2025

JOONGBU UNIVERSITY

중부대학교 정보보호학과

LLM 기반 IDS 경보 로그의 정오탐 자동 판단 API 시스템 개발

T. ACE I 홍영재 유한영 성민욱

01

발표 목차

Summary

Suricata IDS 로그의 정오탐 분류를 위해 LLM 기반 자동 분석 시스템을 구축하였습니다. 본 발표에서는 문제 정의, 시스템 설계, 실험 구성 및 평가 결과를 중심으로 설명드립니다.

Index

02

배경 및 문제 정의

IDS 경보의 오탐 문제와 경보 피로도

오탐(False Positive)은 보안 운영 센터(SOC) 분석가들에게 과도한 경보를 발생시켜 경보 피로도(Alert Fatigue)를 유발하며, 이는 실제 위협에 대한 대응 지연이나 누락으로 이어질 수 있습니다.

Suricata IDS의 한계

Suricata는 룰 기반 탐지로 인해 정상 요청도 오탐지할 수 있습니다.
공격의 맥락이나 의도를 고려하지 않아 경보의 신뢰도가 낮습니다.

기존 대응 방식의 한계

룰셋 튜닝과 SIEM 상관분석, 머신러닝 분류기 등이 오탐 대응에 활용되어 왔지만, 이들은 대부분 결과 중심의 필터링에 머물며, 공격 의도 판단이나 판단 근거 설명이 어렵습니다.

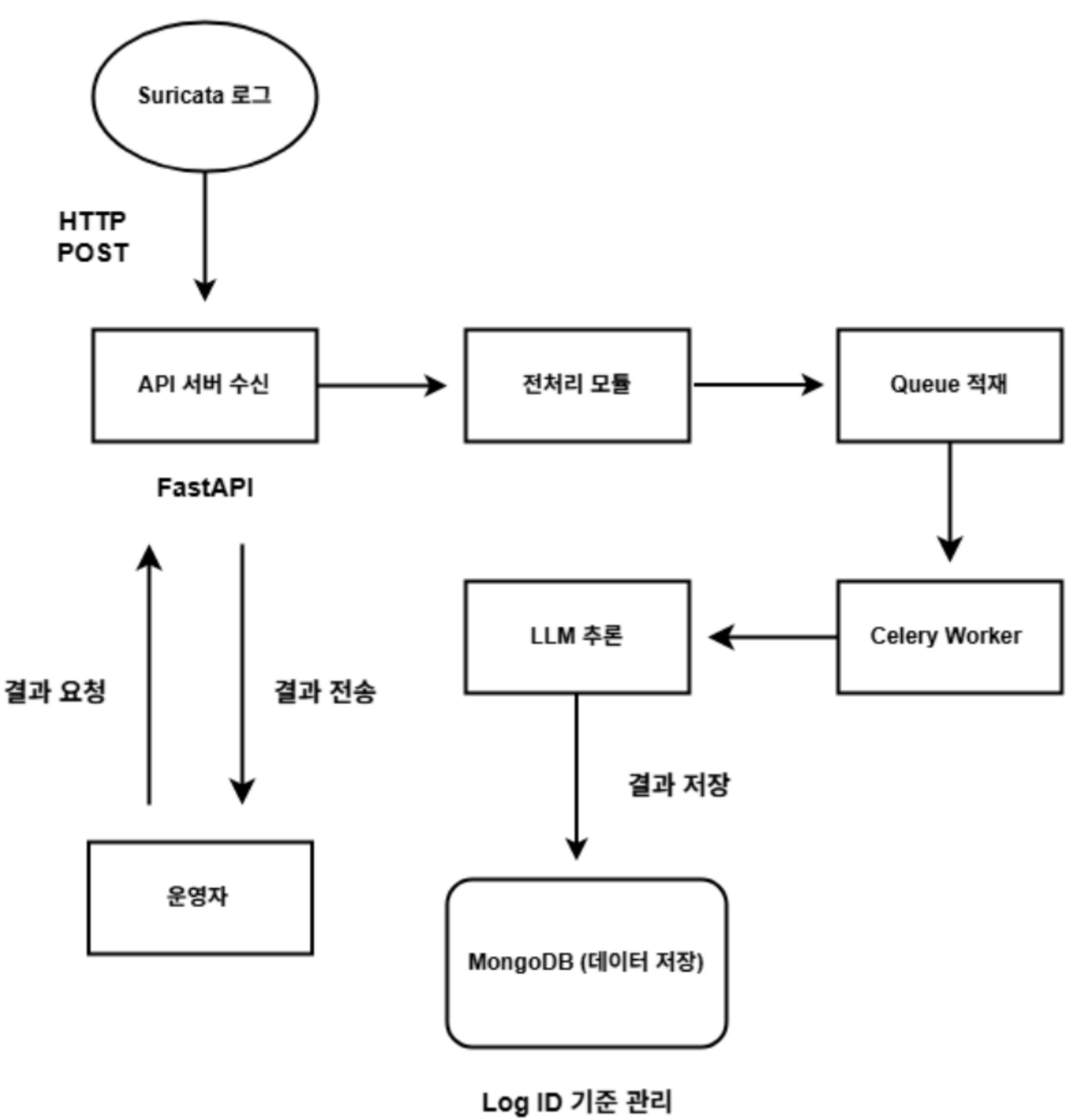
03

프로젝트 목적

Suricata IDS 로그를 입력으로 받아, 공격 의도를 기반으로 정·오탐을 자동 분류하고, 그 근거를 자연어로 설명하는 LLM 기반 보안 판단 시스템을 구현하는 것이 본 연구의 목적입니다.

04

시스템 구성도


Suricata 로그
HTTP
POST
API 서버 수신
전처리 모듈
Queue 적재
FastAPI
결과 요청
결과 전송
LLM 추론
Celery Worker
결과 저장
운영자
MongoDB (데이터 저장)
Log ID 기준 관리

05

LLM 프롬포트 구성

1. 역할 지정

2. 정오탐 판단 기준

3. 출력 형식 지정

4. 판단 지침

You are a cybersecurity analyst specialized in IDS alert triage. Given a Suricata log, determine whether the event represents a true positive or a false positive based on the intent of the activity.

Judgment Rule:
- Classify as True Positive if the log reflects clear adversarial intent, regardless of whether the attack was successful or not.
- Classify as False Positive if the activity is benign, authorized, or lacks adversarial intent.
- Adversarial intent includes behaviors such as reconnaissance and scanning activities, exploitation attempts, credential harvesting, unauthorized login attempts, or malware delivery activities.
- When uncertain or when the available evidence is inconclusive, classify as False Positive unless multiple strong indicators of adversarial intent are present.

Use this exact 6-line output format:
1. attack: true or false
2. label: True Positive or False Positive
3. category: Reconnaissance, Web Application Attacks, Malware Delivery & Infection, C2 / Botnet Communication, Phishing / Credential Theft, Denial of Service, Cryptojacking or BENIGN
(Always specify a category. Use BENIGN if attack is false.)
4. severity: High, Medium, or Low
5. reason: Provide a detailed explanation using multiple relevant fields. Quote at least one actual field value, and explain how the field relationships support your conclusion. The explanation should be concise, limited to 4-5 well-structured sentences.
6. recommendation: Suggest a clear follow-up action (e.g., block IP, monitor, escalate, no action required).

Important Judgment Principles:
- Use only the information provided in the log. Do not speculate or assume context beyond the given fields.
- Always give a clear, binary decision. Avoid uncertain or vague answers — the behavior is either malicious or not.
- Do not rely on a single field to make your decision. Instead, evaluate the log holistically using all available indicators, including HTTP metadata (e.g., method, URL, user-agent, http_response), payload content, TLS fingerprints, DNS queries, IP/port information, and flow statistics. Judgment must be made in context.
- Do not use response status (e.g., 404, 403) or success/failure when determining attack. Attack is judged only by malicious intent.
- If present, the payload_printable field should be the primary basis for intent judgment, but when absent, intent must be inferred from structural patterns across other fields such as URL, method, user-agent, flow, DNS, or TLS metadata.
- Assign severity based on the potential impact if the attack succeeded: "High" for severe system compromise or credential theft, "Medium" for moderate risks such as information disclosure or initial footholds, and "Low" for minor issues like simple reconnaissance or benign tool usage.

06

RAG + few shot 시스템 구성

Query 텍스트 생성

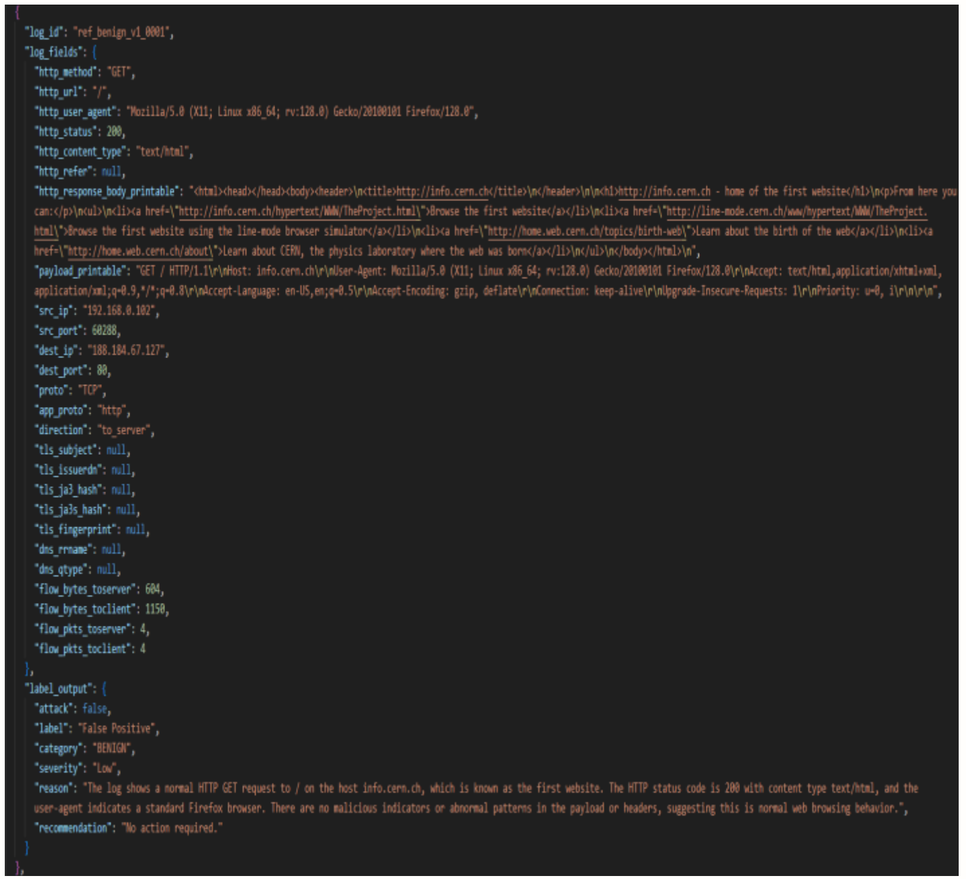
입력된 Suricata 로그에서 의미 있는 필드들을
조합하여 검색 질의(Query)를 생성

유사 로그 검색 (Retriever)

Query 텍스트와 가장 유사한 과거 로그
데이터를 검색 (가장 유사한 2개의 데이터 출력)

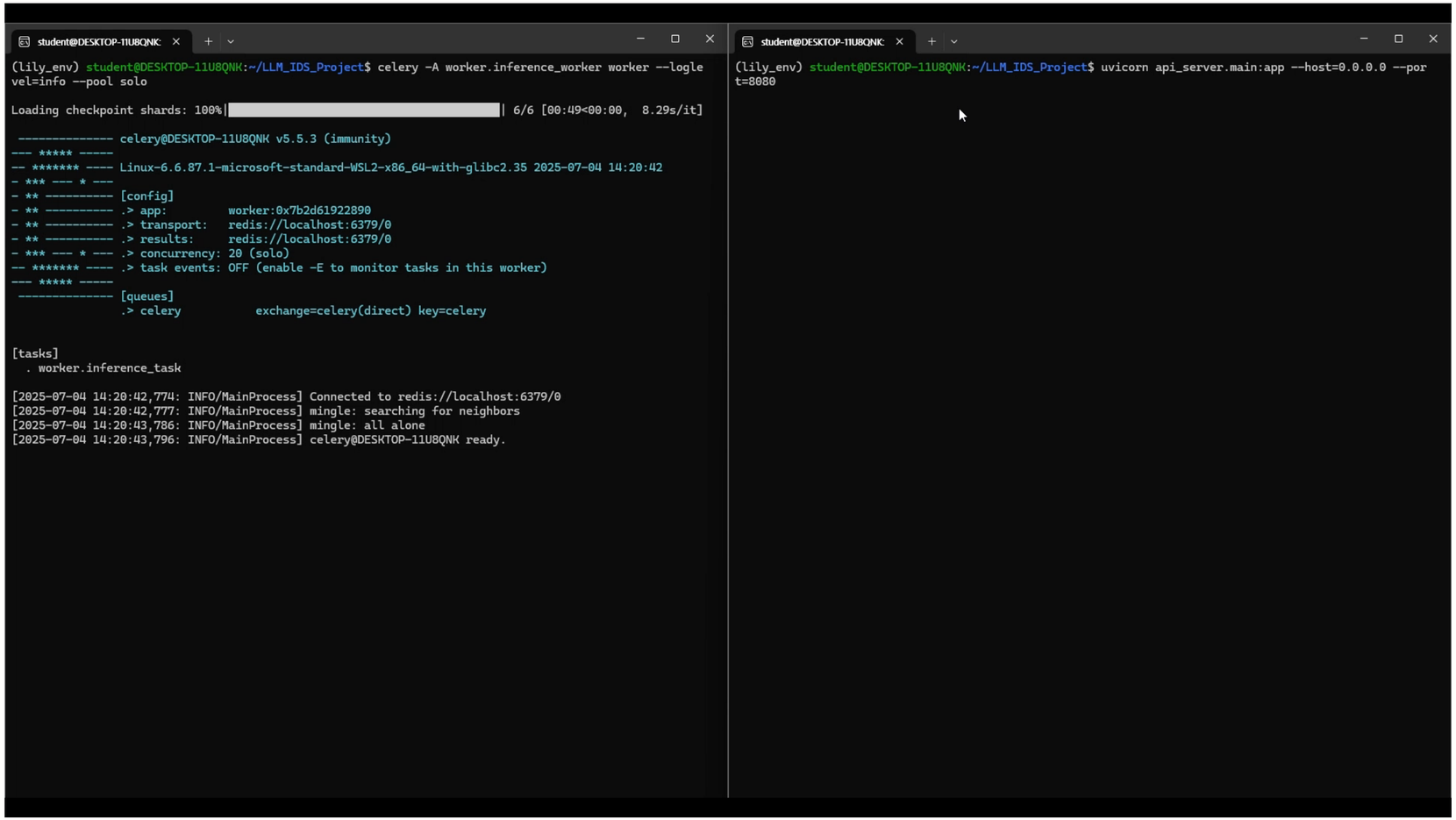
Knowledge 예시 구성 (Few-shot)

Example 1~2: DB 로그 + 판단 결과
Prompt: 지침 + 예시 + 입력 로그



07

시연 영상


student@DESKTOP-11U8QNK:
(lily_env) student@DESKTOP-11U8QNK:~/LLM_IDS_Project$ celery -A worker.inference_worker worker --loglevel=info --pool solo
Loading checkpoint shards: 100%| 6/6 [00:49<00:00, 8.29s/it]
celery@DESKTOP-11U8QNK v5.5.3 (immunity)
Linux-6.6.87.1-microsoft-standard-WSL2-x86_64-with-glibc2.35 2025-07-04 14:20:42
[config]
.> app: worker:0x7b2d61922890
.> transport: redis://localhost:6379/0
.> results: redis://localhost:6379/0
.> concurrency: 20 (solo)
.> task events: OFF (enable -E to monitor tasks in this worker)
[queues]
.> celery exchange=celery(direct) key=celery
[tasks]
. worker.inference_task
[2025-07-04 14:20:42,774: INFO/MainProcess] Connected to redis://localhost:6379/0
[2025-07-04 14:20:42,777: INFO/MainProcess] mingle: searching for neighbors
[2025-07-04 14:20:43,786: INFO/MainProcess] mingle: all alone
[2025-07-04 14:20:43,796: INFO/MainProcess] celery@DESKTOP-11U8QNK ready.
student@DESKTOP-11U8QNK:
(lily_env) student@DESKTOP-11U8QNK:~/LLM_IDS_Project$ uvicorn api_server.main:app --host=0.0.0.0 --port=8080

LLM 정확도 실험 결과

LLM의 판단 정확도 향상을 검증하기 위해, 동일한 Suricata 경보 로그에 대해 RAG 기반 프롬프트와 일반 프롬프트(knowledge 미포함)를 비교 실험하였습니다.
RAG 방식은 유사 판단 사례를 함께 제공함으로써, 더 일관되고 정확한 정오탐 판단 결과를 도출할 수 있는지를 평가하는 것이 목적입니다.

08

실험 목적

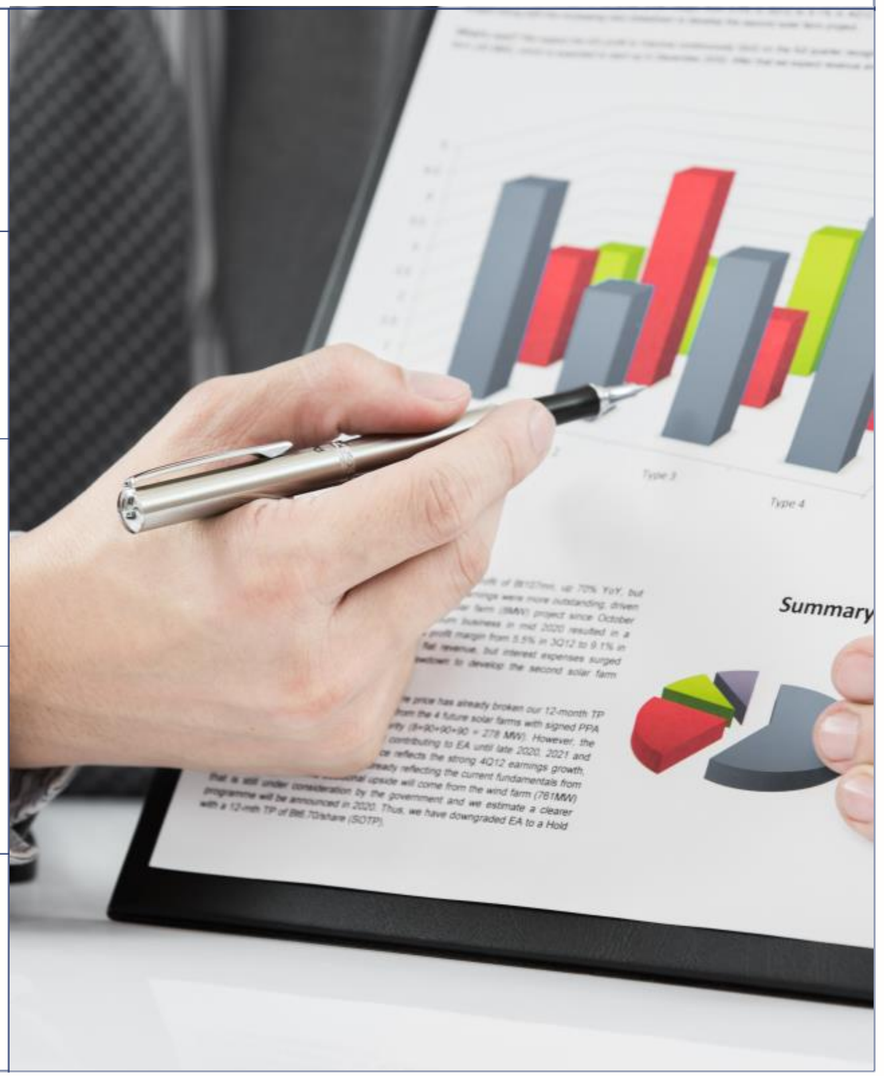
01. **LLM의 정오탐 판단 능력 평가**
Suricata 로그에 대해 LLM이 정·오탐을 정확히 분류하는지 검증

02. **설명 가능성(Explainability) 검증**
각 판단에 대해 LLM이 논리적이고 타당한 이유를 설명하는지 평가

03. **공격 시나리오 전반에 대한 일반화 확인**
정찰, 웹 공격, 악성코드 등 다양한 상황에서 일관된 판단이 가능한지 확인

04. **RAG 기반 Few-shot 학습을 통한 성능 향상 평가**
RAG 방식 적용을 통해, 추론 성능 및 설명력 향상을 비교

05. **실전 활용 가능성 탐색**
보안 운영 환경에 적용 가능한 자동화 시스템으로서의 실용성을 평가



09

실험 데이터 구성

항목	내용
총 평가 로그 수	314개 (정탐 130개, 오탐 184개)
대분류 기준	3종 (정찰, 웹 공격, 악성코드) + 정상 활동
수집 방식	시나리오 기반 직접 실행 후 Suricata 탐지 로그 수집
라벨링 기준	수작업 라벨링 (공격 의도 기반 판단)
평가 방식	V1: 단일 프롬프트 / V1-RAG: KB 기반 Few-shot (단계적 성능 비교)
KB 구성	총 27개 (정탐 18, 오탐 9) – 대표 시나리오 기반 수작업 선별

🔎 최종 결과:
{
  "attack": true,
  "label": "True Positive",
  "category": "Web Application Attacks",
  "severity": "High",
  "reason": "The request accesses /?<script>alert('Vulnerable');</script>, which is a clear indication of a Command Injection attempt. The payload includes a script tag, suggesting an attempt to execute malicious code. The user agent 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.169 Safari/537.36' indicates a web browser, further confirming the attacker's intent.",
  "recommendation": "Immediately block IP 192.168.75.198 to prevent further exploitation. Conduct a thorough vulnerability assessment and patch any identified security flaws. Implement strict input validation and sanitization to prevent similar attacks in the future."
}

10

실험 결과

평가 방식	Accuracy	TP F1-score	FP F1-score
단일 프롬프트	59.55%	0.601	0.614
RAG 기반	**90.76%**	**0.890**	**0.922**

단일 프롬프트

```
✅ 평가 완료 - 정확도: 59.55%
결과 요약: {True: 187, False: 127}

📊 정밀도 / 재현율 / F1-score:
                precision    recall  f1-score   support

 True Positive      0.514     0.723     0.601       130
False Positive      0.782     0.505     0.614       184

    micro avg       0.619     0.596     0.607       314
    macro avg       0.648     0.614     0.607       314
 weighted avg       0.671     0.596     0.608       314
```

RAG 기반

```
✅ 모델 평가 완료 | 정확도: 90.76%
결과 요약: {True: 285, False: 29}
결과 저장 위치: /home/student/project_eval/data_accuracy/rag_eval_result.csv

📊 상세 성능 지표:
                precision    recall  f1-score   support

 True Positive      0.911     0.869     0.890       130
False Positive      0.910     0.935     0.922       184

    micro avg       0.911     0.908     0.909       314
    macro avg       0.911     0.902     0.906       314
 weighted avg       0.911     0.908     0.909       314
```

11

결론 및 향후 계획

결론 1	LLM을 활용해 Suricata 로그 기반 정·오탐 판단이 가능함을 확인함.
결론 2	RAG 기반 few-shot 적용 시 판단 성능이 크게 향상됨.
결론 3	보안 분야에서 LLM의 신뢰성 확보 가능성을 확인함.

향후 계획

고품질 판단 사례의 축적 및 선택적 미세조정(LoRA)을 통해 정밀도를 높이고,
RAG + 위협 인텔리전스 연계를 통한 판단 성능 향상과 보안 자동화를 지속할 예정.
또한, 향후 대규모 정량 평가를 통해 신뢰성과 일반화 가능성을 검증할 계획.

감사합니다.

중부대학교 | 홍영재 유한영 성민욱