 CTI (정부기관, 보안벤더/연구기관, 기술 블로그 등)로 비정형 CTI 데이터셋을 구성하고 (ii) 다국어 데이터(영어 외 중국어, 일본어, 한국어 등)를 포함하여 언어 편향을 완화하며, (iii) 문서 길이(글자 수), Entity 추출 난이도(명시적 vs 암묵적), 구조화 난이도(서술형/표 혼재)에 따라 난이도 등급을 구분해 평가할 수 있도록 해야 한다. 공개 CTI는 기관·지역에 따라 형식과 문체가 다르기 때문에 특정 형식에 치우친 벤치마크가 문서 양식에 대한 편향과 과적합을 유발할 수 있기 때문에 모델의 일반화 성능을 평가하기 위한 핵심 요건이 된다.

이와 같이 본 연구는 CTI 자동 분석 평가 프레임워크를 설계 및 제시함으로써, 서로 다른 데이터셋과 평가 지표로 인한 비교 불가능성을 완화하고자 한다.

5.5. 평가 프로토콜

제안 프레임워크가 연구 간 재현성과 비교 가능성을 확보하기 위해서는 평가 지표뿐 아니라 정답의 정의, 라벨링 기준, 평가자 합의 절차와 같은 구체적 평가 프로토콜이 함께 제시되어야 한다. 본 절에서는 이러한 프로토콜의 핵심 요소를 정리한다.

첫째, 정답 정의 및 라벨링 기준을 명확히 한다. IoC/Entity 추출의 경우 원문에 명시적으로 등장하는 항목만을 정답으로 인정하며, 추론을 통해 도출된 항목은 별도 범주로 분리한다. TTP 매핑은 MITRE ATT&CK의 공식 정의에 부합하는 경우에만 정답으로 인정하고, 하나의 문장에 복수의 Technique에 해당할 경우 모든 Technique를 정답 집합으로 포함한다. Triplet 추출은 주체와 객체가 정답 Entity 집합에 포함되고, 관계가 사전에 합의된 유형 내에서 선택된 경우에만 정답으로 처리한다.

둘째, 평가자 간 합의 절차를 명시한다. 복수의 평가자가 독립적으로 라벨링을 수행한 후, 불일치 항목에 대해 합의를 통해 최종 라벨을 결정하며, 평가자 간 일치도는 Cohen's Kappa 계수로 측정한다 (Artstein, 2017).

이러한 평가 프로토콜을 통해 서로 다른 연구에서 동일한 기준으로 평가가 수행될 수 있으며, LLM 기반 CTI 자동 분석 기술의 발전 정도를 동

일한 수준에서 논의할 수 있을 것으로 기대한다.

Ⅵ. 결론

본 연구는 비정형 CTI 보고서를 보안 운영 환경에서 활용할 수 있는 수준으로 해석·구조화하기 위해 제안된 LLM 기반 CTI 자동 분석 연구들을 수집하고, 연구 별로 접근 방식과 산출물, 평가 방식의 차이를 체계적으로 비교·분석하였다. 이를 통해 LLM이 CTI 자동 분석 과정에서 수행하는 역할이 무엇인지, 그리고 현재 기술이 어느 수준까지 진행되었는지를 정리하고 향후 연구 방향을 도출하고자 하였다.

분석 결과를 종합하면, 초기 연구는 IoC/Entity 추출과 같은 비교적 명확한 정보 단위에 초점을 두었으나, 이후에는 TTP 매핑과 객체 간 관계 추출을 통해 공격 문맥을 구조화하고, 최근에는 탐지 정책 및 쿼리와 같은 즉시 실행 가능한 운영 산출물을 자동 생성하는 방향으로 발전하고 있다. 이러한 변화는 CTI 자동 분석이 단순 정보 추출을 넘어 실제 보안 운영에 활용할 수 있는 형태로 전환되고 있음을 보여준다.

그러나, 기술 확장에도 불구하고 환각과 근거 추적성 부족, 서로 다른 데이터셋과 수동 라벨링 의존, 평가 지표가 상이하다는 한계점은 연구 성과를 동일 기준에서 비교·검증하는 데 제약으로 작용한다. 운영 산출물을 생성하는 연구의 경우, 단순히 문법적으로 실행 가능한지를 확인하는 수준을 넘어, 생성된 정책이 CTI 문맥을 충분히 반영하고 있는지까지 함께 검증할 필요가 있다.

이에 본 연구는 이러한 한계를 극복하고 객관적 성능 검증을 지원하기 위해, 단계별 태스크 번들과 산출물 표준화, 근거 추적성 및 운영 효율성 평가 그리고 표준 데이터셋의 요구사항을 포함하는 CTI 자동 분석 평가 프레임워크를 제안하였다. 본 연구가 제시한 프레임워크는 향후 LLM 기반 CTI 자동 분석 기술을 동일한 기준에서 비교·평가할 수 있는 토대를 마련하고, 나아가 조직이 신뢰할 수 있는 자동화된 위협 대응 체계를 구축하는 데 기여할 것으로 기대된다.

참고문헌

안랩 (2025). 2025년 사이버 위협 동향 및 2026년 전망 보고서. AhnLab 콘텐츠 센터. https://www.ahnlab.com/ko/contents/content-center/36017

OASIS Open. (2021). *STIX™ Version 2.1.* OASIS. https://www.oasis-open.org/standard/6426/

이완형 (2024). AI / Post-AI 시대 기업의 경영전략으로서 '공존경쟁력'에 관한 연구: 시대별 기업 경쟁력의 비교와 변화. 미래사회, 15(3), 181-200.

Alam, M. T., Bhusal, D., Nguyen, L., & Rastogi, N. (2024). CTIBench: A benchmark for evaluating LLMs in cyber threat intelligence. *arXiv preprint arXiv:2406.07599.*

Cheng, Y., Bajaber, O., Tsegai, S. A., Song, D., & Gao, P. (2024). CTINEXUS: Leveraging optimized LLM in-context learning for constructing cybersecurity knowledge graphs under data scarcity. *arXiv preprint arXiv: 2410.21060.*

Dong, F., Jiang, Z., Ma, C., He, Q., Yang, P., Yao, Y., & Wang, J. (2025). From threat report to ATT&CK: Automated extraction and reasoning of TTPs using large language models. *In Proceedings of the IEEE International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE.*

Fieblinger, R., Alam, M. T., & Rastogi, N. (2024). Actionable cyber threat intelligence using knowledge graphs and large language models. *arXiv preprint arXiv:2407.02528.*

Hamzić, D., Skopik, F., Landauer, M., Wurzenberger, M., & Rauber, A. (2025). TTP classification with minimal labeled data: A retrieval-based few-shot learning approach. *In M. Dalla Preda et al. (Eds.), Availability, Reliability and Security. ARES 2025. Lecture Notes in Computer Science 15993,* 387-408.

Hu, Y., Zou, F., Han, J., Sun, X., & Wang, Y. (2024). LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model. *Computers & Security, 145,* 103999.

Ji, H., Yang, J., Chai, L., Wei, C., Yang, L., Duan, Y., Wang, Y., Sun, T., Guo, H., Li, T., Ren, C., & Li, Z. (2024). SEvenLLM: Benchmarking, eliciting, and enhancing abilities of large language models in cyber threat intelligence. *arXiv preprint arXiv:2405.03446.*

Liu, J., & Zhan, J. (2023). Constructing knowledge graph from cyber threat intelligence using large language model. *In 2023 IEEE International Conference on Big Data (BigData)* 516-521.

Mitra, S., Bazarov, A., Duclos, M., Mittal, S., Piplai, A., Rahman, M. R., Zieglar, E., & Rahimi, S. (2025). FALCON: Autonomous cyber threat intelligence mining with LLMs for IDS rule generation. *arXiv preprint arXiv:2508.18684.*

Rani, N., Saha, B., Maurya, V., & Shukla, S. K. (2023). TTPHunter: Automated extraction of actionable intelligence as TTPs from narrative threat reports. *In Proceedings of the 2023 Australasian Computer Science Week (ACSW '23),* 126-134.

Schwartz, Y., Benshimol, L., Mimran, D., Elovici, Y., & Shabtai, A. (2025). LLMCloudHunter: Harnessing LLMs for automated extraction of detection rules from cloud-based CTI. *In Proceedings of the ACM Web Conference 2025 (WWW '25),* 1922-1941.

Siracusano, G., Sanvito, D., Gonzalez, R., Srinivasan, M., Kamatchi, S., Takahashi, W., Kawakita, M., Kakumaru, T., & Bifulco, R. (2023). Time for aCTIon: Automated analysis of cyber threat intelligence in the wild. *arXiv preprint arXiv:2307.10214.*

Xu, M., Wang, H., Liu, J., Lin, Y., Xu, C., Liu, Y., Lim, H. W., & Dong, J. S. (2024). IntelEX: A LLM-driven attack-level threat intelligence extraction framework. *arXiv preprint arXiv: 2412.10872.*

MITRE. (2020). *MITRE ATT&CK®: Design and philosophy.* https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf

MITRE Engenuity. (2023). *Threat Report ATT&CK Mapper (TRAM).* Center for Threat-Informed Defense. https://github.com/center-for-threat-informed-defense/tram

Artstein, R. (2017). Inter-annotator agreement. *Handbook of Linguistic Annotation,* 297-313.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language

models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903.*

투고일자: 2025. 12. 31.
심사일자: 2026. 1. 27.
게재확정일자: 2026. 2. 2.

미래사회
Journal of Future Society
2026. Vol. 17, No. 1, 176~193.
https://doi.org/10.22987/jifso.2026.17.1.176

Evaluation Framework for Large Language Model-based Automated Cyber Threat Intelligence Analysis

JinHo Yu HyoJin Jo

Yonsei University

The recent growth in data breaches and ransomware has increased the need for effective cyber threat intelligence (CTI) analysis. However, most CTI reports are unstructured, rendering manual analysis inefficient. Consequently, recent studies have employed large language models (LLMs) to automate CTI analysis. This paper reviews representative works published between 2023 and 2025, highlighting a shift from basic entity extraction to knowledge structuring and the generation of operational outputs, such as detection rules. Despite this progress, inconsistent datasets, evaluation methods, and hallucinations have limited fair comparisons and real-world use. Consequently, we propose a CTI automated analysis evaluation framework for LLM-based CTI analysis that includes task bundles, standardized outputs, evidence-based evaluation, operational efficiency metrics, and guidelines for diverse dataset construction. This framework supports reliable comparison and practical deployment of automated threat analysis.

Key words: Cyber Threat Intelligence, Large Language Model, Structuring Threat Knowledge, Detection Rule Generation, Benchmark Framework