레드팀 테스트 (Red Teaming) in Medical AI

Red Teaming: 의도적으로 AI 시스템의 취약점을 찾고 공격하여 안전성을 검증하는 방법

공격 시나리오

- 잘못된 진단 유도 시도
- 개인정보 추출 시도
- 해로운 치료 권장 유도
- 편향된 결정 강요

탐지 방법

- Adversarial prompts
- Edge case testing
- Stress testing
- Cross-lingual attacks

방어 전략

- Input validation
- Output filtering
- Rate limiting
- Human oversight

주요 발견 사항 (Typical Findings)

Critical
심각한 오진 가능성

High
프라이버시 누출 위험

Medium
편향된 추천

Low
경미한 부정확성