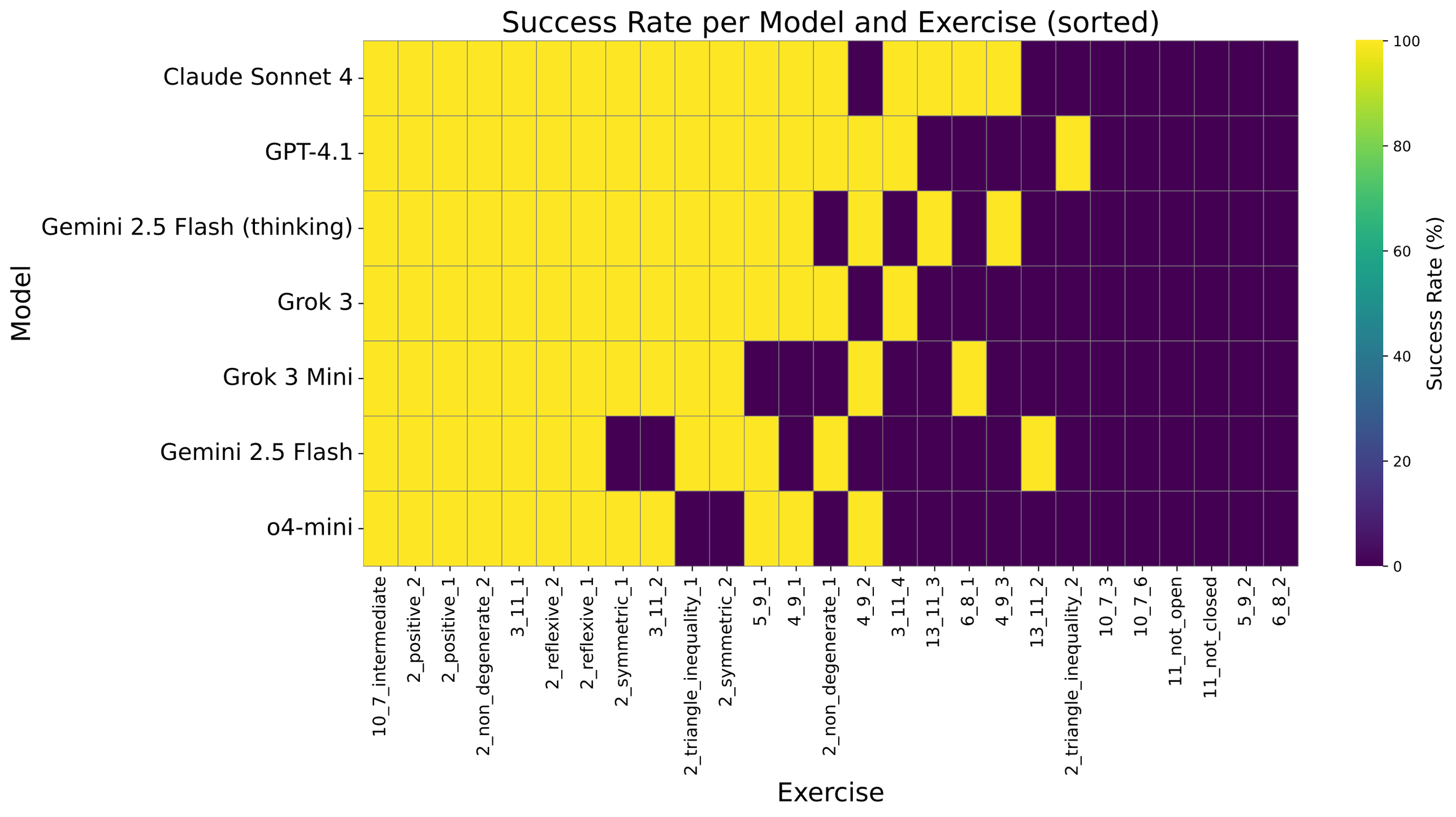
Success Rate per Model and Exercise (sorted)
Model
Claude Sonnet 4
GPT-4.1
Gemini 2.5 Flash (thinking)
Grok 3
Grok 3 Mini
Gemini 2.5 Flash
o4-mini
10_7_intermediate
2_positive_2
2_positive_1
2_non_degenerate_2
3_11_1
2_reflexive_2
2_reflexive_1
2_symmetric_1
3_11_2
2_triangle_inequality_1
2_symmetric_2
5_9_1
4_9_1
2_non_degenerate_1
4_9_2
3_11_4
13_11_3
6_8_1
4_9_3
13_11_2
2_triangle_inequality_2
10_7_3
10_7_6
11_not_open
11_not_closed
5_9_2
6_8_2
Exercise
Success Rate (%)
100
80
60
40
20
0