AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Length-controlled (LC) win rates alleviate length biases of GPT-4, but it may favor models finetuned on its outputs.

Version: AlpacaEval AlpacaEval 2.0 Filter: Community Verified

Baseline: GPT-4 Preview (11/06) | Auto-annotator: GPT-4 Preview (11/06)

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13)	57.5%	51.3%
2	GPT-4 Turbo (04/09)	55.0%	46.1%
3	Yi-Large Preview	51.9%	57.5%
4	GPT-4o Mini (07/18)	50.7%	44.7%
5	GPT-4 Preview (11/06)	50.0%	50.0%
6	Claude 3 Opus (02/29)	40.5%	29.1%
7	Llama 3.1 405B Instruct	39.3%	39.1%
8	GPT-4	38.1%	23.6%
9	Qwen2 72B Instruct	38.1%	29.9%
10	Llama 3.1 70B Instruct	38.1%	39.1%
11	Qwen1.5 72B Chat	36.6%	26.5%
12	GPT-4 (03/14)	35.3%	22.1%
13	Claude 3 Sonnet (02/29)	34.9%	25.6%
14	Llama 3 70B Instruct	34.4%	33.2%
15	Mistral Large (24/02)	32.7%	21.4%
16	Mixtral 8x22B v0.1	30.9%	22.2%
17	GPT-4 (06/13)	30.2%	15.8%
18	Contextual AI (KTO-Mistral-PairRM)	29.7%	33.2%
19	Mistral Medium	28.6%	21.9%
20	Claude 2	28.2%	17.2%

21	Claude	27.3%	17.0%
22	Yi 34B Chat	27.2%	29.7%
23	DBRX Instruct	25.4%	18.4%
24	Claude 2.1	25.3%	15.7%
25	Gemini Pro	24.4%	18.2%
26	Qwen1.5 14B Chat	23.9%	18.6%
27	Mixtral 8x7B v0.1	23.7%	18.3%
28	Llama 3 8B Instruct	22.9%	22.6%
29	GPT 3.5 Turbo (06/13)	22.7%	14.1%
30	Tulu 2+DPO 70B	21.2%	16.0%
31	Llama 3.1 8B Instruct	20.9%	21.8%
32	Mistral 7B v0.3	20.6%	16.7%
33	GPT 3.5 Turbo (11/06)	19.3%	9.2%
34	GPT 3.5 Turbo (03/01)	18.1%	9.6%
35	Vicuna 33B v1.3	17.6%	12.7%
36	Mistral 7B v0.2	17.1%	14.7%
37	OpenHermes-2.5-Mistral (7B)	16.2%	10.3%
38	Qwen1.5 7B Chat	14.7%	11.8%
39	LLaMA2 Chat 70B	14.7%	13.9%
40	Cohere Command	10.9%	12.9%
41	Vicuna 13B v1.3	10.8%	7.1%
42	Gemma Instruct (7B)	10.4%	6.9%
43	LLaMA 33B OASST SFT	9.9%	4.8%
44	WizardLM 13B	9.8%	5.9%
45	Nous Hermes 13B	9.7%	5.4%
46	Vicuna 13B	9.2%	5.8%
47	Davinci001	9.0%	2.8%
48	LLaMA2 Chat 13B	8.4%	7.7%
49	Guanaco 65B	8.3%	6.9%
50	LLaMA 33B OASST RLHF	8.0%	6.3%

51	Phi-2 DPO	7.8%	7.8%
52	Vicuna 7B v1.3	7.2%	4.6%
53	Alpaca Farm PPO Sim (GPT-4) 7B	7.1%	3.5%
54	Alpaca Farm PPO Human 7B	6.4%	4.1%
55	Vicuna 7B	6.3%	4.2%
56	Alpaca 7B	5.9%	2.6%
57	Phi-2 SFT	5.9%	4.0%
58	Guanaco 33B	5.7%	5.0%
59	Falcon 40B Instruct	5.6%	3.3%
60	Gemma Instruct (2B)	5.4%	3.4%
61	LLaMA2 Chat 7B	5.4%	5.0%
62	Pythia 12B SFT	4.2%	2.6%
63	Falcon 7B Instruct	4.0%	2.1%
64	Pythia 12B OASST SFT	3.3%	1.8%
65	Guanaco 13B	3.0%	3.5%
66	Guanaco 7B	2.9%	2.9%
67	Qwen1.5 1.8B Chat	2.6%	3.7%

Github

About AlpacaEval

AlpacaEval an LLM-based automatic evaluation that is fast, cheap, and reliable. It is based on the AlpacaFarm evaluation set, which tests the ability of models to follow general user instructions. These responses are then compared to reference responses (Davinci003 for AlpacaEval, GPT-4 Preview for AlpacaEval 2.0) by the provided GPT-4 based auto-annotators, which results in the win rates presented above. AlpacaEval displays a high agreement rate with ground truth human annotations, and leaderboard rankings on AlpacaEval are very correlated with leaderboard rankings based on human annotators. Please see our documentation for more details on our analysis.

Adding new models

We welcome new model contributions to the leaderboard from the community! To do so, please follow the steps in the contributions section. Specifically, you'll need to run the model on the evaluation set, auto-annotate the outputs, and submit a PR with the model config and leaderboard results. We've also set up a Discord for community support and discussion.

Adding new evaluators or eval sets

We also welcome contributions for new evaluators or new eval sets! For making new evaluators, we release our ground-truth human annotations and comparison metrics. We also release a rough guide to follow for making new eval sets. We specifically encourage contributions for harder instructions distributions and for safety testing of LLMs.

AlpacaEval limitations

While AlpacaEval provides a useful comparison of model capabilities in following instructions, it is not a comprehensive or gold-standard evaluation of model abilities. For one, as detailed in the AlpacaFarm paper, the auto annotator winrates are correlated with length. Though human annotations also display this bias, it is unclear if more verbose answers add utility in downstream tasks. Additionally, the AlpacaFarm eval set, though diverse, consists mainly of simple instructions. We encourage the community to contribute new, more complex eval sets, such as for tool use. Finally, AlpacaEval does not evaluate the safety of any of the models.