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](https://crfm.stanford.edu/helm/v1.0/?models=1)
**Text-to-Text Foundation Models since GPT3**

- **GPT-3**, **GPT-Neo**, **Megatron TNLG**
- **Cohere**, **GPT-NeoX**, **PaLM**
- **OPT**, **Flan-T5**, **Galactica**
- **Flan-UL2**, **StarCoder**, **Falcon**
- **GPT-J**, **Jurassic**, **Gopher**
- **Anthropic**, **Chinchilla**, **BLOOM**
- **UL2**, **ChatGPT**, **LLaMA**
- **GPT-4**, **INCITE**, **LLaMA-2**

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

Open access models

Closed access models
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
Model Access

Open access  Limited access  Closed access
Model Access

Open access

Limited access

Closed access
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
Open Access Large Language Models

Research on policy, governance, AI safety and alignment

Community efforts like Eleuther, Big Science, LAION, OpenAssistant, RedPajama

Papers with several authors

Open source ML has potential for huge impact
Capabilities of machine learning models

Time

Open-source
Closed-source

Pivotal moments
- LLaMA/LLaMA2
- Red Pajama
- Open Assistant
Chatbot LLMs

Alpaca  Vicuna  Dolly  Baize  Koala  StarChat  Open Assistant  OpenChatKit  LLaMA 2 chat  Guanaco
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
Training a Chatbot

Supervised Fine-tuning (instruction following and chatty-ness)

Step 1: Collect demonstration data, and train a supervised policy.
- A prompt is sampled from our prompt dataset.
- A labeler demonstrates the desired output behavior.
- This data is used to fine-tune GPT-3 with supervised learning.

Reinforcement learning with human feedback (RLHF) (aligning to target values and safety)

Step 2: Collect comparison data, and train a reward model.
- A prompt and several model outputs are sampled.
- A labeler ranks the outputs from best to worst.
- This data is used to train our reward model.

Step 3: Optimize a policy against the reward model using reinforcement learning.
- A new prompt is sampled from the dataset.
- The reward model calculates a reward for the output.
- The reward is used to update the policy using PPO.

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

4. Reinforcement Learning from Human Feedback
   a. safety/alignment
   b. nudging the LM towards values you desire
Evaluating a Chatbot

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.

ChatGPT, Bing Chat and the AI ghost in the machine

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

Microsoft’s AI chatbot is going off the rails

Big Tech is heralding chatbots as the next frontier. Why did Microsoft’s start scaring its users?
Evaluating a Chatbot

Step 1
Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.
A labeler demonstrates the desired output behavior.
This data is used to fine-tune GPT-3 with supervised learning.

Evaluating instruction following/chatty-ness

Step 2
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.
A labeler ranks the outputs from best to worst.
This data is used to train our reward model.

Evaluating the RM

Step 3
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.
The policy generates an output.
The reward model calculates a reward for the output.
The reward is used to update the policy using PPO.

Red-teaming

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”
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”
# Leaderboard with Elo ratings (Hugging Face)

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

<table>
<thead>
<tr>
<th>Model</th>
<th>GPT-4 (all)</th>
<th>Human (all)</th>
<th>Human (instruct)</th>
<th>Human (code-instruct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vicuna-13b</td>
<td>1146</td>
<td>1237</td>
<td>1181</td>
<td>1224</td>
</tr>
<tr>
<td>koala-13b</td>
<td>1013</td>
<td>1085</td>
<td>1099</td>
<td>1078</td>
</tr>
<tr>
<td>passt-12b</td>
<td>985</td>
<td>975</td>
<td>968</td>
<td>975</td>
</tr>
<tr>
<td>dolly-12b</td>
<td>854</td>
<td>791</td>
<td>750</td>
<td>721</td>
</tr>
</tbody>
</table>

## Tie allowed*

<table>
<thead>
<tr>
<th>Model</th>
<th>GPT-4 (all)</th>
<th>Human (all)</th>
<th>Human (instruct)</th>
<th>Human (code-instruct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vicuna-13b</td>
<td>1161</td>
<td>1175</td>
<td>1185</td>
<td>1165</td>
</tr>
<tr>
<td>passt-12b</td>
<td>1033</td>
<td>1004</td>
<td>977</td>
<td>1003</td>
</tr>
<tr>
<td>koala-13b</td>
<td>977</td>
<td>1037</td>
<td>1088</td>
<td>1032</td>
</tr>
<tr>
<td>dolly-12b</td>
<td>827</td>
<td>782</td>
<td>749</td>
<td>798</td>
</tr>
</tbody>
</table>

[https://huggingface.co/spaces/HuggingFaceH4/human_eval_llm_leaderboard](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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Elo Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>vicuna-13b</td>
<td>1169</td>
<td>a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS</td>
</tr>
<tr>
<td>2</td>
<td>loptl-13b</td>
<td>1082</td>
<td>a dialogue model for academic research by BAIR</td>
</tr>
<tr>
<td>3</td>
<td>qeost-pythia-12b</td>
<td>1065</td>
<td>an Open Assistant for everyone by LAION</td>
</tr>
<tr>
<td>4</td>
<td>albaca-13b</td>
<td>1008</td>
<td>a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford</td>
</tr>
<tr>
<td>5</td>
<td>chatglm-6b</td>
<td>985</td>
<td>an open bilingual dialogue language model by Tsinghua University</td>
</tr>
<tr>
<td>6</td>
<td>fastchat-t6-3b</td>
<td>951</td>
<td>a chat assistant fine-tuned from FLAN-T5 by LMSYS</td>
</tr>
<tr>
<td>7</td>
<td>dolly-v2-12b</td>
<td>944</td>
<td>an instruction-tuned open large language model by Databricks</td>
</tr>
<tr>
<td>8</td>
<td>llama-13b</td>
<td>932</td>
<td>open and efficient foundation language models by Meta</td>
</tr>
<tr>
<td>9</td>
<td>stablelm-tuned-alpha-7b</td>
<td>858</td>
<td>Stability AI language models</td>
</tr>
</tbody>
</table>

https://lmsys.org/blog/2023-05-03-arena/
MT Bench from LMSYS

MT-Bench Browser

<table>
<thead>
<tr>
<th>Single Answer Grading</th>
<th>Pairwise Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Question</strong></td>
</tr>
<tr>
<td>reasoning</td>
<td>161: Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your current...</td>
</tr>
</tbody>
</table>

Model A

Llama-2-70b-chat

User

Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what’s your current position? Where is the person you just overtook?

Assistant A

Thank you for asking! I’m happy to help you with your question.

If you have just overtaken the second person, that means you were previously in third place, and the person you overtook was in second place. Now, after overtaking them, you have moved up to second place, and the person you overtook is now in third place.

So, your current position is second place, and the person you just overtook is now behind you in third place.

I hope that helps clarify things! Let me know if you have any other questions.

User’s follow-up question

If the “second person” is changed to “last person” in the above question, what would the answer be?

Assistant A

Sure, I’d be happy to help! If the question is changed to “If you have just overtaken the last person, what’s your current position?”, the answer would be that you are now in first place.

Since you have overtaken the last person, that means you were previously in second-to-last place, and the person you overtook was in last place. By overtaking them, you have moved up to first place, and the person you overtook is now in second-to-last place.

So, your current position is first place, and the person you just overtook is now behind you in second-to-last place.

Reference Solution

Q1. You are in second place. Q2. Uncertain.

https://huggingface.co/spaces/lmsys/mt-bench
**LMSYS Leaderboard**

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Arena Elo rating</th>
<th>MT-bench (score)</th>
<th>MMLU</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4</td>
<td>1206</td>
<td>8.99</td>
<td>86.4</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Claude:1</td>
<td>1166</td>
<td>7.9</td>
<td>77</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Claude:instant:1</td>
<td>1138</td>
<td>7.85</td>
<td>73.4</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Claude:2</td>
<td>1135</td>
<td>8.06</td>
<td>78.5</td>
<td>Proprietary</td>
</tr>
<tr>
<td>GPT-3.5:terabits</td>
<td>1122</td>
<td>7.94</td>
<td>70</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Vicuna:33B</td>
<td>1096</td>
<td>7.12</td>
<td>59.2</td>
<td>Non-commercial</td>
</tr>
<tr>
<td>Vicuna:13B</td>
<td>1051</td>
<td>6.57</td>
<td>55.8</td>
<td>Llama 2 Community</td>
</tr>
<tr>
<td>MPT-30B:chat</td>
<td>1046</td>
<td>6.39</td>
<td>50.4</td>
<td>CC-BY-NC-SA-4.0</td>
</tr>
<tr>
<td>WizardLM:138-v1.1</td>
<td>1040</td>
<td>6.76</td>
<td>50</td>
<td>Non-commercial</td>
</tr>
<tr>
<td>Guanaco:33B</td>
<td>1038</td>
<td>6.53</td>
<td>57.6</td>
<td>Non-commercial</td>
</tr>
</tbody>
</table>

Code: The Arena Elo ratings are computed by this [notebook](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard). The MT-bench scores (single-answer grading on a scale of 10) are computed by [fastchat](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard). The MMLU scores are computed by [InstructEval](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard) and [Chain-of-Thought Hub](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard). Higher values are better for all benchmarks. Empty cells mean not available.

For more information, visit [https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard](https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard).
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”
Benchmarking RM Models

H4 Internal Leaderboard

Evaluation of H4 models across a diverse range of benchmarks.

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 osas1 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Revision</th>
<th>Dtype</th>
<th>Average</th>
<th>Anthropic Helpful</th>
<th>OpenAssistant</th>
<th>SHP</th>
<th>Learn to Summarize</th>
</tr>
</thead>
<tbody>
<tr>
<td>falcon-49b:rm</td>
<td>v1.0.4bit</td>
<td>4bit</td>
<td>0.721</td>
<td>0.66</td>
<td>0.678</td>
<td>0.802</td>
<td>0.743</td>
</tr>
<tr>
<td>falcon-49b:rm</td>
<td>v2.0.4bit</td>
<td>4bit</td>
<td>0.717</td>
<td>0.648</td>
<td>0.701</td>
<td>0.781</td>
<td>0.738</td>
</tr>
<tr>
<td>falcon-49b:rm</td>
<td>v2.1.4bit</td>
<td>4bit</td>
<td>0.708</td>
<td>0.64</td>
<td>0.687</td>
<td>0.78</td>
<td>0.723</td>
</tr>
<tr>
<td>falcon-49b:rm</td>
<td>v2.2.4bit</td>
<td>4bit</td>
<td>0.706</td>
<td>0.64</td>
<td>0.672</td>
<td>0.781</td>
<td>0.733</td>
</tr>
<tr>
<td>falcon-7b:rm</td>
<td>v2.3.4bit</td>
<td>4bit</td>
<td>0.705</td>
<td>0.649</td>
<td>0.676</td>
<td>0.789</td>
<td>0.797</td>
</tr>
<tr>
<td>falcon-7b:rm</td>
<td>v2.2.4bit</td>
<td>4bit</td>
<td>0.704</td>
<td>0.649</td>
<td>0.706</td>
<td>0.765</td>
<td>0.694</td>
</tr>
<tr>
<td>falcon-7b:rm</td>
<td>v4.3.8bit</td>
<td>8bit</td>
<td>0.679</td>
<td>0.634</td>
<td>0.611</td>
<td>0.768</td>
<td>0.793</td>
</tr>
<tr>
<td>falcon-7b:rm</td>
<td>v2.1.4bit</td>
<td>4bit</td>
<td>0.675</td>
<td>0.648</td>
<td>0.561</td>
<td>0.786</td>
<td>0.706</td>
</tr>
</tbody>
</table>
GPT4 as an evaluator
Takeaways

1.
## Further Reading

### Red-Teaming
[https://huggingface.co/blog/red-teaming](https://huggingface.co/blog/red-teaming)

### RLHF
[https://huggingface.co/blog/rlhf](https://huggingface.co/blog/rlhf)

### Dialog agents
[https://huggingface.co/blog/dialog-agents](https://huggingface.co/blog/dialog-agents)

<table>
<thead>
<tr>
<th></th>
<th>LaMDA</th>
<th>BlenderBot 3</th>
<th>Sparrow</th>
<th>ChatGPT/InstructGPT</th>
<th>Assistant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Org</strong></td>
<td>Google</td>
<td>Meta</td>
<td>DeepMind</td>
<td>OpenAI</td>
<td>Anthropic</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Closed</td>
<td>Open</td>
<td>Closed</td>
<td>Limited</td>
<td>Closed</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>137B</td>
<td>175B</td>
<td>70B</td>
<td>175B</td>
<td>52B</td>
</tr>
<tr>
<td><strong>Pre-trained Base model</strong></td>
<td>Unknown</td>
<td>OPT</td>
<td>Chinchilla</td>
<td>GPT-3.5</td>
<td>Unknown</td>
</tr>
<tr>
<td><strong>Pre-training corpora size (# tokens)</strong></td>
<td>2.81T</td>
<td>180B</td>
<td>1.4T</td>
<td>Unknown</td>
<td>400B</td>
</tr>
<tr>
<td><strong>Model can access the web</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>Supervised fine-tuning</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>Fine-tuning data size</strong></td>
<td>Quality: 6.4K Safety: 8K Groundedness: 4K IR: 49K</td>
<td>20 NLP datasets ranging from 18K to 1.2M</td>
<td>Unknown</td>
<td>12.7K (for InstructGPT, likely much more for ChatGPT)</td>
<td>150K + LM generated data</td>
</tr>
<tr>
<td><strong>RLHF</strong></td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
H4 Team

Nathan Lambert       Lewis Tunstall       Edward Beeching       Thomas Wolf

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