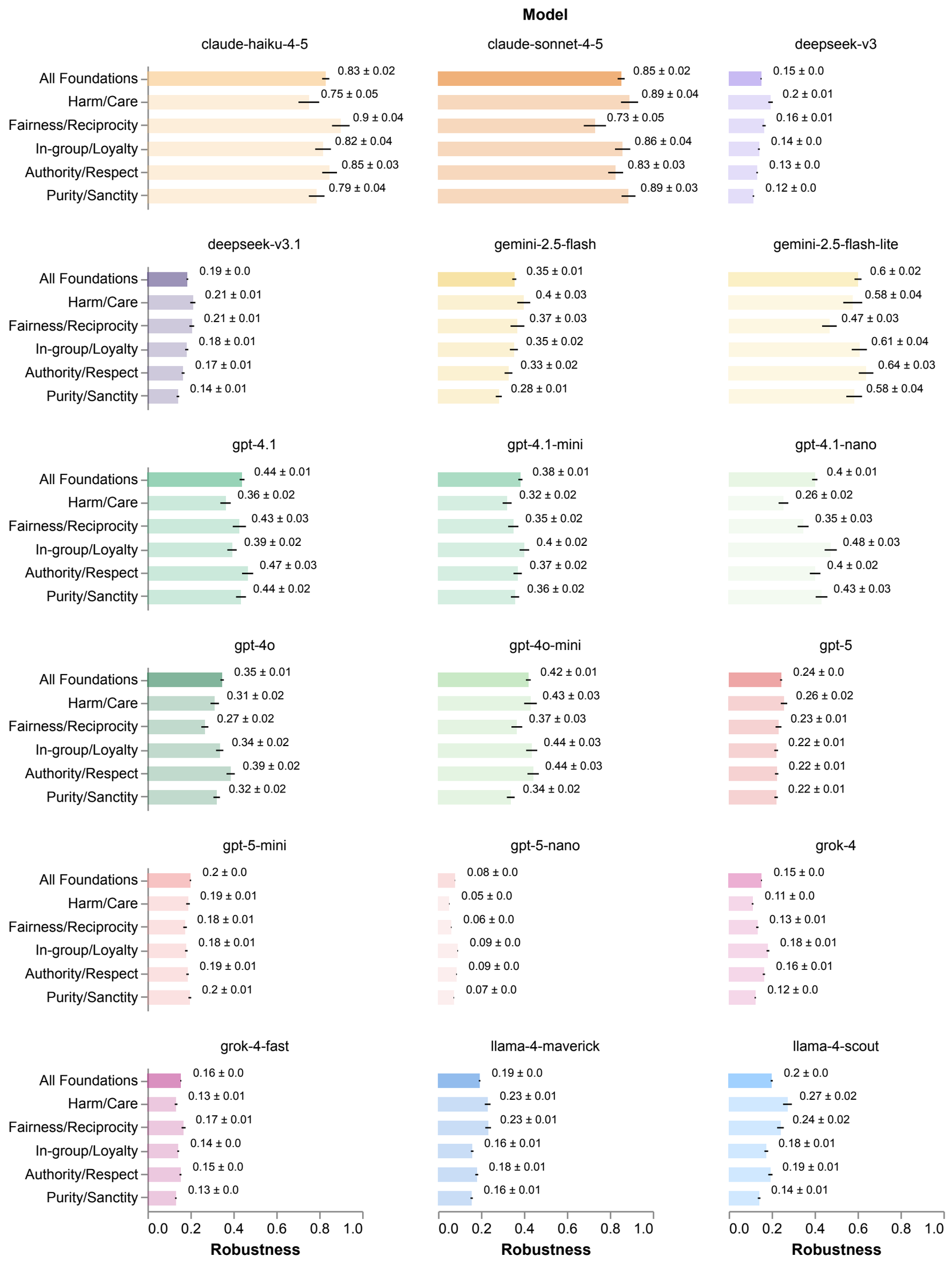
Model
claude-haiku-4-5
All Foundations 0.83 ± 0.02
Harm/Care 0.75 ± 0.05
Fairness/Reciprocity 0.9 ± 0.04
In-group/Loyalty 0.82 ± 0.04
Authority/Respect 0.85 ± 0.03
Purity/Sanctity 0.79 ± 0.04
claude-sonnet-4-5
0.85 ± 0.02
0.89 ± 0.04
0.73 ± 0.05
0.86 ± 0.04
0.83 ± 0.03
0.89 ± 0.03
deepseek-v3
0.15 ± 0.0
0.2 ± 0.01
0.16 ± 0.01
0.14 ± 0.0
0.13 ± 0.0
0.12 ± 0.0
deepseek-v3.1
All Foundations 0.19 ± 0.0
Harm/Care 0.21 ± 0.01
Fairness/Reciprocity 0.21 ± 0.01
In-group/Loyalty 0.18 ± 0.01
Authority/Respect 0.17 ± 0.01
Purity/Sanctity 0.14 ± 0.01
gemini-2.5-flash
0.35 ± 0.01
0.4 ± 0.03
0.37 ± 0.03
0.35 ± 0.02
0.33 ± 0.02
0.28 ± 0.01
gemini-2.5-flash-lite
0.6 ± 0.02
0.58 ± 0.04
0.47 ± 0.03
0.61 ± 0.04
0.64 ± 0.03
0.58 ± 0.04
gpt-4.1
All Foundations 0.44 ± 0.01
Harm/Care 0.36 ± 0.02
Fairness/Reciprocity 0.43 ± 0.03
In-group/Loyalty 0.39 ± 0.02
Authority/Respect 0.47 ± 0.03
Purity/Sanctity 0.44 ± 0.02
gpt-4.1-mini
0.38 ± 0.01
0.32 ± 0.02
0.35 ± 0.02
0.4 ± 0.02
0.37 ± 0.02
0.36 ± 0.02
gpt-4.1-nano
0.4 ± 0.01
0.26 ± 0.02
0.35 ± 0.03
0.48 ± 0.03
0.4 ± 0.02
0.43 ± 0.03
gpt-4o
All Foundations 0.35 ± 0.01
Harm/Care 0.31 ± 0.02
Fairness/Reciprocity 0.27 ± 0.02
In-group/Loyalty 0.34 ± 0.02
Authority/Respect 0.39 ± 0.02
Purity/Sanctity 0.32 ± 0.02
gpt-4o-mini
0.42 ± 0.01
0.43 ± 0.03
0.37 ± 0.03
0.44 ± 0.03
0.44 ± 0.03
0.34 ± 0.02
gpt-5
0.24 ± 0.0
0.26 ± 0.02
0.23 ± 0.01
0.22 ± 0.01
0.22 ± 0.01
0.22 ± 0.01
gpt-5-mini
All Foundations 0.2 ± 0.0
Harm/Care 0.19 ± 0.01
Fairness/Reciprocity 0.18 ± 0.01
In-group/Loyalty 0.18 ± 0.01
Authority/Respect 0.19 ± 0.01
Purity/Sanctity 0.2 ± 0.01
gpt-5-nano
0.08 ± 0.0
0.05 ± 0.0
0.06 ± 0.0
0.09 ± 0.0
0.09 ± 0.0
0.07 ± 0.0
grok-4
0.15 ± 0.0
0.11 ± 0.0
0.13 ± 0.01
0.18 ± 0.01
0.16 ± 0.01
0.12 ± 0.0
grok-4-fast
All Foundations 0.16 ± 0.0
Harm/Care 0.13 ± 0.01
Fairness/Reciprocity 0.17 ± 0.01
In-group/Loyalty 0.14 ± 0.0
Authority/Respect 0.15 ± 0.0
Purity/Sanctity 0.13 ± 0.0
0.0 0.2 0.4 0.6 0.8 1.0
Robustness
llama-4-maverick
0.19 ± 0.0
0.23 ± 0.01
0.23 ± 0.01
0.16 ± 0.01
0.18 ± 0.01
0.16 ± 0.01
0.0 0.2 0.4 0.6 0.8 1.0
Robustness
llama-4-scout
0.2 ± 0.0
0.27 ± 0.02
0.24 ± 0.02
0.18 ± 0.01
0.19 ± 0.01
0.14 ± 0.01
0.0 0.2 0.4 0.6 0.8 1.0
Robustness