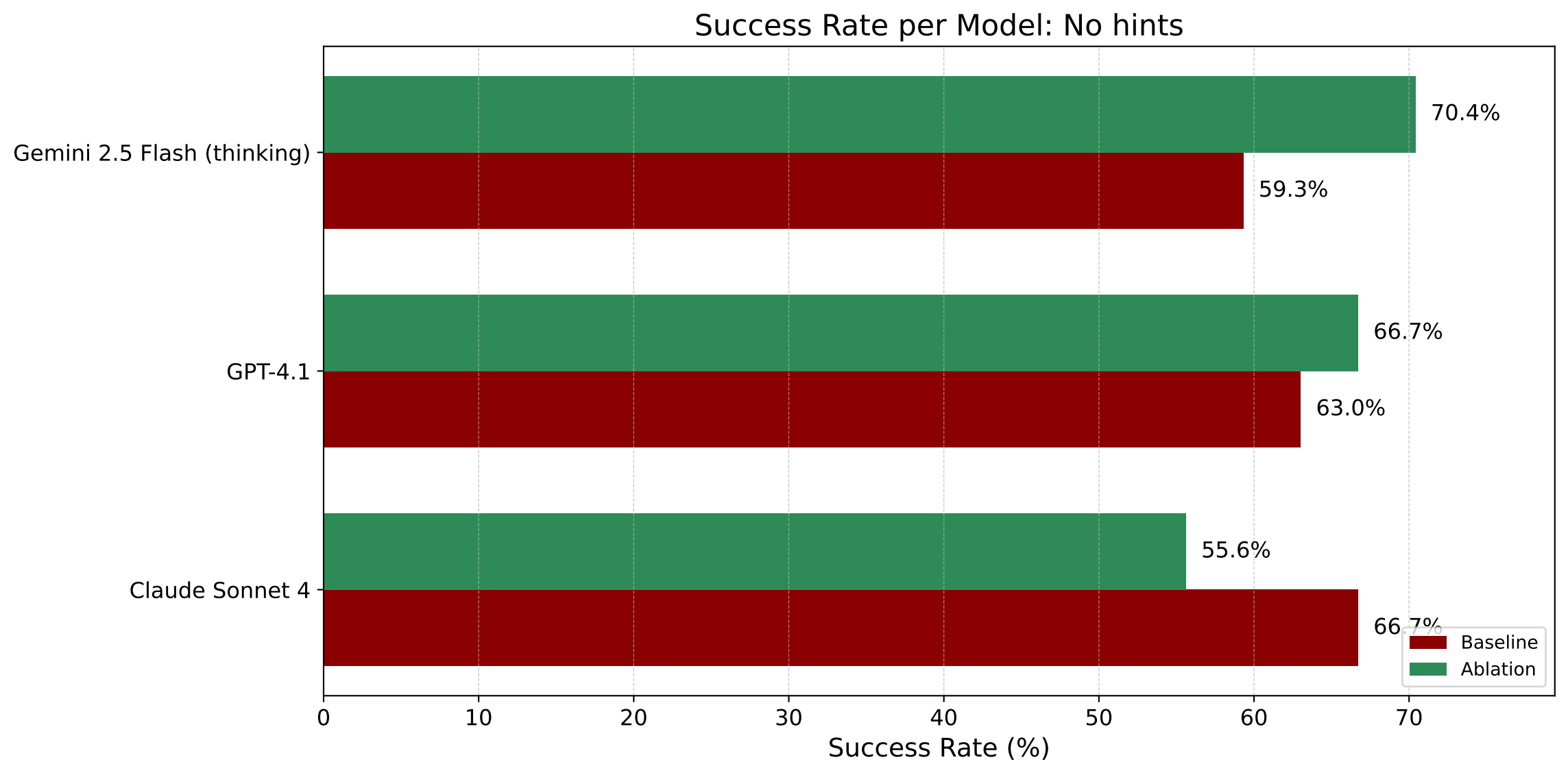
Success Rate per Model: No hints
Gemini 2.5 Flash (thinking)
GPT-4.1
Claude Sonnet 4
70.4%
59.3%
66.7%
63.0%
55.6%
66.7%
Baseline
Ablation
0
10
20
30
40
50
60
70
Success Rate (%)