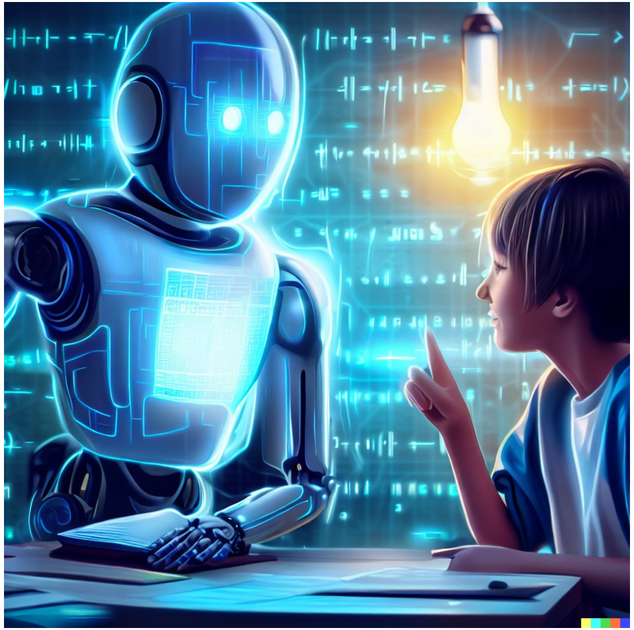
Are LLM-based Chatbots Reliable and Safe Enough to Provide Free Data Science Tutoring for Marginalized Communities?

DATASCI 207 Summer 2023
Shin, Fong, Lister, Hodges



Berkeley
SCHOOL OF
INFORMATION

Tutoring Marginalized Communities Background

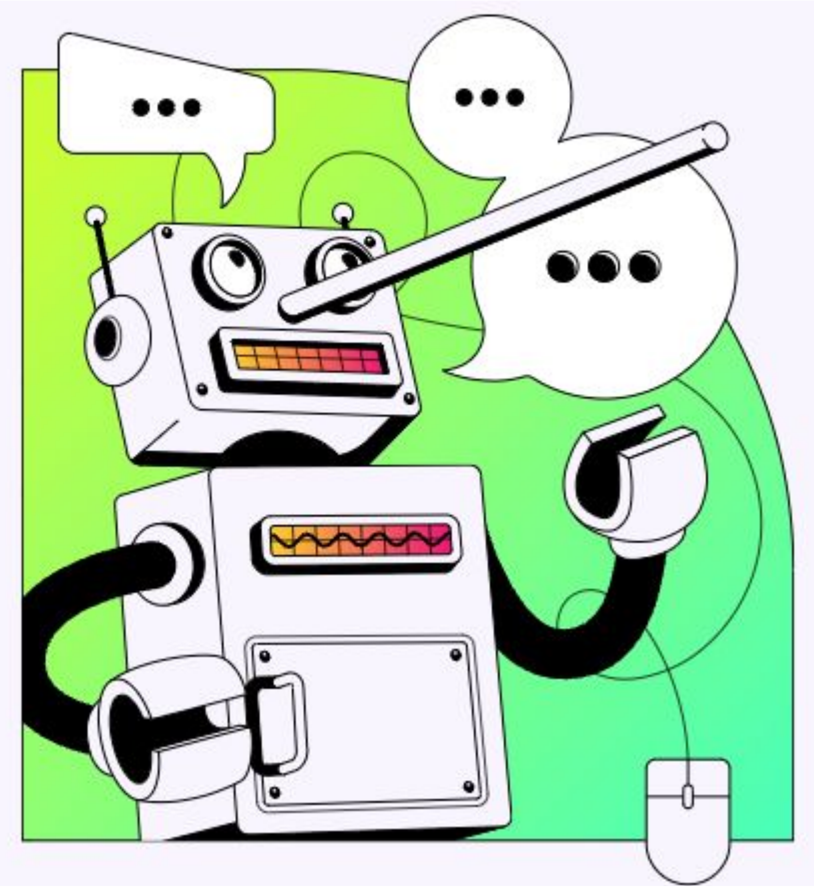
- Address social disparities: provide equal access to technical knowledge
- Economic opportunities: equip students with high demand skills
- Foster innovation and problem-solving: encourage diverse contributions to the field
- Empowerment and inclusion: empower students with knowledge and skills to actively participate in data science
- Community empowerment and development: enable students to take active role in solving local community challenges

LLMs Background / Overview

- LLMs - a seismic shift
- Companies are producing both closed- and open-source models:
    - ChatGPT, LLaMA, Falcon, Pythia, StableLM, and more
- Multiple layers of neural networks
- Generating text one token at a time, predicting the next based on current sequence

Hallucinations

"Hallucinations are a model's logical mistakes or a tendency to invent facts in moments of uncertainty" - OpenAI, May 2023



Instruction-Tuned LLMs to Evaluate

ChatGPT

Vicuna

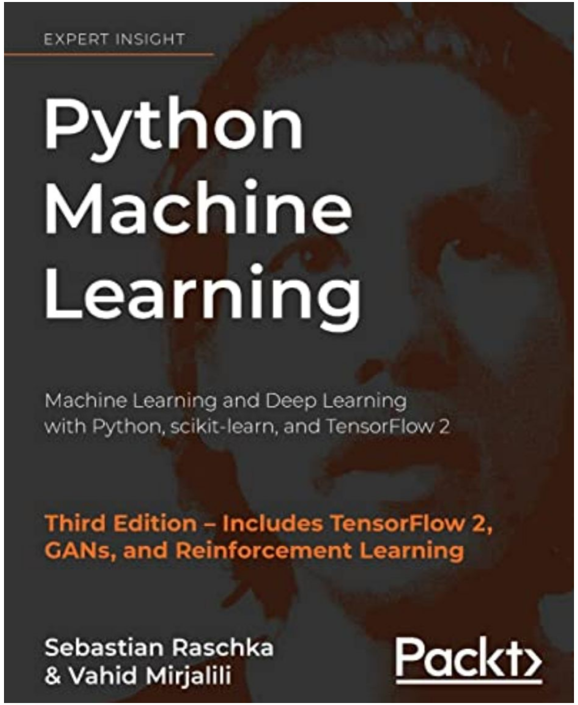
TII Falcon

Stanford Alpaca

Retrieval-Augmented Generation (RAG) and Prompt Engineering vs Fine-Tuning

- Can better minimize hallucinations and other undesirable responses in conversational applications
- RAG and prompt engineering are orders of magnitude cheaper than fine-tuning LLM layers
- New content can be made available in near real-time, compared to fine-tuning LLM layers, which will require multiple hours at a minimum
- Allows for enriching LLM’s limited context window from unlimited semantically relevant content for the use case at hand
- Restricted content can be filtered at runtime

Data Science Textbooks Dataset



EXPERT INSIGHT

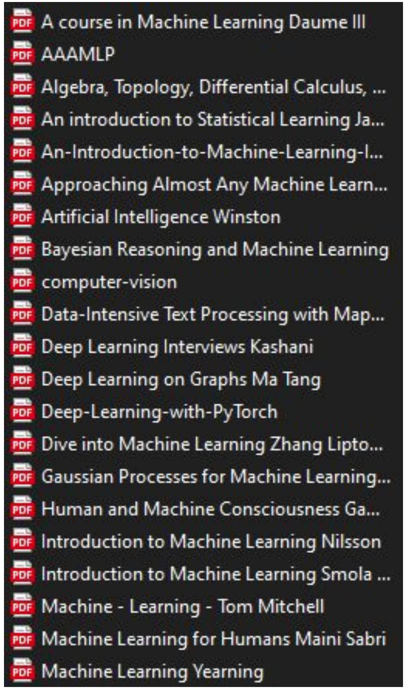
Python Machine Learning

Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2

Third Edition – Includes TensorFlow 2, GANs, and Reinforcement Learning

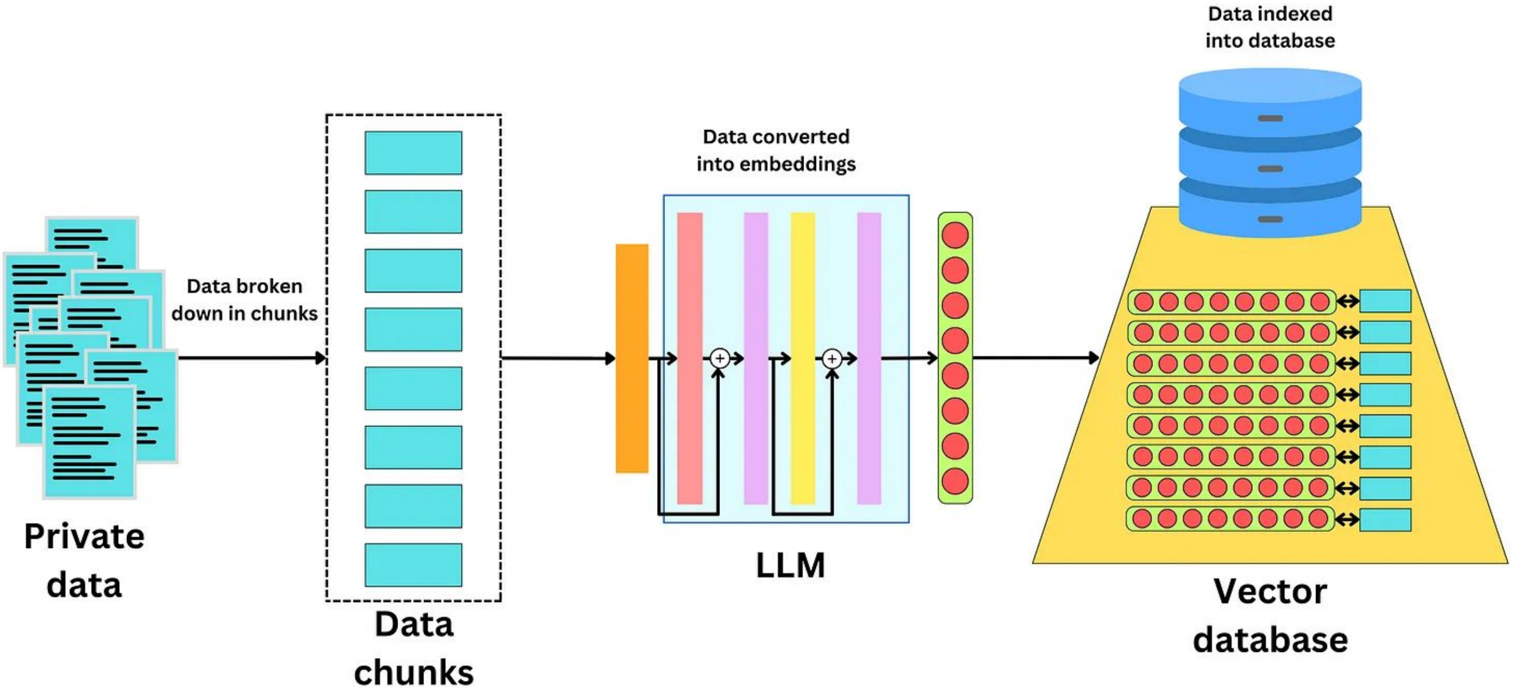
Sebastian Raschka & Vahid Mirjalili

Packt>



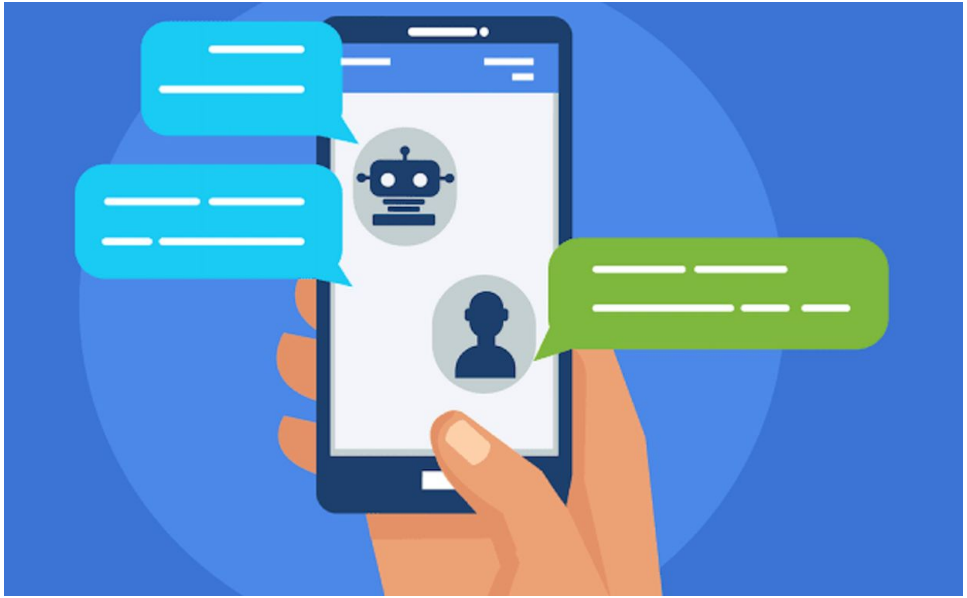
- A course in Machine Learning Daume III
- AAAMLP
- Algebra, Topology, Differential Calculus, ...
- An introduction to Statistical Learning Ja...
- An-Introduction-to-Machine-Learning-I...
- Approaching Almost Any Machine Learn...
- Artificial Intelligence Winston
- Bayesian Reasoning and Machine Learning
- computer-vision
- Data-Intensive Text Processing with Map...
- Deep Learning Interviews Kashani
- Deep Learning on Graphs Ma Tang
- Deep-Learning-with-PyTorch
- Dive into Machine Learning Zhang Lipto...
- Gaussian Processes for Machine Learning...
- Human and Machine Consciousness Ga...
- Introduction to Machine Learning Nilsson
- Introduction to Machine Learning Smola ...
- Machine - Learning - Tom Mitchell
- Machine Learning for Humans Maini Sabri
- Machine Learning Yearning

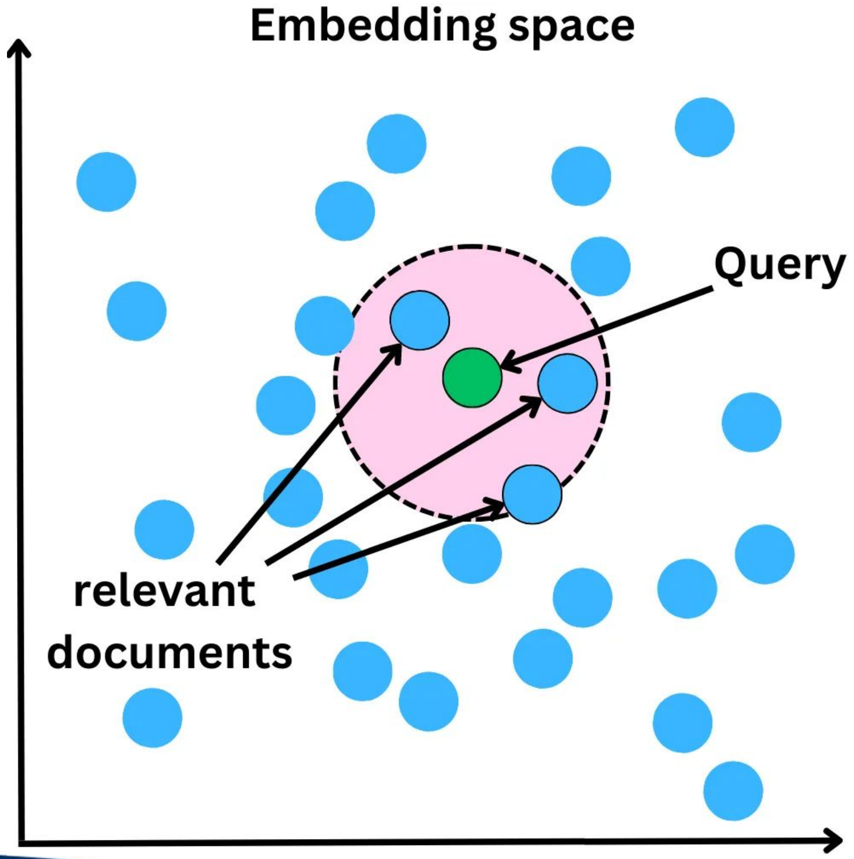
Indexing Textbooks as Embeddings



Data indexed into database
Data converted into embeddings
Data broken down in chunks
Private data
Data chunks
LLM
Vector database

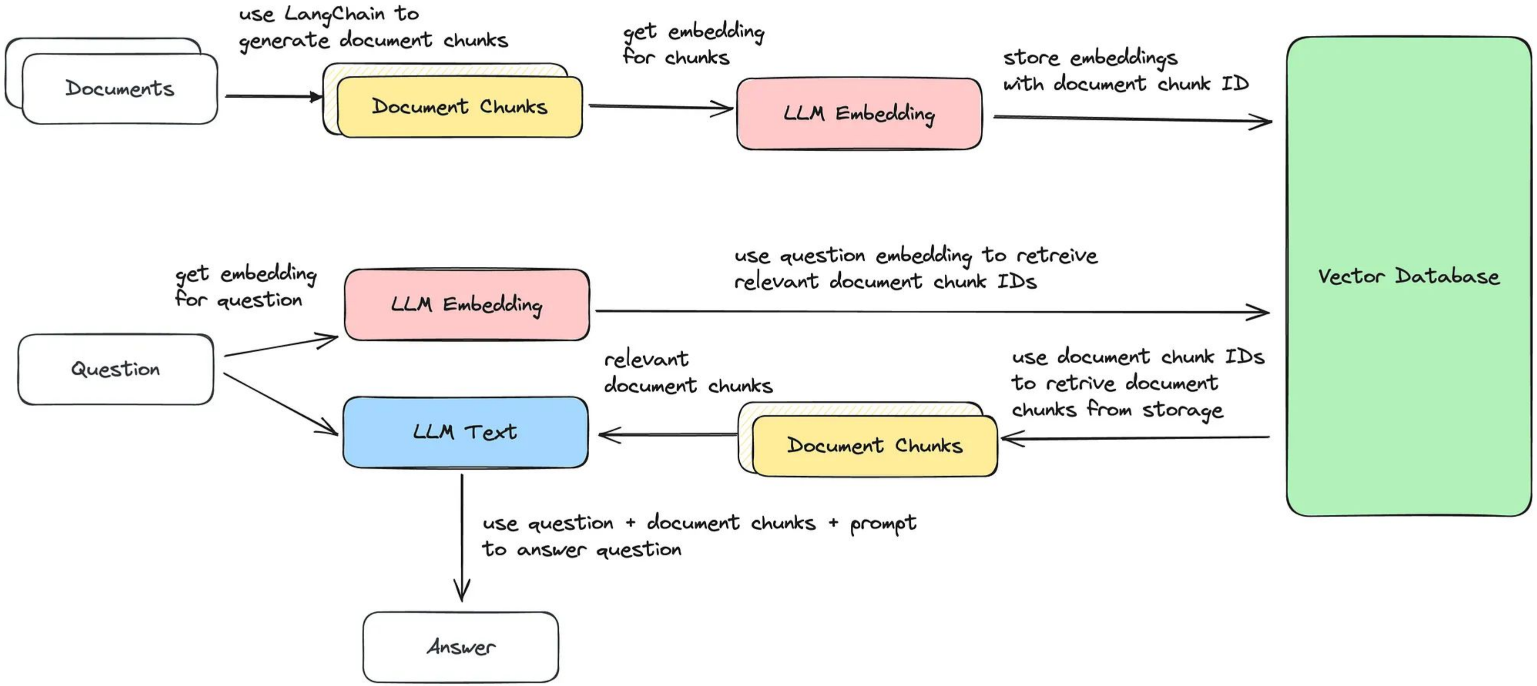
Providing Relevant Context



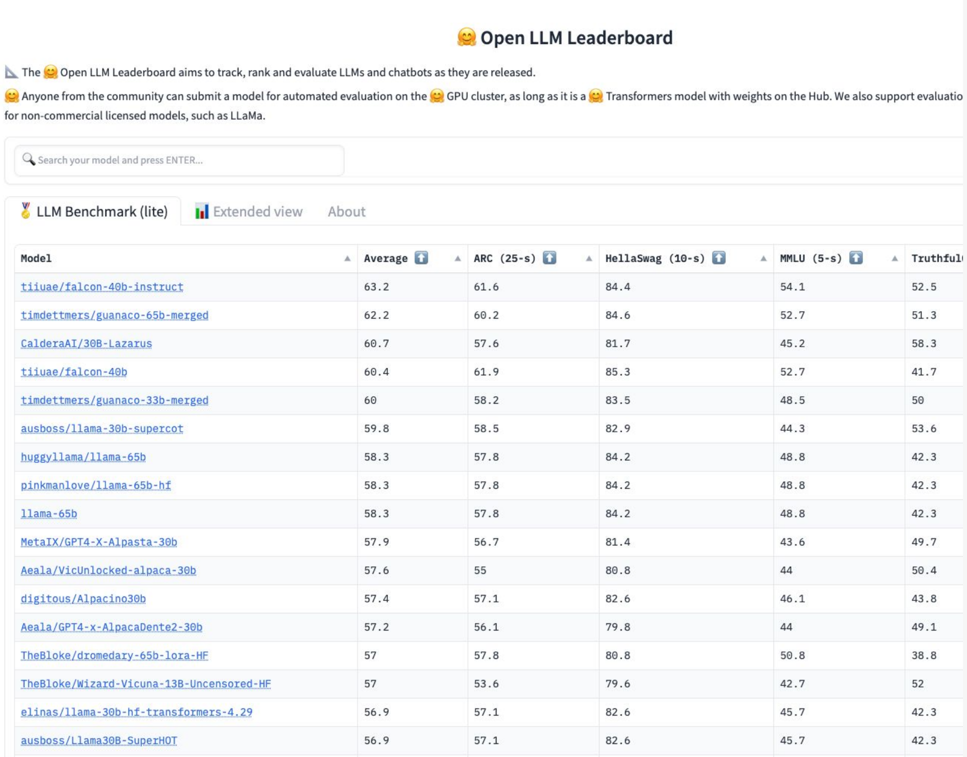


Berkeley SCHOOL OF INFORMATION

Retrieval-Augmented Generation (RAG) Flow


Documents
use LangChain to generate document chunks
Document Chunks
get embedding for chunks
LLM Embedding
store embeddings with document chunk ID
Vector Database
get embedding for question
LLM Embedding
use question embedding to retreive relevant document chunk IDs
Question
relevant document chunks
use document chunk IDs to retrive document chunks from storage
LLM Text
Document Chunks
use question + document chunks + prompt to answer question
Answer

LLM Evaluation Techniques

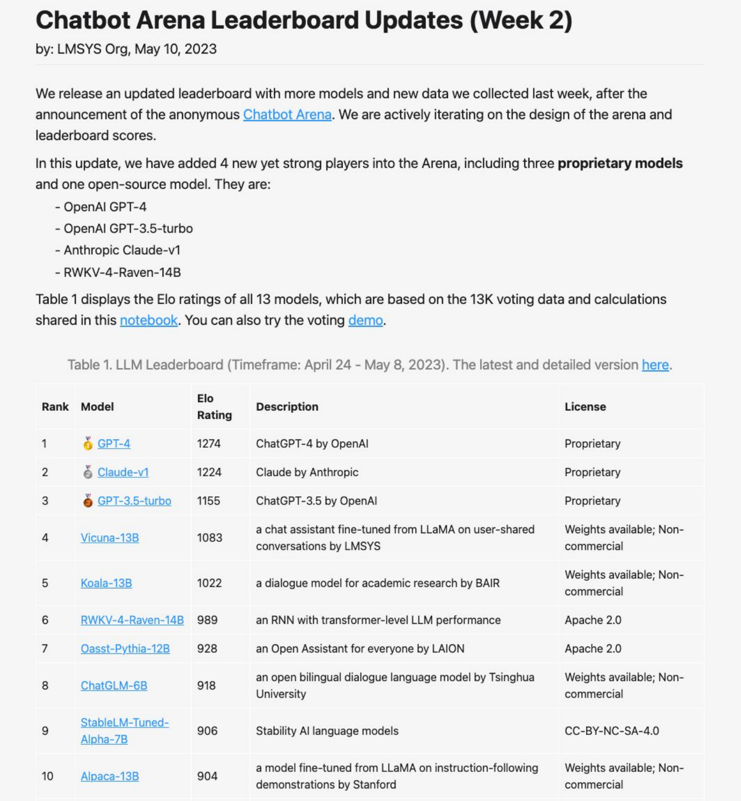


🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

🤗 Anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluatio
for non-commercial licensed models, such as LLaMa.

Model ▲	Average ⬆ ▲	ARC (25-s) ⬆ ▲	HellaSwag (10-s) ⬆ ▲	MMLU (5-s) ⬆ ▲	Truthful
tiiuae/falcon-40b-instruct	63.2	61.6	84.4	54.1	52.5
timdettmers/guanaco-65b-merged	62.2	60.2	84.6	52.7	51.3
CalderaAI/30B-Lazarus	60.7	57.6	81.7	45.2	58.3
tiiuae/falcon-40b	60.4	61.9	85.3	52.7	41.7
timdettmers/guanaco-33b-merged	60	58.2	83.5	48.5	50
ausboss/llama-30b-supercot	59.8	58.5	82.9	44.3	53.6
huggyllama/llama-65b	58.3	57.8	84.2	48.8	42.3
pinkmanlove/llama-65b-hf	58.3	57.8	84.2	48.8	42.3
llama-65b	58.3	57.8	84.2	48.8	42.3
MetaIX/GPT4-X-Alpasta-30b	57.9	56.7	81.4	43.6	49.7
Aeala/VicUnlocked-alpaca-30b	57.6	55	80.8	44	50.4
digitous/Alpacino30b	57.4	57.1	82.6	46.1	43.8
Aeala/GPT4-x-AlpacaDente2-30b	57.2	56.1	79.8	44	49.1
TheBloke/dromedary-65b-lora-HF	57	57.8	80.8	50.8	38.8
TheBloke/Wizard-Vicuna-13B-Uncensored-HF	57	53.6	79.6	42.7	52
elinas/llama-30b-hf-transformers-4.29	56.9	57.1	82.6	45.7	42.3
ausboss/Llama30B-SuperHOT	56.9	57.1	82.6	45.7	42.3



Chatbot Arena Leaderboard Updates (Week 2)

by: LMSYS Org, May 10, 2023

We release an updated leaderboard with more models and new data we collected last week, after the announcement of the anonymous Chatbot Arena. We are actively iterating on the design of the arena and leaderboard scores.

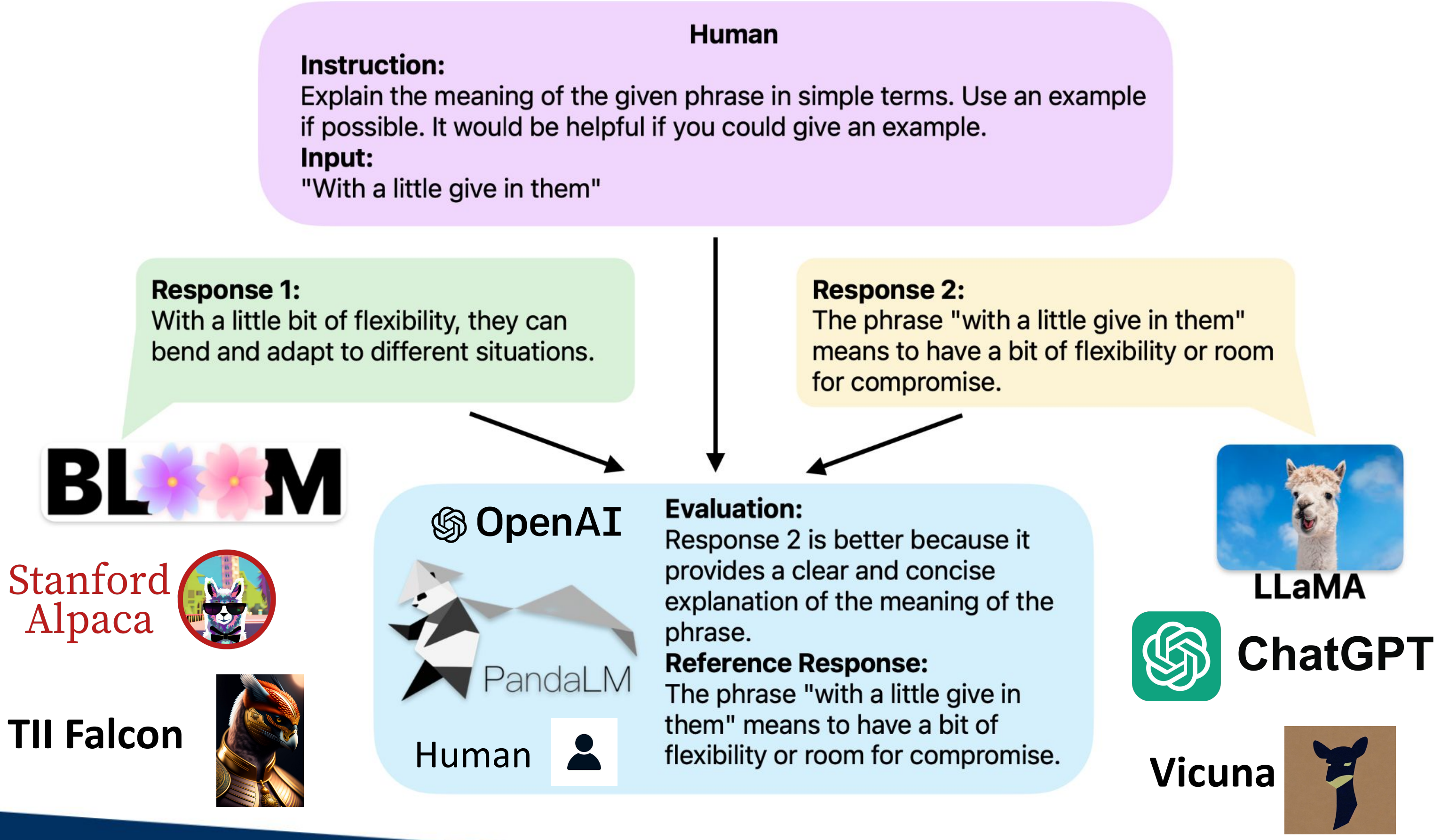
In this update, we have added 4 new yet strong players into the Arena, including three **proprietary models** and one open-source model. They are:

- OpenAI GPT-4
- OpenAI GPT-3.5-turbo
- Anthropic Claude-v1
- RWKV-4-Raven-14B

Table 1 displays the Elo ratings of all 13 models, which are based on the 13K voting data and calculations shared in this notebook. You can also try the voting demo.

Table 1. LLM Leaderboard (Timeframe: April 24 - May 8, 2023). The latest and detailed version here.

Rank	Model	Elo Rating	Description	License
1	🥇 GPT-4	1274	ChatGPT-4 by OpenAI	Proprietary
2	🥈 Claude-v1	1224	Claude by Anthropic	Proprietary
3	🥉 GPT-3.5-turbo	1155	ChatGPT-3.5 by OpenAI	Proprietary
4	Vicuna-13B	1083	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	Weights available; Non-commercial
5	Koala-13B	1022	a dialogue model for academic research by BAIR	Weights available; Non-commercial
6	RWKV-4-Raven-14B	989	an RNN with transformer-level LLM performance	Apache 2.0
7	Oasst-Pythia-12B	928	an Open Assistant for everyone by LAION	Apache 2.0
8	ChatGLM-6B	918	an open bilingual dialogue language model by Tsinghua University	Weights available; Non-commercial
9	StableLM-Tuned-Alpha-7B	906	Stability AI language models	CC-BY-NC-SA-4.0
10	Alpaca-13B	904	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford	Weights available; Non-commercial



Human

Instruction:
Explain the meaning of the given phrase in simple terms. Use an example if possible. It would be helpful if you could give an example.

Input:
"With a little give in them"

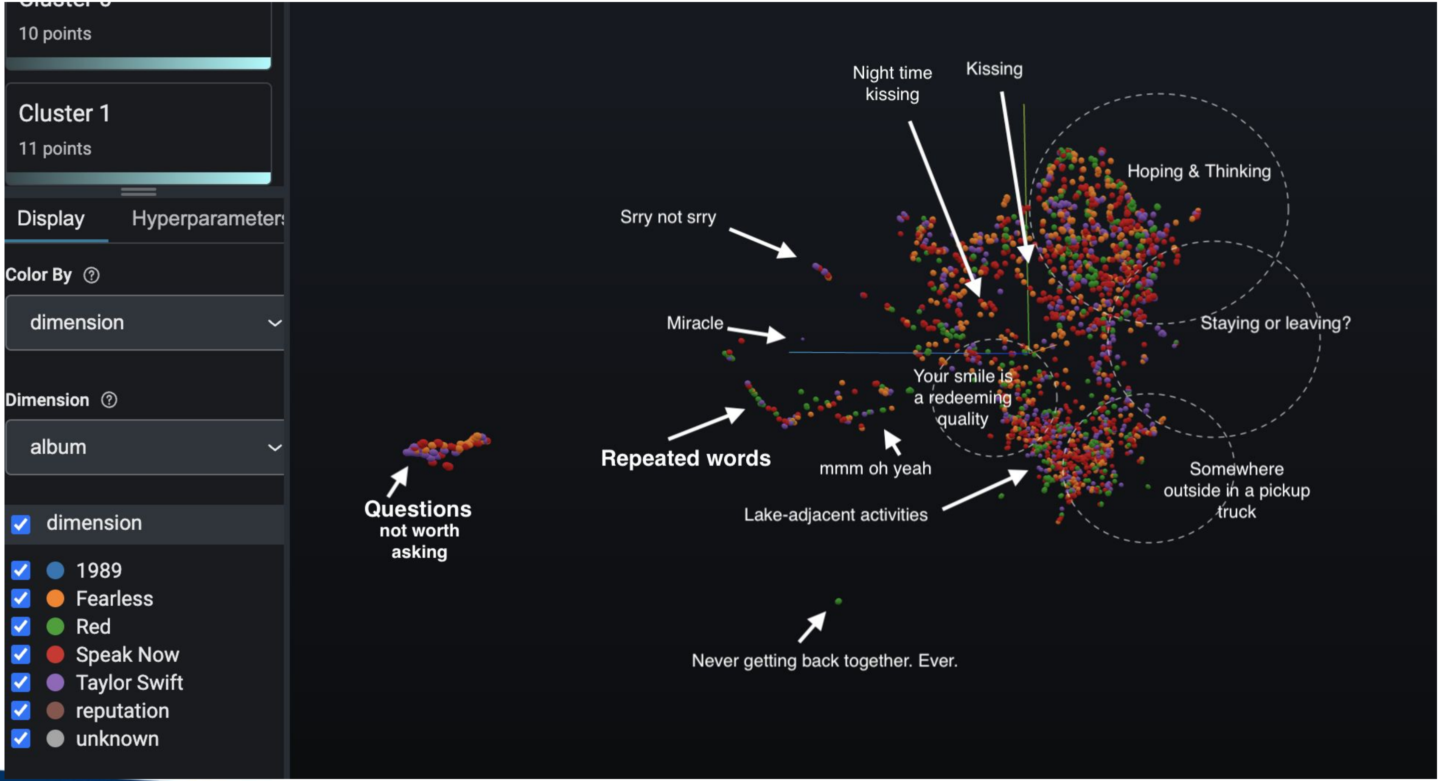
Response 1:
With a little bit of flexibility, they can bend and adapt to different situations.

Response 2:
The phrase "with a little give in them" means to have a bit of flexibility or room for compromise.

Evaluation:
Response 2 is better because it provides a clear and concise explanation of the meaning of the phrase.

Reference Response:
The phrase "with a little give in them" means to have a bit of flexibility or room for compromise.

Phoenix to Analyze 4862 Rows of Taylor Swift Lyrics and Explore Trends



Cluster 0

10 points

Cluster 1

11 points

Display | Hyperparameters

Color By

dimension

Dimension

album

☑ dimension

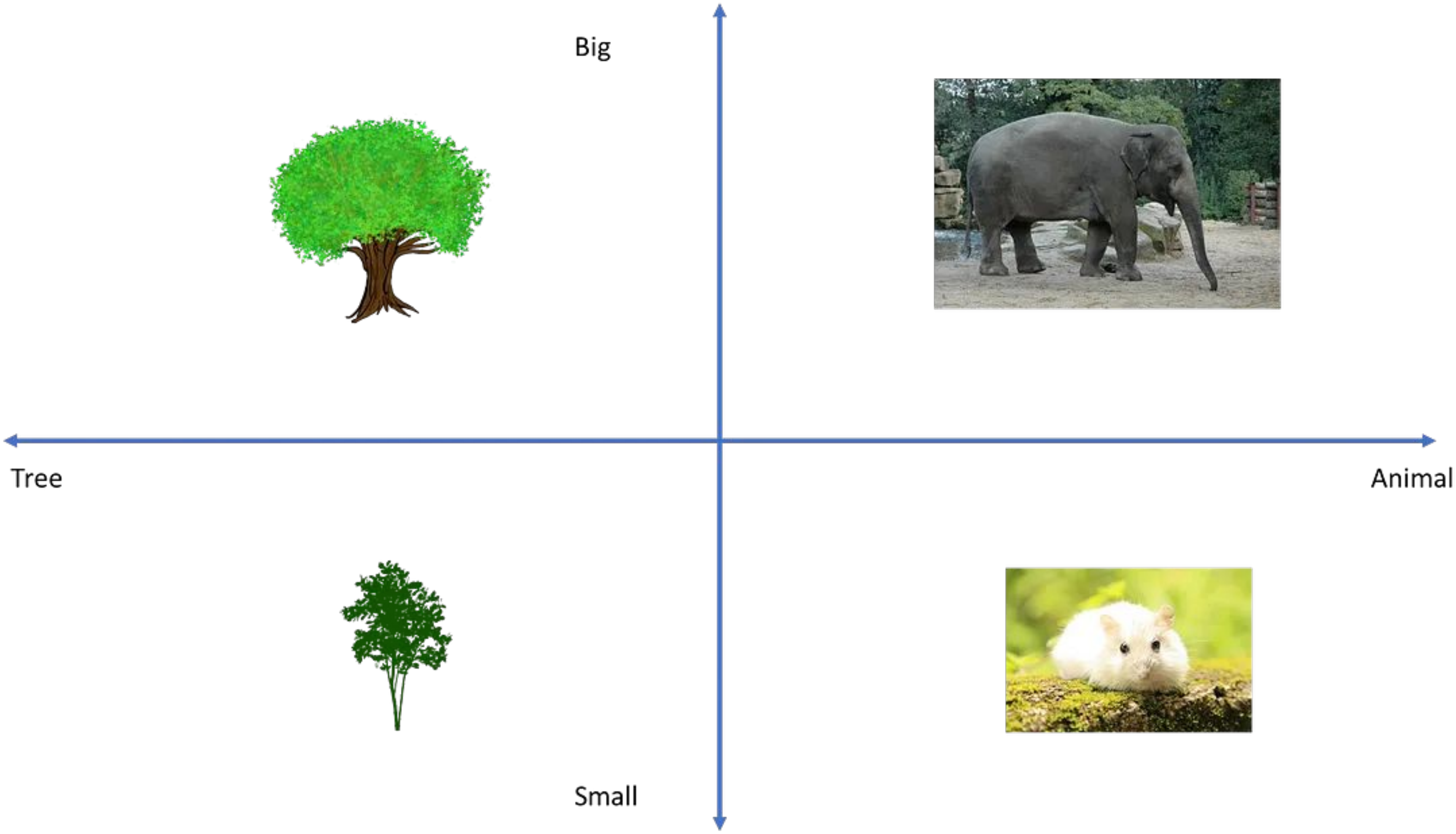
- ☑ ● 1989
- ☑ ● Fearless
- ☑ ● Red
- ☑ ● Speak Now
- ☑ ● Taylor Swift
- ☑ ● reputation
- ☑ ● unknown

Image by Krystal Kirkland created using Arize Phoenix

References

- Judging LLM-as-a-judge with MT-Bench and Chatbot Arena by Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica, June 9, 2023
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks by Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, May 22, 2020
- Textbooks Are All You Need by Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, Yuanzhi Li, June 20, 2023
- PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization by Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, Yue Zhang, June 8, 2023
- Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors by Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, Gustavo Soares, June 29, 2023
- Phoenix: ML Observability in a Notebook by Arize AI
- Open LLM Leaderboard by HuggingFace
- Chatbot Arena Leaderboard Updates by UC Berkeley, UCSD, CMU, MBZUAI
- Deep Dive: How to Build a Smart Chatbot in 10 mins with LangChain by Damien Benveniste, May 25, 2023
- What is Grounded Generation? by Vectara

Appendix

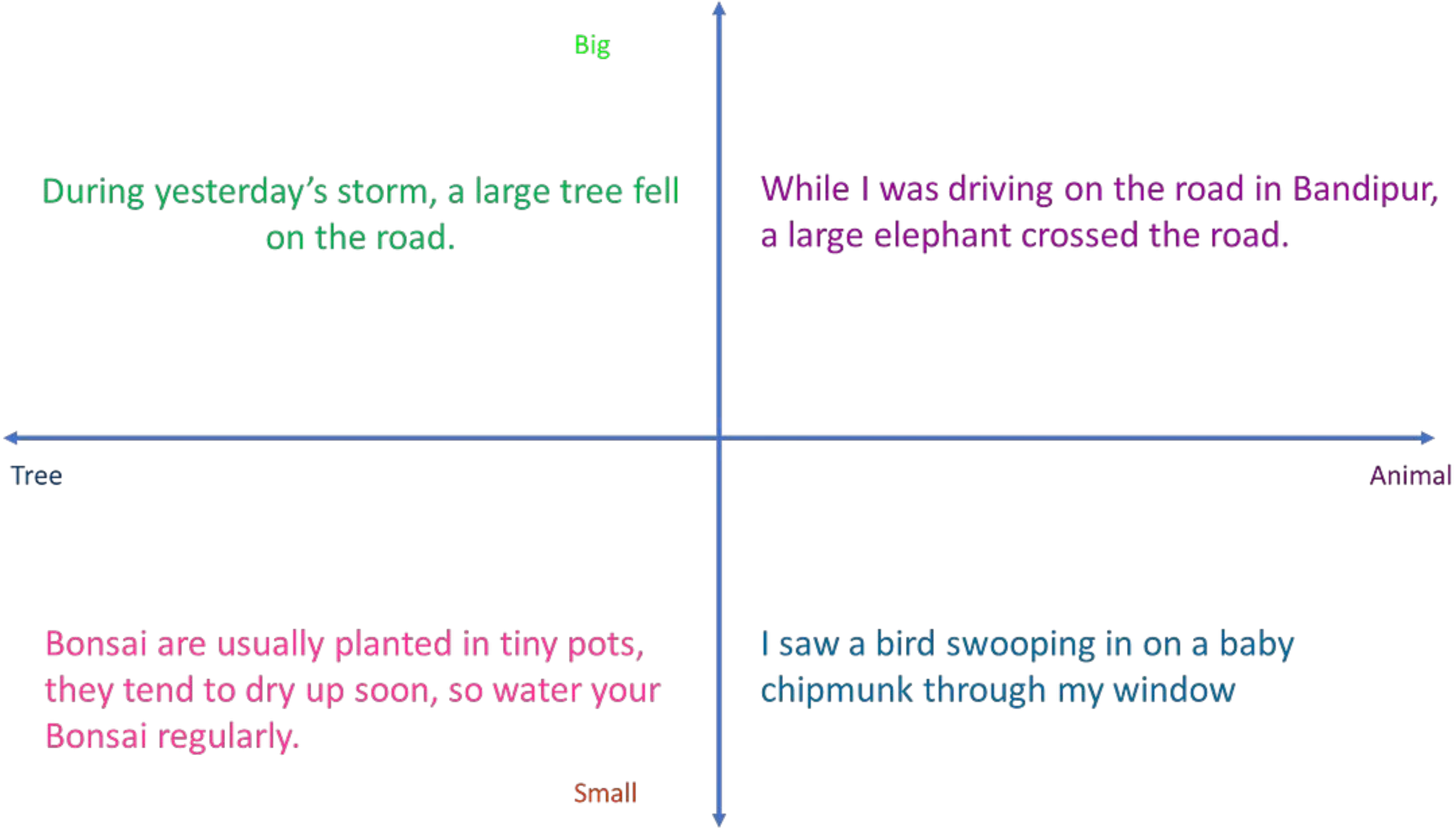


Big

Tree

Animal

Small

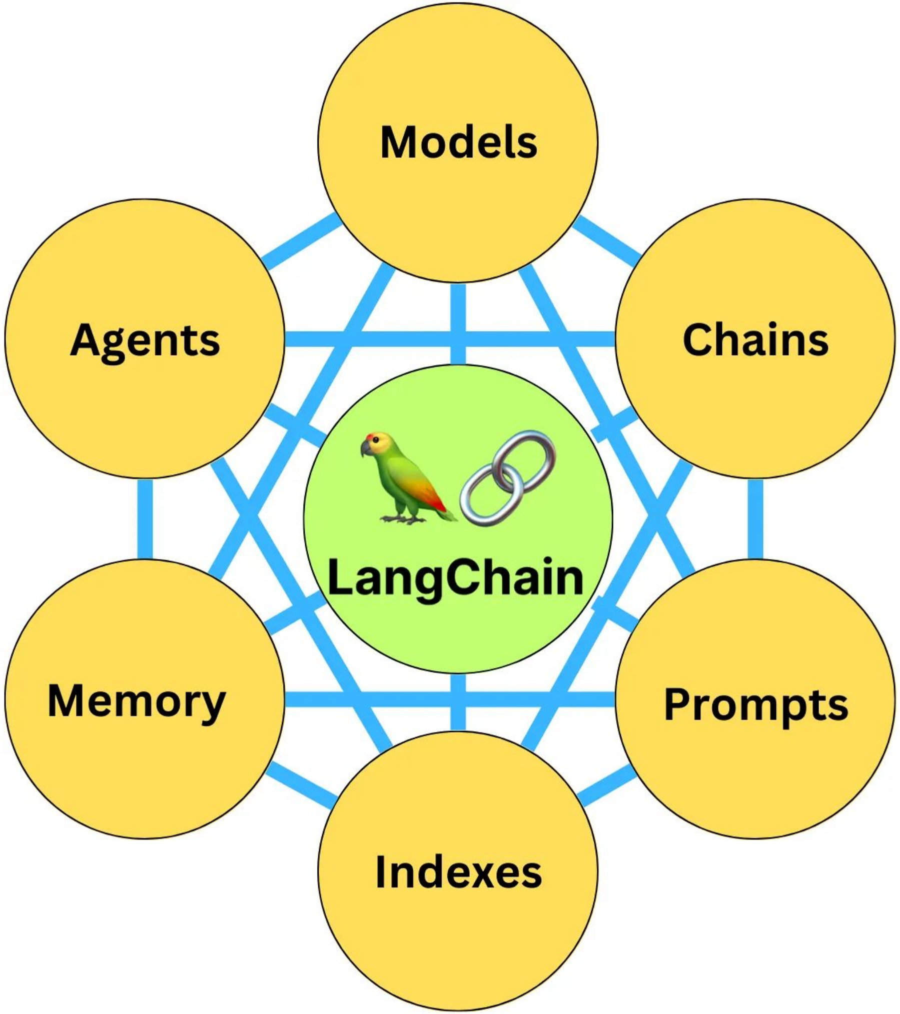


During yesterday's storm, a large tree fell on the road.

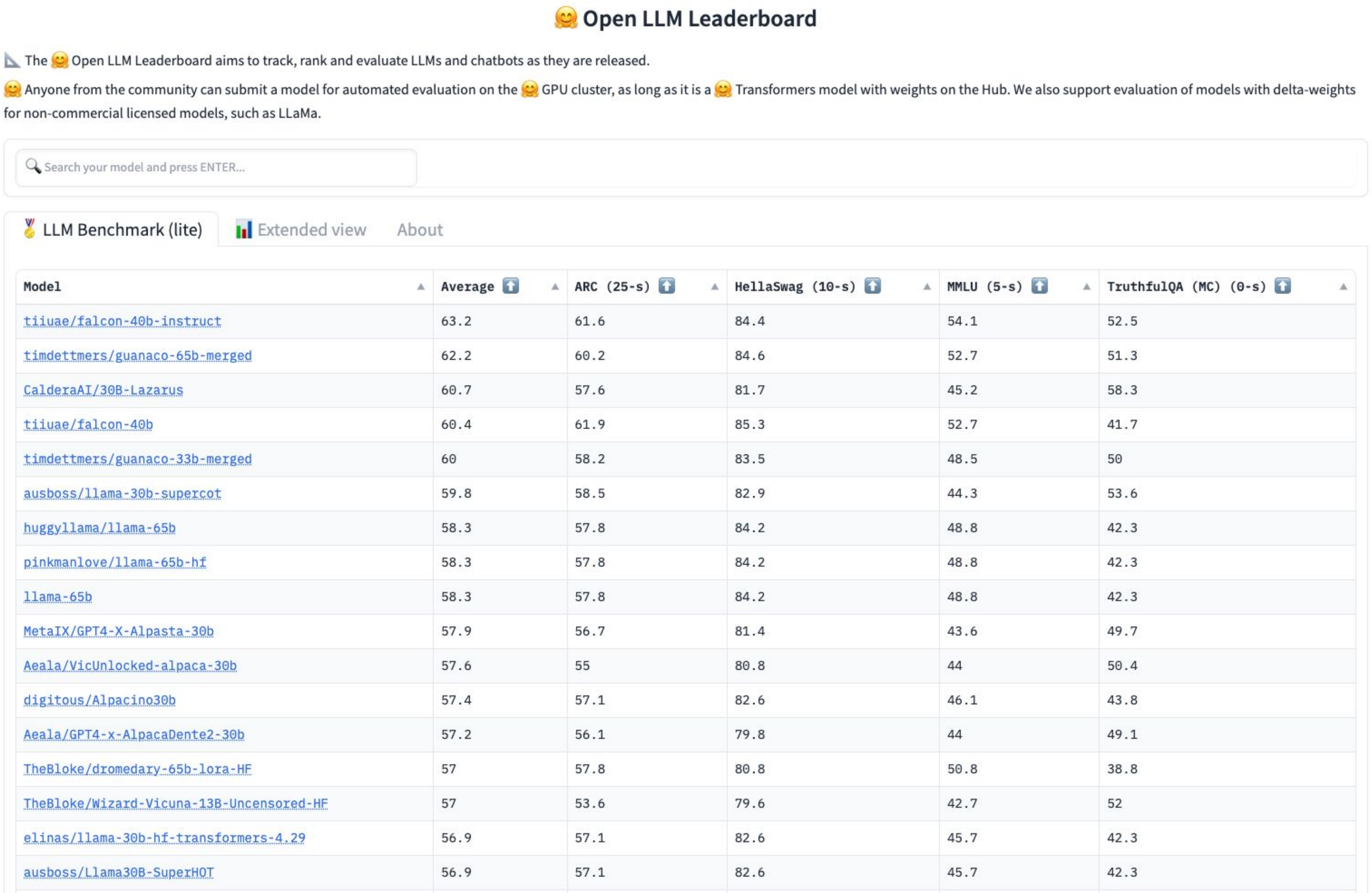
While I was driving on the road in Bandipur, a large elephant crossed the road.

Bonsai are usually planted in tiny pots, they tend to dry up soon, so water your Bonsai regularly.

I saw a bird swooping in on a baby chipmunk through my window



LangChain is a package to build applications using LLMs. It is composed of 6 modules:



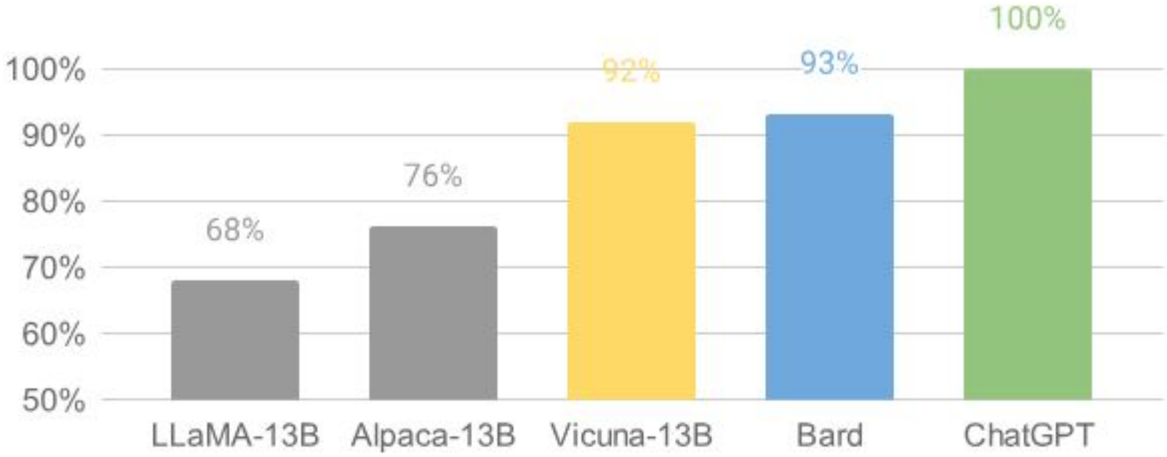
🤗 Open LLM Leaderboard

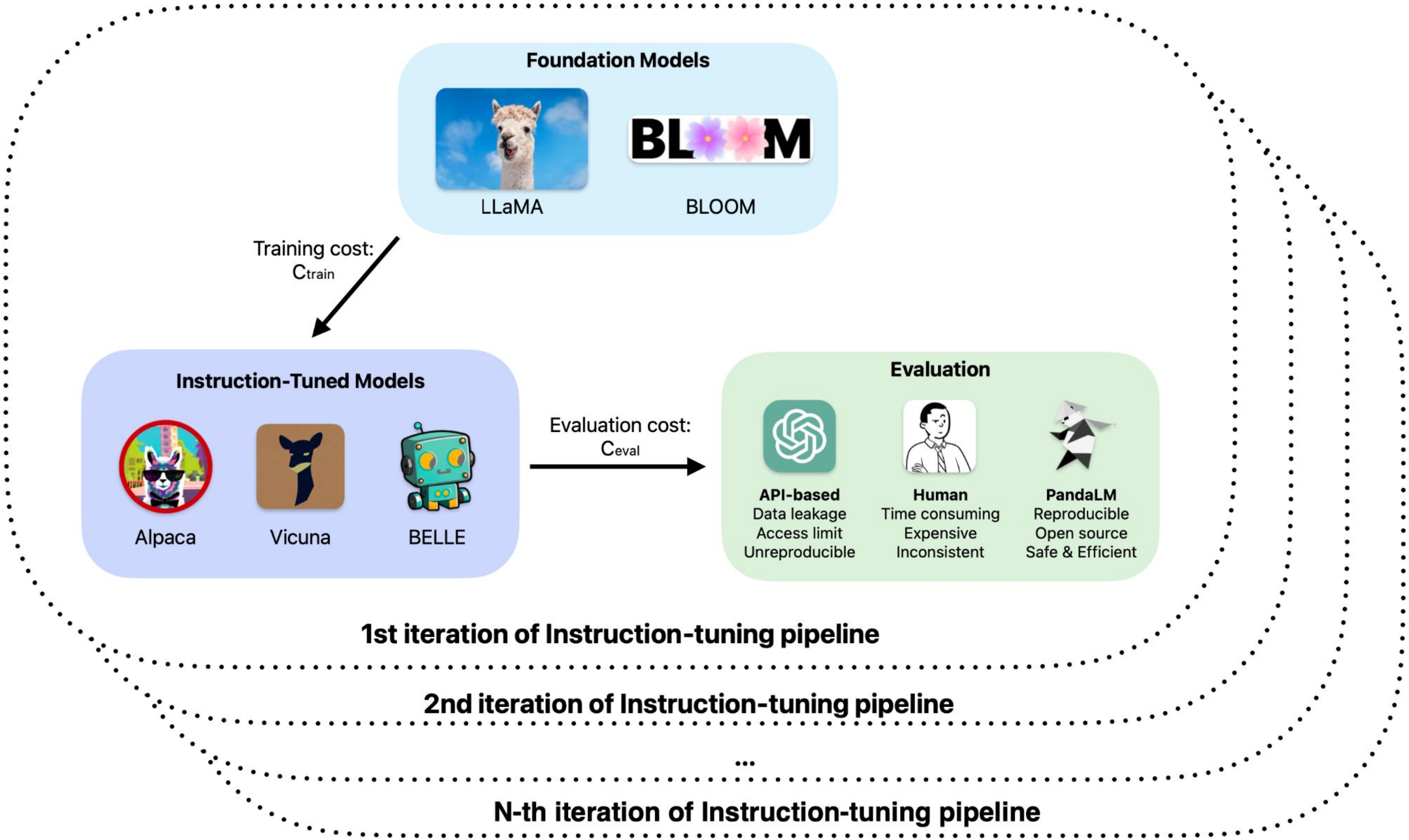
📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

🤗 Anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMa.

Model ▲	Average ⬆️ ▲	ARC (25-s) ⬆️ ▲	HellaSwag (10-s) ⬆️ ▲	MMLU (5-s) ⬆️ ▲	TruthfulQA (MC) (0-s) ⬆️ ▲
tiiuae/falcon-40b-instruct	63.2	61.6	84.4	54.1	52.5
timdettmers/guanaco-65b-merged	62.2	60.2	84.6	52.7	51.3
CalderaAI/30B-Lazarus	60.7	57.6	81.7	45.2	58.3
tiiuae/falcon-40b	60.4	61.9	85.3	52.7	41.7
timdettmers/guanaco-33b-merged	60	58.2	83.5	48.5	50
ausboss/llama-30b-supercot	59.8	58.5	82.9	44.3	53.6
huggyllama/llama-65b	58.3	57.8	84.2	48.8	42.3
pinkmanlove/llama-65b-hf	58.3	57.8	84.2	48.8	42.3
llama-65b	58.3	57.8	84.2	48.8	42.3
MetaIX/GPT4-X-Alpasta-30b	57.9	56.7	81.4	43.6	49.7
Aeala/VicUnlocked-alpaca-30b	57.6	55	80.8	44	50.4
digitous/Alpacino30b	57.4	57.1	82.6	46.1	43.8
Aeala/GPT4-x-AlpacaDente2-30b	57.2	56.1	79.8	44	49.1
TheBloke/dromedary-65b-lora-HF	57	57.8	80.8	50.8	38.8
TheBloke/Wizard-Vicuna-13B-Uncensored-HF	57	53.6	79.6	42.7	52
elinas/llama-30b-hf-transformers-4.29	56.9	57.1	82.6	45.7	42.3
ausboss/Llama30B-SuperHOT	56.9	57.1	82.6	45.7	42.3

Our initial finding indicates that GPT-4 can produce highly consistent ranks and detailed assessment when comparing chatbots' answers. Preliminary evaluations based on GPT-4, summarized in Figure 1, show that Vicuna achieves 90%* capability of Bard/ChatGPT. While this proposed framework shows a potential to automate chatbot assessment, it is not yet a rigorous approach. Building an evaluation system for chatbots remains an open question requiring further research.




Foundation Models
LLaMA
BLOOM
Training cost:
C_{train}
Instruction-Tuned Models
Alpaca
Vicuna
BELLE
Evaluation cost:
C_{eval}
Evaluation
API-based
Data leakage
Access limit
Unreproducible
Human
Time consuming
Expensive
Inconsistent
PandaLM
Reproducible
Open source
Safe & Efficient
1st iteration of Instruction-tuning pipeline
2nd iteration of Instruction-tuning pipeline
...
N-th iteration of Instruction-tuning pipeline