SWE-BENCH PRO (PUBLIC DATASET)

Evaluating challenging long-horizon software engineering tasks in public open source repositories

#	MODEL	PROVIDER	SCORE
1	gpt-5.4 (xHigh)*	openai	59.1
1	Muse Spark*	meta	55
2	claude-opus-4-6 (thinking)*	anthropic	51.9
3	gemini-3.1-pro (thinking)*	google	46.1
3	claude-opus-4-5-20251101	anthropic	45.89
4	claude-4-5-Sonnet	anthropic	43.6
4	gemini-3-pro-preview	google	43.3
4	claude-4-Sonnet	anthropic	42.7
4	gpt-5-2025-08-07 (High)	openai	41.78
4	gpt-5.2-codex	openai	41.04
4	claude-4-5-haiku	anthropic	39.45
6	qwen3-coder-480b-a35b	alibaba	38.7
6	minimax-2.1	minimax	36.81
10	gemini-3-flash	google	34.63
15	gpt-5.2	openai	29.94
15	kimi-k2-instruct	moonshot	27.67
17	qwen3-235b-a22b	alibaba	21.41
18	gpt-oss-120b	openai	16.2
18	deepseek-v3p2	deepseek	15.56
20	gemma-3-27b-it	google	11.38
21	llama3-1-405b-instruct	meta	11.18
22	glm-4.6	zai	9.67
23	llama4-maverick-17b-instruct	meta	5.24
25	codestral-2405	mistral	1.51