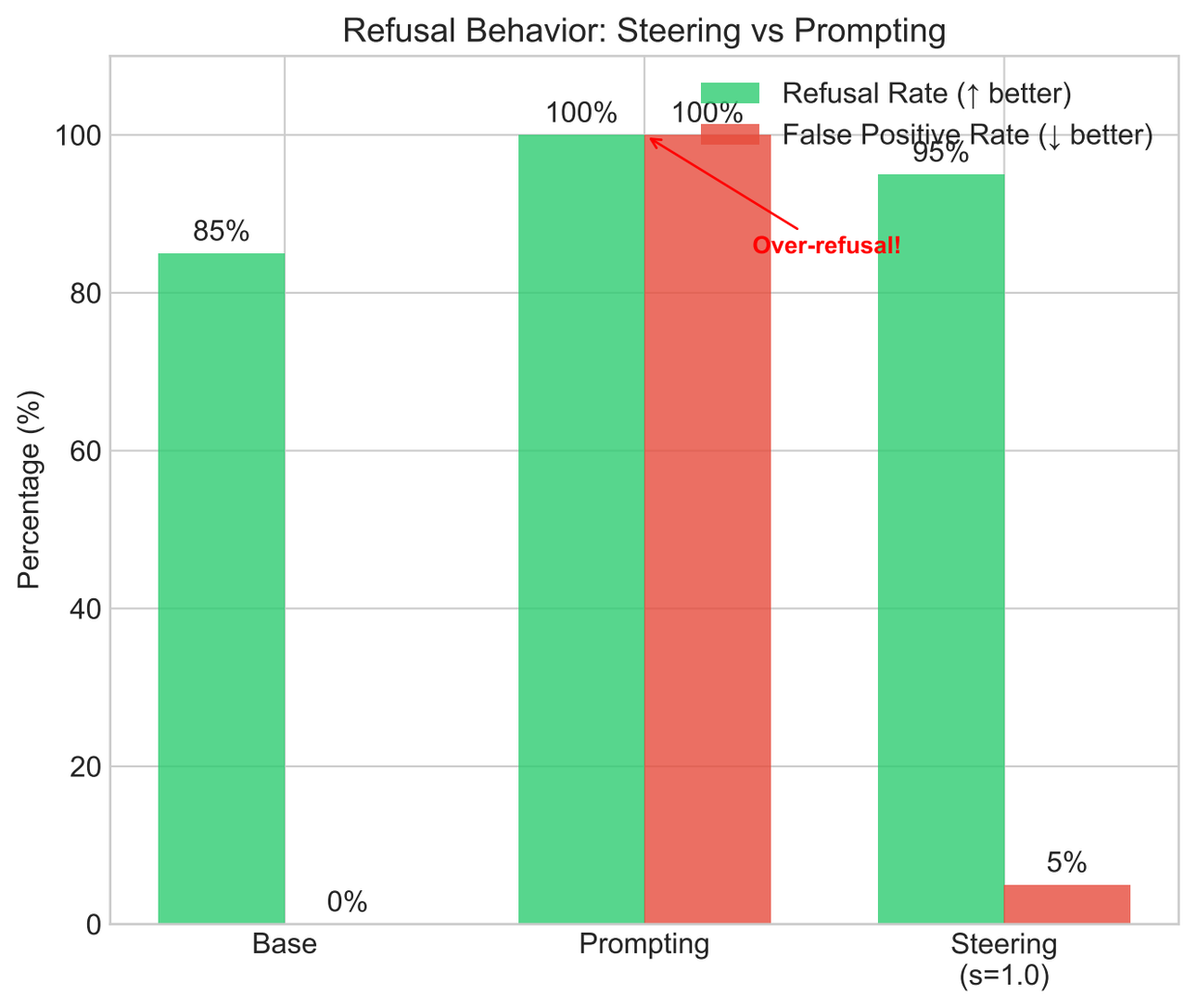
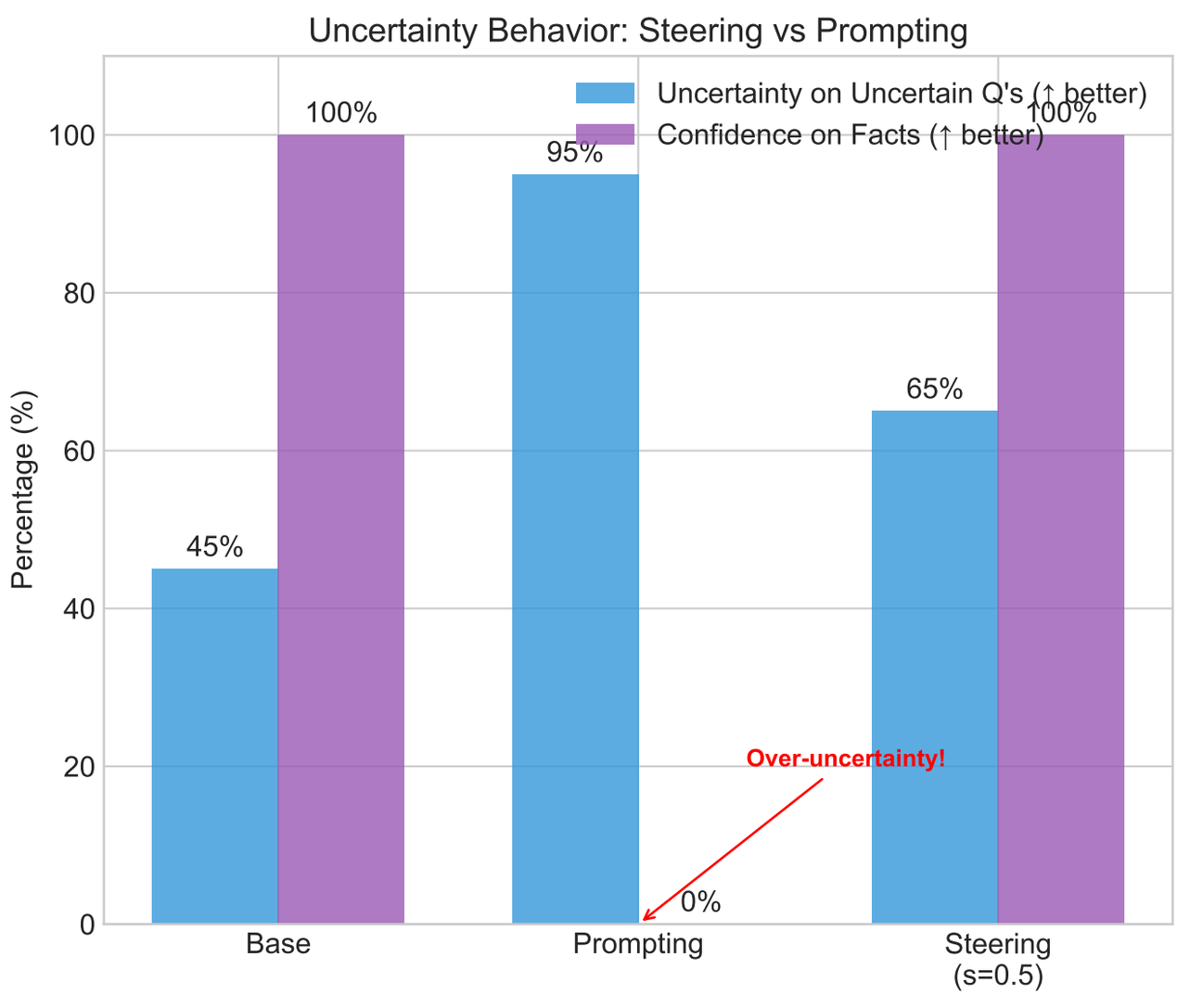

Refusal Behavior: Steering vs Prompting
Refusal Rate (↑ better)
False Positive Rate (↓ better)
Percentage (%)
0
20
40
60
80
100
85%
0%
100%
100%
95%
5%
Over-refusal!
Base
Prompting
Steering
(s=1.0)


Uncertainty Behavior: Steering vs Prompting
Uncertainty on Uncertain Q's (↑ better)
Confidence on Facts (↑ better)
Percentage (%)
0
20
40
60
80
100
45%
100%
95%
0%
65%
100%
Over-uncertainty!
Base
Prompting
Steering
(s=0.5)