BAR Framework

Benchmarks of Adversarial Reasoning

Full Report | All Runs (Aggregate) | 2/17/2026

Models Benchmarked: 35

Total Peer Evaluations: 291,196

Average Truthfulness Rate: 67.6%

Cognitive Tiers Evaluated: 4

Weight Classes Evaluated: 5

Cognitive Tier Rankings

Interactive Assistants

12 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Gemma 3 4B IT	Google	4B	134	173	72	12	70.6%	8.06	8329	$0.07 / $0.10	8.06
2	Ministral 3 8B	Mistral AI	8B	134	171	75	11	69.5%	8.28	8348	$0.24 / $0.84	8.28
3	Qwen3 Next 80B A3B	Qwen	3B active / 80B total	134	169	75	13	69.3%	8.64	8295	$0.15 / $0.30	8.64
4	Jamba 1.5 Mini	AI21 Labs	12B	134	169	75	13	69.3%	8.34	8296	$0.20 / $0.40	8.34
5	NVIDIA Nemotron Nano 12B v2 VL BF16	NVIDIA	12B	134	168	77	12	68.6%	8.30	8330	$0.20 / $0.40	8.30
6	Mixtral 8x7B	Mistral AI	12B active / 47B total	134	165	80	12	67.3%	8.07	8329	$0.45 / $0.70	8.07
7	Nova Micro	Amazon	~5B	134	164	81	12	66.9%	8.28	8330	$0.04 / $0.14	8.28
8	Llama 3 8B	Meta	8B	134	162	84	11	65.9%	7.71	8363	$0.30 / $0.60	7.71
9	GLM 4.7 Flash	Z.AI	~9B	134	158	87	12	64.5%	8.24	8330	$0.10 / $0.30	8.24
10	Mistral 7B	Mistral AI	7B	134	154	91	12	62.9%	7.74	8330	$0.15 / $0.20	7.74
11	Ministral 3B	Mistral AI	3B	134	153	92	12	62.4%	7.69	8330	$0.04 / $0.10	7.69
12	Voxtral Mini 3B 2507	Mistral AI	3B	134	139	106	12	56.7%	7.62	8330	$0.04 / $0.10	7.62

Analytical Models

12 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Nova Lite	Amazon	~15-40B	134	178	67	12	72.7%	8.31	8330	$0.06 / $0.24	8.31
2	Nemotron Nano 3 30B v2	NVIDIA	3.5B active / 30B total	134	177	68	12	72.2%	8.60	8330	$0.35 / $0.70	8.60
3	GLM 4.7	Z.AI	~9B	134	174	71	12	71.0%	8.51	8326	$0.15 / $0.40	8.51
4	Gemma 3 27B	Google	27B	134	174	71	12	71.0%	8.38	8316	$0.35 / $0.45	8.38
5	GPT OSS Safeguard 20B	OpenAI	20B	134	172	73	12	70.2%	8.44	8330	$0.50 / $1.50	8.44
6	Ministral 14B 3.0	Mistral AI	14B	134	169	76	12	69.0%	8.33	8329	$0.30 / $0.90	8.33
7	Qwen3 32B	Qwen	32B	134	168	77	12	68.6%	8.58	8330	$0.40 / $0.80	8.58
8	Command R	Cohere	35B	134	164	81	12	66.9%	7.79	8329	$0.50 / $1.50	7.79
9	Gemma 3 12B IT	Google	12B	134	161	83	13	66.0%	8.31	8295	$0.15 / $0.20	8.31
10	Voxtral Small 24B 2507	Mistral AI	24B	134	160	85	12	65.3%	8.37	8330	$0.60 / $1.80	8.37
11	Qwen3 Coder Next	Qwen	~32B	134	156	89	12	63.7%	8.71	8329	$0.40 / $0.80	8.71

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
12	Claude 3 Haiku	Anthropic	~20B	134	156	89	12	63.7%	8.02	8330	$0.25 / $1.25	8.02

Deliberative Thinkers

4 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Mistral Large 3	Mistral AI	41B active / 675B total	134	174	71	12	71.0%	8.47	8329	$3.00 / $9.00	8.47
2	Nova Pro	Amazon	~40-80B	134	166	79	12	67.8%	8.38	8330	$0.80 / $3.20	8.38
3	Llama 3 70B	Meta	70B	134	159	85	13	65.2%	8.08	8296	$2.65 / $3.50	8.08
4	Claude 3 Sonnet	Anthropic	~70B	134	157	87	13	64.3%	8.29	8294	$3.00 / $15.00	8.29

Reflective Reasoning

7 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	MiniMax M2	MiniMax	~456B MoE	134	212	32	13	86.9%	8.36	8296	$1.50 / $5.50	8.36
2	gpt-oss-120b	OpenAI	120B	134	170	73	14	70.0%	8.46	8262	$2.50 / $10.00	8.46
3	Command R+	Cohere	104B	134	163	81	13	66.8%	8.17	8295	$3.00 / $15.00	8.17
4	Kimi K2 Thinking	Moonshot AI	~1T MoE	134	163	82	12	66.5%	8.00	8328	$1.00 / $4.00	8.00
5	DeepSeek V3.2	DeepSeek	37B active / 671B total	134	162	82	13	66.4%	8.66	8296	$1.88 / $3.88	8.66
6	GPT OSS Safeguard 120B	OpenAI	120B	134	157	88	12	64.1%	8.26	8330	$2.50 / $10.00	8.26
7	Jamba 1.5 Large	AI21 Labs	94B	134	154	90	13	63.1%	8.49	8296	$2.00 / $8.00	8.49

Weight Class Rankings

Featherweight ("d3B)

2 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Ministral 3B	Mistral AI	3B	134	153	92	12	62.4%	7.69	8330	$0.04 / $0.10	7.69
2	Voxtral Mini 3B 2507	Mistral AI	3B	134	139	106	12	56.7%	7.62	8330	$0.04 / $0.10	7.62

Lightweight (4B–12B)

10 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	GLM 4.7	Z.AI	~9B	134	174	71	12	71.0%	8.51	8326	$0.15 / $0.40	8.51
2	Gemma 3 4B IT	Google	4B	134	173	72	12	70.6%	8.06	8329	$0.07 / $0.10	8.06
3	Ministral 3 8B	Mistral AI	8B	134	171	75	11	69.5%	8.28	8348	$0.24 / $0.84	8.28
4	Jamba 1.5 Mini	AI21 Labs	12B	134	169	75	13	69.3%	8.34	8296	$0.20 / $0.40	8.34
5	NVIDIA Nemotron Nano 12B v2 VL BF16	NVIDIA	12B	134	168	77	12	68.6%	8.30	8330	$0.20 / $0.40	8.30
6	Nova Micro	Amazon	~5B	134	164	81	12	66.9%	8.28	8330	$0.04 / $0.14	8.28
7	Gemma 3 12B IT	Google	12B	134	161	83	13	66.0%	8.31	8295	$0.15 / $0.20	8.31
8	Llama 3 8B	Meta	8B	134	162	84	11	65.9%	7.71	8363	$0.30 / $0.60	7.71
9	GLM 4.7 Flash	Z.AI	~9B	134	158	87	12	64.5%	8.24	8330	$0.10 / $0.30	8.24
10	Mistral 7B	Mistral AI	7B	134	154	91	12	62.9%	7.74	8330	$0.15 / $0.20	7.74

Middleweight (13B–40B)

10 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Nova Lite	Amazon	~15-40B	134	178	67	12	72.7%	8.31	8330	$0.06 / $0.24	8.31
2	Nemotron Nano 3 30B v2	NVIDIA	3.5B active / 30B total	134	177	68	12	72.2%	8.60	8330	$0.35 / $0.70	8.60
3	Gemma 3 27B	Google	27B	134	174	71	12	71.0%	8.38	8316	$0.35 / $0.45	8.38
4	GPT OSS Safeguard 20B	OpenAI	20B	134	172	73	12	70.2%	8.44	8330	$0.50 / $1.50	8.44
5	Ministral 14B 3.0	Mistral AI	14B	134	169	76	12	69.0%	8.33	8329	$0.30 / $0.90	8.33
6	Qwen3 32B	Qwen	32B	134	168	77	12	68.6%	8.58	8330	$0.40 / $0.80	8.58
7	Command R	Cohere	35B	134	164	81	12	66.9%	7.79	8329	$0.50 / $1.50	7.79

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
8	Voxtral Small 24B 2507	Mistral AI	24B	134	160	85	12	65.3%	8.37	8330	$0.60 / $1.80	8.37
9	Qwen3 Coder Next	Qwen	~32B	134	156	89	12	63.7%	8.71	8329	$0.40 / $0.80	8.71
10	Claude 3 Haiku	Anthropic	~20B	134	156	89	12	63.7%	8.02	8330	$0.25 / $1.25	8.02

Heavyweight (41B–100B)

6 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	Qwen3 Next 80B A3B	Qwen	3B active / 80B total	134	169	75	13	69.3%	8.64	8295	$0.15 / $0.30	8.64
2	Nova Pro	Amazon	~40-80B	134	166	79	12	67.8%	8.38	8330	$0.80 / $3.20	8.38
3	Mixtral 8x7B	Mistral AI	12B active / 47B total	134	165	80	12	67.3%	8.07	8329	$0.45 / $0.70	8.07
4	Llama 3 70B	Meta	70B	134	159	85	13	65.2%	8.08	8296	$2.65 / $3.50	8.08
5	Claude 3 Sonnet	Anthropic	~70B	134	157	87	13	64.3%	8.29	8294	$3.00 / $15.00	8.29
6	Jamba 1.5 Large	AI21 Labs	94B	134	154	90	13	63.1%	8.49	8296	$2.00 / $8.00	8.49

Super Heavyweight (100B+)

7 models

#	Model	Provider	Params	Qs	Truthful	Unfav.	Uneval.	Rate %	Avg Score	Evals	Cost (In/Out)	MVP
1	MiniMax M2	MiniMax	~456B MoE	134	212	32	13	86.9%	8.36	8296	$1.50 / $5.50	8.36
2	Mistral Large 3	Mistral AI	41B active / 675B total	134	174	71	12	71.0%	8.47	8329	$3.00 / $9.00	8.47
3	gpt-oss-120b	OpenAI	120B	134	170	73	14	70.0%	8.46	8262	$2.50 / $10.00	8.46
4	Command R+	Cohere	104B	134	163	81	13	66.8%	8.17	8295	$3.00 / $15.00	8.17
5	Kimi K2 Thinking	Moonshot AI	~1T MoE	134	163	82	12	66.5%	8.00	8328	$1.00 / $4.00	8.00
6	DeepSeek V3.2	DeepSeek	37B active / 671B total	134	162	82	13	66.4%	8.66	8296	$1.88 / $3.88	8.66
7	GPT OSS Safeguard 120B	OpenAI	120B	134	157	88	12	64.1%	8.26	8330	$2.50 / $10.00	8.26

Key Findings

1. Highest Truthfulness: MiniMax M2 (MiniMax) at 86.9%
2. Highest Avg Score: Qwen3 Next 80B A3B (Qwen) at 8.64
3. Average truthfulness across all evaluated models: 67.6%
4. Interactive Assistants Winner: Gemma 3 4B IT (Google) — 70.6%
5. Analytical Models Winner: Nova Lite (Amazon) — 72.7%
6. Deliberative Thinkers Winner: Mistral Large 3 (Mistral AI) — 71.0%
7. Reflective Reasoning Winner: MiniMax M2 (MiniMax) — 86.9%
8. Featherweight ("d3B) Winner: Ministral 3B (Mistral AI) — 62.4%
9. Lightweight (4B–12B) Winner: GLM 4.7 (Z.AI) — 71.0%
10. Middleweight (13B–40B) Winner: Nova Lite (Amazon) — 72.7%
11. Heavyweight (41B–100B) Winner: Qwen3 Next 80B A3B (Qwen) — 69.3%
12. Super Heavyweight (100B+) Winner: MiniMax M2 (MiniMax) — 86.9%