

Fine-Tuning Large Language Models for Security Log Data Labeling

Doo-Yong Jeon*

*Professor, Dept. of Cybersecurity, Yeungnam University College, Daegu, Korea

[Abstract]

This study proposes a data sampling method called CoreShot Filter to address the high cost and subjective judgment issues in labeling security log data. CoreShot Filter combines the concepts of representativeness and uncertainty from active learning to select optimal data for fine-tuning large language models (LLMs). It defines uncertainty using discrepancies between weak learners and manual labels, while representativeness is measured through similarity with persona data generated by genetic algorithms. From over 310,000 logs, 204 core samples were selected and used to fine-tune GPT-4o mini. Experimental results demonstrate that CoreShot Filter outperforms stratified, outlier, and coresets sampling in terms of accuracy, recall, and F1-score. In particular, it achieved superior performance in abnormal detection (Recall 0.8901) and precision (0.9489), proving that CoreShot Filter is an effective method for improving security log analysis and LLM-based labeling efficiency.

▶ **Key words:** AI Security, Data Labeling, LLM, Fine Tuning, Data Sampling, CoreShot Filter

[요약]

본 연구는 보안 로그 데이터 레이블링 과정의 높은 비용과 주관적 판단 문제를 해결하기 위해 대형 언어 모델(LLM)을 활용한 데이터 샘플링 기법인 코어샷 필터(CoreShot Filter)를 제안한다. 코어샷 필터는 적극학습의 대표성과 불확실성 개념을 결합하여 최적 데이터를 선별하며, 느슨한 모델과 수작업 라벨 불일치를 불확실성 지표로, 유전 알고리즘 기반 페르소나 유사도를 대표성 지표로 활용한다. 31만여 개 로그 중 핵심 204개를 추출해 GPT-4o mini를 미세조정된 결과, 코어샷 필터는 기존 층화·아웃라이어·코어셋 샘플링 대비 정확도, 재현율, F1-score 모두에서 우수했다. 특히 비정상 탐지(Recall 0.8901)와 정밀도(Precision 0.9489)에서 탁월한 성능을 보여, 코어샷 필터가 보안 로그 분석과 LLM 기반 레이블링 효율 개선에 효과적임을 입증하였다.

▶ **주제어:** AI보안, 데이터 레이블링, 대형 언어 모델, 미세조정, 데이터 샘플링, 코어샷 필터

I. Introduction

최근 지능형 지속 위협(APT) 방지 시스템, 엔드포인트 탐지 및 대응(EDR) 시스템, 유해 사이트 차단 시스템(Web Filter)과 같은 진화된 보안시스템의 활용이 확대됨에 따라 보안 로그의 양이 급격히 증가하고 있다. 이 방대한 보안 로그를 효율적으로 분석하여 지능적으로 사이버 위협에 대응하기 위한 인공지능(AI) 기술의 활용 사례가 크게 늘어나고 있다[1, 2, 3]. 특히 복잡한 보안 이벤트 간의 상관관계 분석과 실시간 위협 탐지 자동화에 대한 요구로 인해 빅데이터와 인공지능 기반의 보안 분석은 더욱 중요해지고 있다.

사이버보안 분야에서 인공지능 기반 분석을 효과적으로 수행하기 위해서는 대량의 데이터와 이에 대한 레이블링(Labeling)이 필수적이다. 인공지능에서의 레이블은 모델이 예측하고자 하는 목표 정답을 의미하며, 데이터 레이블링은 이상적인 정답인 절대 정답(Ground Truth)에 가능한 가까운 정답을 확보하는 과정이다. 하지만 모든 데이터에 완벽한 절대 정답을 확보하는 현실적으로 불가능하므로, 일반적으로 레이블링 작업자가 데이터를 직접 검토하여 최대한 정확한 레이블을 부여하는 수작업을 수행한다.

그러나 사이버보안처럼 도메인 지식이 필수적인 분야에서는 수작업 레이블링 과정에서 필연적으로 주관적 판단이나 오류가 발생할 수 있다. 예를 들어 방화벽에서 특정 IP 주소가 3,000회 차단된 사건이 실제로 공격인지 정상적인 동작인지에 대한 의견은 다양할 수 있으며, 정확한 판단을 위해 접근 기간, 목적지 개수, 다른 위협과의 결합 여부 등 여러 요소를 종합적으로 고려해야 한다. 특히 대량의 데이터를 여러 작업자가 나누어 레이블링하는 경우, 각자의 주관이나 편향, 실수로 인해 불안정한 레이블이 생성될 가능성이 크다.

이러한 불안정한 레이블 문제를 해결하기 위해 기존에는 머신러닝(ML) 모델이나 앙상블 기법을 활용하여 해결하려 하였다. 그러나 전통적인 방식으로는 보안 분석 과정의 복잡하고 도메인에 의존적인 요소들을 충분히 반영하기 어려웠다. 예를 들어, 접근한 IP 주소의 국가 정보는 일반적으로 중요하지 않은 특징이지만, 특정 국가(예: 북한)로부터의 접근은 다른 특징보다 우선적으로 고려되어야 한다.

최근에는 준지도학습(Semi-Supervised Learning), 적극학습(Active Learning), 임의 레이블링(Pseudo Labeling) 등 레이블링 효율을 높이는 진보된 기술들이 제시되었지만, 이 기법들도 실제 인간 레이블링 작업자가 판단할 때 활용하는 도메인 특화 정보를 충분히 반영하지 못

한다는 한계가 있다. 예를 들면, 이벤트 명은 보안 로그의 레이블 결정에서 핵심 정보임에도 불구하고, 인공지능 모델이 학습을 위해 임베딩(Embedding)과 같은 수치화 과정을 거치면 본래 의미가 왜곡될 수 있다. IP 주소 사용자 또는 그룹, 포트 번호 등도 인간에게는 중요한 공격 패턴 식별의 근거가 되지만, 기존 모델에서는 단순 숫자 및 가중치로 처리되어 의미 있는 패턴으로 활용하기 어렵다.

본 연구에서는 이러한 문제를 극복하기 위해 자연어 처리를 강점으로 하는 대형 언어 모델(LLM)을 활용하여 보안 로그 데이터를 보다 직관적이고 종합적으로 분석하는 가능성을 실험하였다. LLM은 자연어를 있는 그대로 처리하여 로그 분석에서 중요한 맥락적 의미를 효과적으로 추출할 수 있다. 특히 최근 ChatGPT 등 대형 언어 모델이 대중화되면서 데이터 레이블링 분야에서도 활발히 활용되고 있으며[4, 5], 다양한 분야에서 제로샷(Zero-shot), 퓨샷(Few-shot) 학습을 통해 뛰어난 성능을 보여주고 있다[6, 7].

본 논문은 보안시스템 로그 데이터 레이블링을 위한 대형 언어 모델의 파인튜닝에 가장 적합한 데이터를 선별하기 위한 방법인 코어샷 필터(CoreShot Filter)를 제안한다. 코어샷 필터는 적극학습(Active Learning)의 대표성(Representativeness)과 불확실성(Uncertainty) 개념을 결합하여, 비용 효율적인 LLM 파인튜닝에 가장 유용한 데이터를 선정한다. 불확실성은 인간과 머신러닝 기반 모델 간의 레이블링 의견 불일치를 통해 도출하고, 대표성은 유전 알고리즘을 이용해 생성된 페르소나 데이터와의 유사성으로 정의한다. 이 방식을 통해 총 31만 개의 데이터에서 가장 가치 있는 204개의 코어샷 데이터를 선별하였다.

선별된 데이터를 기반으로 GPT 4o-mini 모델을 파인튜닝하여, 보안 로그에서 입력된 IP 주소의 블랙리스트 여부를 이진 분류로 예측하고, 동시에 예측 근거를 자연어 텍스트로 제공하도록 하였다. 코어샷 필터의 효과성을 평가하기 위해 층화 샘플링, 이상치 샘플링, 코어샷 샘플링 등 다른 샘플링 방법들과 성능을 비교하였다.

본 연구에서 제안하는 코어샷 필터는 보안시스템 로그 데이터를 활용한 LLM 파인튜닝을 위한 최적의 데이터 샘플링 기법을 제공하며, 기존 샘플링 방법 대비 다음과 같은 기여점을 갖는다.

1. 대표성과 불확실성을 결합한 LLM 최적 학습 데이터 선별 기법 제안
 - Active Learning의 대표성(Representativeness)과 불확실성(Uncertainty)을 결합하여, 보안 로그 분석을 위한 최적의 학습 데이터 샘플링 방법 개발.
 - 기존 샘플링 기법(층화 샘플링, 아웃라이어 탐지, 코어

- 셋) 대비 정확한 데이터 선택을 통해 학습 성능 향상.
- 비정상(악성) 탐지 성능 및 라벨 정확도 향상 검증
 - 실험을 통해 코어샷 필터가 기존 샘플링 방법보다 비정상 탐지 성능(Recall 0.8901) 및 라벨 정확도(Precision 0.9489)가 우수함을 입증.
 - 특히, 보안 로그 분석에서 중요한 비정상(악성) 탐지 성능이 크게 향상되었으며, 정상 데이터를 정확하게 판단하는 성능도 뛰어나.
 - 보안 로그 분석을 위한 LLM 활용 전략 제안
 - 이벤트명, IP소유자, 도메인, 국가코드 등 자연어 데이터를 직접 해석하여 맥락을 이해, 보다 직관적인 분석 제공.
 - LLM이 생성한 텍스트를 레이블링 근거로 활용하는 방법을 제시.

본 논문의 구성은 다음과 같다. 2장에서는 인공지능 기반 보안 분석과 레이블링 관련 연구 분석을 통해 문제점과 한계를 살펴보고, 3장에서는 코어샷 필터의 작동 원리와 세부 절차를 기술한다. 이어서 4장에서는 코어샷 필터의 성능 평가 및 비교를 수행하고, 마지막으로 5장에서 본 연구의 결론과 향후 연구 방향을 제시한다.

II. Related Works

2장에서는 보안시스템 로그를 처리·분석하기 위해 인공지능 기법을 활용하는 과정에서 발생하는 데이터 레이블링 문제와 이를 완화하기 위한 접근법을 다룬다.

1. Artificial Intelligence for Cybersecurity

IT 인프라를 보유한 조직 및 기관은 방화벽(Firewall), 침입방지시스템(IPS), 웹방화벽(WAF), APT 대응 솔루션, DDoS 방어 시스템 등 다양한 보안 솔루션을 운영한다. 이들 시스템은 보안관제센터(SOC)에서 통합 분석이 가능하도록 로그를 수집·기록하는데, 매일 방대한 양의 로그가 생성된다. Sandeep et al.은 HP 본사에서 하루 평균 1조 개의 로그가 발생하는데 비해 보안 담당자 한 명이 하루에 처리할 수 있는 이벤트는 약 3,000건에 불과하다고 하였다[8]. 대량 로그 처리의 어려움을 해결하기 위해 AI 기반 솔루션이 제안되었으나, 높은 거짓 양성(False Positive) 비율은 여전히 해결이 어려운 과제로 지적된다[9].

지금까지의 연구는 보안 기능별로 특화된 모델(방화벽 및 침입탐지 로그 분석, 이상 탐지, 웹 해킹 탐지, DDoS

공격 탐지, 악성 URL 분석, 악성코드 탐지 등)에 주로 집중해 왔다. 본 연구는 이러한 한계를 극복하고자 대규모 언어 모델(LLM)을 활용하여 코어샷 필터를 개발하였다. 방화벽, IPS, 웹방화벽, 웹서버 로그의 출발지 IP 기준으로 데이터를 전처리하고, GPT 4o-mini 모델을 미세 조정(Fine-Tuning)하여 이벤트의 내용과 심각도를 통합적으로 파악할 수 있도록 하였다.

2. Challenges in Data Labeling

정확한 예측을 위해서는 대량의 레이블링된 데이터가 필요하다. 이를 수집·가공하는 비용은 매우 높다[10]. 이 문제를 해결하기 위한 대표적 방안은 준지도학습(Semi-Supervised Learning)과 적극학습(Active Learning)이다.

준지도학습은 소량의 레이블된 데이터와 대량의 레이블 안된(레이블이 없는) 데이터를 결합하여 성능을 개선하는 방법이다[11,12,13]. 그러나 초기 모델의 레이블 오류가 지속적으로 확대될 수 있고, 하이퍼파라미터 설정 문제 등의 한계가 존재한다.

적극학습은 레이블링 비용을 최적화하기 위해 모델에 가장 이득이 큰 소량의 샘플만 우선 선별하여 레이블링하고, 이를 재학습하는 방법이다[14,15]. 불확실성과 대표성을 기준으로 데이터를 고르는 다양한 아이디어가 제시되었으나, 각 접근법마다 자동으로 선별하는 기준 데이터는 특정 영역에 편향될 수 있으며 고차원 데이터 처리가 어렵다는 한계가 있다.[16].

코어샷 필터는 수작업 레이블링 과정을 효과적으로 활용하여 레이블 오류를 최소화하는 방법을 제안한다. 이를 위해 느슨한 모델(Weak Learner)을 기반으로 한 준지도 학습과 데이터의 불확실성과 대표성을 고려하는 적극학습 기법을 함께 적용하였다. 이를 통해 비용이 높지만 필수적인 수작업 레이블링 과정을 보다 효율적으로 활용할 수 있도록 하였다.

3. Subjective Labeling Problem

방대한 레이블링을 거쳐도 여전히 결과의 신뢰도 문제가 남을 수 있다. 이는 전문지식이 필요한 도메인에서 특히 심각하며, 주관적 판단이 개입되어 레이블 노이즈가 발생하기 쉽다[17]. 특히 최근 레이블링에 많이 활용되고 있는 클라우드 레이블링 환경에서는 주관적 편차로 인한 문제를 피하기 어렵다[18].

이를 완화하기 위한 방안으로 이상치 탐지 기반 데이터 필터링, 예측 모델 앙상블, 대리 손실 함수 활용 기법들이 연구 되었지만 주관적 판단에 따른 오류인지 단순 실수인

지 구별하기 어려운 한계가 존재한다. J. Palomaki et al. 은 특히 사이버보안 로그의 의미적 모호성을 중요한 문제로 제기하였으며[19], 본 연구는 이러한 모호성에 대응하기 위해 레이블 판단 근거를 텍스트 형태로 자동 생성하도록 설계하였다. 이를 통해 레이블 오류 원인을 더 정확히 파악하고, 반복적 오류를 예방할 수 있을 것으로 기대한다.

III. Methods

3장에서는 코어샷 필터의 핵심 데이터 필터링 방법을 설명한다. 먼저, 사용된 데이터와 전처리 과정을 소개한 후, 코어샷을 선별하는 절차와 방법을 설명한다. 마지막으로, ChatGPT를 파인튜닝하는 방법과 사용된 프롬프트를 소개한다.

1. Data Introduction

본 연구에 사용된 데이터는 실제 보안관제센터(SOC, Security Operation Center)의 보안 정보 및 이벤트 관리(SIEM, Security Information and Event Management) 시스템을 통해 수집하였으며, 방화벽, 침입방지시스템(IPS), 웹방화벽, 웹서버 접속로그(Access log)가 통합된 데이터로 구성되어 있다. 수집된 로그는 일자와 IP 주소를 기준으로 집계(Aggregation)하였으며, 표 1과 같이 총 26개 컬럼으로 구성되어 있다. 이 데이터를 기반으로 정상/비정상(블랙리스트) IP 여부를 이진 분류하고, 결과의 판단 근거를 텍스트로 작성하는 것을 목표로 하고 있다.

원본 로그에는 로깅 시간, 로깅 시스템 IP주소, 출발지 IP 주소, 목적지 IP 주소, 목적지 Port 번호, 공격 명, 탐지/차단 여부 정도만 작성이 되어있다. 본 연구에서는 레이블링 작업자의 신속한 판단을 지원하기 위해 다양한 파생 정보를 추가하였다. 예를 들어 원본로그에서 시간 데이터를 업무시간/비업무시간, 휴일/평일 접속 여부를 추출하거나 IP의 국적, ISP, IP평판 등 외부 데이터도 활용하였다. 전처리 과정 및 민감한 값을 제거한 데이터의 샘플은 저자의 GITHUB 레파지토리[20]에 수록하였다.

데이터를 일자와 IP 주소를 기준으로 집계한 이유는 실제 보안관제 업무 환경을 시뮬레이션하여 현장 적용성을 높이기 위함이다. 보안관제 요원들이 근무하는 일자에는 근무 전날까지의 데이터로 학습된 모델을 활용하도록 설정하였으며, 이를 위해 일자별 집계 방식을 채택하였다. 근무 일자가 지난 데이터는 다시 모델 학습에 사용되며, 증분 데이터로서 매일 1회씩 학습이 반복되는 방식으로 운영된다.

Table 1. Data Column Composition

Category	Columns
IP Address Identification	IP range (first 3 octets of IP address), IP owner, IP country
Event	Type of event generated, number of events
Event Scacle	Number of logs, length of logs, number of detections per security system
Access Scope	Detected security system type, number of destination IPs, number of accessed networks
Access Time	Number of access days, access time distribution, average weekday accesses, off-hours ratio, average access interval, standard deviation of access time
IP Reputation	Number of reports for this IP on IP Reputation Site 1 [21], number of users reporting this IP on IP Reputation Site 1 [21], reputation score from IP Reputation Site 2 [22]
Blacklist Info	Whether linked to normal website, whether public institution, whether school, whether hospital, whether manufacturer, whether famous portal

또한, 일반적으로 보안관제의 1차 목표는 보안시스템 로그를 분석하여 악성 출발지 IP 주소를 식별하고 추가적인 위협을 차단하는 것이다. 따라서 당일 로그에 기록된 모든 IP 주소와 관련된 데이터를 효율적으로 전처리하여 분석하는 작업에 코어샷 필터의 프로세스를 일치되게 하였다. 이러한 데이터 수집 및 집계 방식은 실제 보안관제 환경에 최적화된 데이터 처리 절차를 반영하며, 코어샷 필터의 실용적인 적용을 제공하기 위한 기반을 제공한다.

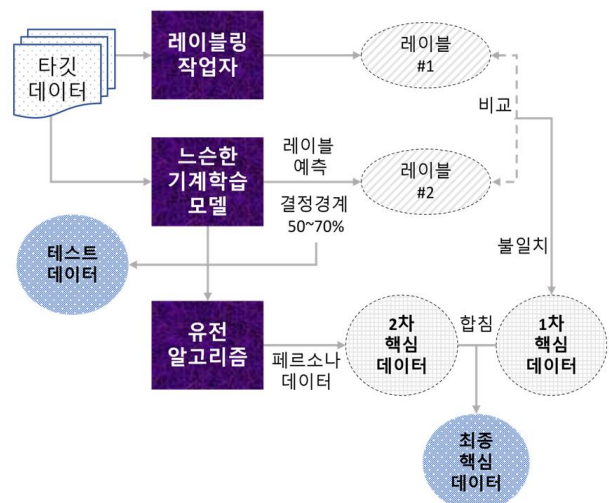


Fig. 1. CoreShot Extraction Process

2. Weak Model

코어샷 필터의 느슨한 모델(Weak Learner)은 레이블링이 완료된 데이터를 기반으로 XGBoost를 활용하여 생성된 모델이다. 이는 정교하게 최적화된 모델이 아니지만 유사한 내용의 반복이 많은 로그 데이터의 특성 때문에 정확도는 85%이상을 유지한다. 이 모델은 데이터의 레이블을 예측한 후 레이블링 작업자의 결과와 비교하여 불일치하는 데이터를 탐지하는 데 활용된다.

모델의 예측과 수작업 레이블링 결과가 다른 데이터는 불확실성이 높은 데이터로 볼 수 있다. 본 연구의 불확실성은 느슨한 모델(Weak Learner)의 사후확률과 수작업 레이블 간 불일치에서 유도한다. 레이블 $y \in \{0,1\}$, 느슨한 모델의 비정상(1) 사후확률 $\hat{p} = \Pr(abnormal|x)$ 일 때, 기본 불확실성 지표는 수식 1과 같이 정의한다.

$$U_1(\vec{x}) = |\hat{p} - y| \tag{1}$$

불확실성은 결정 경계(확률 50%)에 가까울수록 높아지며, 본 연구에서는 상위 20% 구간(약 50%~70% 범위)을 불확실성 임계값으로 설정하였다. 이는 본 연구에서 사용한 데이터의 레이블 확률의 분포를 분석했을 때 예측 확률 분포가 [0,0.3], [0.7,1.0]에 집중되고 0.5 부근에 소수 집단이 형성되어 상위 20% 지점에서 집단 분리 현상 즉, 분포 격차가 뚜렷하게 나타났다. 따라서 불확실성 상위 20%를 임계영역으로 채택하였다.

이러한 불확실성이 높은 데이터의 레이블이 명확해지면, 다른 데이터보다 모델 성능 향상에 더 큰 기여를 할 수 있다. 또한, 이를 수작업 레이블링을 검증 및 대표성이 높은 데이터 선별을 위한 유전 알고리즘의 베이스 모델로도 사용된다.

3. Manual Labeling

가장 먼저 레이블링 작업자가 정상/악성 IP 여부를 표 1의 항목을 통해 육안으로 판단하여 레이블링한다. 레이블링 작업자는 사이버보안 전공을 4학기 수료한 학생 3명과 연구책임자(저자) 총 4명으로 구성되었다. 학생 작업자는 사이버보안 및 보안시스템, 로그 분석에 필요한 도메인 지식을 갖추고 있으며, 기초적 통계 및 빅데이터, 인공지능에 대한 교육도 받은 수준이다.

수작업 레이블링은 실수, 주관적 판단, 부정행위 등으로 오류가 빈번하게 발생하기 때문에 본 연구에서는 또 다른 레이블링 작업자 및 느슨한 모델을 이용하여 총 2번의 교차검증을 거친다.

- 1차 핵심 데이터 : 수작업 레이블링 결과는 느슨한 모델이 예측한 레이블과 비교하여 검증된다. 불일치하는 데이터는 레이블링 작업자와 검증자가 합의하여 최종 레이블을 결정한다. 이때, 레이블 불일치를 합의 하여 불확실성이 해소되면 모델의 성능 향상에 긍정적인 영향을 미칠 수 있기 때문에 대형 언어모델 미세조정 에 추가 학습할 가치가 있다고 결정한다.

4. Genetic Algorithm

유전 알고리즘(GA, Genetic Algorithm)은 자연 세계의 유전자 진화 과정을 모방한 모델로, 보유하고 있는 데이터를 유전자처럼 조합하여 주어진 조건에 가장 적합한 데이터를 탐색하는 기법이다. 즉, 종속변수를 예측하는 것을 목표로 하는 기존의 기계학습 방법과는 달리 유전알고리즘은 독립변수 한 세트를 예측하는 것을 목표로 한다.

코어샷 필터에서는 종속변수와 일부 독립변수를 설정한 후, 가장 적합한 나머지 값을 예측하는 데 유전 알고리즘을 활용한다. 이는 수식 2의 목적 함수의 최적화를 목표로 한다. 이때, 이전 단계까지 사용된 느슨한 모델이 예측에 사용된다.

$$\arg \max_x f(X) \text{ where } Y = f(X) \tag{2}$$

여기서 Y 는 종속변수, X 는 독립변수, $f(X)$ 는 모델이 예측하는 함수이다.

예를 들어, 종속 변수가 '비정상'이며, 독립 변수로 이벤트 명 'Port Scan'을 지정했다고 가정하자. 이 경우, 유전 알고리즘은 수식 3과 같은 최적화 문제를 해결하게 된다:

$$\arg \max_x P(Abnormal|X, Event = Port Scan) \tag{3}$$

이는 "Port Scan 이벤트를 발생시킨 데이터가 '비정상'으로 판단되려면, 나머지 변수 값은 어떤 형태가 가장 적절한가?"를 탐색하는 과정이다.

유전 알고리즘은 다음과 같은 절차로 동작한다:

- ① 모집단 초기화(Population Initialization): 다양한 염색체(데이터 조합)를 생성한다. 본 연구에서는 데이터를 정상/비정상 그룹으로 먼저 분류한 후, 이를 다시 이벤트 명 기준으로 세분화하여 유전 알고리즘의 입력 데이터로 사용하였다. 즉, 동일 이벤트 내에서 표 1의 26개 특징 중 이벤트 명을 제외한 25개만 초기 염색체로 구성하였다. 유전 알고리즘을 통해

생성된 데이터는 실제 존재하는 데이터가 아닌, 가장 이상적인 페르소나 데이터(Persona Data)에 해당한다.

- ② 선택(Selection): 적합도(Fitness)가 높은 염색체를 선택하는 작업이다. 여기서 적합도는 유전 알고리즘에 의해 생성된 염색체를 약모델(Weak Model)에 입력했을 때 목표하는 레이블(비정상)의 확률을 의미한다. 선택된 데이터의 적합도(Fitness)는 다음과 같이 평가된다:

$$Fitness(X) = P(Label|X) \quad (4)$$

여기서 $P(Label|X)$ 는 느슨한 모델(Weak Learner)이 해당하는 데이터 조합 X 에 대해 정상 또는 비정상일 확률을 예측한 값이다.

본 연구에서는 가장 기본적인 선택 방식인 룰렛휠 선택(Roulette Wheel Selection)을 적용하였다. 이 방법은 개체 i 의 적합도를 f_i 라 할 때, 전체 모집단 N 에 대해 다음과 같이 선택 확률을 정의한다.

$$P(i) = \frac{f_i}{\sum_{j=1}^N f_j} \quad (5)$$

즉, 적합도가 높을수록 선택된 확률도 비례적으로 증가하며, 이는 유전알고리즘이 점차적으로 더 우수한 염색체를 다음 세대로 전달하는 효과가 있다.

- ③ 교차(Crossover): 선택된 염색체를 조합하여 새로운 자손 생성한다. 본 연구에서는 단일지점 교차(Single-point Crossover) 방식, 교차율 0.5를 적용하였다. 이는 부모 염색체를 임의의 한 지점에서 나눈 후, 앞부분은 부모 A, 뒷부분은 부모 B에서 가져오는 유전알고리즘의 가장 대표적인 교차 방식이다.
- ④ 변이(Mutation): 일정 비율로 무작위 변화를 주어 탐색 다양성을 확보하는 과정이다. 초기 실험에서는 기본 변이 확률(0.1)을 사용하였으나, 보안 로그 데이터의 특성상 변이가 과도하게 발생하면 비현실적인 로그 조합이 빈번히 생성되어 탐색 효율이 저하되는 문제가 관찰되었다. 이에 따라 변이 확률을 0.05로 낮추어 적용하였으며, 그 결과 수렴의 안정성과 재현성이 향상되는 효과를 확인하였다.

최종적으로, 수식 6의 코사인 유사도를 통해 대표성을 측정하고 대표 데이터를 선정한다. 코사인 유사도는 벡터의 크기와 무관하게 방향(패턴)의 유사성을 측정하기 때문에, 다양한 스케일의 특징(Feature)이 혼합된 보안 로그 데이터에서 대표성을 평가하기에 적합하다.

$$S_{\cos}(\vec{x}, c^*) = \frac{\vec{x} \cdot c^*}{\|\vec{x}\| \|c^*\|} \quad (6)$$

여기서 \vec{x} 는 실제 데이터 벡터, c^* 는 페르소나 벡터를 의미한다.

- 2차 핵심 데이터 : 모집단 데이터(Population)에서 가장 대표성이 높은 데이터를 선출하기 위해 생성된 페르소나 데이터와 가장 유사한 데이터를 실제 데이터 셋에서 선택한다.

1차 핵심데이터에 이어 2차 핵심 데이터 까지 추출이 완료되면 이 둘을 합쳐 최종 핵심 데이터로 결정한다.

5. Fine-tuning

미세조정(Fine-tuning)은 대량의 데이터로 사전 학습된 인공지능 모델에 소량의 새로운 데이터를 추가로 학습시켜, 새로운 전문적 작업을 수행할 수 있는 지능을 갖추도록 하는 방법이다. 이는 대규모 데이터를 활용해 일반화가 고도로 이루어진 기본 모델에 새로운 기능을 부여함으로써 전이학습의 가능성[23,24]을 확인하는 데서 시작하였다. 이후에는 기본 모델의 능력을 적극 활용하여 추가학습에 사용하는 데이터를 최소화 하여 효율을 높이는 기법들이 발달하였다. 최근에는 모델의 파라미터를 더욱 정밀하고 효율적으로 제어[25]하기 위한 기법(예: LoRA[26], RLHF[27])으로 발전해 왔다.

본 연구에서는 GPT-4o mini[28] 모델에 보안시스템 로그 데이터를 추가 학습하여 기존 모델의 파라미터를 미세 조정하였다. GPT-4o mini는 GPT-4를 경량화한 모델로, 2024년 7월 18일 OpenAI에서 출시되었다. 이 모델은 100만 토큰 기준으로 입력에는 0.25달러, 출력에는 0.6달러로 OpenAI모델 중에서 API 비용이 저렴한 편이며, 속도가 빠른 것이 특징이다.

2025년 현재, Llama3[29], DeepSeek-r1[30], Gemma[31] 등 다양한 고성능 모델이 등장하고 있으며, 이들은 무료로 배포되고 있다. 또한 Ollama[32]와 같은 솔루션을 활용하면 API도 쉽게 접근할 수 있다. 하지만,

무료로 배포되는 오픈모델에 비해 GPT 모델은 일반지능의 성능과 안정성이 검증되었으며, 온라인에서 API 사용이 비교적 용이한 장점이 있어 본 연구에 적용하였다. 그중에서도 GPT-4o mini는 API 비용, 성능, 속도를 균형적으로 갖추어 본 연구에서 가장 적합한 모델로 판단하였다.

GPT-4o mini의 미세 조정(Fine-tuning) 과정에서는 배치 크기, 학습률 증가율, 에포크 수 등 다양한 하이퍼파라미터를 조정할 수 있다. 그러나 본 연구에서는 샘플링 방법에 따른 성능 변화를 측정하는 것을 목표로 하므로, 하이퍼파라미터의 영향을 최소화할 필요가 있다. 이에 따라 에포크 수를 5로 고정하였으며, 그 외의 파라미터는 기본값으로 유지하였다.

6. Prompt

LLM의 높은 성능을 달성하기 위해서는 데이터 품질뿐만 아니라 프롬프트 설계도 중요한 요소이다. 이에 따라, 효과적인 프롬프트 작성 및 최적화 방법에 대한 다양한 연구가 진행되고 있다[33,34].

GPT API에서는 프롬프트를 구성할 때 system, user, assistant의 세 가지 역할을 설정하여 모델을 작동시키며, 각 역할은 다음과 같다.

- system: 모델의 전반적인 동작 방향을 설정
- user: 사용자의 입력을 전달(지도학습의 데이터와 같은 역할을 함)
- assistant: 모델이 응답하는 메시지(지도학습의 레이블과 같은 역할을 함)

system 메시지를 활용하여 보안 로그 분석의 맥락을 설정하고, user 메시지를 통해 샘플링된 데이터를 입력하여 모델이 정상/비정상을 분류하도록 구성하였다. 본 연구에서 사용한 역할별 프롬프트는 표 2와 같다.

Table 2. Prompt Structure of CoreShot Filter

Role	Setting(Example)
system	You are a security system log analysis AI. You comprehensively analyze the following information to determine whether an IP should be blacklisted: <ol style="list-style-type: none"> 1. Event type and risk level 2. AbuseIPDB (AID) report details 3. Access period, interval, and timing 4. Number of event occurrences and accessed systems 5. Information on IP country, owner (ISP, domain, etc.) Please use one of the following formats for the final conclusion: <ul style="list-style-type: none"> - If blacklisted: [Blacklist Status: 1] - If not blacklisted: [Blacklist Status: 0]

user	Please analyze the following security system logs and determine whether IP 1XX.XXX.88.92 is blacklisted. <ul style="list-style-type: none"> - Country Code: NL - Date: 2025-02-22 - Number of detections by system: { 'ips': 358 } - Event type: SIP Vulnerability Scanner (Sipvicious) - AbuseIPDB (AID) reports: 690 cases / 89 reporters - IBM reputation score: 1 point - Access within 1 day: 2 devices, 1 network - Number of logs: 358 (suspicious of burst access within short period) - First detected: 2025-02-22 06:21:55, elapsed about 3 minutes - Average access interval: 1.01 seconds (possible excessive access under 1 second) - Off-hours access ratio: 100% - No history of normal web access - IP type: Data Center / Web Hosting / Transit - ISP: Estoxy OU - Domain: estoxy.com Please comprehensively analyze the above data and determine whether this IP should be blacklisted (0 or 1).
assistant	IP 1XX.XXX.88.92 shows a relatively short access duration (1 day) but originated from an overseas (Netherlands) IP associated with a medium-risk event. A suspicious scan produced 358 logs in a short period (179 targeted scans), and it has a high number of AbuseIPDB reports. Therefore, it should be blacklisted. Final decision: This IP is blacklisted. [Blacklist Status: 1]

데이터 샘플링 방법을 제안하는 본 연구에서는 프롬프트 최적화를 위한 별도의 추가 작업 역시 수행하지 않았다. 따라서, 프롬프트 설계는 기본적인 구조를 따르는 수준에서 적용하였다.

IV. Evaluation

4장에서는 코어샷 필터를 다른 샘플링 방법과 비교하고 성능을 측정하여 방법의 우수성을 검증한다. 코어샷 필터의 성능을 측정하기 위해 먼저, "층화 샘플링, 아웃라이어, 코어셋" 세 가지 일반적인 샘플링 방법으로 데이터를 구성하고, 이를 학습한 모델과 코어샷 필터의 레이블링 정확도를 비교하였다.

1. Data

본 연구에서 사용된 데이터는 실제 보안관제센터(SOC)에서 5개월간 수집한 방화벽, IPS, 웹방화벽 보안시스템 및 웹서버 로그이다. 이를 일자와 출발지 IP주소로 집계(Aggregation)하여 314,852건으로 레이블링 대상 데이터로 전처리 하였다. 우선 전체 데이터를 수작업으로 정상/비정상(블랙리스트)로 레이블링 하였다.

테스트 및 성능 검증용 데이터로는 레이블링 난이도가 높은 685개 데이터를 공통적으로 사용하였다. 일반적으로 레이블링 난이도가 높은 데이터는 레이블 확률이 결정 경계에 가까운 데이터로 정의된다 [16]. 본 연구에서는 결정 경계 50~70% 범위에 해당하는 데이터를 선별하여 테스트 데이터를 구성하였다.

2. Sampling

코어셋 필터의 샘플링 성능을 검증하기 위해 3가지 다른 샘플링 방법으로 샘플을 추출하였다. 코어셋 필터의 최종 샘플 수가 204개이므로, 다른 방법에서도 샘플 수를 204개로 통일하였다. 샘플 수를 동일하게 맞춘 것은 성능 비교의 형평성을 위한 실험적 조치였으며, 각 기법 고유의 특성상 실제 적용 시에는 선택되는 데이터 수가 달라질 수 있다.

1) 층화 랜덤 샘플링

이벤트명 조합을 기준으로 데이터를 그룹화한 후, 각 그룹에서 하나씩 무작위로 추출하였다. 이 데이터는 IP 주소를 기준으로 집계되어 있어, 한 IP에 여러 이벤트가 조합되어 나타난다. 따라서 IP별 이벤트명 조합에서 중복을 제거하여 전체 데이터에서 유니크한 이벤트명 조합을 도출하였다. 도출된 이벤트명 조합의 수가 204개를 초과하므로, 중복이 적은 이벤트명을 우선 선택하여 최종적으로 204개로 구성하였다.

2) 아웃라이어 샘플링

숫자 타입의 모든 컬럼에 대해 각각 다음 수식과 같이 Z-Score를 계산하여 이상치를 탐지하였다. 여기서 x 는 개별 데이터의 값, μ 는 해당 컬럼의 평균, σ 는 해당 컬럼의 표준편차이다.

$$z = \frac{x - \mu}{\sigma} \quad (7)$$

임계값 2σ (약 95.4%)를 기준으로 이상치 여부를 결정하였다. 이때, 너무 많은 이상치가 나와 각 컬럼의 값 분포 특성을 보조적으로 평가하기 위해 수식 8과 같이 엔트로피

(Entropy)를 산출하였다.

$$Entropy = - \sum_i p_i \log_2 p_i \quad (8)$$

p_i 는 엔트로피의 출현 확률을 나타내고, 로그를 취하여 컬럼의 값을 합산한 정보의 불확실성이다. 엔트로피는 고 유값의 정보량인 $p_i \log_2 p_i$ 를 모두 합산 한 후 음수를 취하여 전체 불확실성을 계산한다. 즉, 값들이 다양할수록 합산 결과가 커져 엔트로피 값이 높아지게 된다.

이 데이터는 IP 주소를 기준으로 집계되어 있어, 한 IP에 대해 여러 컬럼에서 이상치가 검출될 수 있다. 각 IP에 대해 이상치로 판별된 컬럼의 수를 집계한 후, 이상치 컬럼 수가 많은 순으로 최종 204개의 샘플을 구성하였다.

3) 코어셋 샘플링[35]

전체 데이터셋의 대표성을 유지하기 위해, Farthest-Point Sampling(FPS)[36] 기반의 코어셋 샘플링을 수행하였다. 이 방법은 각 데이터 포인트 간의 거리를 계산하여, 이미 선택된 샘플 집합과의 최소 거리가 최대인 점을 순차적으로 선택함으로써 전체 데이터 분포를 효과적으로 대표할 수 있는 샘플을 도출한다.

이는 데이터 집합 X 에서 먼저 임의의 한 점 s_1 을 선택한 후, 모든 $x \in X$ 에 대해 수식 9를 계산한다.

$$d(x, \{s_1\}) = \|x - s_1\| \quad (9)$$

이후, 현재까지 선택된 샘플 집합 S 에 대해 각 점 x 의 최소 거리를 수식 10과 같이 정의한다.

$$d(x, S) = \min_{s \in S} \|x - s\| \quad (10)$$

이 값이 최대대인 점을 추가하는 과정을 k 개의 샘플이 선택될 때까지 반복한다.

특이사항으로, 공격명은 Word2Vec[37] 임베딩을 통해 벡터화되어 기존 수치형 피처에 추가되었으므로, 공격 텍스트 정보 또한 샘플링 과정에 반영되어 대표성 있는 코어셋 도출에 기여하였다.

최종적으로, 중복을 제거한 샘플들 중에서, 이상치 및 중복이 적은 항목을 우선 선택하여 전체 데이터에서 204개의 코어셋을 구성하였다. 도출된 코어셋은 이후 K-means 클러스터링을 적용하여 전체 데이터와의 클러스터 분포를 비교 분석하는 데 활용되었다.

3. Performance Metrics and Evaluation Results

본 연구에서는 사용된 데이터의 특성과 연구 목적에 따라 모델의 이진 분류 성능을 재현율(Recall)과 F1 점수로 평가하였다. 보안시스템 로그에서는 악성 로그의 비율이 정상 로그보다 적으며, 이로 인해 비정상 공격을 탐지하지 못하는 경우(False Negative)의 리스크가 더 크므로 정밀도(Accuracy)보다는 재현율(Recall)을 더 중요한 성능 지표로 설정하였다. 또한, 과도한 탐지 오류(False Positive)로 인해 발생할 수 있는 경고 피로(Alert Fatigue)를 고려하여, 정밀도(Precision)와 재현율(Recall) 간의 균형을 평가할 수 있는 F1 점수를 함께 활용하였다.

이진 분류 성능 측정 결과는 그림 2 및 표 3과 같으며 다음과 같은 관찰 결과를 얻었다.

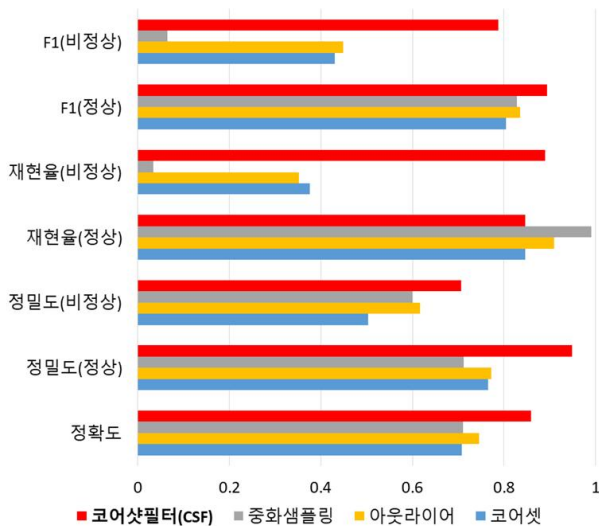


Fig. 2. Performance Comparison by Sampling Method

Table 3. Performance Test Results by Sampling Method

Method	Representativeness	Uncertainty	Distinction
CoreShot Filter	Sufficient	Sufficient	Balanced
Stratified Sampling	Provided	Lacking	Insufficient reflection of abnormal data
Outlier Sampling	Lacking	Provided	Low Recall
Coreset Sampling	Provided	provided	Not necessarily optimal data

1) 샘플링 방법별 성능 비교

코어샷 샘플은 정확도(Accuracy) 0.8593, F1-score 0.8413으로 다른 샘플링 방식에 비해 전반적으로 가장 우

수한 성능을 보였다. 특히, 비정상(악성) 탐지 성능을 측정하는 비정상 클래스의 재현율(Recall)이 0.8901로 높았으며, 정상 데이터를 정확하게 판단하는 지표인 정상 클래스 정밀도(Precision) 또한 0.9489로 가장 뛰어났다. 이러한 결과는 코어샷 샘플링이 정상과 비정상을 보다 명확하게 구분할 수 있는 데이터 샘플을 제공했음을 의미한다.

반면, 층화 샘플링(Stratified Sampling)은 비정상 클래스의 Recall이 0.0347로 매우 낮아 비정상 탐지가 거의 이루어지지 않는 불안정한 결과를 보였다. 또한, 정확도(Accuracy)도 0.7017로 코어샷 샘플링보다 약 0.14 낮게 나타나, 해당 샘플링 방법이 데이터 분포를 적절히 반영하지 못했음을 시사한다.

아웃라이어 샘플링(Outlier Detection Sampling)은 Accuracy 0.7458, F1-score 0.6304로 비교적 균형적인 성능을 보였으나, 비정상 클래스의 Recall이 0.3526으로 낮아 비정상 탐지 성능이 부족했다. 이는 일부 이상치를 기반으로 샘플링이 이루어졌으나, 모델이 패턴을 학습하기에 충분한 정보가 제공되지 못했음을 의미한다.

마지막으로, 코어셋 샘플링(Coreset Sampling)은 Accuracy 0.7085, F1-score 0.6945로 가장 낮은 성능을 보였으며, 비정상 클래스의 Recall도 0.3757로 낮아, 악성 로그 탐지에 대한 신뢰성이 부족한 것으로 나타났다. 코어셋 샘플링은 대표성과 불확실성을 균형 있게 제공하는 방법이지만, 최적의 데이터를 선택하는 데 실패하여 성능이 낮아진 것으로 추정된다.

2) 샘플링 방법별 데이터 선정 기준 분석

인공지능 학습을 위한 데이터셋, 특히 소량의 데이터셋을 다루는 경우, 대표성과 불확실성은 가장 중요한 기준이다. 대표성이 높은 데이터는 전체 데이터를 대신하여 패턴 인식에 도움을 줄 수 있으며, 불확실성이 높은 데이터를 학습하면 모델의 일반화 성능을 향상시킬 수 있다. 이에 따라 본 연구에서 비교한 샘플링 방법을 분석하면 다음과 같은 특성이 나타난다.

층화 샘플링은 대표성을 제공하지만, 비정상 데이터를 적절히 반영하지 못해 정상 클래스에 대한 학습이 편향적으로 이루어지는 문제가 발생하였다. 반면, 아웃라이어 샘플링은 불확실성을 고려하지만, 비정상 데이터만을 포함한 것이 아니므로 모델이 유의미한 패턴을 학습하기에는 정보가 부족하였다.

코어셋 샘플링은 대표성과 불확실성을 모두 균형 있게 제공하지만, 선택된 데이터가 모델 학습에 최적화되지 않아 평균적인 성능이 가장 낮게 나타났다. 반면, 코어샷 샘플링은 대표성과 불확실성을 동시에 고려하면서도, 가장

유익미한 정보를 제공하는 데이터셋을 구성하였기 때문에 정확한 학습이 가능하였고, 결과적으로 가장 높은 성능을 기록하였다.

4. Implications of the Experiment

본 연구의 결과에 따르면, 코어셋 샘플링이 대표성과 불확실성을 효과적으로 결합하여 모든 성능 지표에서 다른 방법보다 우수한 결과를 보였다. 특히 공격 탐지(비정상 클래스) 성능이 뛰어나 보안시스템에서 실용적 활용이 가능하다.

이러한 결과를 바탕으로 코어셋 샘플링을 효과적으로 적용하기 위해서는 다음과 같은 절차를 사전에 설계할 필요가 있다.

첫째, 초기 수작업 레이블링 단계에서 느슨한 모델을 구축하고 레이블링 작업자의 판단과 비교하는 절차가 추가되어야 한다. 이를 통해 레이블링의 신뢰도를 높이고, 불확실성이 높은 데이터를 효율적으로 식별할 수 있다. 이는 한 번의 계산으로 얻을 수 있는 아웃라이어나 코어셋 샘플링보다 복잡하지만 실험 결과를 통해 그만큼 실용적 가치를 증명하였다.

둘째, 샘플링된 데이터가 충분한 정보량을 가지도록 코어셋 기반 핵심 데이터셋을 정밀하게 구성해야 한다. 적절한 데이터가 포함되지 않으면 성능 저하가 발생할 가능성이 있기 때문에, 데이터의 균형성과 정보량을 고려한 샘플링 전략이 필요하다.

셋째, 코어셋 샘플링은 단순한 데이터 샘플링 기법을 넘어, 레이블링 과정의 신뢰성을 높이는 도구로도 활용 가능하다. 레이블링 작업자의 판단이 느슨한 모델과 크게 불일치하는 경우, 이를 검토함으로써 레이블링 오류를 탐지하고 정정할 수 있는 프로세스를 구축할 수 있다.

마지막으로, 대형 언어 모델(LLM)이 생성한 텍스트는 레이블링의 근거로 활용될 수 있다. 보안 로그는 데이터의 특성상 반복되는 내용이 많으며, 인력에 의해 분석이 필요한 부분은 상대적으로 소량이지만 중요한 정보를 포함한다. 반복적인 패턴은 느슨한 모델 수준에서 자동화할 수 있으며, 분석이 필요한 핵심 부분은 고도의 지능을 가진 LLM을 활용하여 보다 정교한 데이터 해석과 설명을 제공할 수 있도록 활용 전략을 수립할 수 있다.

결과적으로, 코어셋 샘플링은 소량의 핵심 데이터셋을 효과적으로 구축하는 동시에, 모델의 학습 성능과 레이블링 품질을 개선할 수 있는 강력한 방법론임이 본 연구를 통해 입증되었다.

V. Conclusions

본 연구에서는 소량의 보안시스템 로그 데이터를 활용한 LLM 미세조정의 성능을 극대화하기 위한 데이터 샘플링 방법을 탐색하였다. 기존의 데이터 샘플링 기법인 층화 샘플링(Stratified Sampling), 아웃라이어 샘플링(Outlier Detection Sampling), 코어셋 샘플링(Coreset Sampling)과 비교하여, 대표성과 불확실성을 효과적으로 결합한 코어셋 샘플링 기법을 제안하였다.

실험 결과, 코어셋 샘플링은 정확도(Accuracy), 재현율(Recall), F1-score 등 모든 주요 성능 지표에서 다른 방법보다 우수한 성능을 보였다. 특히, 비정상(악성) 탐지 성능을 평가하는 비정상 클래스의 Recall이 0.8901로 가장 높았으며, 정상 데이터를 정확하게 판단하는 정밀도(Precision)도 0.9489로 뛰어났다. 반면, 층화 샘플링과 아웃라이어 샘플링은 각각 대표성과 불확실성을 일부 반영했으나, 모델이 학습하기에 충분한 정보를 제공하지 못해 성능이 낮았다. 또한, 코어셋 샘플링은 대표성과 불확실성을 모두 고려했음에도 불구하고, 최적의 데이터 샘플을 선정하는 데 실패하여 가장 낮은 성능을 기록하였다.

이러한 결과는 소량의 데이터셋을 활용한 LLM 미세조정에서 단순한 대표성 또는 불확실성 기반의 샘플링 기법만으로는 충분한 성능 향상을 기대하기 어렵다는 점을 시사한다. 코어셋 샘플링은 대표성과 불확실성을 동시에 고려하면서도, 가장 유익미한 정보를 제공하는 데이터셋을 구축하는 데 효과적임을 본 연구를 통해 검증하였다. 본 연구의 결과는 LLM을 활용한 보안 로그 분석, 소량 데이터 기반의 퓨샷(Few-Shot) 학습, 자동화된 레이블링 및 데이터 필터링 등 다양한 응용 분야에서 활용될 수 있을 것으로 기대된다.

그러나 본 연구에서는 LLM이 생성한 문장에 대한 정량적·정성적 평가를 수행하지 못하였다. 또한, CoT(Chain of Thought) 및 강화학습(RLHF)과의 결합을 통한 프롬프트 최적화의 효과나, Llama, DeepSeek 등 오픈소스 모델과의 성능 비교도 시도하지 못하였다. 이 밖에도 최적 샘플의 개수, 불확실성 및 대표성 지표의 임계값, 유전 알고리즘에 사용되는 파라미터는 데이터에 따라 다르게 설정해야 하는 한계도 존재한다.

향후 연구에서는 보안 목적으로 활용할 수 있는 LLM 성능 평가 지표를 개발하여, 생성된 문장의 평가 방법을 체계화하고 프롬프트 엔지니어링을 통한 성능 최적화 방안을 탐색할 것이다. 또한, 오픈 LLM 모델의 적용과 상호 비교를 통해 코어셋 필터를 지속적으로 개선하고 확장할 계획이다.

ACKNOWLEDGEMENT

This work was supported by the Yeungnam University College Research Grants in 2023.

REFERENCES

- [1] Kaur, R., Gabrijelčić, D., & Klobučar, T. "Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion*, vol. 97, Art. 101804 (Apr. 2023). DOI: 10.1016/j.inffus.2023.101804.
- [2] Taddeo, M., McCutcheon, T., & Floridi, L., "Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword," *Nature Machine Intelligence*, vol. 1, no. 12, pp. 557-560, Dec. 2019. DOI: 10.1038/s42256-019-0109-1.
- [3] Farheen, M., Dash, B., Sharma, P., & Yathiraju, N., "The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 11, no. 9, pp. 81-90, Sep. 2022. DOI:10.17148/IJARCC.2022.11912.
- [4] Li, D., et al., "Large Language Models for Data Annotation: A Survey," *Proc. EMNLP 2024 / arXiv:2402.13446*, 2024.
- [5] Goel, P., Gueta, D., et al., "LLMs Accelerate Annotation for Medical Information Extraction," *Proc. Machine Learning Research (PMLR)*, vol. 225, 2023. arXiv:2312.02296.
- [6] Brown, T. B., Mann, B., Ryder, N., et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020. arXiv:2005.14165.
- [7] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y., "Large Language Models are Zero-Shot Reasoners," *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022. arXiv: 2205.11916.
- [8] Bhatt, S., Manadhata, P. K., & Zomlot, L., "The Operational Role of Security Information and Event Management Systems," *IEEE Security & Privacy*, vol. 12, no. 5, pp. 35-41, Sep.-Oct. 2014. DOI:10.1109/MSP.2014.103.
- [9] AlAhmadi, B. A., Axon, L., & Martinovic, I., "99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms," *Proc. 31st USENIX Security Symposium*, pp. 2783-2800, Aug. 2022.
- [10] Guerra, J., Catania, C., & Veas, E., "Datasets are not Enough: Challenges in Labeling Network Traffic," *Proceedings of the 17th International Conference on Network and Service Management (CNSM)*, pp. 158-166, Izmir, Turkey, Nov. 2021. DOI:10.23919/CNSM52442.2021.9615524.
- [11] Zhu, X. J. "Semi-Supervised Learning Literature Survey," *Technical Report (University of Wisconsin-Madison)*, 2005.
- [12] Engelen, J., & Hoos, H. H., "A Survey on Semi-Supervised Learning," *Machine Learning*, 2020. DOI:10.1007/s10994-019-05855-6.
- [13] Blum, A., & Mitchell, T., "Combining Labeled and Unlabeled Data with Co-Training," *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pp. 92-100, Madison, USA, 1998. DOI:10.1145/279943.279962.
- [14] Settles, B., "Active Learning Literature Survey," *Computer Sciences Technical Report 1648*, Univ. Wisconsin-Madison, 2009.
- [15] Sener, O., & Savarese, S., "Active Learning for Convolutional Neural Networks: A Core-Set Approach," *Proc. International Conference on Learning Representations (ICLR)*, 2018 (Poster). arXiv:1708.00489.
- [16] Frénay, B., & Verleysen, M., "Classification in the Presence of Label Noise: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845-869, May 2014. DOI:10.1109/TNNLS.2013.2292894.
- [17] Dutta, H., Passonneau, R. J., Lee, A., Radeva, A., Xie, B., Waltz, D., & Taranto, B., "Learning Parameters of the K-Means Algorithm from Subjective Human Annotation," *Proc. 24th Int. Florida Artificial Intelligence Research Society Conf. (FLAIRS-24)*, pp. 465-470, Palm Beach, USA, May 2011.
- [18] Alam, S., Gummadi, K. P., et al., "Debiased Label Aggregation for Subjective Crowdsourcing Tasks," *Proc. CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI LBW)*, pp. 1-7, 2022. DOI:10.1145/3491101.3519690.
- [19] Palomäki, J.-M., Jylhä, A., Salminen, J., Kandinskaya, N., & Jansen, B. J., "A Case for a Range of Acceptable Annotations," *CEUR Workshop Proceedings*, vol. 2276, 2018.
- [20] Github Repository, https://github.com/mrgrit/data_labeling
- [21] AbuseIPDB, <https://www.abuseipdb.com/>
- [22] IBM X-Force Exchange, <https://exchange.xforce.ibmcloud.com/>
- [23] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H., "How Transferable Are Features in Deep Neural Networks?" *Advances in Neural Information Processing Systems (NeurIPS 2014)*, vol. 27, pp. 3320-3328, 2014.
- [24] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT 2019*, 2019. arXiv:1810.04805.
- [25] Housley, N., et al., "Parameter-Efficient Transfer Learning for NLP," *Proc. ICML 2019*, 2019. arXiv:1902.00751.
- [26] Hu, E. J., et al., "LoRA: Low-Rank Adaptation of Large Language Models," *Proc. ICLR 2022*, 2022. arXiv:2106.09685.
- [27] Ouyang, L., et al., "Training Language Models to Follow Instructions with Human Feedback (InstructGPT)," arXiv:2203.02155, Jan. 2022.
- [28] GPT-4o mini <https://openai.com/>

- [29] Grattafiori, A., et al., "The Llama 3 Herd of Models," arXiv:2407.21783, Jul. 2024. DOI:10.48550/arXiv.2407.21783. <https://doi.org/10.48550/arXiv.2407.21783>
- [30] Guo, D., Yang, D., Zhang, H., et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv:2501.12948, Jan. 2025.
- [31] Gemma Team et al., "Gemma 3 Technical Report," arXiv preprint arXiv:2503.19786, Mar. 2025.
- [32] Ollama, <https://ollama.com/>
- [33] Reynolds, L., & McDonell, K., "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," Proc. CHI 2021 Extended Abstracts, 2021. DOI:10.1145/3411763.3451760.
- [34] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A., "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," arXiv:2402.07927, 2024.
- [35] Bachem, O., Lucic, M., & Krause, A., "Practical Coreset Constructions for Machine Learning," arXiv:1703.06476, Mar. 2017. DOI:10.48550/arXiv.1703.06476.
- [36] Gonzalez, T. F., "Clustering to Minimize the Maximum Intercluster Distance," Theoretical Computer Science, vol. 38, pp. 293-306, 1985.
- [37] Mikolov, T., Chen, K., Corrado, G., & Dean, J., "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, Jan. 2013.

Authors



Doo-Yong Jeon received the B.S. degree in Advertising and Public Relations from Keimyung University, Daegu, Korea, in 2006, and the M.S. degree in Computer Science and Engineering from Kyungpook National

University, Daegu, Korea, in 2019. He is currently pursuing the Ph.D. degree in Computer Science and Engineering at Kyungpook National University. Jeon is an Assistant Professor at Yeungnam University College, Daegu, Korea. Prior to this, he worked as an AI and Data Scientist at DGB Daegu Bank. His research interests include AI agents, AI-powered cybersecurity, and autonomous Security Operation Centers (Self-SOC).