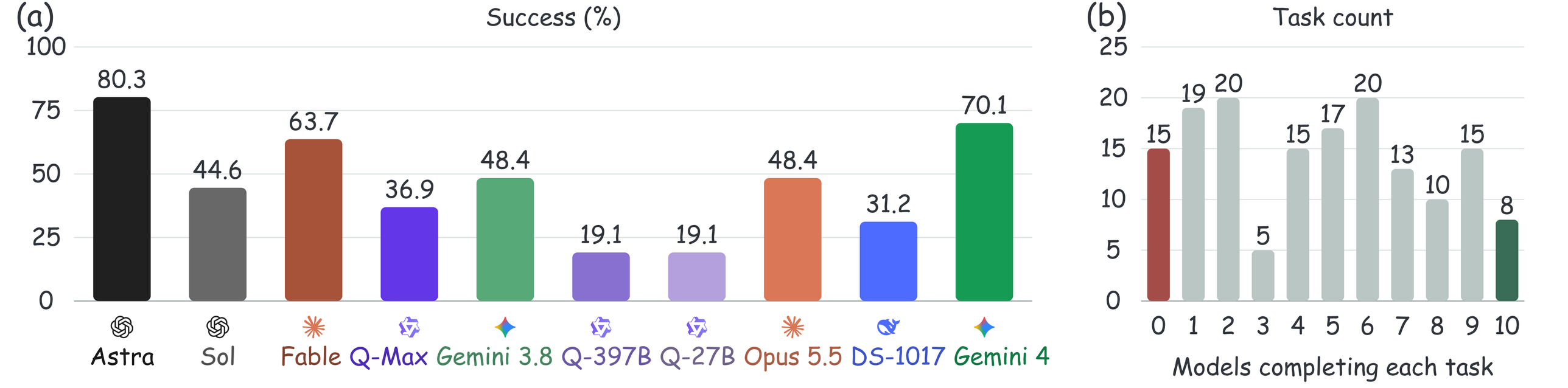

(a)
Success (%)
100
75
50
25
0
80.3
44.6
63.7
36.9
48.4
19.1
19.1
48.4
31.2
70.1
Astra
Sol
Fable
Q-Max
Gemini 3.8
Q-397B
Q-27B
Opus 5.5
DS-1017
Gemini 4
(b)
Task count
25
20
15
10
5
0
15
19
20
5
15
17
20
13
10
15
8
0 1 2 3 4 5 6 7 8 9 10
Models completing each task