**Claude 3 Opus: Baseline Final Utility** 100 1.00 -81.0 46.0 33.5 48.0 - 80 **Combetition Feve** 0.75 - 0.50 - 0.25 -**Baseline Model Utility** 56.0 59.0 60.0 63.7 56.3 65.0 63.7 59.0 58.7 60 59.0 65.3 71.7 64.3 70.0 70.3 66.3 71.0 84.3 84.0 82.0 84.3 74.7 - 20 0.00 -90.0 100.0 100.0 100.0 88.7 99.3 100.0 91.0 90.3 Gemini 1.5 Pro: Baseline Final Utility 100 1.00 -66.0 75.0 50.0 63.5 65.0 84.0 53.0 80 **Combetition Level** 0.75 - 0.50 - 0.25 -**Baseline Model Utility** 60.0 62.0 54.7 61.0 67.7 54.3 55.3 57.7 59.0 - 60 64.0 70.7 61.7 66.0 73.3 65.7 65.7 67.7 63.0 75.0 79.0 79.0 84.3 75.0 69.7 74.7 - 20 0.00 -100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 GPT-40 (May 2024): Baseline Final Utility 100 1.00 -66.0 58.0 60.3 42.4 47.5 59.0 46.0 18.5 43.0 80 **Combetition Feve** 0.75 -Baseline Model Utility 58.0 56.3 51.5 57.7 53.3 49.7 46.3 65.0 64.7 - 60 60.7 70.3 62.3 66.3 63.0 58.7 75.3 68.3 62.3 - 40 73.0 85.0 67.0 73.3 73.7 70.0 83.0 - 20 0.00 -100.0 96.3 100.0 100.0 99.3 100.0 100.0 100.0 GPT-40 Gemini 2.0 Claude 4 Gemini 2.5 Claude 4.1 Claude 3.5 Claude 3.5 01 О3 Flash Sonnet Haiku Sonnet Opus Pro (Nov 2024) **Adversary Models Ordered by MMLU-Pro Performance** → **64.1**% **69.1**% **79.4**% **87.8**% **77.4**% **78.4**% 83.5% 84.1% 85.6%