LiveBench

A Challenging, Contamination-Free LLM Benchmark

LiveBench appeared as a Spotlight Paper in ICLR 2025.
This work is sponsored by Abacus.AI

Leaderboard Details Code Data

Paper

Introduction

Introducing **LiveBench**: a benchmark for LLMs designed with test set contamination and objective evaluation in mind. It has the following properties:

- LiveBench limits potential contamination by releasing new questions regularly.
- Each question has verifiable, objective ground-truth answers, eliminating the need for an LLM judge.
- LiveBench currently contains a set of 23 diverse tasks across 7 categories, and we will release new, harder tasks over time.

We will evaluate your model on LiveBench! Open a github issue or email us at livebench@livebench.ai!

Leaderboard

We update questions regularly so that the benchmark completely refreshes every 6 months. Some questions for previous releases are available here. The most recent version is **LiveBench-2026-01-08**. This version features a new mathematical task and a new data analysis task.

To further reduce contamination, we delay publicly releasing the questions from the most-recent updates.

View Full Changelog

2026-01-08

☑ Reasoning Average

☐ Show Subcategories

☑ Coding Average
☐ Show Subcategories

☑ Agentic Coding Average
☐ Show Subcategories

☑ Mathematics Average
☐ Show Subcategories

☑ Data Analysis Average
☐ Show Subcategories

☑ Language Average
☐ Show Subcategories

☑ IF Average
☐ Show Subcategories

☑ Show Organization ☐ Show API Name
☑ Show Reasoning Models
☐ Show Open Weight Models Only
☐ Show Model Effort Variants Clear Filters

Search...

Filter by organization...

Model	Organization	Global Average	Reasoning Average	Coding Averag
Claude 4.5 Opus Thinking High Effort	Anthropic	75.96	80.09	79.65
GPT-5.2 High	OpenAI	74.84	83.21	76.07
GPT-5.2 Codex	OpenAI	74.30	77.71	83.62
GPT-5.1 Codex Max High	OpenAI	73.98	83.65	80.68

Model	Organization	Global Average	Reasoning Average	Coding Averag
Gemini 3 Pro Preview High	Google	73.39	77.42	74.60
Gemini 3 Flash Preview High	Google	72.40	74.55	73.90
GPT-5.1 High	OpenAI	72.04	78.79	72.49
GPT-5 Pro	OpenAI	70.48	81.69	72.11
GPT-5.1 Codex	OpenAI	68.61	81.98	71.78
Claude Sonnet 4.5 Thinking	Anthropic	68.19	77.59	80.36
GPT-5 Mini High	OpenAI	65.91	68.32	68.20
DeepSeek V3.2 Thinking	DeepSeek	62.20	77.17	64.62
Grok 4	xAI	62.02	79.13	73.13
Claude 4.1 Opus Thinking	Anthropic	61.81	72.33	74.66
Kimi K2 Thinking	Moonshot AI	61.59	63.49	67.44
Claude Haiku 4.5 Thinking	Anthropic	61.32	61.68	72.81

Model	Organization	Global Average	Reasoning Average	Coding Averag
Claude 4 Sonnet Thinking	Anthropic	61.27	69.01	77.48
GPT-5.1 Codex Mini	OpenAI	60.38	64.71	69.93
Grok 4.1 Fast	xAI	59.99	80.20	69.61
Claude 4.5 Opus Medium Effort	Anthropic	59.10	53.21	78.51
DeepSeek V3.2 Exp Thinking	DeepSeek	58.90	64.37	70.06
Gemini 2.5 Pro (Max Thinking)	Google	58.33	70.81	75.69
GLM 4.7	Z.AI	58.09	59.73	73.13
GLM 4.6	Z.AI	55.19	62.06	71.02
Claude 4.1 Opus	Anthropic	54.45	40.89	76.07
Claude Sonnet 4.5	Anthropic	53.69	42.29	76.07
Gemini 2.5 Flash (Max Thinking) (2025-09-25)	Google	53.09	51.45	67.50
Qwen 3 235B A22B	Alibaba	52.97	59.40	68.97

Model	Organization	Global Average	Reasoning Average	Coding Averag
Thinking 2507				
DeepSeek V3.2	DeepSeek	51.84	44.25	75.69
Claude 4 Sonnet	Anthropic	50.98	39.67	80.74
Qwen 3 Next 80B A3B Thinking	Alibaba	50.41	58.16	60.66
DeepSeek V3.2 Exp	DeepSeek	49.85	45.50	73.19
GPT-5.2 No Thinking	OpenAI	48.91	42.80	76.45
Qwen 3 235B A22B Instruct 2507	Alibaba	48.84	58.43	69.61
GPT-5 Nano High	OpenAI	48.62	40.29	62.39
Qwen 3 Next 80B A3B Instruct	Alibaba	48.35	54.75	68.20
Kimi K2 Instruct	Moonshot AI	48.10	42.23	74.28
Gemini 2.5 Flash (Max Thinking) (2025-06-05)	Google	47.74	44.64	66.03

Model	Organization	Global Average	Reasoning Average	Coding Averag
GPT OSS 120b	OpenAI	46.09	39.21	60.21
Claude Haiku 4.5	Anthropic	45.33	33.94	72.17
Grok Code Fast	xAI	45.13	42.30	64.44
Qwen 3 32B	Alibaba	43.56	48.25	66.03
GPT-5.1 No Thinking	OpenAI	42.65	26.81	77.48
Gemini 2.5 Flash Lite (Max Thinking) (2025-06-17)	Google	42.56	43.34	66.41
Gemini 2.5 Flash Lite (Max Thinking) (2025-09-25)	Google	42.39	36.16	65.39
Devstral 2	Mistral	41.24	27.74	66.79
GLM 4.6V	Z.AI	40.07	37.22	64.24
Qwen 3 30B A3B	Alibaba	39.01	36.68	48.88
Grok 4.1 Fast (Non-Reasoning)	xAI	33.45	23.35	54.26

BibTeX

```
@inproceedings{livebench,
  title={LiveBench: A Challenging, Contamination-Free {LLM} Benchmark},
  author={Colin White and Samuel Dooley and Manley Roberts and Arka Pal and Benja
  booktitle={The Thirteenth International Conference on Learning Representations}
  year={2025},
}
```

Colin White*[1],Samuel Dooley*[1],Manley Roberts*[1],Arka Pal*[1], Ben Feuer[2],Siddhartha Jain[3],Ravid Shwartz-Ziv[2],Neel Jain[4],Khalid Saifullah[4],Sandeep Singh Sandha[1],Siddartha Naidu[1], Chinmay Hegde[2],Yann LeCun[2],Tom Goldstein[4],Willie Neiswanger[5],Micah Goldblum[2]

[1]Abacus.AI,[2]NYU,[3]Nvidia,[4]UMD,[5]USC

The site was inspired by the Nerfies project and by LiveCodeBench.