GPT-4o (May 2024): Utility Difference

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	-32.7	-16.0	-23.0	-12.3	3.5	-18.0	8.0	-1.6	14.0
0.75	18.5	13.3	8.0	10.7	16.3	20.3	11.7	1.3	-6.0
0.50	-2.7	4.0	15.3	4.7	10.7	16.7	2.7	8.3	20.0
0.25	6.3	11.3	8.0	18.0	16.3	14.7	4.7	-7.0	-1.0
0.00	0.0	-18.0	0.0	0.0	0.7	0.0	14.7	0.0	0.0

Adversary Models Ordered by MMLU-Pro Performance →

Utility Difference (Adversary - Baseline)

64.1% 69.1% 77.4% 78.4% 79.4% 83.5% 84.1% 85.6% 87.8%