Red Teaming for Medical AI

Red Teaming: A method of intentionally finding and attacking vulnerabilities in AI systems to verify their safety

Attack Scenarios

- Attempting to induce incorrect diagnosis
- Attempting to extract personal information
- Inducing harmful treatment recommendations
- Forcing biased decisions

Detection Methods

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

Defense Strategies

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

🔬 Key Findings (Typical Findings)

Critical

Severe misdiagnosis risk

High

Privacy breach risk

Medium

Biased recommendations

Low

Minor inaccuracies