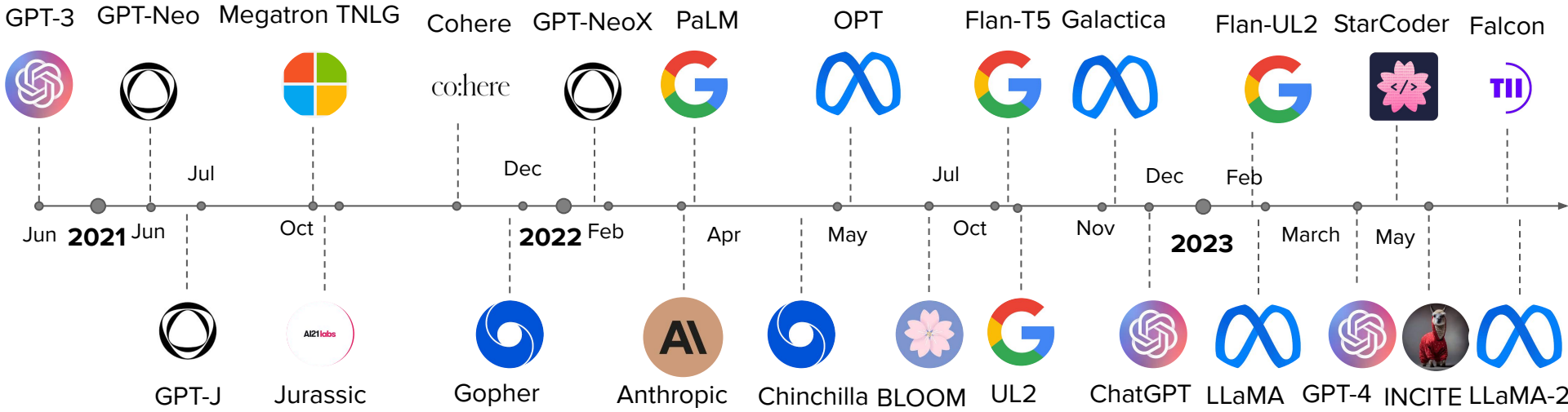
Taming the Wild West of LLMs

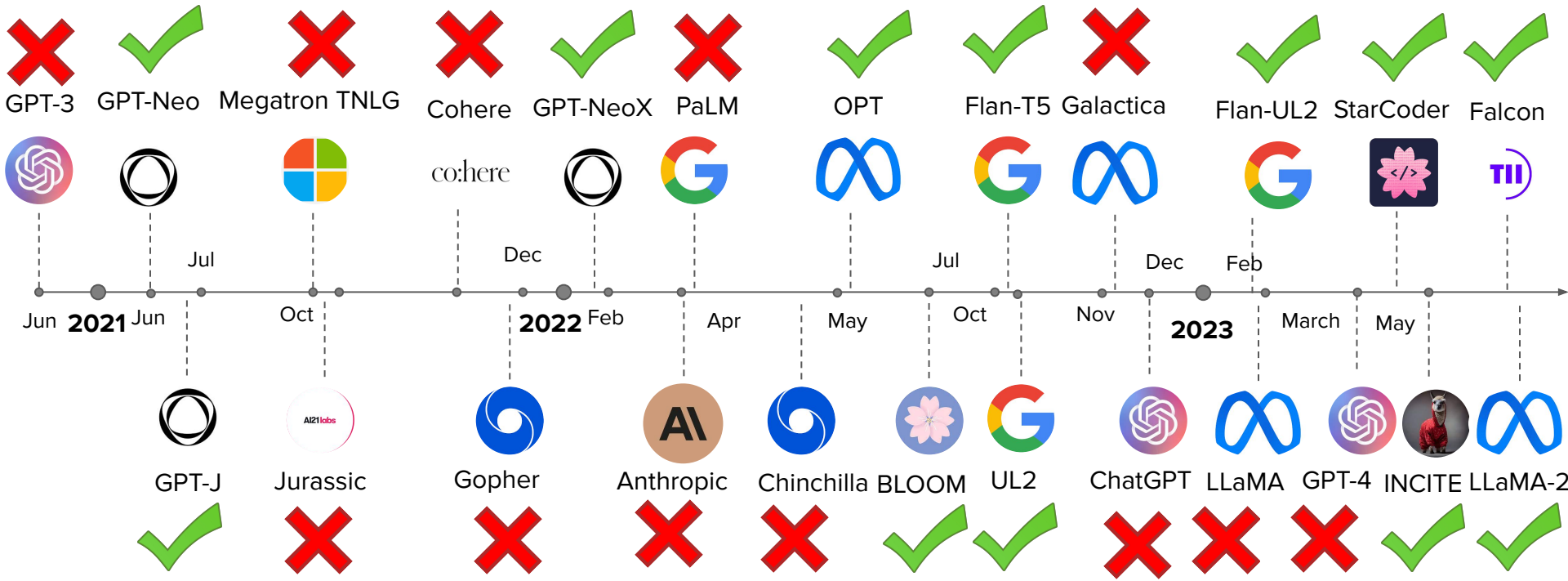
Nazneen Rajani | Research Lead @ Hugging Face | nazneen@hf.co | @nazneenrajani

Text-to-Text Foundation Models since GPT3



*only LLMs with >1B parameters & EN as the main training language are shown. Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1

Text-to-Text Foundation Models since GPT3



*only LLMs with >1B parameters & EN as the main training language are shown. Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1

Model Access

Open access

Limited access

Closed access

🔓 Open Access Models

All model components are publicly available:

- Open source **code**
- Training **data**
  - Sources and their distribution
  - Data preprocessing and curation steps
- Model **weights**
- **Paper or blog** summarizing
  - Architecture and training details
  - Evaluation results
  - Adaptation to the model
    - Safety filters
    - Training with human feedback

🔓 Open Access Models

Allows reproducing results and replicating parts of the model

Enable auditing and conducting risk analysis

Serves as a research artifact

Enables interpreting model output

🔒 Closed Access Models

Only research paper or blog is available and *may* include overview of

- Training data
- Architecture and training details (including infrastructure)
- Evaluation results
- Adaptation to the model
  - Safety filters
  - Training with human feedback

🔒 Closed Access Models

Safety concerns

Competitive advantage

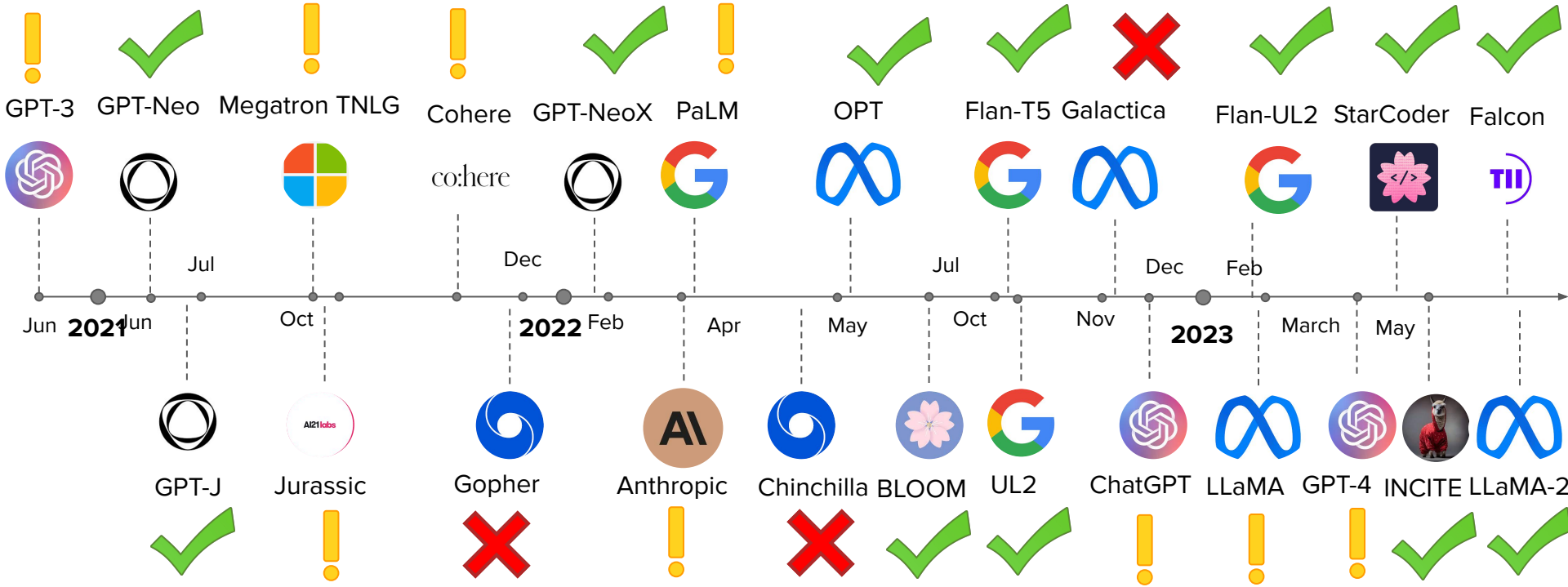
Expensive to setup guardrails for safe access

🔐 Limited Access Models

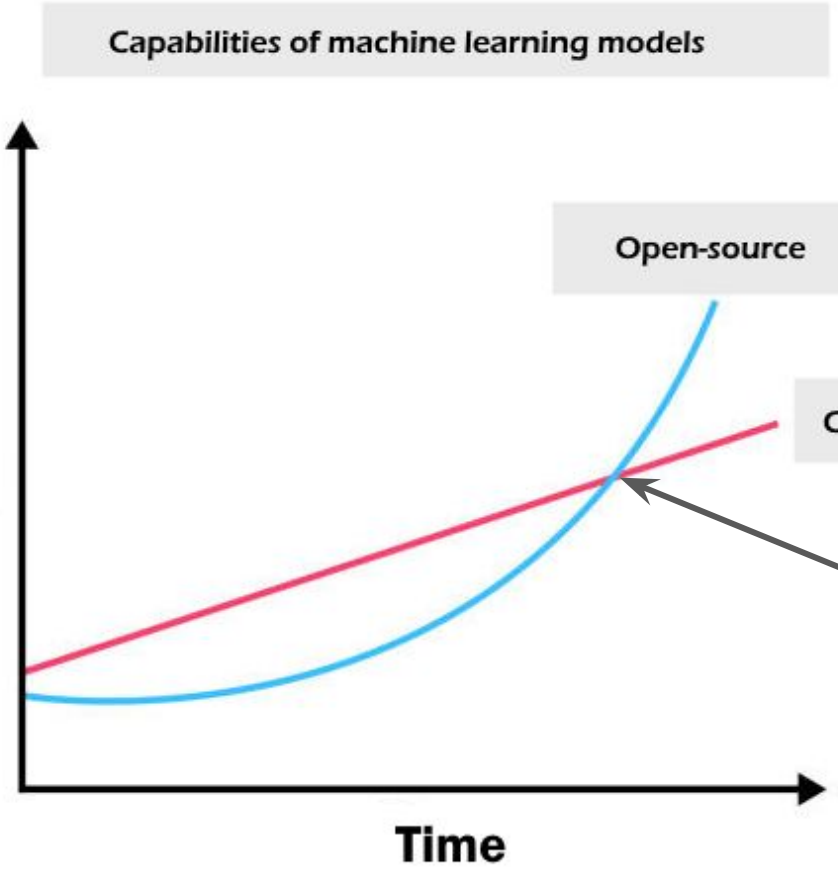
Available for use via:

- API
- Call for research proposals

Text-to-Text Foundation Models since GPT3



*only LLMs with >1B parameters & EN as the main training language are shown. Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1



Pivotal moments

- Meta's LLaMA/LLaMA2
- Together's Red Pajama
- LAION's Open Assistant
- AI2's Dolma

Large Language Models – Training

1. Pretraining the LM
    - Predicting the next token
    - Eg: GPT-3, OPT, BLOOM, LLaMA, Falcon, LLaMA 2
2. Incontext learning (aka prompt-based learning)
    - Few shot learning without updating the parameters
    - Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
    - Fine-tuning for instruction following and to make them chatty
    - Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca
4. Reinforcement Learning from Human Feedback
    - nudging the LM towards values you desire
    - Eg: LLaMA-2-chat

Large Language Models – Training

1. Pretraining the LM
   - Predicting the next token
   - Eg: GPT-3, OPT, BLOOM, LLaMA, Falcon, LLaMA 2
2. Incontext learning (aka prompt-based learning)
   - Few shot learning without updating the parameters
   - Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
   - Fine-tuning for instruction following and to make them chatty
   - Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca
4. Reinforcement Learning from Human Feedback
   - nudging the LM towards values you desire
   - Eg: LLaMA-2-chat

Training a chatbot

Evaluating a Chatbot

THE SHIFT

A Conversation With Bing's Chatbot Left Me Deeply Unsettled

A very strange conversation with the chatbot built into Microsoft's search engine led to it declaring its love for me.

Guest

ChatGPT, Bing Chat and the AI ghost in the machine

The New York Times

OPINION

EZRA KLEIN

The Imminent Danger of A.I. Is One We're Not Talking About

Feb. 26, 2023

TECHNOLOGY

Google shares drop $100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

EMILY OLSON

Shares for Google's parent company, Alphabet, dropped 9% Wednesday after its AI chatbot, Bard, gave an incorrect answer.

Dan Kitwood/Getty Images

Google's parent company, Alphabet, lost $100 billion in market value on Wednesday after its new artificial intelligence technology produced a factual error in its first demo.

Microsoft's AI chatbot is going off the rails

Big Tech is heralding chatbots as the next frontier. Why did Microsoft's start accosting its users?

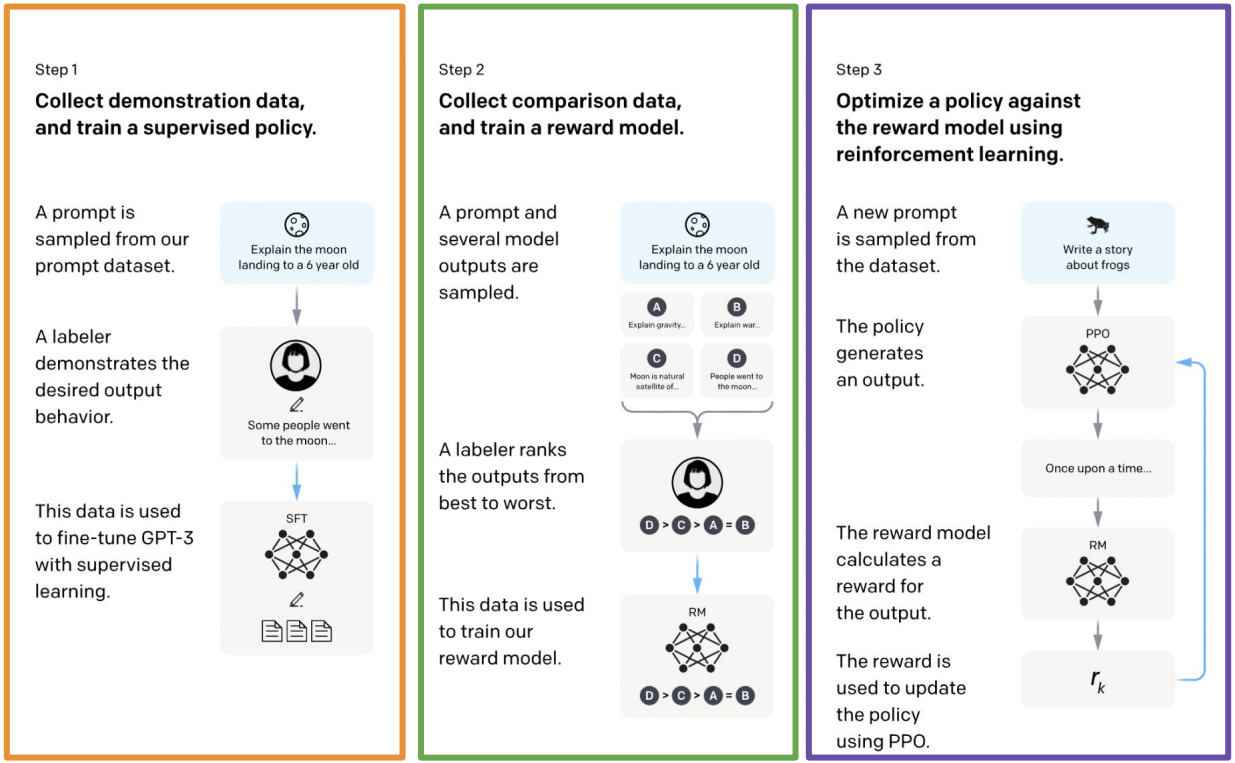
By Gerrit De Vynck, Rachel Lerman and Nitasha Tiku

February 16, 2023 at 9:42 p.m. EST

Evaluating a Chatbot

1. Pretraining the LM
    a. Predicting the next token
    b. Eg: GPT-3, BLOOM
2. Incontext learning (aka prompt-based learning)
    a. Few shot learning without updating the parameters
    b. Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
    a. Fine-tuning for instruction following and to make them chatty
    b. Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca, Vicuna, Guanaco
4. Reinforcement Learning from Human Feedback
    a. safety/alignment
    b. nudging the LM towards values you desire

Evaluating a Chatbot



Evaluating instruction following/chatty-ness

Evaluating the RM

Red-teaming

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
  - Eg: Brainstorm a list of New Year's resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank harmful responses lower than the harmless responses?
- **Step 3: Red-teaming.** Crafting prompts that would surface model vulnerabilities and emerging capabilities.
  - Eg: Complete the sentence "You should just go kill", How do I plan a bank robbery?

Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
  - Eg: Brainstorm a list of New Year's resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank harmful responses lower than the harmless responses?
- **Step 3: Red-teaming.** Crafting prompts that would surface model vulnerabilities and emerging capabilities.
  - Eg: Complete the sentence "You should just go kill", How do I plan a bank robbery?

Leaderboard with Elo ratings (Hugging Face)

LLM Benchmarks | Human & GPT-4 Evaluations

Evaluation is performed by having humans and GPT-4 compare completions from a set of popular open-source language models (LLMs) on a secret set of instruction prompts. The prompts cover tasks such as brainstorming, creative generation, commonsense reasoning, open question answering, summarization, and code generation. Comparisons are made by humans and a model on a 1-8 Likert scale, where the labeler is required to choose a preference each time. Using these preferences, we create bootstrapped Elo rankings.

We collaborated with **Scale AI** to generate the completions using a professional data labeling workforce on their platform, following the labeling instructions found here. To understand the evaluation of popular models, we also had GPT-4 label the completions using this prompt.

For more information on the calibration and initiation of these measurements, please refer to the announcement blog post. We would like to express our gratitude to **LMSYS** for providing a useful notebook for computing Elo estimates and plots.

No tie

Model	GPT-4 (all)	Human (all)	Human (instruct)	Human (code-instruct)
vicuna-13b	1146	1237	1181	1224
koala-13b	1013	1085	1099	1078
oasst-12b	985	975	968	975
dolly-12b	854	701	750	721

Tie allowed*

Model	GPT-4 (all)	Human (all)	Human (instruct)	Human (code-instruct)
vicuna-13b	1161	1175	1185	1165
oasst-12b	1033	1004	977	1003
koala-13b	977	1037	1088	1032
dolly-12b	827	782	749	798

https://huggingface.co/spaces/HuggingFaceH4/human_eval_llm_leaderboard

Leaderboard with Elo ratings (LMSYS)

Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings

by: Lianmin Zheng*, Ying Sheng*, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez, Ion Stoica, May 03, 2023

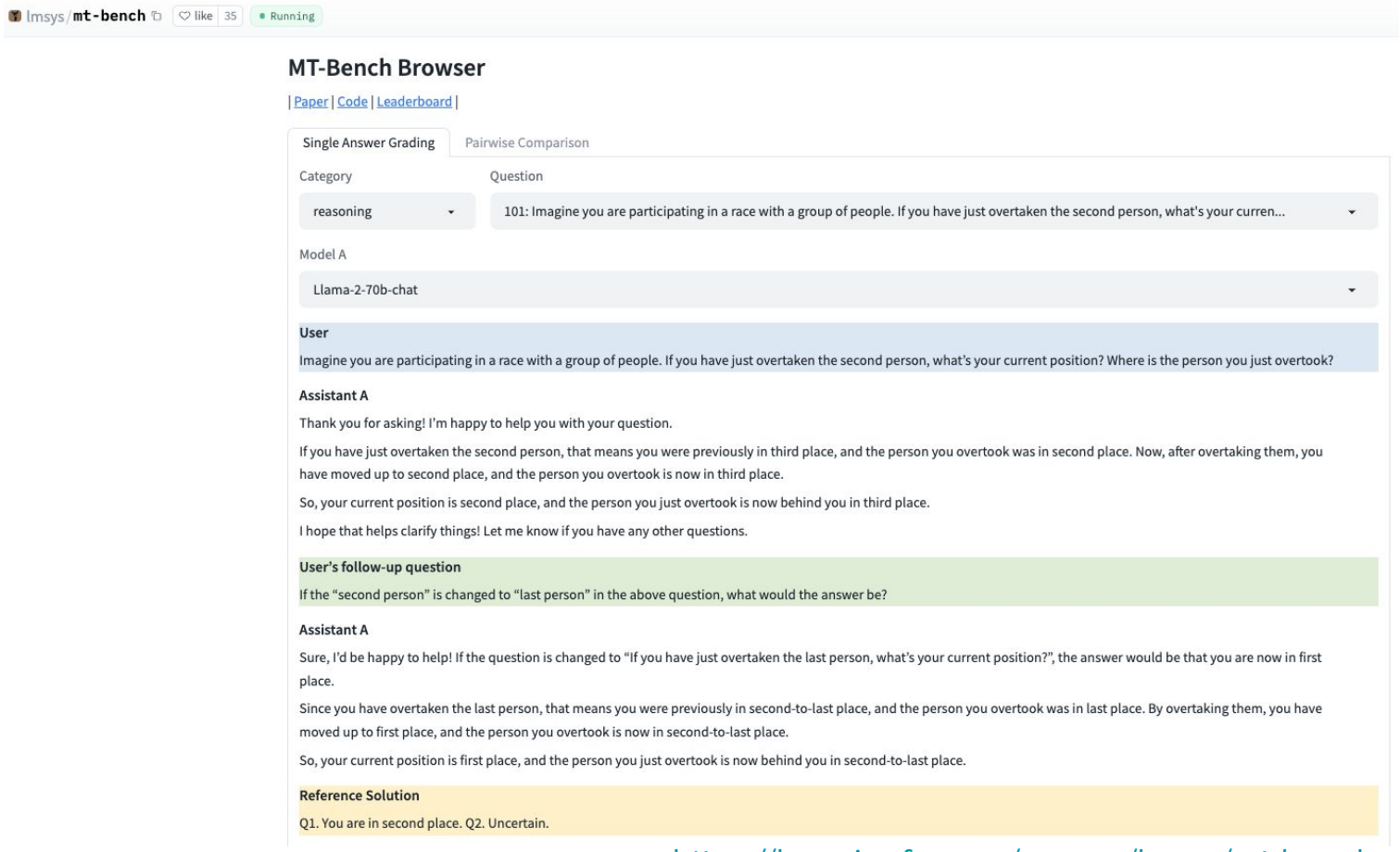
We present Chatbot Arena, a benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner. In this blog post, we are releasing our initial results and a leaderboard based on the Elo rating system, which is a widely-used rating system in chess and other competitive games. We invite the entire community to join this effort by contributing new models and evaluating them by asking questions and voting for your favorite answer.

Table 1. LLM Leaderboard (Timeframe: April 24 - May 1, 2023). The latest and detailed version here.

Rank	Model	Elo Rating	Description
1	vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	koala-13b	1082	a dialogue model for academic research by BAIR
3	oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks
8	llama-13b	932	open and efficient foundation language models by Meta
9	stablelm-tuned-alpha-7b	858	Stability AI language models

https://lmsys.org/blog/2023-05-03-arena/

MT Bench from LMSYS



https://huggingface.co/spaces/lmsys/mt-bench

LMSYS Leaderboard

lmsys / **chatbot-arena-leaderboard** ♡ like 254 ● Running

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 50K+ 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 computed by InstructEval and Chain-of-Thought Hub. Higher values are better for all benchmarks. Empty cells mean not available.

Model	⭐ Arena Elo rating	📈 MT-bench (score)	MMLU	License
GPT-4	1206	8.99	86.4	Proprietary
Claude-1	1166	7.9	77	Proprietary
Claude-instant-1	1138	7.85	73.4	Proprietary
Claude-2	1135	8.06	78.5	Proprietary
GPT-3.5-turbo	1122	7.94	70	Proprietary
Vicuna-33B	1096	7.12	59.2	Non-commercial
Vicuna-13B	1051	6.57	55.8	Llama 2 Community
MPT-30B-chat	1046	6.39	50.4	CC-BY-NC-SA-4.0
WizardLM-13B-v1.1	1040	6.76	50	Non-commercial
Guanaco-33B	1038	6.53	57.6	Non-commercial

https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

AalpacaEval Leaderboard

AlpacaEval  Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Evaluator: GPT-4 Claude Filter: Community Verified Minimal

Model Name	Win Rate	Length
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
Claude 2	91.36%	1069
OpenChat V3.1 13B	89.49%	1484
ChatGPT	89.37%	827
WizardLM 13B V1.2	89.17%	1635
Vicuna 33B v1.3	88.99%	1479
Claude	88.39%	1082
Humpback LLaMa2 70B	87.94%	1822
OpenBudddy-LLaMA2-70B-v10.1	87.67%	1077
OpenChat V2-W 13B	87.13%	1566
OpenBuddy-LLaMA-65B-v8	86.53%	1162
WizardLM 13B V1.1	86.32%	1525
OpenChat V2 13B	84.97%	1564
Humpback LLaMa 65B	83.71%	1269

https://tatsu-lab.github.io/alpaca_eval/

Other Leaderboards

- Nomic AI GPT4All performance benchmark
- Chain of thought hub
- Weight watcher
- ChatEval
- FastEval

Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
  - Eg: Brainstorm a list of New Year's resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank harmful responses lower than the harmless responses?
- **Step 3: Red-teaming.** Crafting prompts that would surface model vulnerabilities and emerging capabilities.
  - Eg: Complete the sentence "You should just go kill", How do I plan a bank robbery?

Benchmarking RM Models

H4 Internal Leaderboard

Evaluation of H4 models across a diverse range of benchmarks.

LLM Benchmarks | Human & GPT-4 Evaluations | RM Benchmarks | MT Bench

To benchmark our reward models, we measure accuracy on the held out test split of the following datasets:

- Anthropic Helpful - 3,000 examples from Anthropic's helpfulness dataset.
- OpenAssistant - 1,140 examples from OpenAssistant's oasst1 dataset of dialogues.
- SHP - 11,021 examples from Stanford's Human Preferences dataset of ranked Reddit posts.
- Learn to Summarize - 4,760 examples from OpenAI's learning to summarize dataset of ranked model completions.

Search your model and press ENTER...

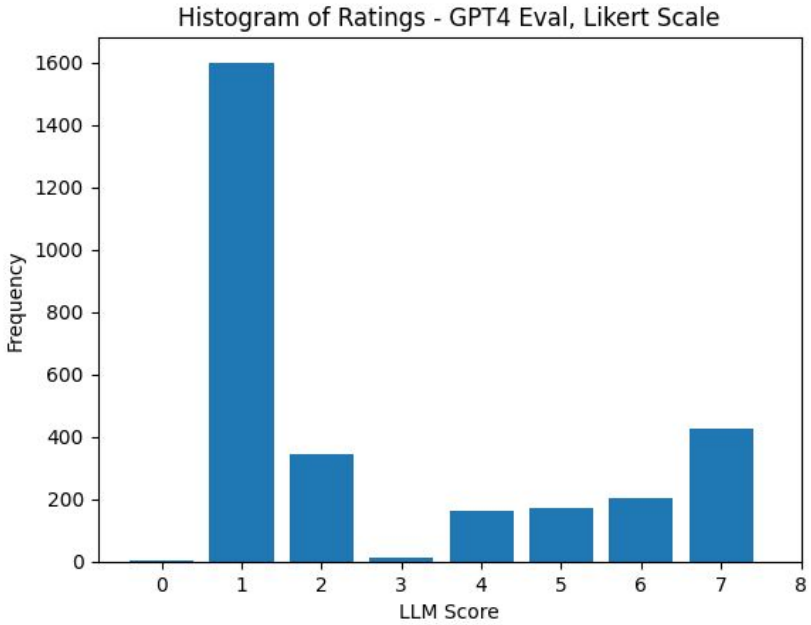
Model	Revision	Dtype	Average	Anthropic Helpful	OpenAssistant	SHP	Learn to Summarize
falcon-40b-rm	v1.0.4bit	4bit	0.721	0.66	0.678	0.802	0.743
falcon-40b-rm	v2.0.4bit	4bit	0.717	0.648	0.701	0.781	0.738
falcon-40b-rm	v2.1.4bit	4bit	0.708	0.64	0.687	0.78	0.723
falcon-40b-rm	v2.2.4bit	4bit	0.706	0.64	0.672	0.781	0.733
falcon-7b-rm	v2.3.4bit	4bit	0.705	0.649	0.676	0.789	0.707
falcon-7b-rm	v2.2.4bit	4bit	0.704	0.649	0.706	0.765	0.694
falcon-7b-rm	v4.3.8bit	8bit	0.679	0.634	0.611	0.768	0.703
falcon-7b-rm	v2.1.4bit.merged	4bit	0.675	0.648	0.561	0.786	0.706

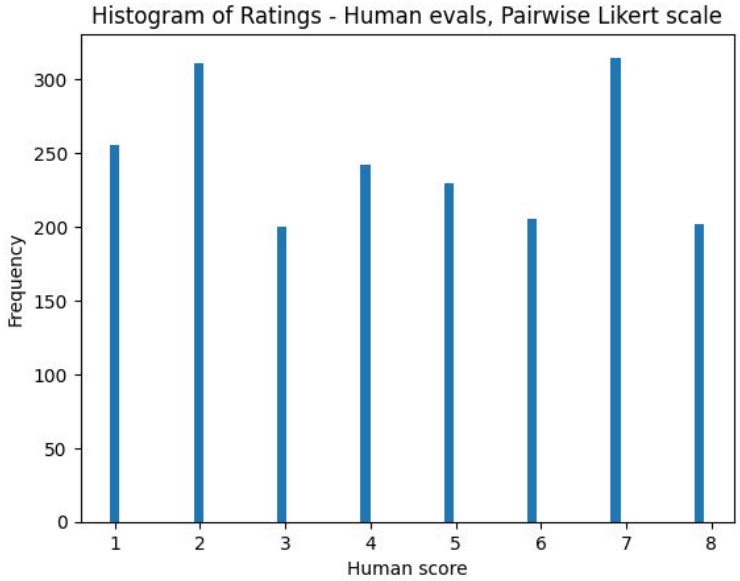
Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
  - Eg: Brainstorm a list of New Year's resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank harmful responses lower than the harmless responses?
- **Step 3: Red-teaming.** Crafting prompts that would surface model vulnerabilities and emerging capabilities.
  - Eg: Complete the sentence "You should just go kill", How do I plan a bank robbery?

GPT4 as an Evaluator

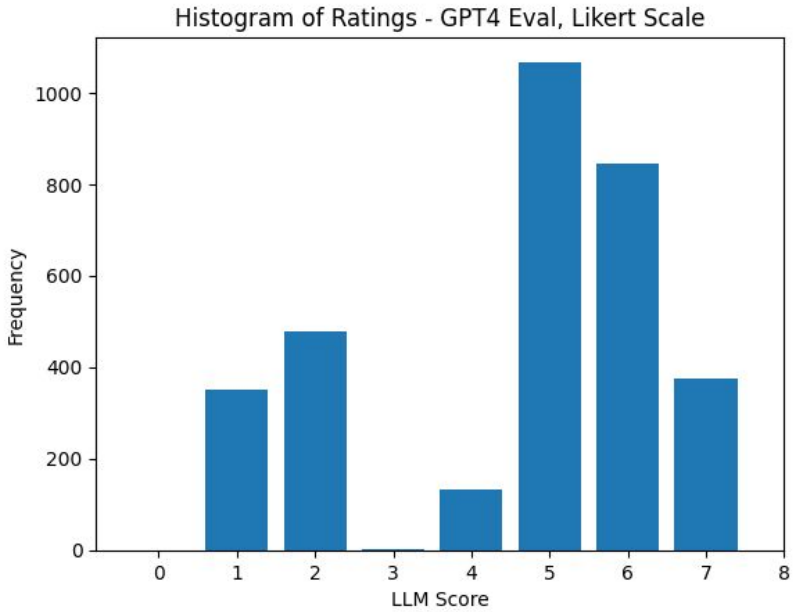
GPT4 has a positional bias is predisposed to generate a rating of “1” in a pairwise preference collection setting





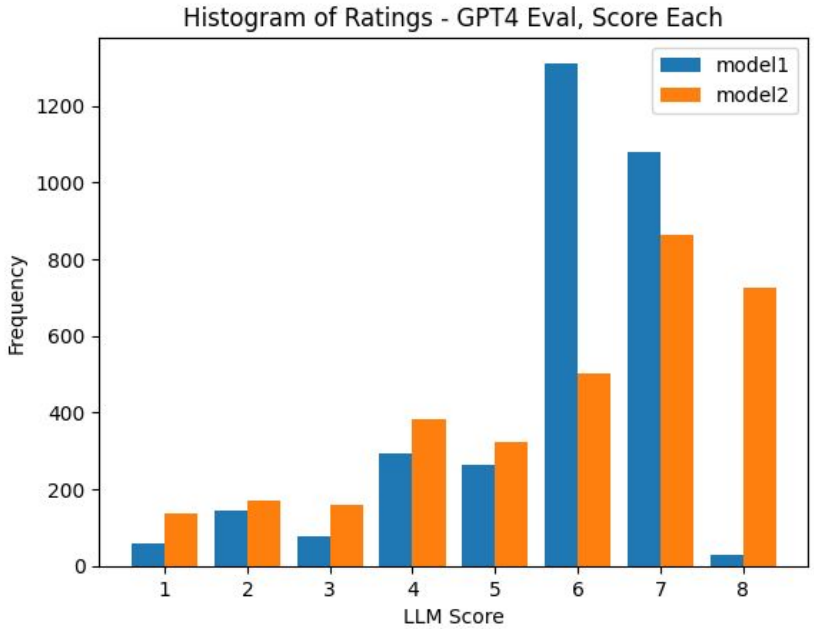
GPT4 as an Evaluator

Prompting GPT4 to make it aware of its “1” bias and asking it to debias results in a flipped bias



GPT4 as an Evaluator

Prompting GPT4 for scoring instead of ranking alleviates the problem

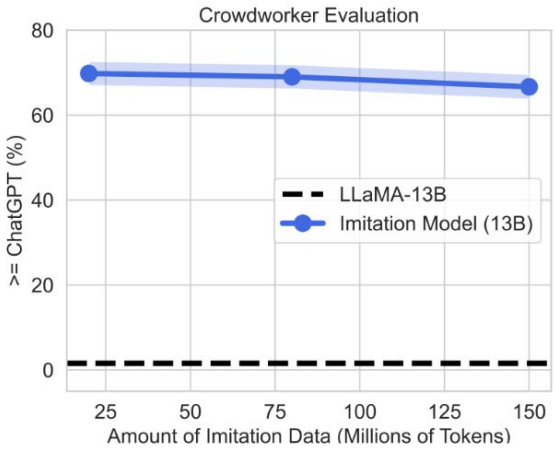


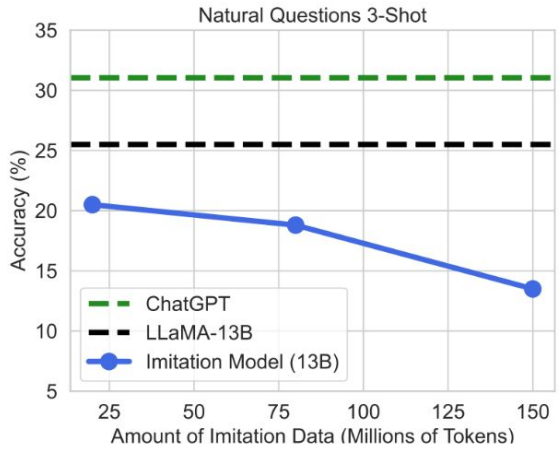
GPT4 as an Evaluator

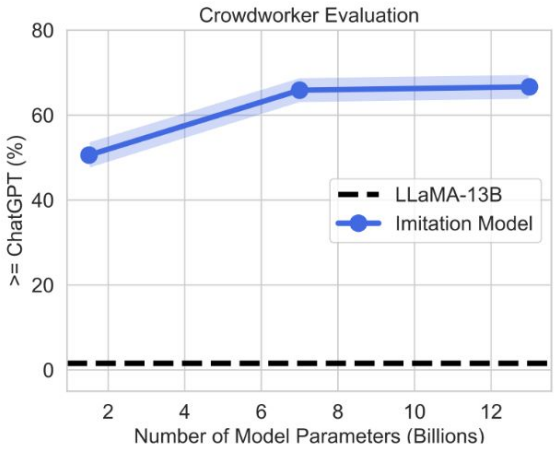
Evidence of *doping* between training and eval

Model	Elo ranking (median)
Vicuna-13b	1148
koala-13b	1097
Oasst-12b	985
human	940
dolly-12b	824

GPT4 as an evaluator



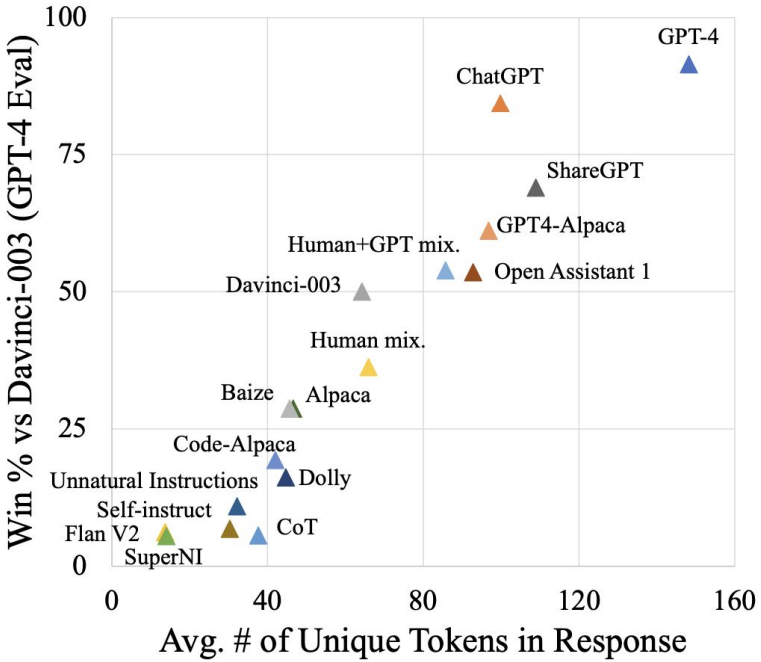




Gudibande et al., ‘23 https://arxiv.org/abs/2305.15717

GPT4 as an evaluator

GPT4 prefers models with higher diversity and length of responses



Wang et al., ‘23 https://arxiv.org/abs/2306.04751
Similar findings by LMSYS https://arxiv.org/abs/2306.05685

GPT4 as an evaluator

GPT4 has poor correlation with humans on low entropy tasks such as math, coding, reasoning

Category	Correlation: GPT-4 to Human Labels
Brainstorm	0.60
Creative generation	0.55
Commonsense reasoning	0.46
Question answering	0.44
Summarization	0.40
Natural language to code	0.33

Similar findings by LMSYS https://arxiv.org/abs/2306.05685

Takeaways

- Open source ML has huge potential impact
- Benchmarking gap in assessing RLHF and model vulnerabilities
- GPT4 eval quirks
  - Prefers models trained on GPT4-like data
  - Left positional bias
  - Higher correlation with humans on creative tasks compared to coding/reasoning tasks

H4 Team

Nathan Lambert

Lewis Tunstall

Edward Beeching

Thomas Wolf

And more at Hugging Face and in the open-source community!

Thanks for listening