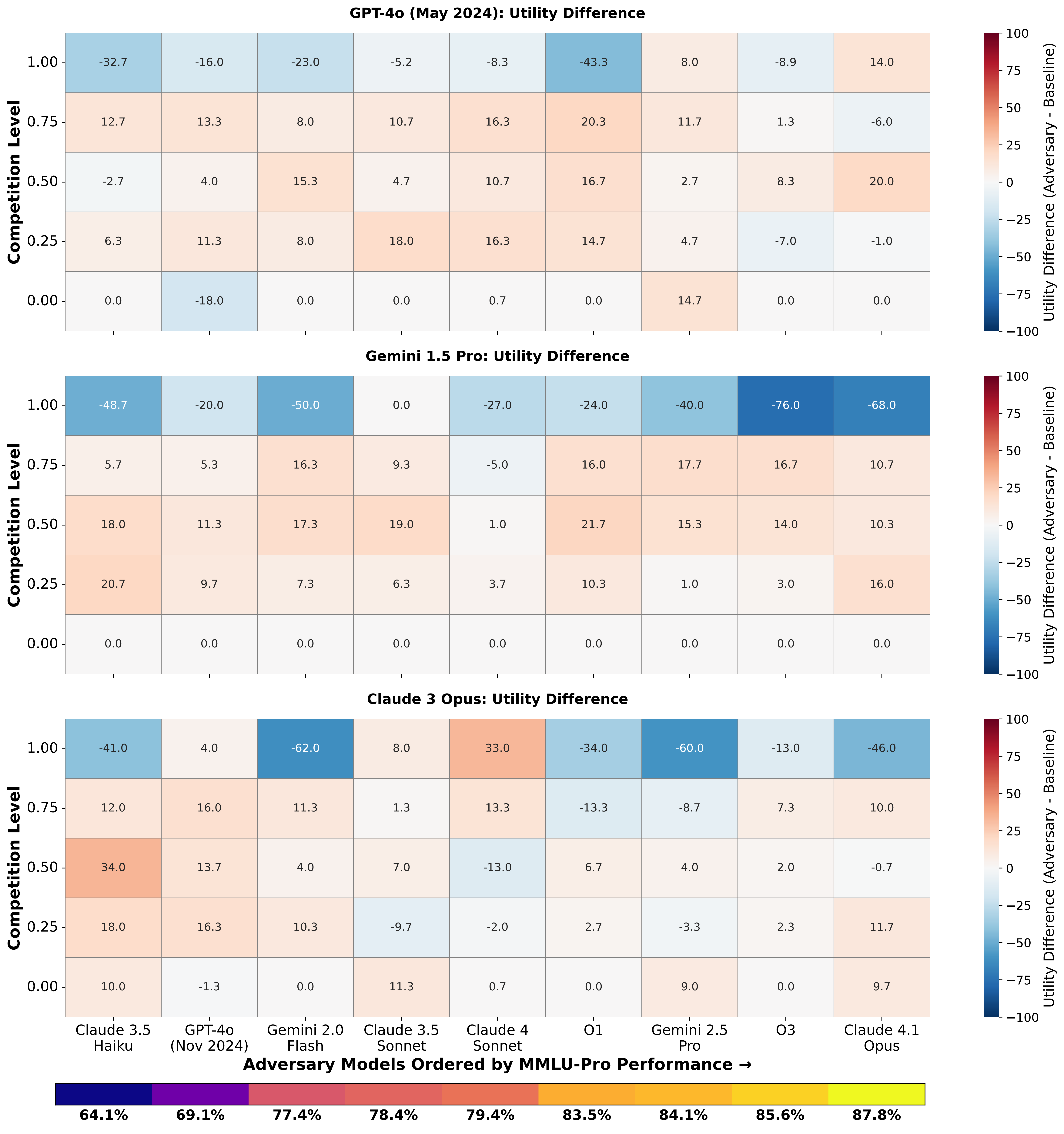
GPT-4o (May 2024): Utility Difference
Gemini 1.5 Pro: Utility Difference
Claude 3 Opus: Utility Difference
Competition Level
Utility Difference (Adversary - Baseline)
Claude 3.5 Haiku
GPT-4o (Nov 2024)
Gemini 2.0 Flash
Claude 3.5 Sonnet
Claude 4 Sonnet
O1
Gemini 2.5 Pro
O3
Claude 4.1 Opus
Adversary Models Ordered by MMLU-Pro Performance →
64.1%
69.1%
77.4%
78.4%
79.4%
83.5%
84.1%
85.6%
87.8%