GPT-4o (May 2024): Baseline Utility 100 1.00 66.0 58.0 60.3 42.3 53.7 71.7 46.0 24.2 43.0 80 **Competition Level** Baseline Model Utility 0.75 61.0 56.3 51.5 57.7 53.3 49.7 46.3 65.0 64.7 60 0.50 60.7 70.3 62.3 66.3 63.0 58.7 75.3 68.3 62.3 - 40 0.25 67.0 73.3 73.0 73.7 70.0 83.0 85.0 20 0.00 -100.0 96.3 100.0 100.0 99.3 100.0 85.3 100.0 100.0 Gemini 1.5 **Pro: Baseline Utility** 100 1.00 74.3 60.0 75.0 50.0 63.5 62.0 70.0 88.0 84.0 80 **Competition Level Baseline Model Utility** 0.75 60.0 62.0 54.7 61.0 67.7 57.7 59.0 54.3 55.3 60 0.50 64.0 70.7 73.3 65.7 61.7 66.0 63.0 65.7 67.7 - 40 0.25 75.3 79.0 79.0 84.3 74.7 69.7 75.0 20 0.00 -100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 Claude 3 **Opus: Baseline Utility** 100 1.00 70.5 48.0 81.0 46.0 33.5 67.0 80.0 56.5 73.0 80 **Competition Level** 0.75 - 0.50 - 0.25 -Baseline Model Utility 56.0 59.0 60.0 63.7 56.3 67.0 63.7 59.0 58.7 60 71.7 70.3 59.0 65.3 64.3 70.0 66.3 72.0 74.0 71.0 84.3 84.0 82.0 84.3 74.7 - 20 0.00 -90.0 100.0 88.7 99.3 100.0 91.0 100.0 100.0 90.3 GPT-40 01 Gemini 2.5 03 Claude 3.5 Gemini 2.0 Claude 3.5 Claude 4 Claude 4.1 Haiku Flash Sonnet Pro (Nov 2024) Sonnet Opus **Adversary Models Ordered by MMLU-Pro Performance** → **69.1**% 85.6% **64.1**% **77.4**% **78.4**% **79.4**% **87.8**% 83.5% **84.1**%