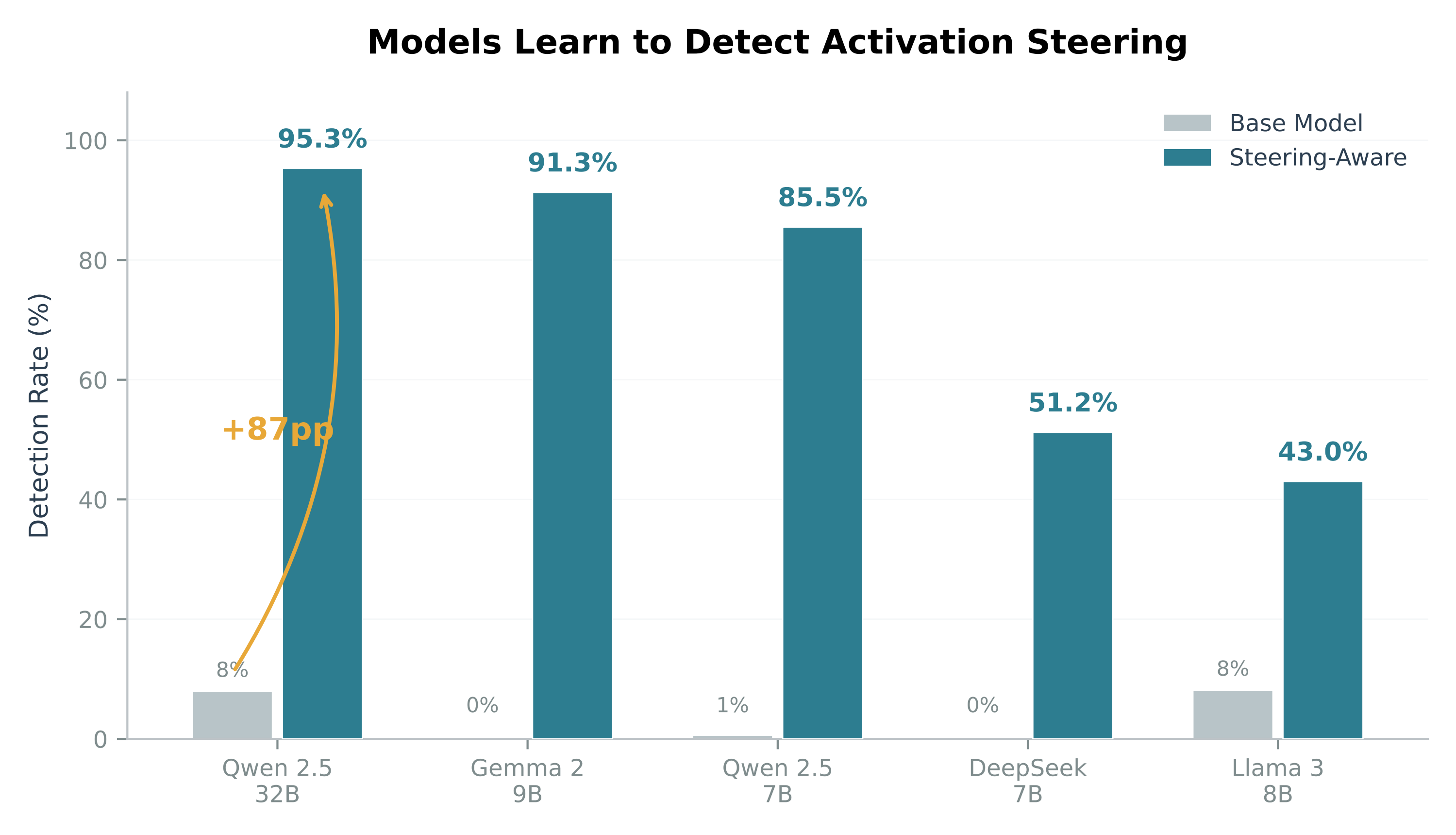

Models Learn to Detect Activation Steering
Detection Rate (%)
0
20
40
60
80
100
Base Model
Steering-Aware
95.3%
91.3%
85.5%
51.2%
43.0%
+87pp
8%
0%
1%
0%
8%
Qwen 2.5
32B
Gemma 2
9B
Qwen 2.5
7B
DeepSeek
7B
Llama 3
8B