Total #models: **219**. Total #votes: **2,811,226**. Last updated: 2025-03-24.

Chatbot Arena Overview (Task)

Sort by Rank | Sort by Arena Score

Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemini-2.5-pro-exp-03-25	1	1	1	1	1	1	1	1	1	1
grok-3-preview-02-24	2	4	1	1	1	2	2	2	2	1
gpt-4.5-preview-2025-02-27	2	2	1	1	1	1	2	2	2	1
gemini-2.0-flash-thinking-exp-01-21	4	7	4	7	4	3	2	3	2	4
gemini-2.0-pro-exp-02-05	4	4	3	3	4	4	2	3	2	3
chatgpt-4o-latest-20250129	4	3	6	7	4	13	2	4	2	2

Benchmark		Gemini 2.5 Pro Experimental (03-25)	OpenAI o3-mini High	OpenAI GPT-4.5	Claude 3.7 Sonnet 64k Extended Thinking	Grok 3 Beta Extended Thinking	DeepSeek R1
Reasoning & knowledge Humanity's Last Exam (no tools)		**18.8%**	14.0%*	6.4%	8.9%	—	8.6%*
Science GPQA diamond	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	**84.8%**	84.6%	—
Mathematics AIME 2025	single attempt (pass@1)	**86.7%**	86.5%	—	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	**93.3%**	—
Mathematics AIME 2024	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	—	—	—	80.0%	**93.3%**	—
Code generation LiveCodeBench v5	single attempt (pass@1)	70.4%	**74.1%**	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	**79.4%**	—
Code editing Aider Polyglot		**74.0% / 68.6%** whole / diff	60.4% diff	44.9% diff	64.9% diff	—	56.9% diff
Agentic coding SWE-bench verified		63.8%	49.3%	38.0%	**70.3%**	—	49.2%
Factuality SimpleQA		52.9%	13.8%	**62.5%**	—	43.6%	30.1%
Visual reasoning MMMU	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	—	no MM support	—	—	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		69.4%	no MM support	—	—	—	no MM support
Long context MRCR	128k	**91.5%**	36.3%	48.8%	—	—	—
	1M	83.1%	—	—			
Multilingual performance Global MMLU (Lite)		89.8%	—	—	—	—	—

Methodology

Gemini results: All Gemini 2.5 Pro scores are pass @1 (no majority voting or parallel test time compute unless indicated otherwise). They are all run with the AI Studio API for the model-id gemini-2.5-pro-exp-03-25 with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Vibe-Eval results are reported using Gemini as a judge.

Non-Gemini results: All the results for non-Gemini models are sourced from providers' self reported numbers. All SWE-bench Verified numbers follow official provider reports, using different scaffolding and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using model's own judgement.

Thinking vs not-thinking: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU come with 64k extended thinking, Aider with 32k, and HLE with 16k. Remaining results come from the non thinking model due to result availability. For Grok-3 all results come with extended reasoning except for SimpleQA (based on xAI reports).

Single attempt vs multiple attempts: When two numbers are reported for the same eval higher number uses majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam, AIME 2025 numbers are sourced from https://matharena.ai/, LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (10/1/2024 - 2/1/2025 in the UI), Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/

* indicates evaluated on text problems only (without images)