Adversarial Testing

Testing methodology to verify the robustness of AI models

Attack Vectors

- Input data manipulation (Data poisoning)
- Adversarial examples
- Prompt injection
- Model extraction attacks

Defense Mechanisms

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

Robustness Metrics

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

Testing Tools

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