FabAcademy AI Lab

Recitations - AI - FabAcademy 2024

César García Sáez - @lahoramaker - 12 Feb 2024

FabAcademy AI Lab

Workgroup

- Launched at Instructor’s Bootcamp Leon (2024)
- Explore transversal use of AI for FabAcademy and digital fabrication in general
- Communication channel: Mattermost FabAcademy AI Lab
- Bi-weekly meetings (alternate weeks when recitations are not happening)
- Currently instructors and enthusiasts, but open to FabAcademy students

FabAcademy AI Lab

Current research lines

- Create an AI student that produces FabAcademy results using only generative AI tools
- Offer customized agents (GPTs) to assist in every week assignment
- Map COTS / NOTS solutions for generative AI in different use cases (create 3D printing pieces, vector (SVG), etc).
- Gershenfeld-Turing test: have Neil test a “remote student” based on his assignment and chat conversations

Custom GPTs

Custom GPTs

Powered by OpenAI GPT-4

- You can create them and customize the behavior via conversational interface
- You can add reference documents
- You can share them with the world
- But…

?

Custom GPTs

There are limitations

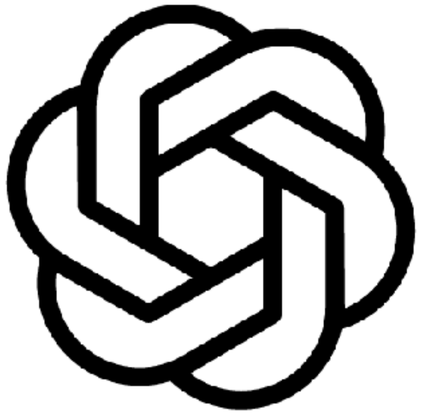
- Up to 20 documents, needs preprocessing
- Maximum of 512 MB per doc
- No more that 100 Gb in total
- Reports show the system is barely usable with very large context: https://community.openai.com/t/gpts-knowledge-capacity-limits/492955/28

COTS / NOTS Approach

Commercial Off The Shelf

AI as a Service

- OpenAI GPT-4
- Google Gemini Ultra / Pro
- Anthropic Claude 2
- Azure AI Studio models



Open Weights (NOTS?)

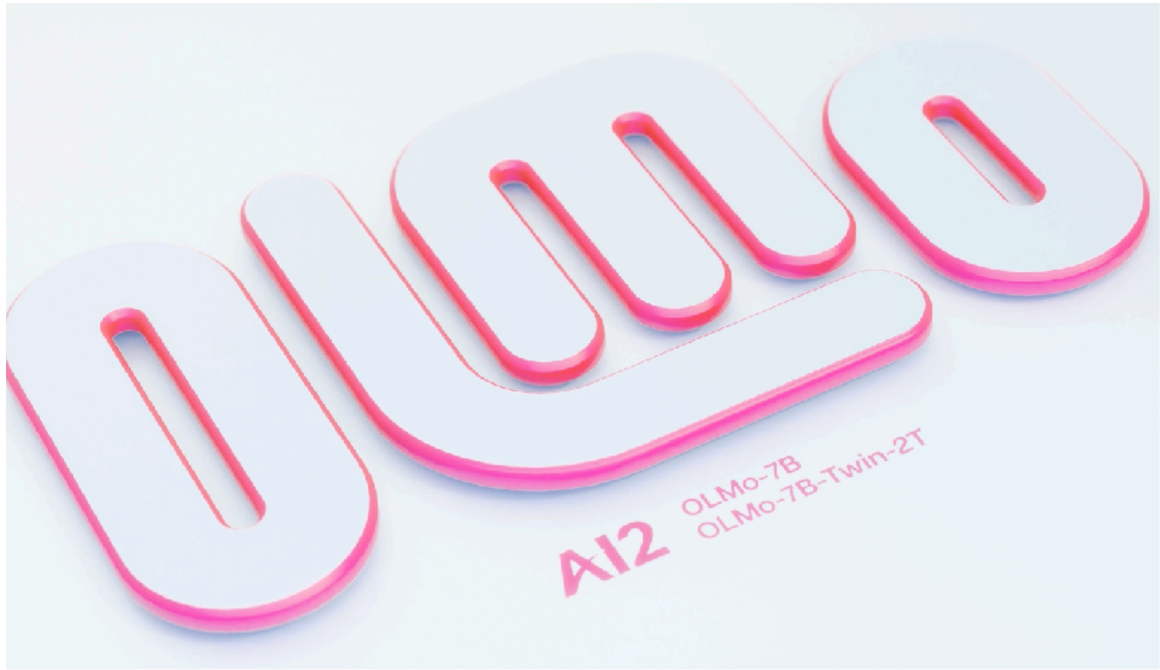
AI model weights available

- Meta Llama-2 (available as a service too)
- Yi 01
- Alibaba Qwen
- DeepSeeker Coder

Open Source Models

Weights, data, code open license

- OLMo - Allen AI institute
- Flor - Barcelona Supercomputing Center (weights, data)



Retrieval Augmented Generation

Use a vector store to find related contents

- Enables you to provide domain knowledge documents
- Each documents is chunked and context semantically vectorized
- This enables you to search external information and pass it to your LLM context
- Requires curation of the original data
- Documents should not be enough to generalize new “skills”

Fine-tuning techniques

LoRA, QLoRA, etc

- Enable you to customize the model, adding domain knowledge
- Can be offered as a thin layer on top of LLM
- Cost to customize is very small compared to LLM (100s USD vs several million USD)
- Size: hundred of Mb vs multiple Gb

Benchmarking

Best models - LMsys LLM Arena

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **200,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **58**. Total #votes: **268246**. Last updated: Feb 2, 2024.

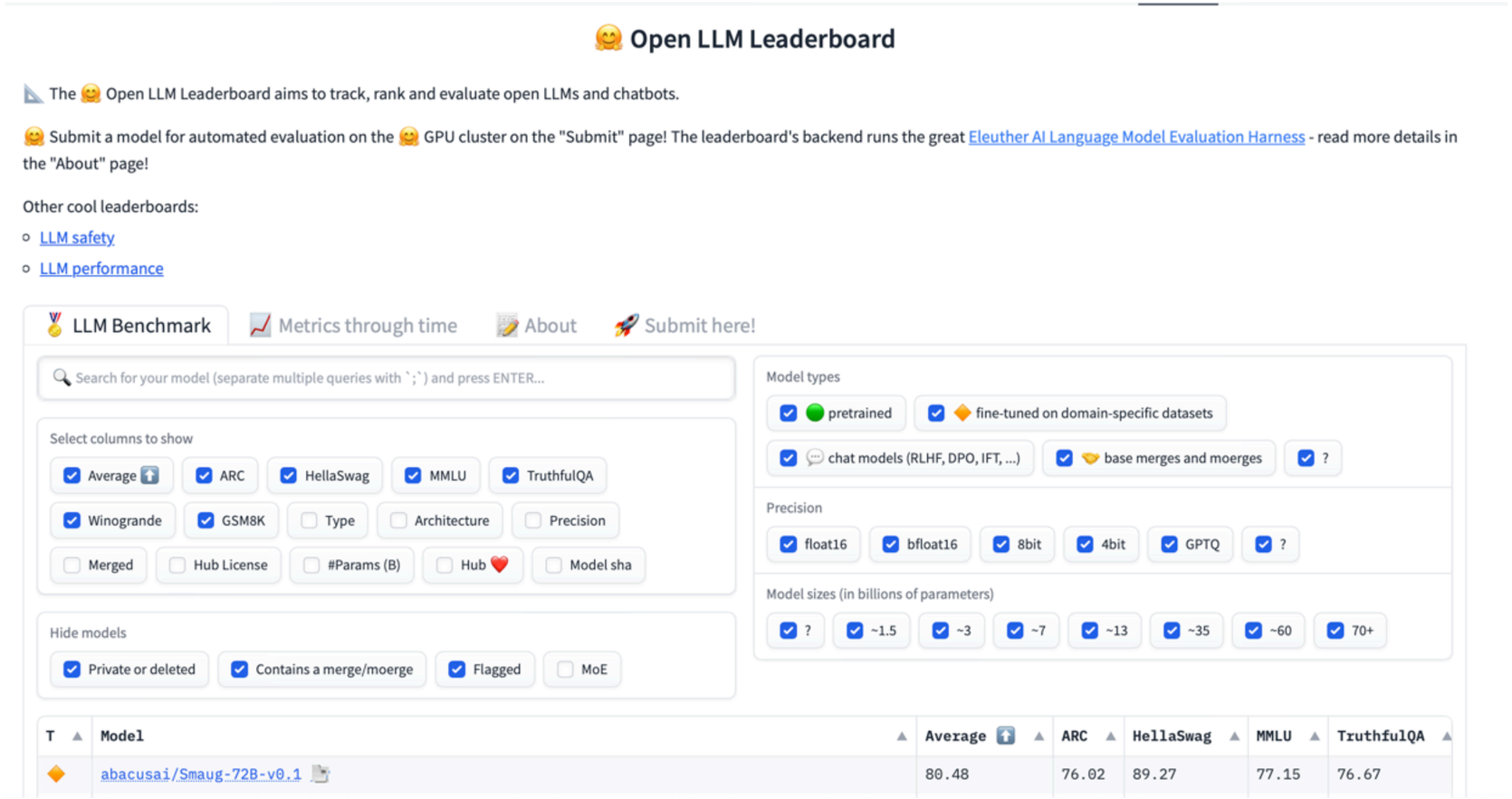
Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	☆ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-0125-preview	1253	+10/-11	3922	OpenAI	Proprietary	2023/4
2	GPT-4-1106-preview	1252	+5/-6	35385	OpenAI	Proprietary	2023/4
3	Bard (Gemini Pro)	1224	+9/-9	9081	Google	Proprietary	Online
4	GPT-4-0314	1190	+5/-6	18945	OpenAI	Proprietary	2021/9
5	GPT-4-0613	1162	+4/-5	29950	OpenAI	Proprietary	2021/9
6	Mistral Medium	1150	+6/-7	15447	Mistral	Proprietary	Unknown
7	Claude-1	1149	+6/-6	18189	Anthropic	Proprietary	Unknown
8	Claude-2.0	1132	+6/-7	12131	Anthropic	Proprietary	Unknown
9	Gemini Pro (Dev API)	1120	+7/-7	7616	Google	Proprietary	2023/4
10	Claude-2.1	1119	+5/-6	25494	Anthropic	Proprietary	Unknown

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

Benchmarking Open Source models

Best models - Open LLM Leaderboard



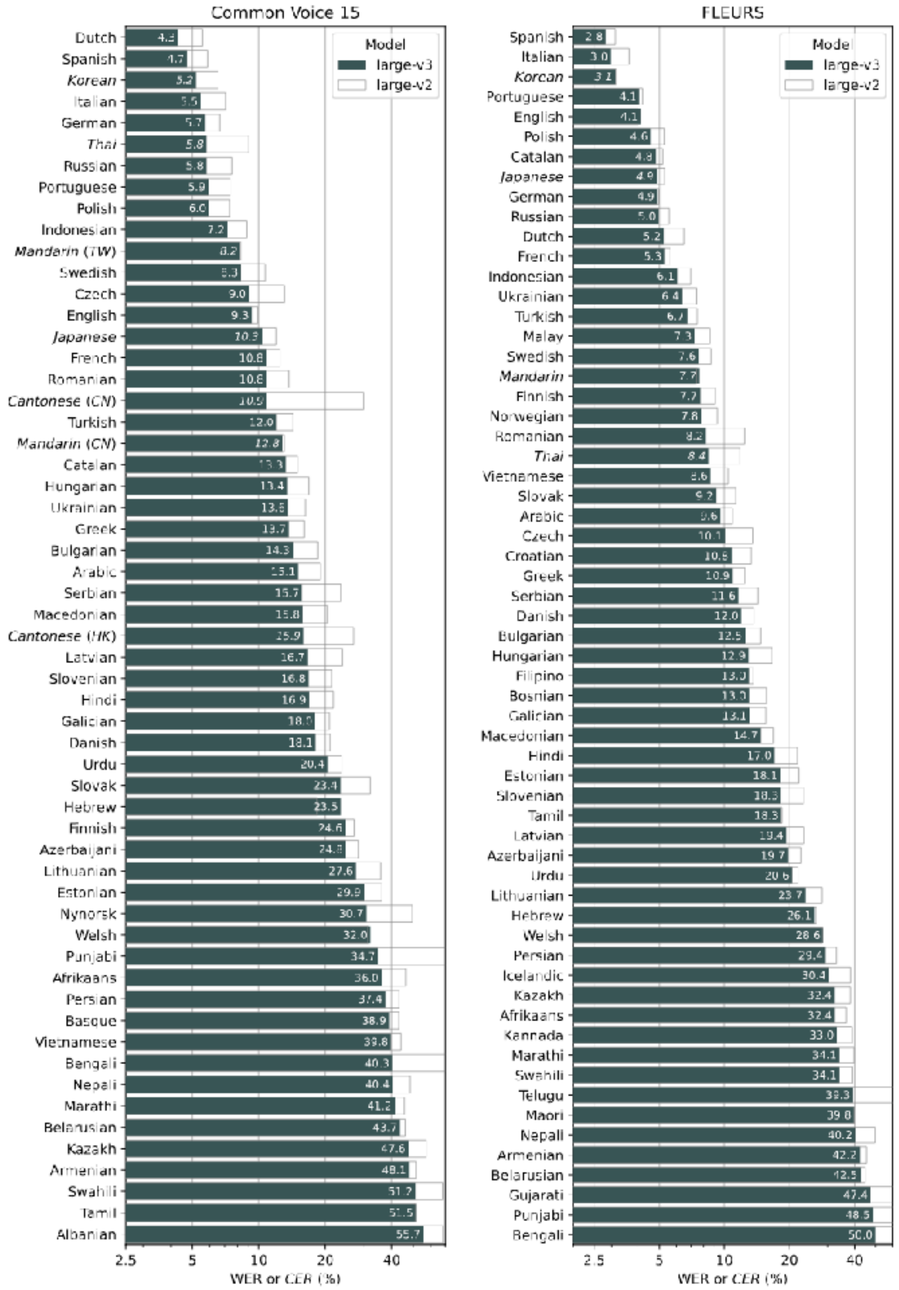
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Open source AI helper tools

Using AI to document

Whisper Model

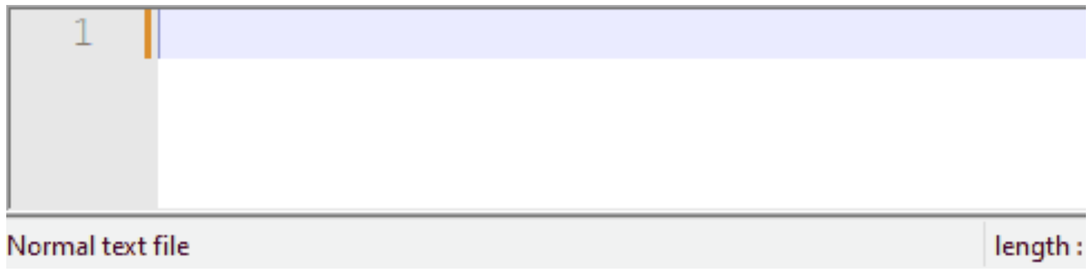
- Whisper is a Speech-to-Text open model (MIT license)
- It can transcribe audio files in 100 languages
- It can also translate from any language to English
- https://github.com/openai/whisper



Whisper Writer

Multiplatform app for Speech to text

- It enables you to record by pressing a key combination
- Transcription is pasted under your cursor
- Link: https://github.com/savbell/whisper-writer
- Mac: MacWhisper
- Linux: SpeechNote



Local LLM Apps

Based on open weight models

- Usually it requires a local GPU to make the most of it
- Best multi platform: GPT4ALL (enables you to consult your documents)
- Best UI: LM Studio (Mac, Win, Linux (Beta))
- Best for light customization: Ollama + Ollama Web UI
- Best for LLM Geeks: Text Generation Web UI

La Hora Maker

@LaHoraMaker · 10 K suscriptores · 325 vídeos

Versión web experimental de La Hora Maker. La Hora Maker te ofrece semanalmente l... >

itunes.apple.com/es/podcast/la-hora-maker/id1056919335?mt=2 **y 3 enlaces más**

Suscribirme

Inicio Vídeos Shorts En directo Pódcasts Listas Comunidad

Para ti

Cómo utilizar tus propios documentos con LLMs - Conceptos fundamentales de sistemas RAG

1,5 K visualizaciones • hace 2 meses

Manos a la obra - Chatea con tus documentos de forma 100% local y privada usando GPT4ALL

3,3 K visualizaciones • hace 2 meses

Novedades LM Studio - Chatea con tus imágenes y modelos mucho más potentes (Llava y Mixtral)

2,2 K visualizaciones • hace 1 mes

Spanish speakers - https://www.youtube.com/lahoramaker

Q&A