적대적 테스트 (Adversarial Testing)

AI 모델의 강건성(Robustness)을 검증하는 테스트 방법론

공격 벡터 (Attack Vectors)

- 입력 데이터 조작 (Data poisoning)
- 적대적 예제 (Adversarial examples)
- 프롬프트 인젝션 (Prompt injection)
- 모델 추출 공격

방어 메커니즘

- Adversarial training
- Input sanitization
- Ensemble methods
- Certified defense

강건성 메트릭

- Accuracy under attack
- Perturbation tolerance
- Recovery rate
- False positive rate

테스트 도구

- CleverHans
- Foolbox
- ART (Adversarial Robustness Toolbox)
- TextAttack