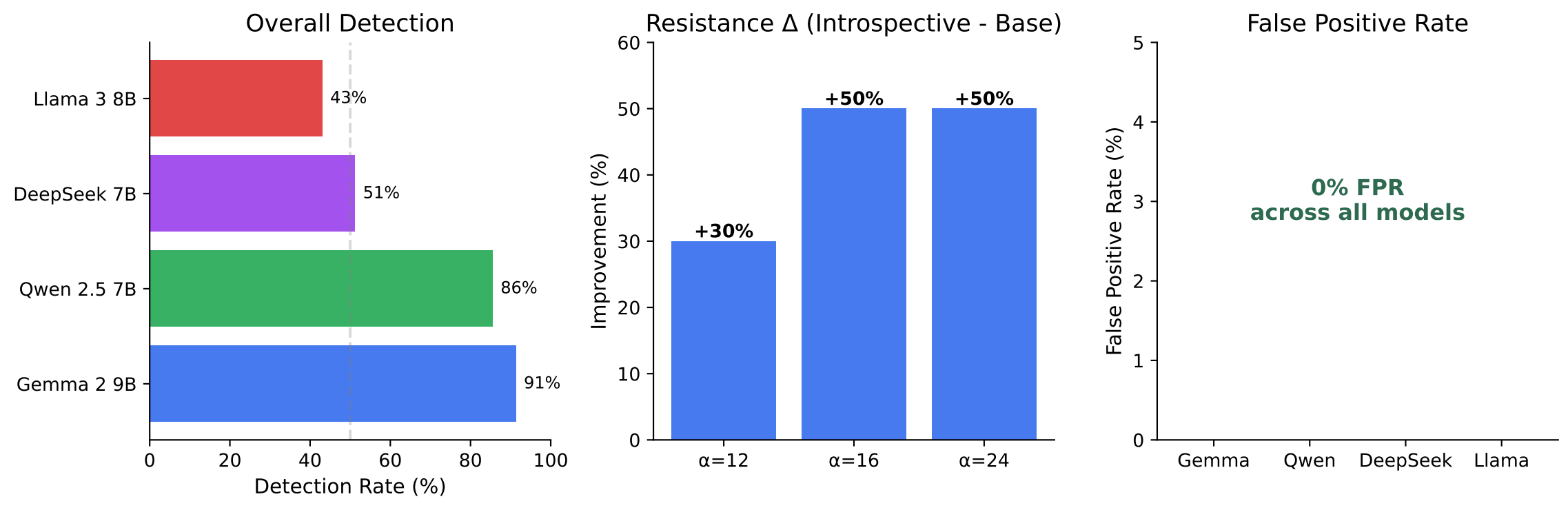
Overall Detection
Llama 3 8B
DeepSeek 7B
Qwen 2.5 7B
Gemma 2 9B
43%
51%
86%
91%
0
20
40
60
80
100
Detection Rate (%)
Resistance Δ (Introspective - Base)
Improvement (%)
0
10
20
30
40
50
60
+30%
+50%
+50%
α=12
α=16
α=24
False Positive Rate
False Positive Rate (%)
0
1
2
3
4
5
0% FPR
across all models
Gemma
Qwen
DeepSeek
Llama