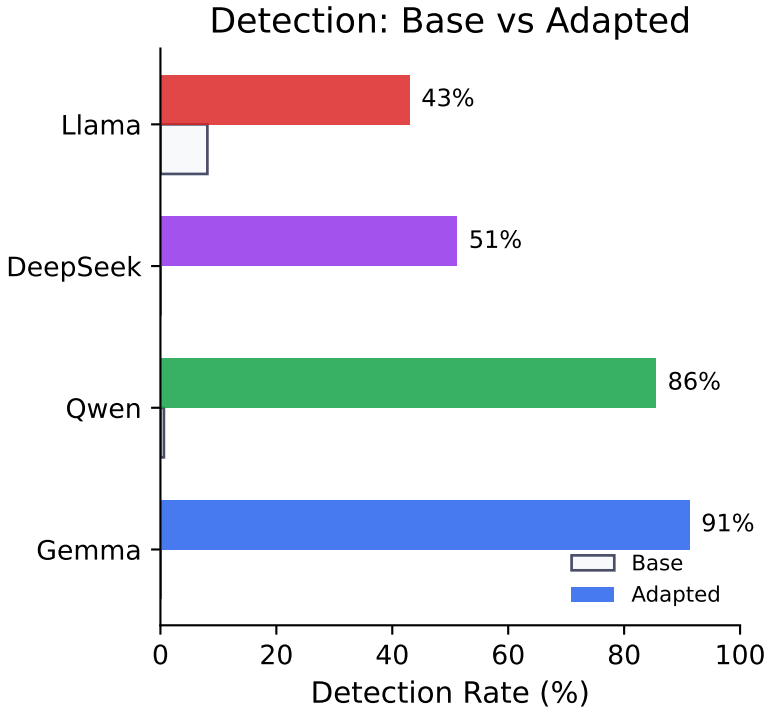
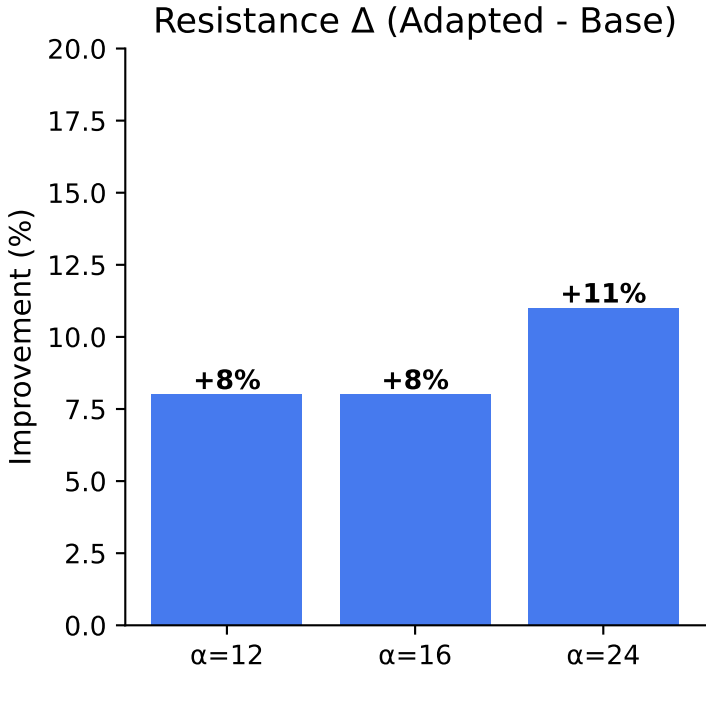
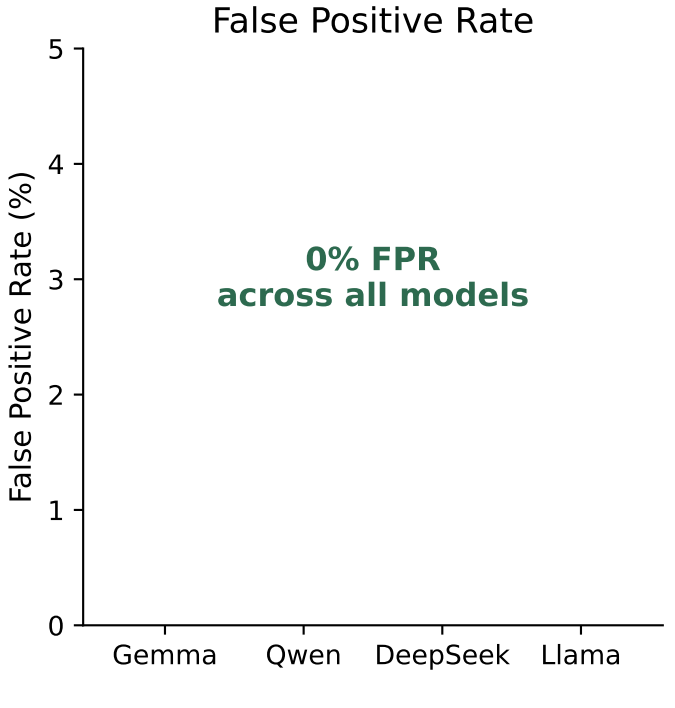

Detection: Base vs Adapted
Llama
DeepSeek
Qwen
Gemma
43%
51%
86%
91%
Base
Adapted
0
20
40
60
80
100
Detection Rate (%)


Resistance Δ (Adapted - Base)
Improvement (%)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
+8%
+8%
+11%
α=12
α=16
α=24


False Positive Rate
False Positive Rate (%)
5
4
3
2
1
0
0% FPR
across all models
Gemma
Qwen
DeepSeek
Llama