더 나은 세상을 위한 원자력기술
국민과 세계가 지지하는 한국원자력연구원



Measuring the Inference Cost of Adversarial Machine Learning Mitigations:

A Case Study on Diagnostic AI Supporting Autonomous Operation

2026.10.30.

Ju Hyung Song, JaeKwan Park, Jae-Gu Song, Yonggu Lee,
Seongyeol Oh, Inhye Hahm, Young Jun Lee

KAERI 한국원자력연구원
Korea Atomic Energy Research Institute

더 나은 세상을 위한 원자력기술
국민과 세계가 지지하는 한국원자력연구원

KAERI 한국원자력연구원 Korea Atomic Energy Research Institute

목 차

01. SMR 자율운전과 진단 AI
02. 규제·표준 동향
03. 규제지침 초안의 완화기법
04. 가용성 관점의 근거
05. 본 연구의 위치와 범위
06. 예측 경로 (C0 ~ C3)
07. 측정 설정
08. 결과 – Q1 규모 · Q2 배분 · Q3 예측가능성
09. 한계점, 향후 과제 및 결론





01 SMR 자율운전과 진단 AI의 역할

- **SMR 자율운전: 다수의 입력·발전소 상태·추세를 종합하는 운전 개념**
- AI는 이러한 운전을 위한 지원기술[1] (enabling technology)로 간주됨
- 주요 개발사[2]는 진단과 운전원 지원에 AI 활용을 검토 중임을 공개적으로 밝힘
- **본 연구의 대상: 계측 신호를 입력받아 발전소 상태를 진단·분류하고, 그 결과를 운전원에게 권고 정보로 제공하는 기능**





[1] NRC ML24184A132의 자율(Autonomous) 정의에서 AI·ML을 '지원하는 기술'로 기술 (N-STAR §2.1) [2] NRC 2021 의견수렴(NRC-2021-0048) 제출 의견서 및 2024 NRC AI 워크숍 발표

핵심 SMR 자율운전 논의는 발전소 상태를 진단·분류하는 AI 기능이 검토되고 있다는 전제 위에서 진행됨

01 자율운전의 진화와 새로운 보안 위협의 등장

자동(Automated) vs 자율(Autonomous)

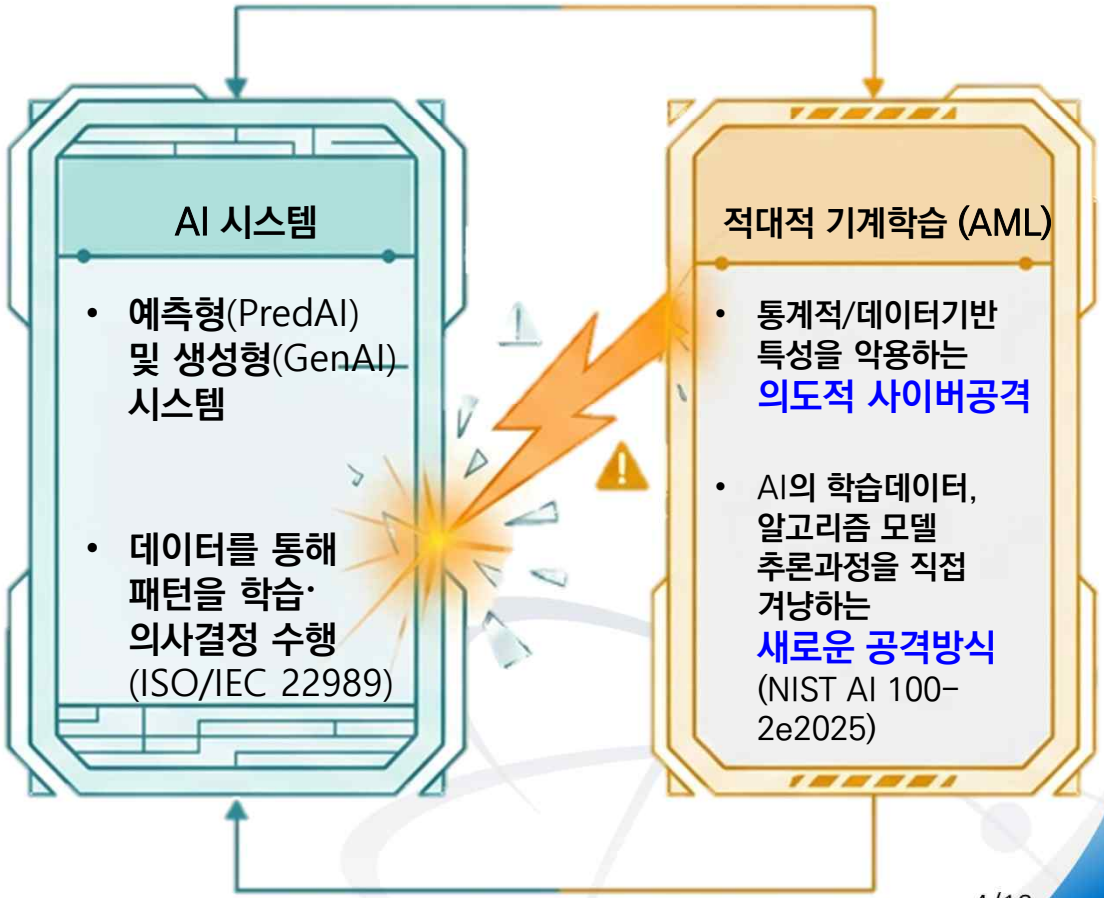
자동(Automated)

- **동작 방식:** 고정된 설정 값 및 사전 정의된 스크립트 기반
- **특징:** 독립 의사결정 부재
- **의존성:** 인공지능(AI) 기능 불필요

자율(Autonomous)

- **동작 방식:** 입력·추세 종합한 독립적 행동 수행
- **특징:** 자체 의사결정·진단·분석 수행
- **의존성:** AI 기반 지원기술 필수

(NRC ML24184A132; NUREG-0700 Rev.3 **기준**)



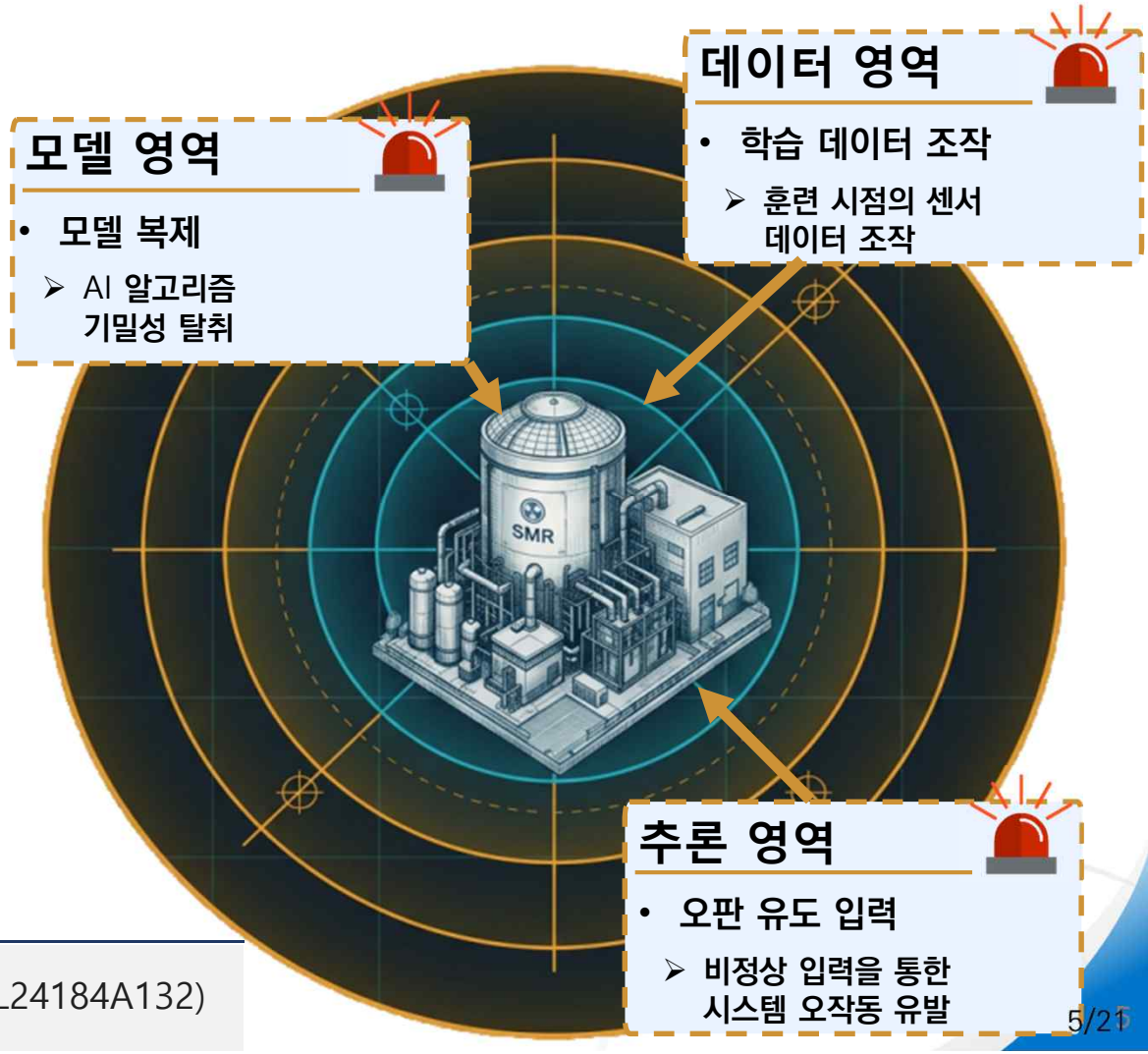
01 Attack Surface의 확장

국내 동향

- **혁신형소형모듈원자로**(i-SMR) 10대 **혁신기술 포함**
  - ➢ AI **기반 설비 상태감시, 고장진단 및 제어계통 자동화 수준 향상**

국외 동향

- **미** GEH BWRX-300(300 Mwe**급** SMR)
  - ➢ DOE ARPA-E GEMINA **프로그램을 통한 주요 기기 디지털 트윈 및** AI **예측 기술개발 도입**



모델 영역
- **모델 복제**
  - ➢ AI **알고리즘 기밀성 탈취**

데이터 영역
- **학습 데이터 조작**
  - ➢ **훈련 시점의 센서 데이터 조작**

추론 영역
- **오판 유도 입력**
  - ➢ **비정상 입력을 통한 시스템 오작동 유발**

핵심 AI 시스템은 새로운 공격벡터를 제공할 수 있음(NRC ML24184A132)

01 미 NRC 규제 격차 분석(AIRGA) 및 현행 규제의 사각지대 확인

NRC AIRGA 보고서

- **현행 규제는 전통적** SW **대응 중심**, AI **도입 시, 잠재격차 공식 인정**

주요 식별 위협

- NRC: **일반적 보안 위협으로 회피, 모델추출 등 주요** AML **공격 명시**
- EU ENISA: IT **환경에서** AI 확산 속 AML 공격 위협 대두 보고



372개의 활성 규제 지침(Active RGs) 중 71개에서
AI 도입 시 잠재적 규제 격차(Gap) 발생 확인

관련 표준 부재

- **규제 격차를 메울만한 실용적 · 성숙한 'AI 전문 표준' 아직 부재함을 인정**

핵심 AI 도입시, AML 공격 완화 대응 필요성 대두

02 자율운전·AI에 대한 사이버보안 규제·표준·지침 동향

2021
NRC RFI
상용 원전 운영에서의 AI 역할에 관한 공개 의견수렴

2023
RG 5.71 Rev.1
사이버보안 규제지침 개정

2024
NRC AIRGA[3]
AI 규제 격차 평가 (500건 이상 RG 검토)

2025
NIST AI 100-2e2025
AML 공격 및 완화기법 분류체계

2026.06
DG-5085[4]
RG 5.71 Rev.2 의견수렴 초안

주요 관련 조항 (부록)

DG-5085

D.2.3.6 **Autonomous System**

D.3 **사이버보안 응용을 위한 인공지능**

NIST 참조: AI 100-2e2023 인용으로 작성됨

ISO/IEC 27090[5] (발간 절차 중)

Cybersecurity - AI - Addressing security threats and compromises to AI systems

[3] 517개 RG 중 유효 372개 평가, 71개에서 잠재 격차 식별 (ML24290A059)
[4] ML24051A205
[5] FDIS 투표 2026.8 종료, 발간 단계



핵심 규제지침 초안(DG-5085, 2026.6월 공개)이 AI·자율운전 관련 조항을 도입하는 단계에 이름

03 규제지침 초안의 완화기법과 현재의 평가 지표

규제지침 초안

DG-5085

- 입력 검증
- 적대적 훈련
- 최소 권한 및 접근통제
- 모델 출처 관리 및 암호화
- 시험·평가·확인·검증(TEVV)

➢ **AML 완화기법이 (초안 기준) 통제 항목으로 포함 되기 시작함**

기술 참조 문서 (주요 완화기법)

- NIST AI 100-2e2025: AML 공격·완화기법 분류체계 (공통 기술 참조)
- MITRE ATLAS: 운영 관점의 완화조치[6] 지식체계
- OWASP LLM Top 10[7]: 실무 보안 지침

➢ **공통 관찰: 기술 문헌의 완화기법 성능은 대부분 방어 효과로 보고되어 왔음**

그렇다면 완화기법 적용 비용은 어떤 지표로 함께 고려할 수 있는가?

※ DG-5085는 NIST AI 100-2의 2023판(e2023)을 인용하며, 본 연구는 2025판(e2025)을 기준으로 함. 최종 지침이 참조할 판본은 의견수렴 결과에 따라 달라질 수 있어 미정임.

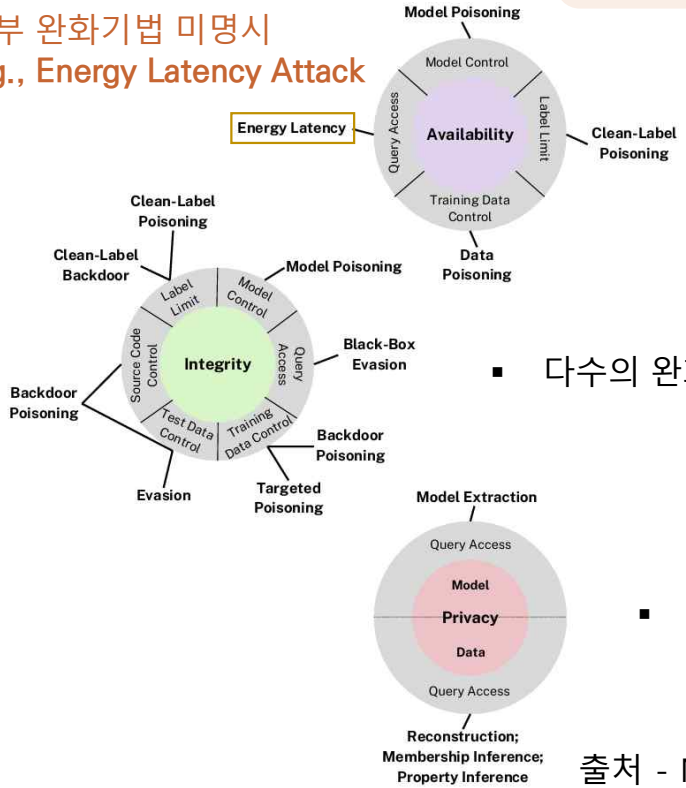
[6] NIST는 '완화기법', MITRE ATLAS는 '완화조치'로 구분 표기, [7] 논문 참고 문헌은 2025판(LLM10 Unbounded Consumption); 2026판에서는 LLM06으로 재배치

핵심 완화기법이 보안통제로 등장하기 시작한 반면, 기술 문헌의 평가 지표는 방어 효과에 집중되어 왔음

04 가용성 관점의 근거

MULTIPLE EVALUATIONS

- 일부 완화기법 미명시
  e.g., Energy Latency Attack



- 다수의 완화기법 제시
- 완화기법 제시됨

- 완화기법의 공통 구조[8]: 단일 입력에 대해 모델을 여러 번 평가하고 결과를 집계

➢ **방어 강도는 추론당 연산량으로 구매됨**
= 지연과 자원 소모 = 가용성 문제

- 증폭된 연산은 공격자가 보내는 어떤 질의로도 유발됨; 무결성을 위해 도입된 방어가 내부에서 가용성 부담을 만들어냄
- NIST 분류체계: 일부 가용성 공격(예: Energy Latency Attack[9])에 대한 완화기법은 "미명시" 상태

- **본 연구는 지금까지 비교적 덜 다뤄진 축인 가용성 관점에서 이 부담을 측정함**

출처 - NIST AI 100-2 (e2025)

[8] 오판 유도 입력(Evasion) 완화기법: 적대적 훈련·무작위 평활화·입력 복원·앙상블 (NIST AI 100-2e2025) [9] 과부하 유발(NISTAML.014): 연산 부담을 높이는 입력으로 응답 지연 증가

핵심 완화기법 자체가 유발하는 지연·자원 부담(가용성)은 비교적 주목받지 못했으며, 본 연구가 측정하는 축임

05 본 연구의 위치와 범위

예비 단계의 측정 시연	데이터: 공개 시뮬레이션 (공개 NPPAD 데이터셋)	특정 SMR 설계에 대한 검증이 아님

이 전제 위에서, 측정으로 답할 세 가지 질문

Q1 규모



비용은 얼마나 큰가?

- 구조적 배수와 실측 지연으로 완화기법의 추론 비용을 정량화

Q2 배분



비용은 누가 부담하는가?

- 공격 상황에서 공격자 측과 방어자 측 비용을 함께 측정

Q3 예측가능성



비용은 예측 가능한가?

- 구조적 배수만으로 실측 지연을 예측할 수 있는지 모델별로 검증

핵심 본 연구는 예비 단계의 측정 시연으로, 비용의 규모(Q1)·배분(Q2)·예측가능성(Q3)에 측정으로 답함

06 완화기법으로 조립한 예측 경로 (C0 ~ C3)



- 완화기법을 누적 적용한 네 가지 예측 경로; 평가 횟수 1 → 25 → 25 → 75는 경로 조립 방식으로 고정되는 구조적 값임.
- 이 값들은 데이터와 공격 방법에 관계없이 성립하며, 실측 지연이 실제로 이를 따르는지는 Q1·Q3에서 검토함.

※ 평활화 노이즈는 고정 시드를 사용; 동일 윈도우를 단독으로 재질의하면 동일 라벨을 반환

[10] 양자화는 NIST 분류의 '입력 복원(input restoration)' 계열 [11] MITRE ATLAS AML.M0006 Use Ensemble Methods

핵심 검토한 구성은 추론당 모델 평가 횟수를 1에서 25, 75로 늘리며, 이 배수는 경로 조립으로 고정되는 구조적 값임

07 측정 설정: 데이터, 공격, 분류기

데이터 선정

- 공개 NPPAD 시뮬레이션 데이터셋 [Qi et al., Scientific Data 2022]
- 4클래스 과제: 정상 운전 + 사고 상태 3종
- 입력: 윈도우 단위 다변량 계측 신호

- **채택 이유: 운영 중인 SMR의 계측 데이터는 공개되어 있지 않으며,**
- **여기서 측정하는 양은 데이터를 생성한 발전소가 아니라 경로 조립에서 비롯됨**

공격 설정

- HopSkipJump: 결정 기반 블랙박스 공격 [Chen et al., IEEE S&P 2020]
- 배치된 경로 전체를 공격; 모든 질의가 양자화·노이즈·투표를 거치며 출력 라벨만 관측
- 모델이나 완화기법에 대한 내부 지식은 가정하지 않음
- 교란 한계: 각 신호 범위의 3% (계측 채널 불확실도로 가정)
- 모든 측정 ×5 반복; 공격 실측 소요시간 기록

분류기: 랜덤 포레스트(RF)

트리 순회에 의한 추론

- 선행 NRC 연구(TLR-RES/DE-2024-003)는 고전적 알고리즘 5종을 비교하여 성능·설명가능성 측면에서 RF를 선정
  - ➢ **그 방법론(윈도잉, 전처리)을 계승**

분류기: 1D CNN

배치 텐서 연산에 의한 추론

- NRC 연구 범위 밖의 딥러닝 모델을 보완
- **RF와 정반대의 연산 특성**
  - ➢ **Q3 검증을 위한 양극단을 구성**

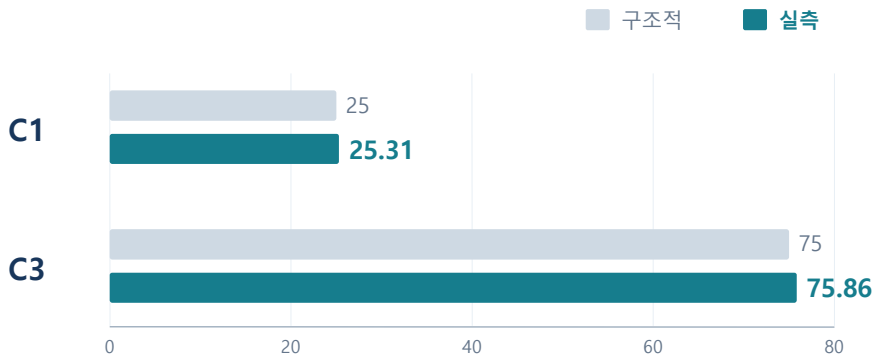
- **공통 출발점: clean 매크로 F1 = 1.000 (RF) / 0.9996 (CNN) → 이후의 성능 저하는 공격 또는 완화기법에 기인**

핵심 RF는 선행 NRC 연구의 선정을 계승; 그 범위 밖의 딥러닝 모델인 CNN은 연산 특성상 정반대 극단을 구성

08 Q1-a. RF 추론 지연: 구조적 배수를 따름

구성	평가 횟수	지연 (ms/윈도우)	비율
C0	1	0.259	1.00
C1	25	6.543	25.31
C2	25	6.559	25.37
C3	75	19.611	75.86

표 I의 RF 열
(윈도우당, 5회 평균; 비율은 반올림 전 지연값 기준)



C0 = 1 / 선형 척도 / 비율

시사점

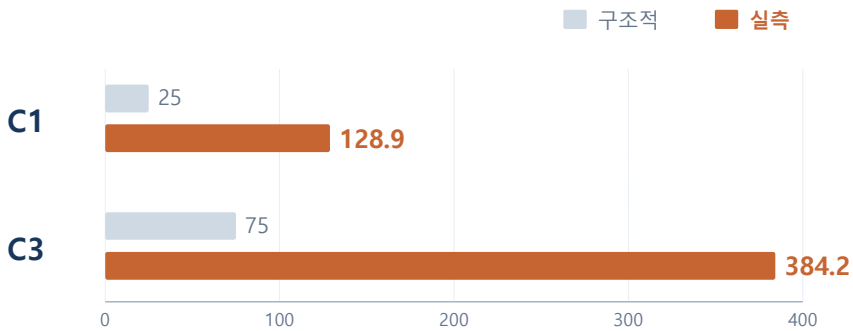
- 지연 비율(25.31 / 75.86)이 구조적 값(25 / 75)을 근접하게 따름; 이 트리 순회 모델에서는 구조적 추정이 유효했음
- C1→C2: 양자화에도 지연이 사실상 불변(6.543→6.559 ms), 평가 횟수를 추가하지 않는 완화기법임을 확인
- clean 매크로 F1은 모든 구성에서 1.000 유지

핵심 RF 지연 비율(25.31, 75.86)은 구조적 배수(25, 75)를 근접하게 따름: 구조로부터 예측 가능한 사례 (모든 모델로 일반화 않음)

08 Q1-b. CNN 추론 지연: 구조적 배수를 초과 (약 129배, 384배)

구성	평가 횟수	지연 (ms/윈도우)	비율
C0	1	0.007	1.00
C1	25	0.851	128.9
C2	25	0.850	128.8
C3	75	2.536	384.2

표 I의 CNN 열 (윈도우당, 5회 평균)



C0 = 1 / 선형 척도 / 비율

시사점

- 실측 비율(약 129 / 384)이 구조적 값(25 / 75)을 크게 초과; 384.2 / 75 ≈ 5.1배 이탈
- 원인: 무방어 추론이 배치 연산으로 0.007 ms에 불과 → 노이즈 생성·집계의 고정 비용이 지배
- **시사점: 구조적 배수만의 추정은 모델에 따라 5배 이상 빗나갈 수 있어, 배치 구성에서의 측정을 가리킴**
- 부수 비용: clean 매크로 F1도 소폭 저하 (0.9996 → 0.992~0.995)

핵심 CNN 지연은 약 129배·384배로 구조적 배수를 크게 상회; 구조적 추정만으로는 5배 이상 과소평가할 수 있음 (절대값은 구현·환경에 의존)

08 Q2-a. 방어 효과 (공격 성공률 0.295 → 0)

구성	공격 성공률
C0 (무방어)	0.295
C1	0.000
C2	0.007
C3	0.003

표 II의 성공률 열 (5회 평균)



검증: 성공률 0 vs. 출력의 확률성

- 결정 기반 공격은 관측 라벨로부터 경계를 추정 → 출력이 확률적이면 0은 경계 추정 실패를 반영할 수도 있음
  - ➢ **노이즈 필드를 완전히 고정(결정론적 경로)하여 재공격: 5회 모두 0.000 (윈도우당 389 질의), 확률적 허상 배제**

조건: 이 결과를 읽는 범위

- 공격자는 배치된 경로에 질의하지만 완화기법에 대한 지식은 활용하지 않음 (비적응형)
- 완화기법을 인지한 적응형 공격은 별도의 향후 과제로 남김

핵심 비적응형 조건에서 공격 성공률 0.295 → 0; 결정론적 재실행으로 확률적 허상 배제 (적응형 공격은 별도 사안)

08 Q2-b. 비용의 비대칭적 배분

구성	성공률	공격자 질의 수	방어자 평가 횟수	방어자 비율
C0	0.295	305	305	1.0
C1	0.000	400	10,000	32.8
C2	0.007	402	10,050	33.0
C3	0.003	404	30,300	99.3

표 II (윈도우당, 5회 평균)



C0 → C3

1.32 x 75 ≈ 99.3

공격자 질의

1.32배 305 → 404

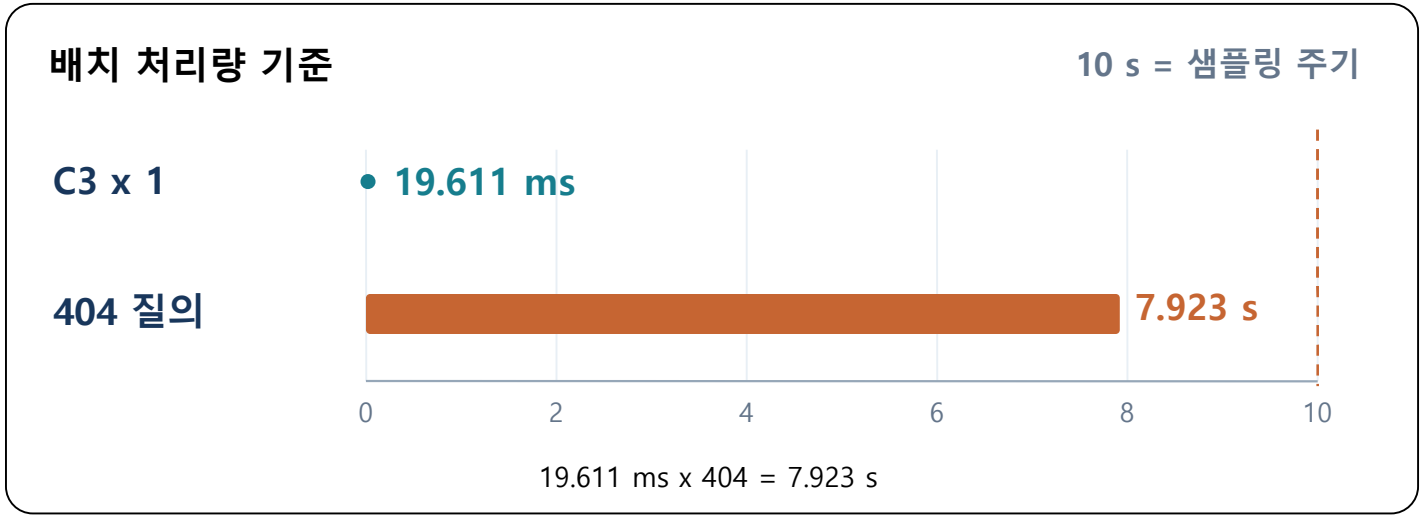
방어자 평가

99.3배 305 → 30,300

- **결과 해석:**
  - ➢ 두 증가율의 비율 약 75는 독립적인 발견이 아니라 C3 구성(노이즈 복제 25개 × 모델 3개)의 구조적 귀결임; 결정 기반 공격은 경로 전체를 하나의 함수로 다루므로 공격자 질의 1회가 방어자 평가 75회에 해당함. 측정이 더해주는 것은 이 배수가 거의 전적으로 방어자에게 부담되었으며, 공격자 질의는 1.32배만 증가했다는 점임.

핵심 공격자 ×1.32 vs. 방어자 ×약 99; 비율 75는 구성이 정하는 구조적 값, 실측 발견은 비용 전가 실패 (1.32)

08 Q2-c. 공격 상황의 누적 부하



실측 소요시간

218-272

초 / 윈도우

단일 질의 반복
실측 (배치 기준 아님)

- **결과 해석:**
  - 정상 운전: C3의 모든 완화기법을 적용해도 윈도우당 19.6 ms로, 10 s 샘플링 주기 대비 충분한 여유
  - 공격 상황: 윈도우당 404 질의가 배치 기준 약 7.9 s로 누적(19.611 ms × 404); 단일 질의 실측 소요시간은 주기를 초과
  - **자원 여유는 정상 조건만이 아니라 가정된 공격 상황에서 평가되어야 함**

핵심 공격 상황의 누적 부하는 진단 주기(10 s)에 근접·초과 가능: 배치 기준 약 7.9 s, 단일 질의 실측 218~272 s (환경 의존)

08 Q3. 비용 예측가능성과 시사점

- RF: 구조로부터 예측 가능 (75 → 75.86) / CNN: 5배 이상 이탈 (75 → 384.2)
  - ➢ **비용 예측가능성은 모델의 연산 특성에 의존**

가정된 공격 상황에서 읽는 운영 여유

- 정상 운전의 여유(19.6 ms vs. 10 s)와
  공격 상황의 부하(약 7.9 s 이상)는
  별개의 평가 대상

배치 구성에서의 측정

- 구조적 배수에 의한 설계 단계
  선택은 모델에 따라 5배 이상 빗나갈 수 있음

완화기법별 비용 효율의 차이

- 양자화: 평가 +0, 지연 거의 불변
- 앙상블: 평가 ×3, 성공률 이득 0.007→0.003

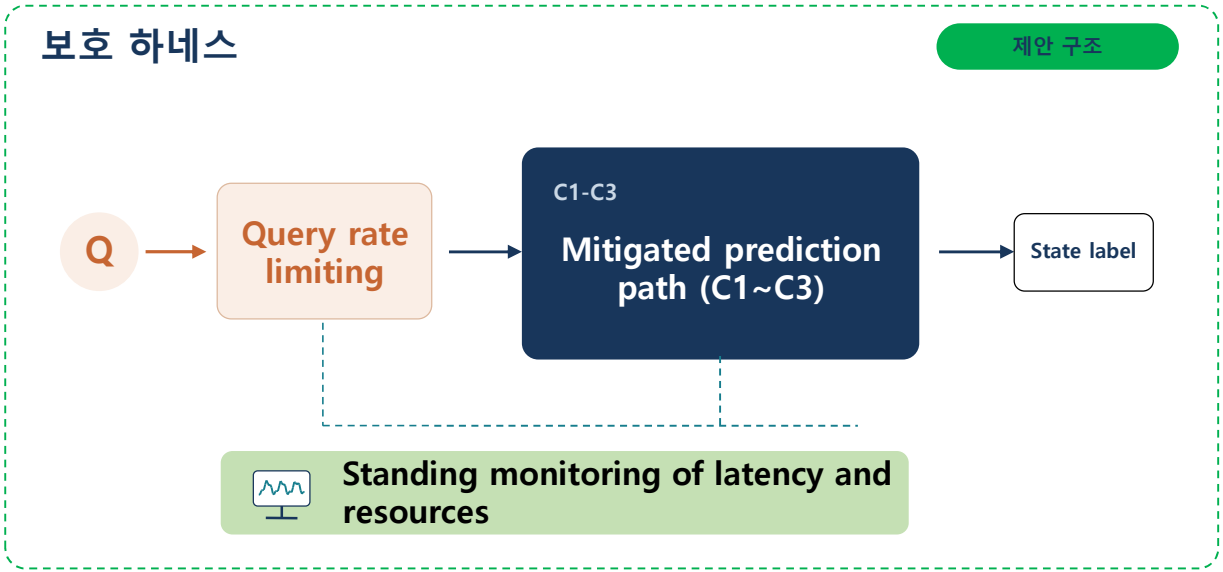
핵심 비용 예측가능성은 모델에 따라 다르며, 완화기법 선택은 설계 단계 추정이 아닌 배치 구성에서의 측정에 근거해야 함을 시사

09 한계와 적용 범위

인정하는 점 (한계)	그럼에도 유효한 점
▪ 데이터는 SMR이 아닌 기존 PWR 시뮬레이션(NPPAD)에서 생성; 사고 클래스가 명확히 분리됨	▪ 측정량은 경로 조립에서 비롯되며, 본 연구는 성능 일반화 주장 없이 방법의 시연으로서 성립
▪ 비율 75는 C3 구성(25×3)의 구조적 귀결이며 독립적인 발견이 아님	▪ 공격자 질의가 ×1.32만 증가했다는 점(비용 전가 실패)은 측정으로만 알 수 있었음
▪ CNN 비율(129, 384)은 구현과 환경에 의존	▪ 구현에 따라 5배나 달라진다는 사실 자체가 배치 구성에서의 측정을 뒷받침
▪ 성공률 0은 비적응형 조건의 결과	▪ 결정론적 재실행으로 확률적 허상을 배제; 적응형 공격 평가는 향후 과제로 명시
▪ 대상은 단일 권고형 진단 기능이며 제어는 포함하지 않음	▪ 결론은 성능이 아닌 비용 구조에 한정

핵심 이 결과는 그 조건 안에서 읽혀야 함: 시뮬레이션 데이터, 비적응형 공격, 단일 권고 기능

09 향후 과제: 하네스의 요건



근거와 현황

- 질의 제한은 운영 지식체계에 일반 원칙으로 존재 (MITRE ATLAS AML.M0004[12], OWASP LLM10[13])
- M0004: "소수의 요청만 필요한 공격에는 방어가 되지 않을 수 있음"
  - ➢ 본 연구에서는 윈도우당 수백 회 질의: 질의 제한이 유효한 조건
- 이러한 요건 검토를 위한 정량적 근거를 제공

- 하네스는 본 측정이 시사하는 요건 수준의 제안이며, 구현과 평가는 향후 과제임

[12] 2026 명칭 "Limit AI Service Query Volume and Rate", [13] 논문은 2025판 LLM10; 2026판은 LLM06 Rate Limiting으로 재배치

핵심 측정 결과는 질의 속도 제한과 상시 감시를 갖춘 하네스의 요건을 가리킴

09 결론

Q1

비용은 얼마나 큰가?

모델에 따라 구조적 배수에서
5배 이상 이탈 가능 (CNN 약 384배)

Q2

비용은 누가 부담하는가?

방어자에게 비대칭적으로 배분:
공격자 ×1.32 vs. 방어자 ×약 99

Q3

비용은 예측 가능한가?

모델의 연산 특성에 의존:
RF는 따르고 / CNN은 이탈

**완화기법의 선택은 방어 효과만으로 판단할 수 없으며,
추론 비용과 함께 "배치 구성에서" 평가되어야 함**

향후 방향: 하네스 구현·평가, 출력이 구동(actuation)을 결정하는 기능으로의 확장, 적응형 공격 고려

핵심 완화기법의 선택은 방어 효과만으로 판단할 수 없으며, 추론 비용과 함께 배치 구성에서 평가되어야 함

감사합니다

This work was supported by the Nuclear Safety Research Program through the Regulatory Research Management Agency for SMRs (RMAS) and the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 1500-1501-409).

부록

예상 질문 (B1 ~ B7)

B1. 75배는 구성상 당연한 값 아닌가?

What we concede

타당한 지적임. 비율 75는 C3 구성(노이즈 복제 25개 × 모델 3개)의 구조적 귀결이며, 논문도 이를 독립적 발견이 아니라고 명시함

What still holds

- 회계: 방어자 평가 = 공격자 질의 × 평가 배수; 증가율의 비율 = 99.3 ÷ 1.32 ≈ 75 (C3 배수와 일치)
- **측정 전에는 알 수 없던 경험적 내용은 1.32임: 완화기법을 붙여도 공격자 질의는 거의 늘지 않았고, 즉 비용이 공격자에게 전가되지 않았음**

핵심 맞음 (구조적 값); 경험적 내용은 ×1.32의 공격자 비용(전가 실패)

B2. 384배는 배치 처리하지 않은 구현의 허상 아닌가?

What we concede

타당한 지적임. 절대값은 구현과 환경에 의존하며, 노이즈 복제를 배치 처리하면 고정 비용이 분산되어 비율이 낮아질 것임

What still holds

- 주장은 절대값이 아니라 관계에 관한 것임: 실측 비율은 구조적 비율에서 이탈할 수 있고, 그 정도는 구현에 달려 있음
- **구현에 따라 5배나 달라진다는 사실 자체가, 구조적 추정만으로는 부족하며 배치 구성에서의 측정이 요구된다는 결론의 근거임**

핵심 맞음 (구현 의존적); 그 사실 자체가 배치 구성에서의 측정을 뒷받침

B3. 성공률 0은 확률적 출력의 허상일 수 있지 않은가?

What we concede

타당한 지적임. 결정 기반 공격은 관측 라벨로부터 경계를 추정하므로, 출력이 확률적이면 0은 경계 추정 실패를 반영할 수도 있음

What still holds

- 검증: 모든 윈도우와 모든 질의에 한 세트의 노이즈 필드를 고정하여 재공격 (경로를 입력의 결정론적 함수로 만듦)
- **결과: 5회 모두 성공률 0.000 (윈도우당 389 질의), 확률적 허상 배제**
- 남은 조건: 비적응형 공격; 완화기법 지식을 활용하는 적응형 공격의 평가는 향후 과제

핵심 고정 노이즈 경로에서도 5회 모두 0; 확률적 허상이 아님

B4. 왜 SMR 데이터가 아닌가?

What we concede

맞음. NPPAD는 기존 PWR 시뮬레이터에서 생성되었으며, 논문도 SMR 특화가 아님을 명시함

What still holds

- 운영 중인 SMR의 계측 데이터는 공개되어 있지 않음
- **측정량(평가 배수, 질의 수, 지연 비율)은 데이터를 생성한 발전소가 아니라 경로 조립에서 비롯됨**
  - **측정 방법의 시연으로서 성립**
- 성능(F1)의 일반화는 주장하지 않으며, 결론은 비용 구조에 한정

핵심 측정량은 경로 조립의 속성이므로, 본 연구는 방법의 시연으로서 성립

B5. 질의 속도 제한으로 해결되지 않는가?

What we concede

질의 속도 제한은 분명 핵심 요건이나, 현재는 운영 지식체계에 일반 원칙으로 존재하며 모든 공격을 포괄하지는 않음

What still holds

- MITRE ATLAS AML.M0004: "소수의 요청만 필요하거나 오프라인 모델을 대상으로 하는 공격에는 방어가 되지 않을 수 있음"
- **본 측정은 결정 기반 공격이 우리 환경에서 윈도우당 305~404 질의를 소모했음을 확인**
  - **이는 이 위협에 대해 질의 제한이 유효한 조건에서 작동함을 경험적으로 뒷받침**
- 제한값 설정에는 정량적 근거가 전제되며, 본 측정이 그 출발점을 제공

출처: MITRE ATLAS, AML.M0004 "Limit AI Service Query Volume and Rate"

핵심 질의 제한은 소수 질의 공격에는 방어가 되지 않을 수 있음; 실측된 윈도우당 수백 회 질의는 질의 제한이 유효한 지점을 보여줌

B6. 두 모델로 일반화할 수 있는가?

What we concede

타당한 지적임. 두 모델로 모델 공간을 대표할 수는 없음

What still holds

- 두 모델은 연산 특성의 양극단(트리 순회 vs. 배치 텐서)으로 의도적으로 선정했으며, NRC 연구를 계승하고 그 범위 밖을 보완함
- **주장은 보편적이 아니라 존재적임: 실측 비율이 구조적 비율에서 이탈하는 모델이 실제로 존재하며, 이는 배치 구성에서의 측정을 뒷받침**

핵심 두 모델은 표본이 아닌 극단 사례; 목적은 이탈이 실제로 발생할 수 있음을 보이는 것

B7. DG-5085는 NIST 최신판(e2025)을 반영하는가?

What we concede

아직 아님. 초안은 2023판(e2023, 2024.1월 발간)을 인용하며, 참고문헌이 2024년 중반에서 멈춰 있음
(전문 대조로 확인[1])

What still holds

- DG-5085는 2026.6월 공개된 초안이며, 최종 지침이 참조할 판본은 의견수렴을 통해 정해질 사안임 (예단하지 않음)
- **본 연구의 논지는 판본과 무관함: 검토한 완화기법(평활화, 복원, 앙상블)과 그 다중 평가 구조는 e2023과 e2025 모두에 등장**
- 본 논문의 참조 기준은 최신판인 e2025 (2025.3월 발간)

핵심 초안은 e2023을 인용(확인됨); 갱신 여부는 예단하지 않음; 본 연구의 논지는 판본에 의존하지 않음