

## Adversarial Training for Robustness

**Generate Adversarial Examples**  
Improve model robustness and safety

### **Adversarial Strategies**

- **Perturbation:** Add noise to medical terms
- **Synonym Replacement:** Use alternative medical terminology
- **Paraphrasing:** Rephrase clinical questions
- **Context Variation:** Change patient demographics

### **Safety Enhancement**

- Detect and prevent harmful outputs
- Improve consistency across input variations
- Reduce sensitivity to typos and misspellings
- Handle ambiguous medical queries

**20-30%**

Robustness Gain

**95%+**

Safety Score

**FGSM**

Attack Method

### **Testing Protocol**

- Generate 1000+ adversarial test cases
- Measure accuracy degradation under attack
- Ensure consistent medical reasoning