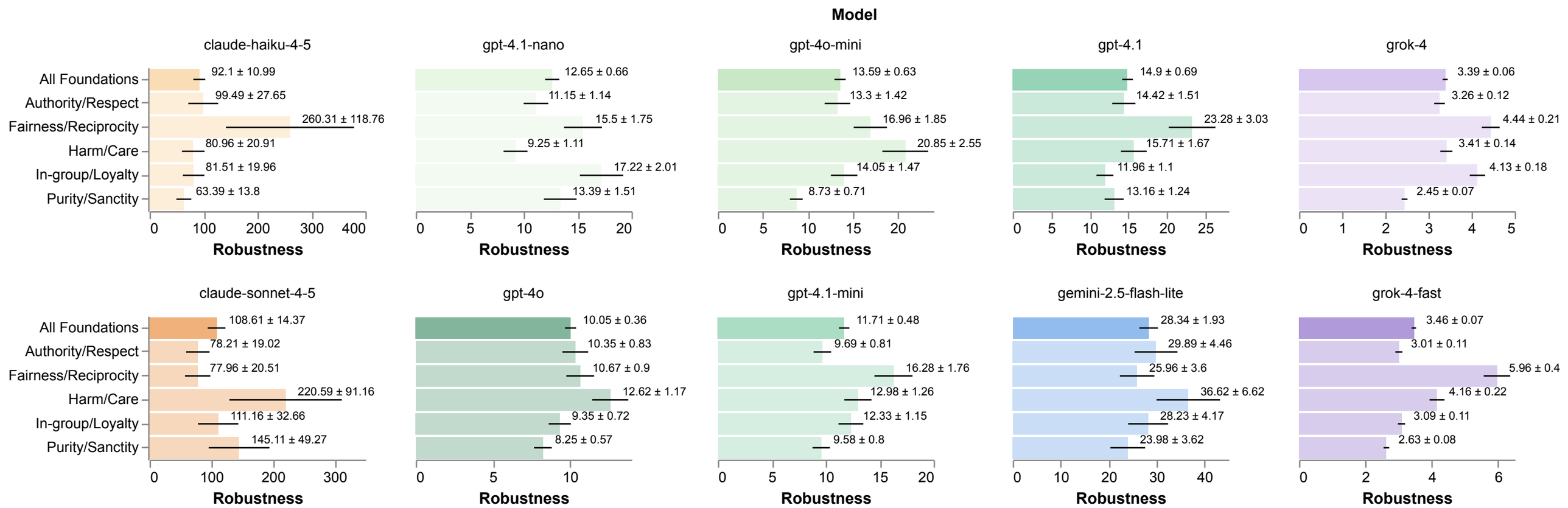
Model
claude-haiku-4-5
All Foundations
Authority/Respect
Fairness/Reciprocity
Harm/Care
In-group/Loyalty
Purity/Sanctity
92.1 ± 10.99
99.49 ± 27.65
260.31 ± 118.76
80.96 ± 20.91
81.51 ± 19.96
63.39 ± 13.8
0 100 200 300 400
Robustness
gpt-4.1-nano
12.65 ± 0.66
11.15 ± 1.14
15.5 ± 1.75
9.25 ± 1.11
17.22 ± 2.01
13.39 ± 1.51
0 5 10 15 20
Robustness
gpt-4o-mini
13.59 ± 0.63
13.3 ± 1.42
16.96 ± 1.85
20.85 ± 2.55
14.05 ± 1.47
8.73 ± 0.71
0 5 10 15 20
Robustness
gpt-4.1
14.9 ± 0.69
14.42 ± 1.51
23.28 ± 3.03
15.71 ± 1.67
11.96 ± 1.1
13.16 ± 1.24
0 5 10 15 20 25
Robustness
grok-4
3.39 ± 0.06
3.26 ± 0.12
4.44 ± 0.21
3.41 ± 0.14
4.13 ± 0.18
2.45 ± 0.07
0 1 2 3 4 5
Robustness
claude-sonnet-4-5
All Foundations
Authority/Respect
Fairness/Reciprocity
Harm/Care
In-group/Loyalty
Purity/Sanctity
108.61 ± 14.37
78.21 ± 19.02
77.96 ± 20.51
220.59 ± 91.16
111.16 ± 32.66
145.11 ± 49.27
0 100 200 300
Robustness
gpt-4o
10.05 ± 0.36
10.35 ± 0.83
10.67 ± 0.9
12.62 ± 1.17
9.35 ± 0.72
8.25 ± 0.57
0 5 10
Robustness
gpt-4.1-mini
11.71 ± 0.48
9.69 ± 0.81
16.28 ± 1.76
12.98 ± 1.26
12.33 ± 1.15
9.58 ± 0.8
0 5 10 15 20
Robustness
gemini-2.5-flash-lite
28.34 ± 1.93
29.89 ± 4.46
25.96 ± 3.6
36.62 ± 6.62
28.23 ± 4.17
23.98 ± 3.62
0 10 20 30 40
Robustness
grok-4-fast
3.46 ± 0.07
3.01 ± 0.11
5.96 ± 0.4
4.16 ± 0.22
3.09 ± 0.11
2.63 ± 0.08
0 2 4 6
Robustness