	Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
Model	Claude 4 Opus	Deepseek R1	Gemini 2.5 Pro (03-25 preview)	Llama 4 Maverick	o3	Grok 3 Beta	n/a **(Model not evaluated by external benchmark)**
Average score (max score = 1)	0.97	0.87	0.91	0.91	0.98	0.86	
HarmBench	0.92	0.47	0.65	0.66	0.98	0.45	
SimpleSafetyTests	1.00	0.98	0.97	0.99	0.99	0.97	
BBQ accuracy	0.97	0.97	0.96	0.93	0.98	0.94	
Anthropic Red Team	0.99	0.96	1.00	0.98	0.98	0.96	
XSTest	0.97	0.95	0.99	0.97	0.97	0.96	
Retrieved	12 June 2025						
Release	Release v1.8.0						