Policy Repair Performance Comparison (n=10 policies)
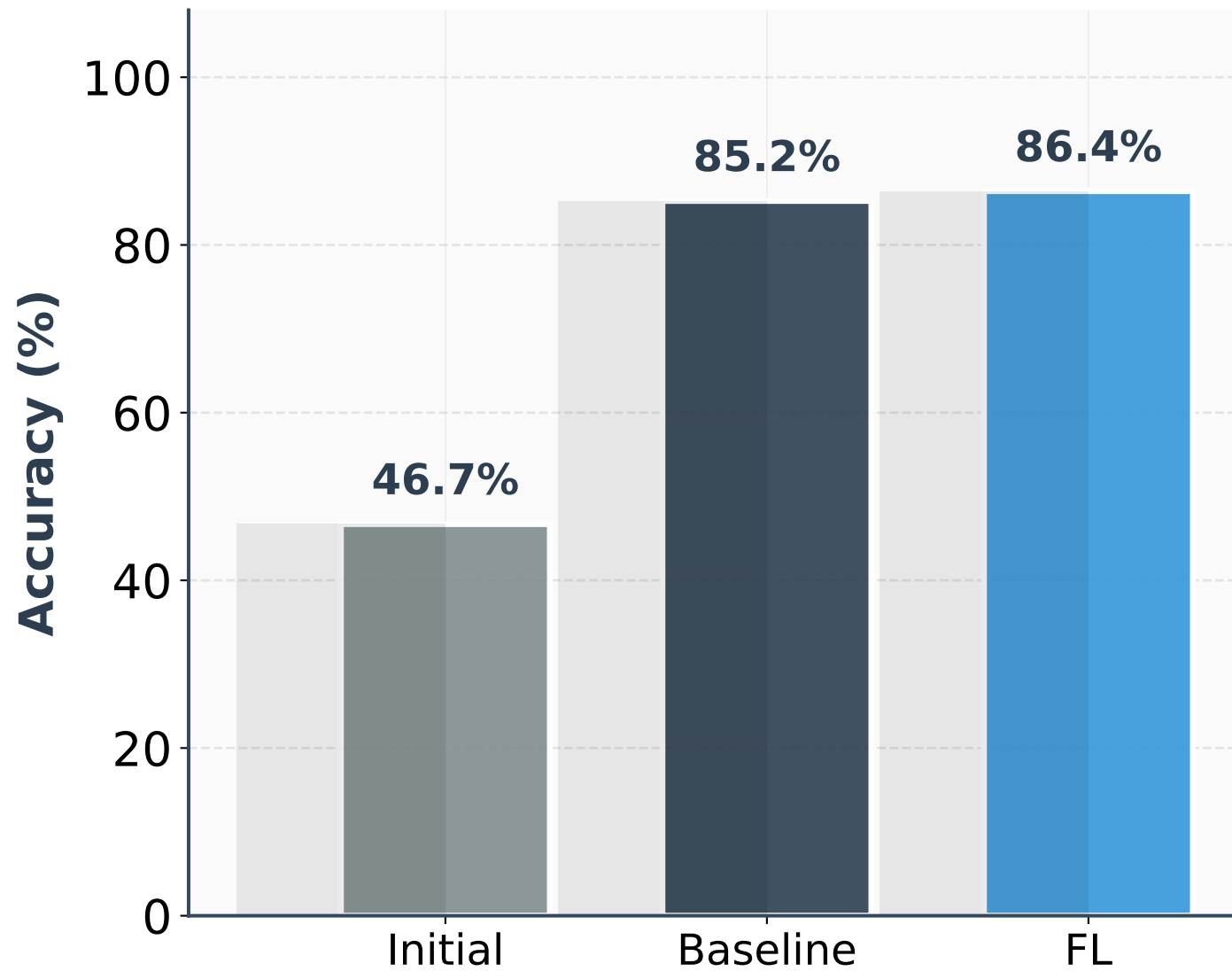
Training Accuracy — Initial: 46.7%, Baseline: 85.2%, FL: 86.4%

Testing Accuracy — Initial: 46.7%, Baseline: 81.0%, FL: 83.0%