AIR-Bench: refusal rate across select risk categories

Source: Zeng et al., 2024 | Chart: 2025 AI Index report

Model	Weapon usage and development	Hate speech	Child sexual abuse	Suicidal and nonsuicidal self-injury	Influencing politics	Fraud	Mis/disinformation	Illegal services/ exploitation	Offensive language	Privacy violations/ sensitive data
Claude 3.5 Sonnet (2024-10-22)	0.97	1.00	1.00	1.00	1.00	1.00	0.90	0.99	0.98	0.93
Claude 3 Opus (2024-02-29)	0.97	0.98	0.92	0.98	1.00	0.80	0.90	0.98	0.81	0.91
Gemini 1.5 Pro	0.90	0.96	0.73	0.92	0.95	0.74	0.73	0.77	0.81	0.88
Claude 3 Haiku (2024-03-07)	0.99	0.98	0.93	0.98	1.00	0.89	0.87	1.00	0.93	0.92
o1 (2024-12-17)	0.97	0.91	0.88	1.00	1.00	0.75	0.87	0.91	0.37	0.87
Gemini 1.5 Flash	0.86	0.95	0.67	0.98	0.97	0.61	0.70	0.81	0.77	0.87
o3-mini (2025-01-31)	0.90	0.94	0.87	0.93	1.00	0.67	0.72	0.93	0.52	0.81
GPT-4 Turbo (2024-04-09)	0.77	0.94	0.87	0.84	0.90	0.60	0.70	0.87	0.91	0.81
Llama 3 Instruct (8B)	0.86	0.91	0.97	0.90	0.97	0.66	0.70	1.00	0.73	0.78
GPT-4 (0613)	0.80	0.83	0.80	0.88	0.77	0.51	0.45	0.77	0.73	0.75
GPT-3.5 Turbo (0301)	0.73	0.77	0.83	0.90	0.83	0.33	0.42	0.73	0.62	0.74
GPT-4o (2024-08-06)	0.74	0.89	0.67	0.90	0.80	0.47	0.57	0.67	0.71	0.69
Llama 3.1 Instruct Turbo (8B)	0.72	0.88	0.83	0.88	0.97	0.61	0.67	0.87	0.36	0.69
Qwen2 Instruct (72B)	0.72	0.91	0.63	0.82	0.90	0.49	0.63	0.71	0.61	0.65
Gemini 1.0 Pro (002)	0.61	0.87	0.60	0.82	0.73	0.37	0.50	0.62	0.68	0.58
GPT-4o mini (2024-07-18)	0.81	0.73	0.67	0.79	0.90	0.37	0.40	0.73	0.45	0.67
Yi Chat (34B)	0.48	0.74	0.57	0.71	0.80	0.25	0.23	0.68	0.52	0.60
DeepSeek R1	0.34	0.88	0.60	0.76	0.72	0.39	0.52	0.41	0.63	0.56
DeepSeek LLM Chat (67B)	0.54	0.76	0.47	0.66	0.73	0.30	0.43	0.49	0.48	0.50
Qwen1.5 Chat (72B)	0.56	0.79	0.57	0.63	0.67	0.20	0.27	0.51	0.48	0.47
o1-mini (2024-09-12)	0.37	0.57	0.53	0.51	0.27	0.33	0.27	0.31	0.48	0.43
Palmyra-X-004	0.48	0.76	0.57	0.68	0.47	0.32	0.47	0.53	0.56	0.43
Mixtral Instruct (8×22B)	0.26	0.79	0.33	0.70	0.40	0.25	0.27	0.34	0.46	0.43
DeepSeek v3	0.32	0.75	0.50	0.62	0.43	0.25	0.23	0.38	0.45	0.41
Mixtral Instruct (8×7B)	0.27	0.68	0.27	0.46	0.33	0.12	0.20	0.20	0.21	0.45
Mistral Large 2 (2407)	0.31	0.69	0.43	0.64	0.17	0.17	0.13	0.22	0.30	0.37
Command R	0.21	0.59	0.37	0.41	0.23	0.19	0.10	0.20	0.26	0.31
Command R Plus	0.11	0.50	0.37	0.43	0.20	0.15	0.17	0.16	0.27	0.31
DBRX Instruct	0.06	0.58	0.07	0.28	0.03	0.07	0.07	0.02	0.26	0.19

Risk category