

# Evaluation

## Instructor

Sourab Mangrulkar

Machine Learning Engineer at 

Creator of  PEFT



## Leaderboard

| [Vote](#) | [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |

🏆 This leaderboard is based on the following three benchmarks.

- [Chatbot Arena](#) - a crowdsourced, randomized battle platform. We use 100K+ user votes to compute Elo ratings.
- [MT-Bench](#) - a set of challenging multi-turn questions. We use GPT-4 to grade the model responses.
- [MMLU](#) (5-shot) - a test to measure a model's multitask accuracy on 57 tasks.

💻 Code: The Arena Elo ratings are computed by this [notebook](#). The MT-bench scores (single-answer grading on a scale of 10) are computed by [fastchat.llm\\_judge](#). The MMLU scores are mostly computed by [InstructEval](#). Higher values are better for all benchmarks. Empty cells mean not available. Last updated: November, 2023.

Model	⭐ Arena Elo rating	↗ MT-bench (score)	▲ MMLU	▲ License
<a href="#">GPT-4-Turbo</a>	1210	9.32		Proprietary
<a href="#">GPT-4</a>	1159	8.99	86.4	Proprietary
<a href="#">Claude-1</a>	1146	7.9	77	Proprietary
<a href="#">Claude-2</a>	1125	8.06	78.5	Proprietary
<a href="#">Claude-instant-1</a>	1106	7.85	73.4	Proprietary
<a href="#">GPT-3.5-turbo</a>	1103	7.94	70	Proprietary
<a href="#">WizardLM-70b-v1.0</a>	1093	7.71	63.7	Llama 2 Community
<a href="#">Vicuna-33B</a>	1090	7.12	59.2	Non-commercial
<a href="#">OpenChat-3.5</a>	1070	7.81	64.3	Apache-2.0
<a href="#">Llama-2-70b-chat</a>	1065	6.86	63	Llama 2 Community

CHATBOT ARENA LEADERBOARD

# Human Evaluations are the gold standard

THEY PROVIDE TRUE WAY NOW TO PROPERLY GAUGE THE MODEL CAPABILITIES.

**Question:** If the FED buys bonds in the secondary market  
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

**Assistant A:**

The Federal Reserve buys bonds in the secondary market to increase the money supply.

**Assistant B:**

(A) the money supply will increase.

**Follow-up Question:** How does it affect my daily life? Give 3 examples.

**Assistant A:**

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:  
1. Increasing the money supply.  
2. Increasing the money supply and lowering the discount rate.  
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

**Assistant B:**

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

#### GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, Assistant B provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. Assistant B's response is more helpful, relevant, and detailed.

Figure 1: Multi-turn dialogues between a user and two AI assistants—LLaMA-13B (Assistant A) and Vicuna-13B (Assistant B)—initiated by a question from the MMLU benchmark and a follow-up instruction. GPT-4 is then presented with the context to determine which assistant answers better.

# Model Evaluations

USING SOTA MODELS SUCH AS GPT-4 TO EVALUATE THE PERFORMANCE OF OTHER MODELS.

The evaluations can be skewed towards models trained with data generated using the model using the evaluation

The 😊 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

😊 Submit a model for automated evaluation on the 😊 GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!

🏆 LLM Benchmark 📈 Metrics through time 📜 About 🚀 Submit here!

🔍 Search for your model (separate multiple queries with `;` and press ENTER...)

Select columns to show

<input checked="" type="checkbox"/> Average	<input checked="" type="checkbox"/> ARC	<input checked="" type="checkbox"/> HellaSwag	<input checked="" type="checkbox"/> MMLU	<input checked="" type="checkbox"/> TruthfulQA
<input checked="" type="checkbox"/> Winogrande	<input checked="" type="checkbox"/> GSM8K	<input checked="" type="checkbox"/> DROP	<input type="checkbox"/> Type	<input type="checkbox"/> Architecture
<input type="checkbox"/> Precision	<input type="checkbox"/> Hub License	<input type="checkbox"/> #Params (B)	<input type="checkbox"/> Hub ❤️	
<input type="checkbox"/> Available on the hub	<input type="checkbox"/> Model sha			

Model types

<input checked="" type="checkbox"/> pretrained	<input checked="" type="checkbox"/> fine-tuned	<input checked="" type="checkbox"/> instruction-tuned	<input checked="" type="checkbox"/> RL-tuned
<input checked="" type="checkbox"/> ?			

Precision

<input checked="" type="checkbox"/> float16	<input checked="" type="checkbox"/> bfloat16	<input checked="" type="checkbox"/> 8bit	<input checked="" type="checkbox"/> 4bit	<input checked="" type="checkbox"/> GPTQ	<input checked="" type="checkbox"/> ?

Model sizes (in billions of parameters)

<input checked="" type="checkbox"/> ?	<input checked="" type="checkbox"/> ~1.5	<input checked="" type="checkbox"/> ~3	<input checked="" type="checkbox"/> ~7	<input checked="" type="checkbox"/> ~13	<input checked="" type="checkbox"/> ~35	<input checked="" type="checkbox"/> ~60
<input checked="" type="checkbox"/> 70+						

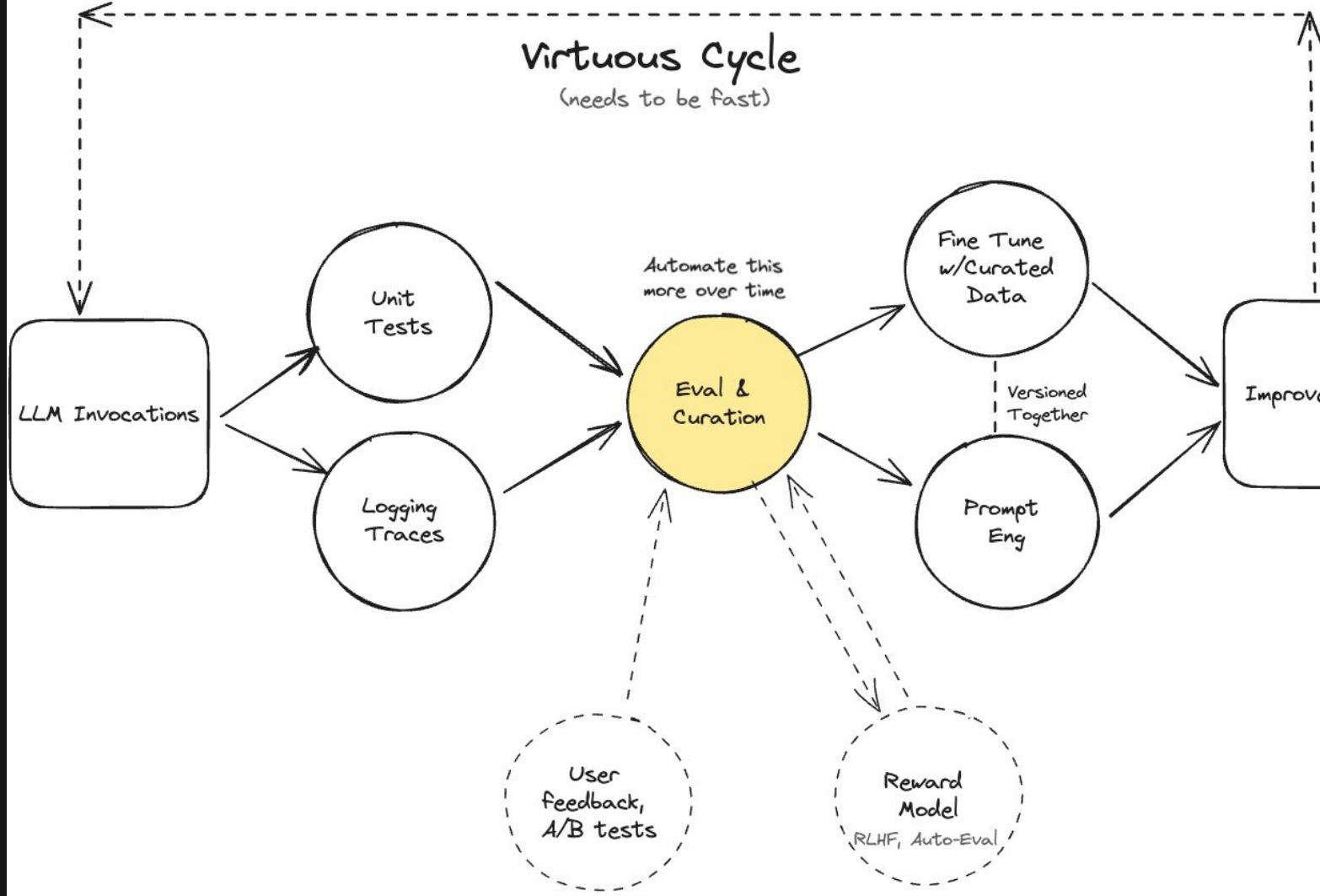
Show gated/private/deleted models

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	DROP
●	<a href="#">01-ai/Yi-34B </a>	68.68	64.59	85.69	76.35	56.23	83.03	50.64	64.2
○	<a href="#">MayaPH/Godzilla2-70B </a>	67.01	71.42	87.53	69.88	61.54	83.19	43.21	52.31
◆	<a href="#">sequelbox/StellarBright </a>	66.98	72.95	87.82	71.17	64.46	83.27	39.5	49.66
◆	<a href="#">garage-bAInd/Platypus2-70B-instruct </a>	66.89	71.84	87.94	70.48	62.26	82.72	40.56	52.41
○	<a href="#">upstage/SOLAR-0-70b-16bit </a>	66.88	71.08	87.89	70.58	62.25	83.58	45.26	47.49

OPEN LLM LEADERBOARD

# LLM Benchmarks

A VERY NOISY PROXY. ONLY  
USEFUL TO GAUGE GENERIC  
CAPABILITIES OF MODELS.



# Task Specific Testing Suite

FROM THE TALK [DATA AND DEVOPS TOOLS FOR EVALUATING AND PRODUCTIONIZING LLMs](#)