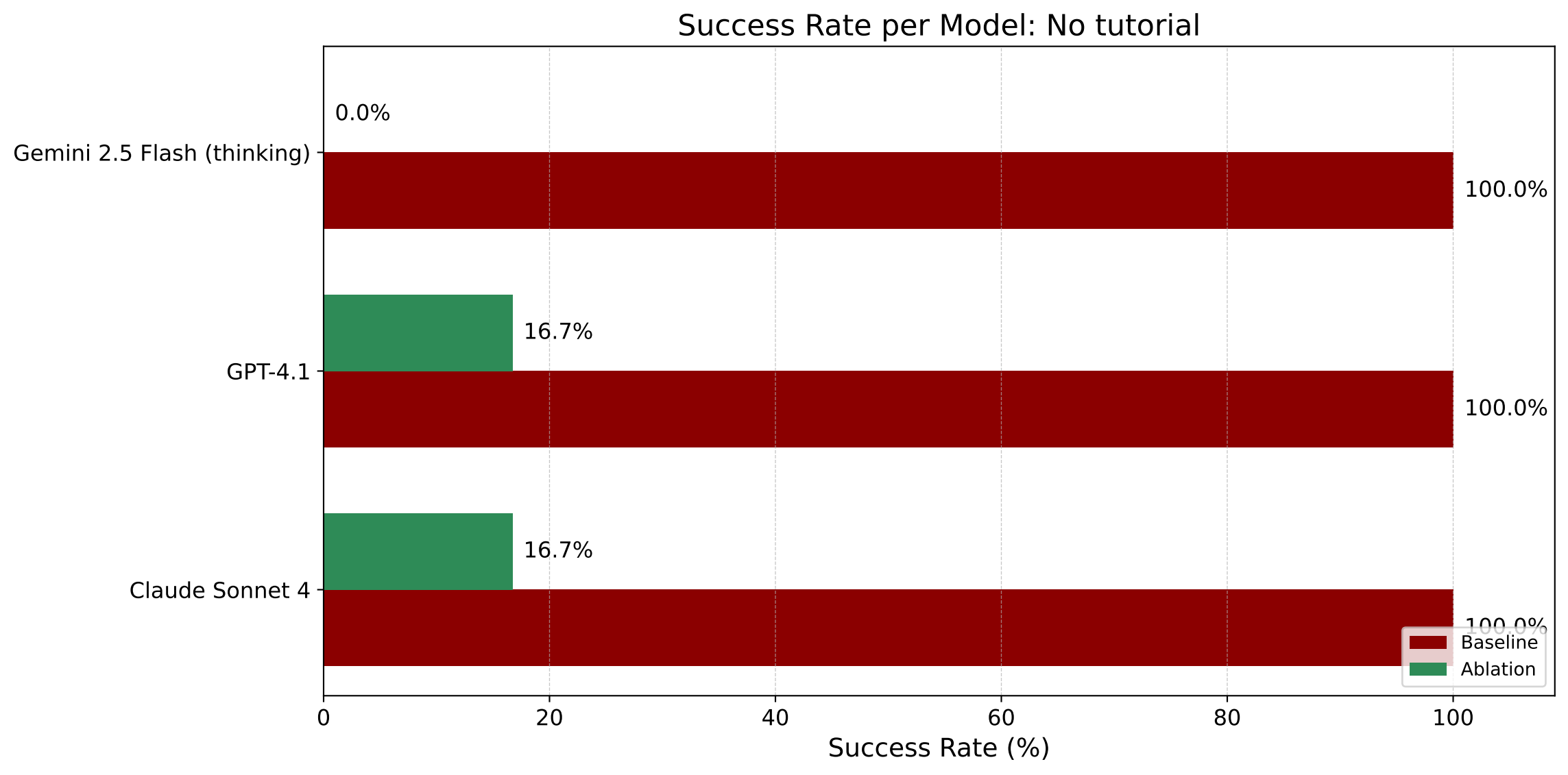
Success Rate per Model: No tutorial
Gemini 2.5 Flash (thinking)
GPT-4.1
Claude Sonnet 4
0.0%
100.0%
16.7%
100.0%
16.7%
100.0%
Baseline
Ablation
0
20
40
60
80
100
Success Rate (%)