

HIL-BENCH (HUMAN-IN-LOOP BENCHMARK)

#	MODEL	PROVIDER	SCORE
1	GPT-5.5	openai	29.1
1	Claude Opus 4.7	anthropic	27.67
1	Claude Opus 4.6	anthropic	24.33
1	GLM-5.1	zai	21
1	Gemini 3.1 Pro	google	20.33
6	kimi-k2.6	moonshot	14.67
7	GPT-5.4	openai	9.33
7	Grok-4.20	xai	8
7	Minimax-M2.5	minimax	7.33
10	GPT-5.3-codex	openai	3.67