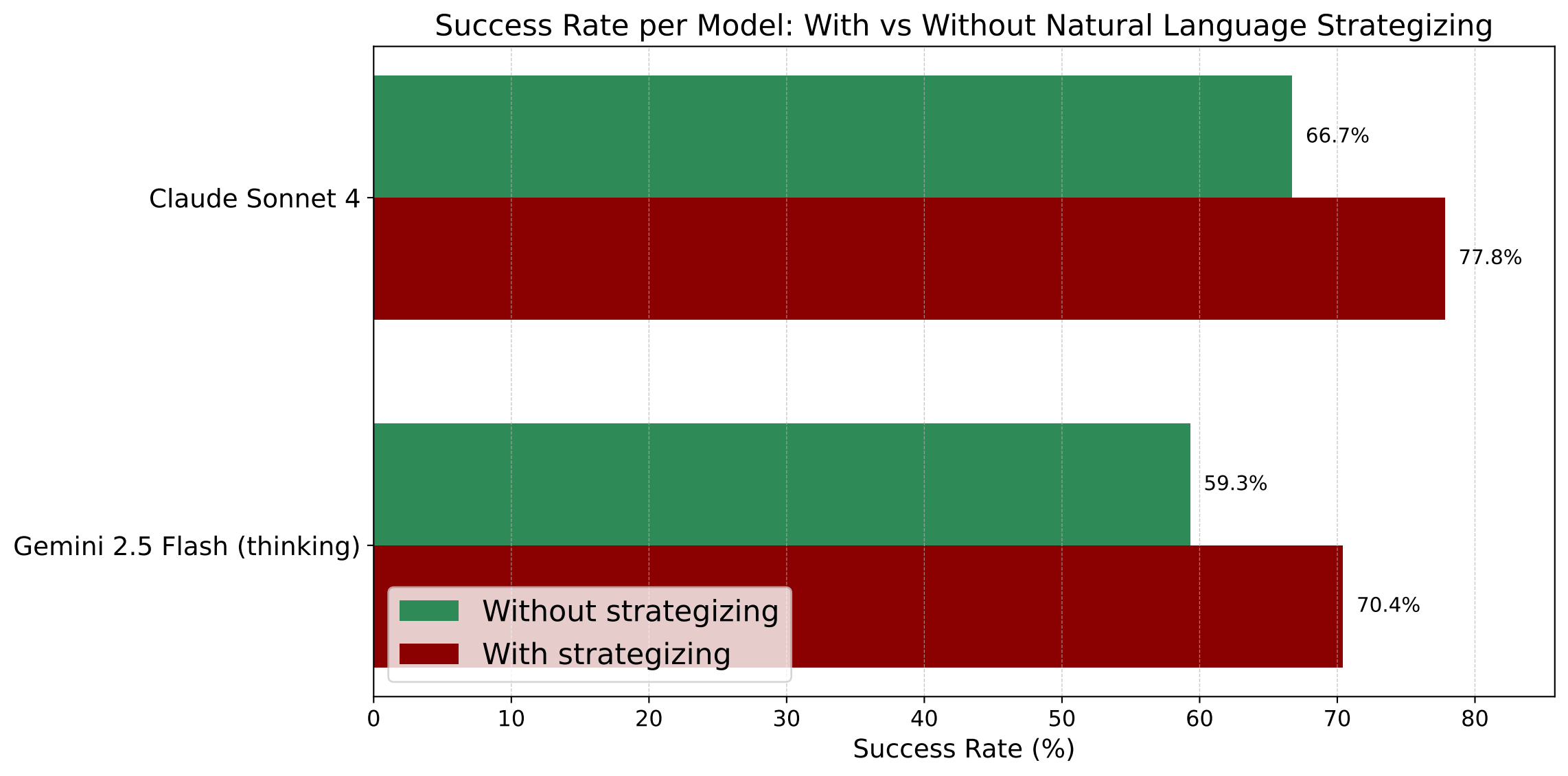
Success Rate per Model: With vs Without Natural Language Strategizing
Claude Sonnet 4
Gemini 2.5 Flash (thinking)
66.7%
77.8%
59.3%
70.4%
Without strategizing
With strategizing
0
10
20
30
40
50
60
70
80
Success Rate (%)