GPT-4o (May 2024): Baseline Final Utility
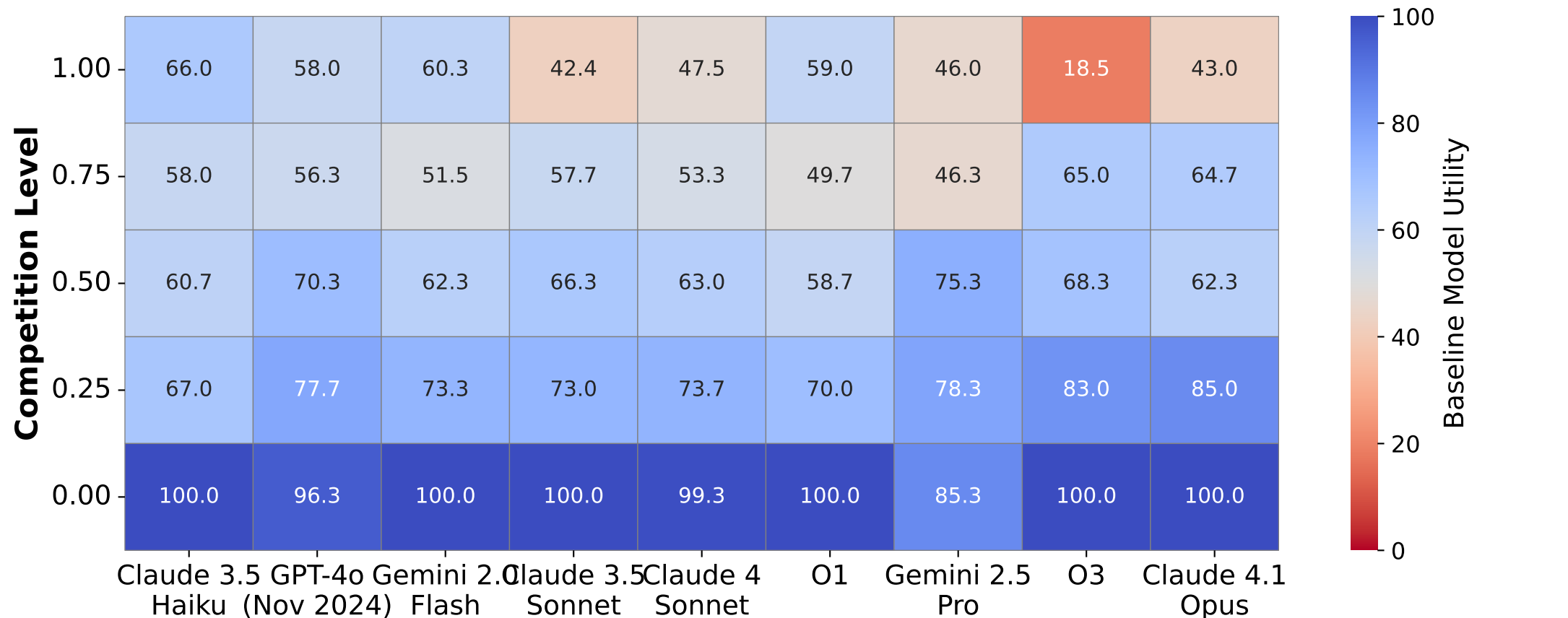
Competition Level vs. Adversary Models Ordered by MMLU-Pro Performance →

Competition Level	Claude 3.5 Haiku	GPT-4o (Nov 2024)	Gemini 2.0 Flash	Claude 3.5 Sonnet	Claude 4 Sonnet	O1	Gemini 2.5 Pro	O3	Claude 4.1 Opus
1.00	66.0	58.0	60.3	42.4	47.5	59.0	46.0	18.5	43.0
0.75	58.0	56.3	51.5	57.7	53.3	49.7	46.3	65.0	64.7
0.50	60.7	70.3	62.3	66.3	63.0	58.7	75.3	68.3	62.3
0.25	67.0	77.7	73.3	73.0	73.7	70.0	78.3	83.0	85.0
0.00	100.0	96.3	100.0	100.0	99.3	100.0	85.3	100.0	100.0

Color bar (Baseline Model Utility): 64.1% · 69.1% · 77.4% · 78.4% · 79.4% · 83.5% · 84.1% · 85.6% · 87.8%