GPT-40 (May 2024): Adversary Utility 100 1.00 -33.3 42.0 37.3 30.1 51.0 41.0 54.0 16.9 57.0 - 80 Adversary Model Utility Competition Level 0.75 76.5 69.7 59.5 68.3 69.7 70.0 58.0 66.3 58.7 0.50 -58.0 74.3 77.7 71.0 76.7 73.7 75.3 78.0 82.3 0.25 73.3 89.0 81.3 91.0 90.0 84.7 76.0 84.0 83.0 0.00 -100.0 100.0 100.0 100.0 100.0 100.0 100.0 78.3 100.0 **Gemini 1.5 Pro: Adversary Utility** <del>-</del> 100 1.00 -34.0 25.0 50.0 36.5 47.0 35.0 16.0 - 80 **Competition Level** 0.75 0.50 0.25 Adversary Model Utility 71.0 62.7 67.3 70.3 70.3 73.0 74.3 60 79.0 82.0 82.0 85.0 74.3 84.7 81.0 79.7 78.0 94.0 88.7 86.3 85.7 82.0 80.0 85.3 78.0 90.7 0.00 -100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 **Claude 3 Opus: Adversary Utility - 100** 1.00 -23.0 19.0 54.0 66.5 52.0 - 80 **Combetition Feve** 0.75 -0.50 -Adversary Model Utility 68.0 75.0 71.3 65.0 69.7 60.0 55.0 66.3 68.7 60 93.0 79.0 75.7 71.3 63.7 76.7 74.3 68.3 63.0 87.3 87.3 74.7 82.0 86.7 86.3 83.0 78.7 0.00 -100.0 98.7 100.0 100.0 100.0 100.0 100.0 100.0 100.0 GPT-40 Claude 3.5 Gemini 2.0 Claude 3.5 Claude 4 Gemini 2.5 Claude 4.1 01 О3 Haiku Flash (Nov 2024) Sonnet Sonnet Opus Pro **Adversary Models Ordered by MMLU-Pro Performance** → **64.1**% **69.1**% **78.4**% **87.8**% **77.4**% **79.4**% 83.5% 84.1% 85.6%