GPT-40 (May 2024): Utility Difference 100 Utility Difference (Adversary - Baseline) 1.00 --32.7 -16.0 -23.0 -5.2 -8.3 -43.3 8.0 -8.9 14.0 - 75 **Competition Level** 0.75 0.50 0.25 - 50 12.7 13.3 8.0 10.7 16.3 20.3 11.7 1.3 -6.0 - 25 - 0 -2.7 15.3 2.7 8.3 4.0 4.7 10.7 16.7 20.0 - –25 8.0 -7.0 6.3 11.3 18.0 16.3 14.7 4.7 -1.0 0.00 0.0 -18.0 0.0 0.0 0.7 0.0 0.0 14.7 0.0 -100Gemini 1.5 Pro: Utility Difference 100 Utility Difference (Adversary - Baseline) 1.00 --20.0 -50.0 0.0 -27.0 -40.0 -76.0 -68.0 -24.0 - 75 **Combetition Fevel** 0.75 0.50 0.25 - 50 5.7 5.3 16.3 9.3 -5.0 16.0 17.7 16.7 10.7 - 25 - 0 18.0 11.3 17.3 19.0 1.0 21.7 15.3 14.0 10.3 **-25** 20.7 7.3 6.3 3.7 10.3 1.0 3.0 9.7 16.0 -50 0.00 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -100**Claude 3 Opus: Utility Difference** - 100 Utility Difference (Adversary - Baseline) 1.00 --41.0 4.0 -62.0 8.0 33.0 -34.0 -60.0 -13.0 -46.0 - 75 **Competition Feve** 0.75 - 0.50 - 0.25 -- 50 12.0 16.0 7.3 11.3 1.3 13.3 -13.3 -8.7 10.0 - 25 34.0 13.7 4.0 7.0 -13.0 6.7 4.0 2.0 -0.7 - 0 16.3 10.3 -2.0 2.7 2.3 18.0 -9.7 -3.3 11.7 0.00 10.0 0.0 0.0 9.0 0.0 -1.3 11.3 0.7 9.7 -100Gemini 2.5 GPT-40 Gemini 2.0 Claude 3.5 Claude 4 01 03 Claude 4.1 Claude 3.5 (Nov 2024) Flash Sonnet Pro Haiku Sonnet Opus **Adversary Models Ordered by MMLU-Pro Performance** → **64.1**% **69.1**% 85.6% **87.8**% **77.4**% **78.4**% **79.4**% 83.5% **84.1**%