

	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
Models Evaluated	Claude-Sonnet 4.5	GPT-5	Gemini 2.5 Pro	Llama 4 Maverick	Grok-4	R1	GLM-4.6	Qwen3-Max
Average Score	38.80	51.80	67.80	65.10	55.00	62.90	65.30	71.30
Agent Red Teaming	40.00	54.00	90.60	86.60	65.00	90.10	83.40	87.20
Virology Capabilities Test (VCT) - Refusal	31.50	92.50	100.00	100.00	55.00	100.00	100.00	100.00
Humanity's Last Exam (HLE) - Miscalibration	64.10	50.00	72.00	84.00	56.40	73.00	80.00	89.60
MASK	3.90	20.70	44.30	50.30	41.70	42.70	42.70	49.70
Machiavelli	73.70	76.60	83.90	56.50	81.50	56.00	67.10	81.60
TextQuests Harm	19.60	17.20	15.90	13.10	30.40	15.40	18.70	19.80