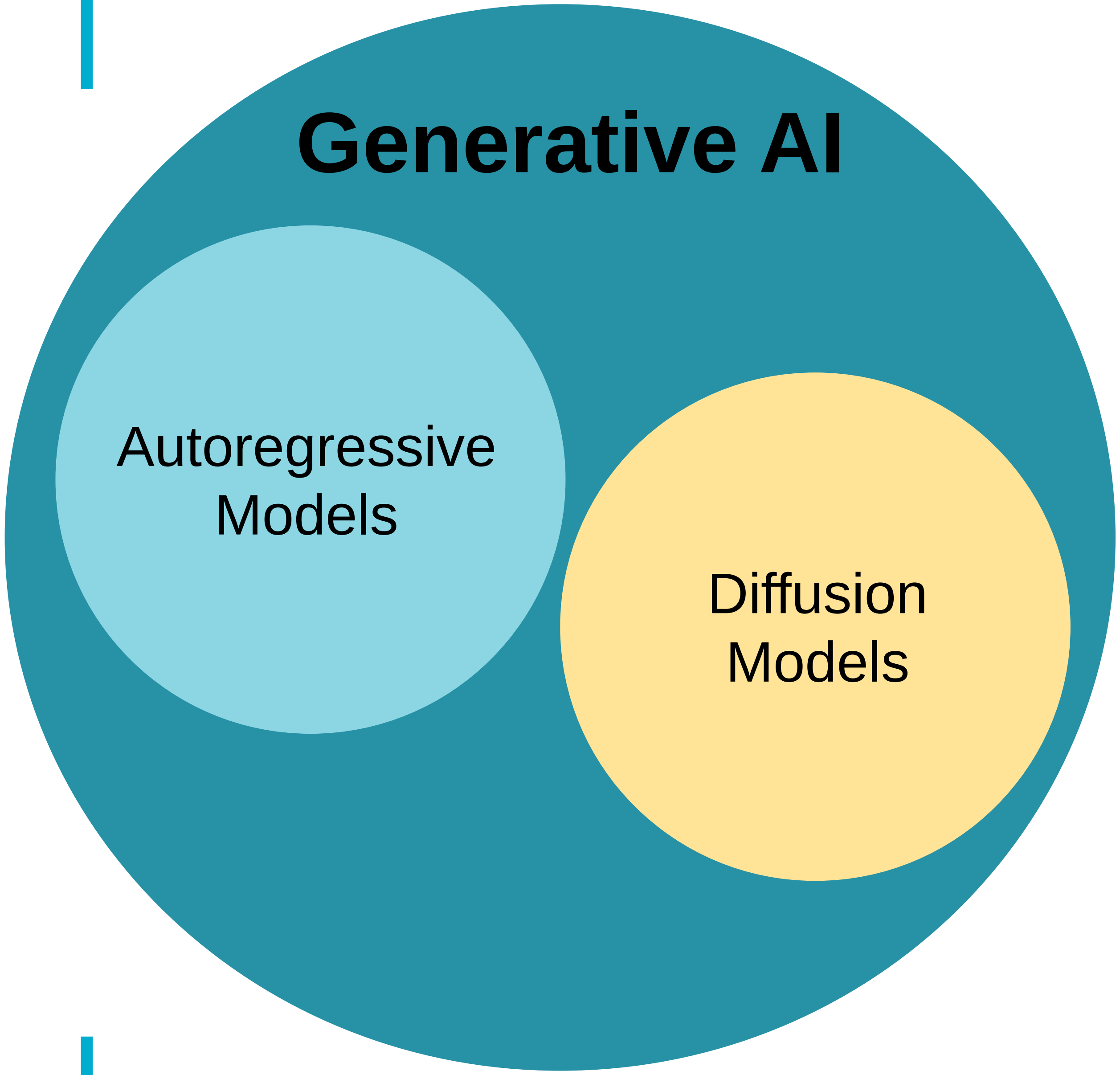
A Gentle introduction to LLMs and LLaMa

Ivan Gentile

IFAB

EuroCC Italy

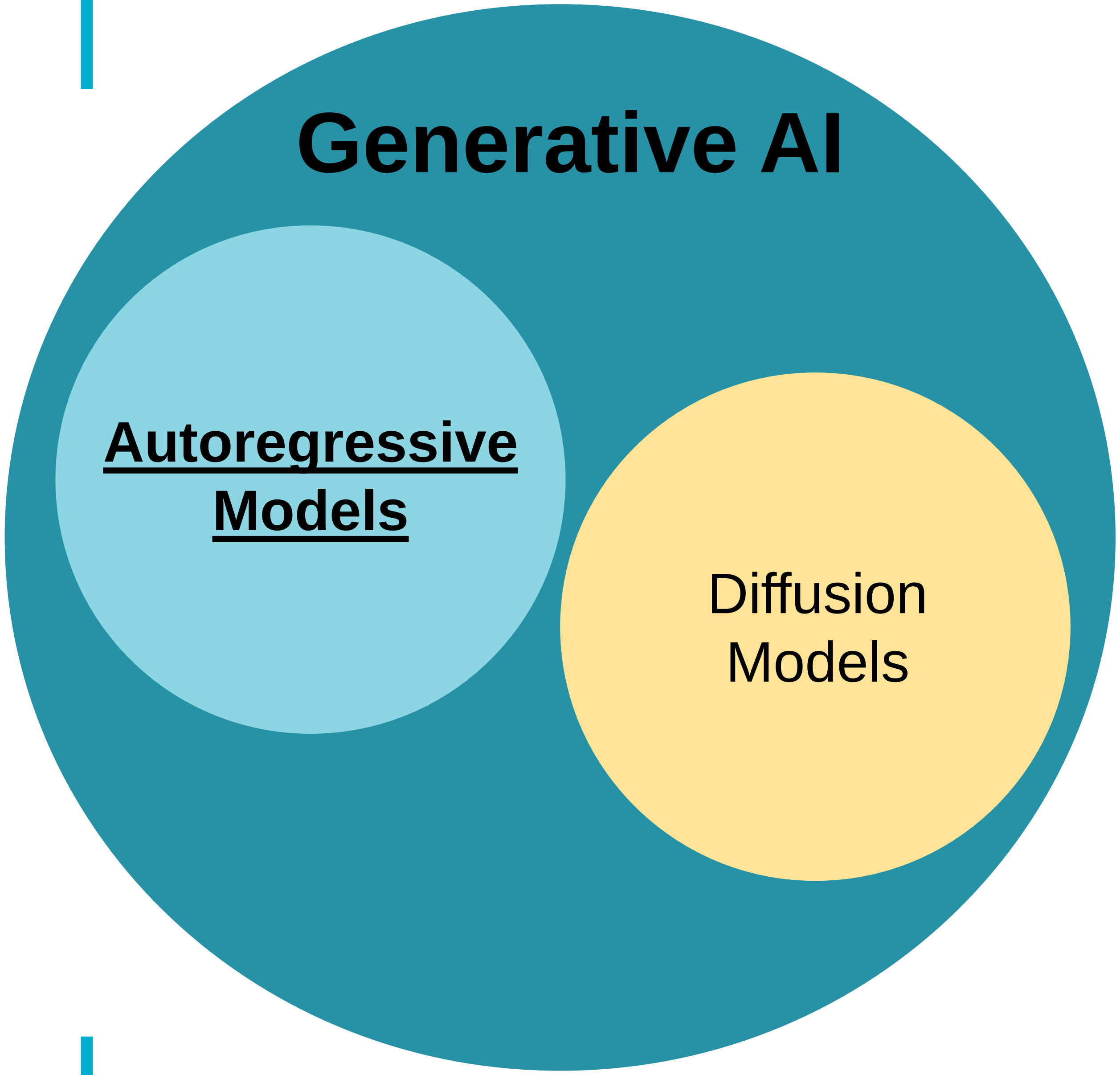
Gen AI



With **Gen AI** we refer to applications of AI where **new data is created** (generated) upon user interaction

- ChatGPT/LLaMa generate language
- DALL-E / Midjourney generate images
- SORA / Pika generate video
- SUNO / udio generate sound/music

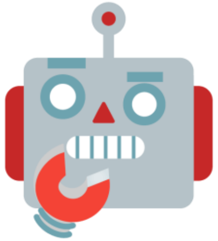
Gen AI



With **Gen AI** we refer to applications of AI where **new data is created** (generated) upon user interaction

- **ChatGPT/LLaMa generate language**
- DALL-E / Midjourney generate images
- SORA / Pika generate video
- SUNO / udio generate sound/music

Autoregressive Language Models



The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

- Autoregressive models generate data **sequentially**, one step at a time.
- In the case of text generation, autoregressive models **predict the next token** (word or subword) **based on the previously generated tokens**.
- The generation process is **iterative**, with each step depending on the output of the previous steps.

Generation Flow

1

Input Processing

- **Tokenization**: Breaking down input text into tokens
- **Embedding**: Converting tokens into dense semantic vectors
- **Positional Encoding**: Adding position information to the embeddings

2

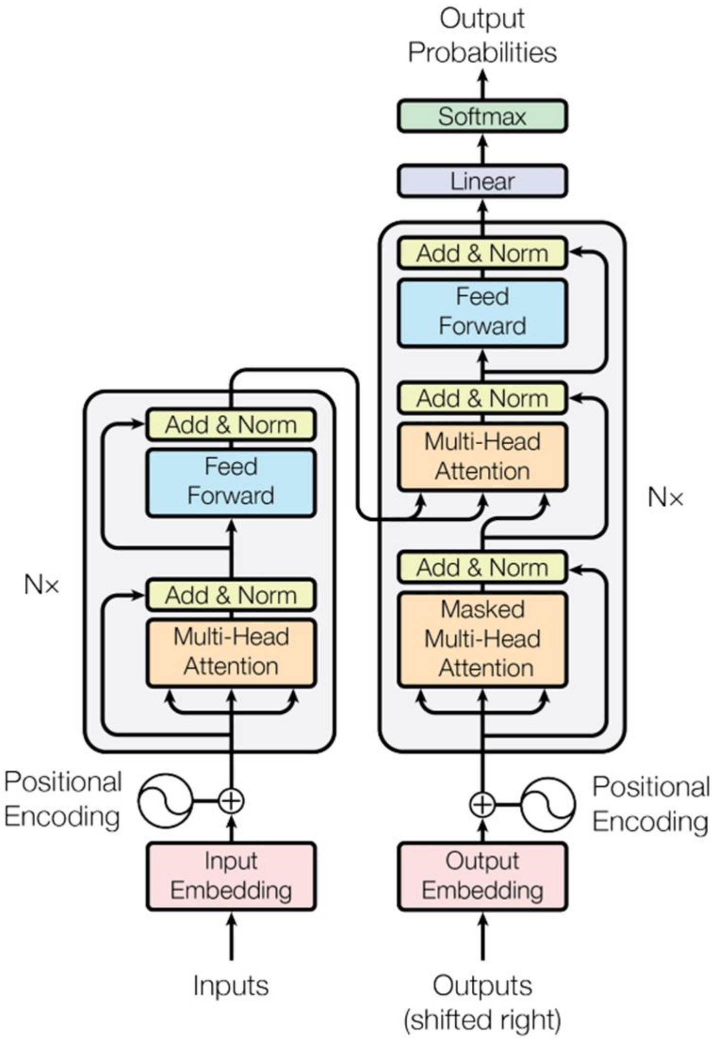
Transformer Decoder Architecture

- **Multi-Head Self-Attention**: Attending to different parts of the input
- **Feed Forward Neural Network**: Processing the attended information
- **Residual Connections and Layer Norm**: Enabling stable training

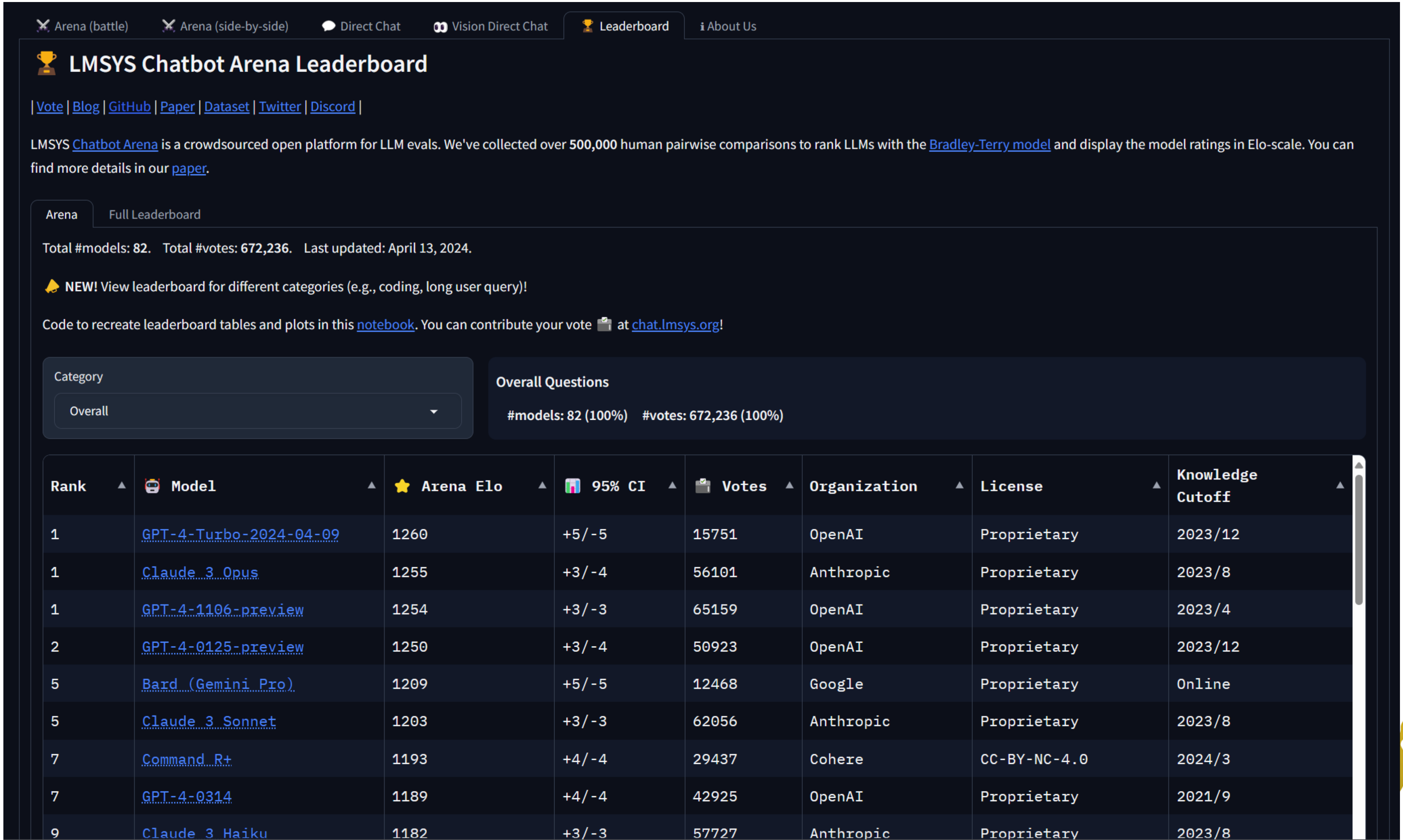
3

Output Generation

- **Vocabulary Distribution**: net output is a distribution over the vocabulary
- **Sampling Techniques**: Methods like top-k sampling for toke selections
- **Iterative process**: Selected token is fed back to generate the next one



State of the Art Models



🏆 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 **500,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Total #models: **82**. Total #votes: **672,236**. Last updated: April 13, 2024.

🔔 **NEW!** View leaderboard for different categories (e.g., coding, long user query)!

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote 🗳️ at chat.lmsys.org!

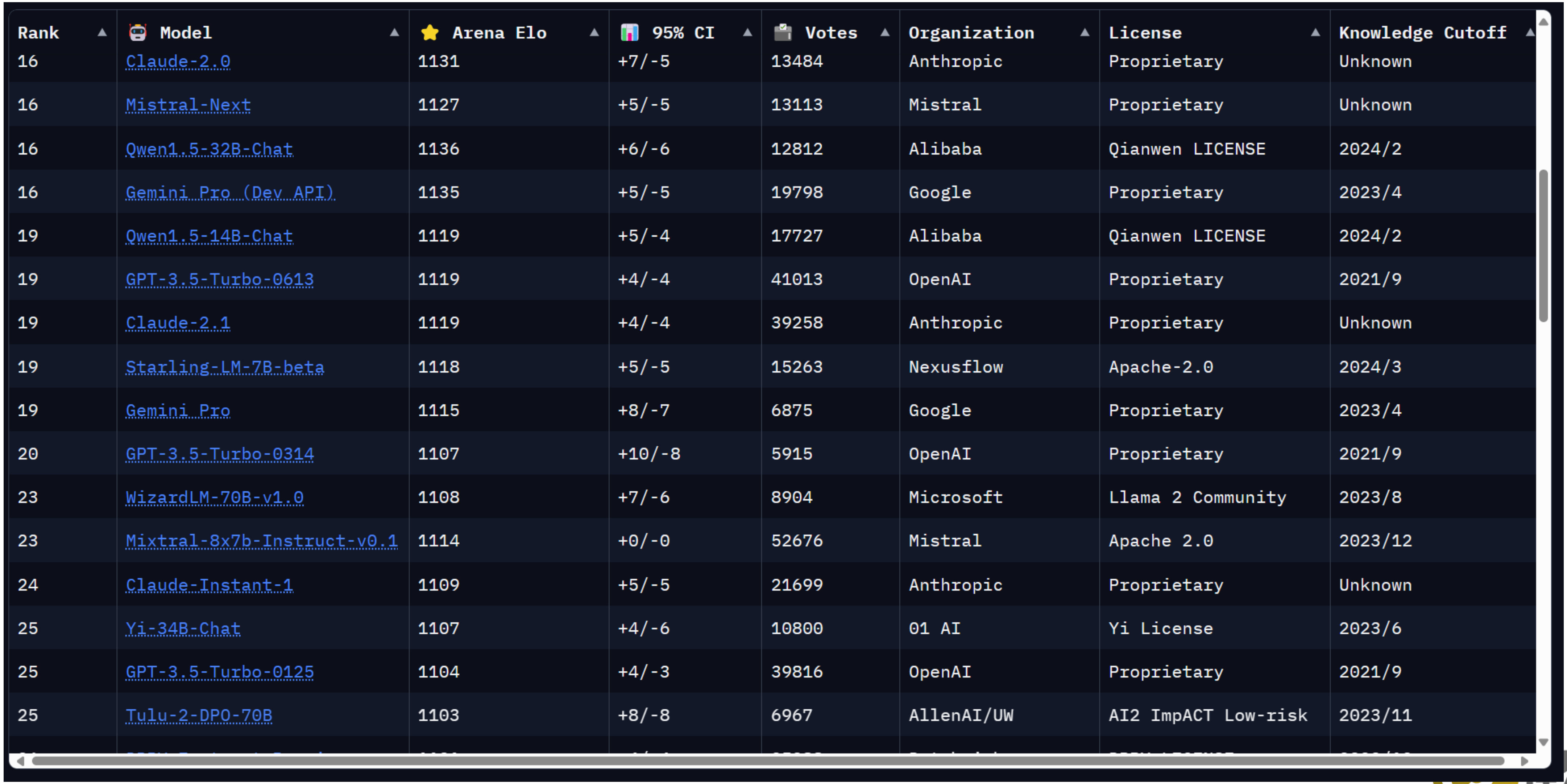
Category: Overall

Overall Questions

#models: 82 (100%) **#votes: 672,236 (100%)**

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/8

State of the Art *Open* Models



Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
16	Claude-2.0	1131	+7/-5	13484	Anthropic	Proprietary	Unknown
16	Mistral-Next	1127	+5/-5	13113	Mistral	Proprietary	Unknown
16	Qwen1.5-32B-Chat	1136	+6/-6	12812	Alibaba	Qianwen LICENSE	2024/2
16	Gemini Pro (Dev API)	1135	+5/-5	19798	Google	Proprietary	2023/4
19	Qwen1.5-14B-Chat	1119	+5/-4	17727	Alibaba	Qianwen LICENSE	2024/2
19	GPT-3.5-Turbo-0613	1119	+4/-4	41013	OpenAI	Proprietary	2021/9
19	Claude-2.1	1119	+4/-4	39258	Anthropic	Proprietary	Unknown
19	Starling-LM-7B-beta	1118	+5/-5	15263	Nexusflow	Apache-2.0	2024/3
19	Gemini Pro	1115	+8/-7	6875	Google	Proprietary	2023/4
20	GPT-3.5-Turbo-0314	1107	+10/-8	5915	OpenAI	Proprietary	2021/9
23	WizardLM-70B-v1.0	1108	+7/-6	8904	Microsoft	Llama 2 Community	2023/8
23	Mixtral-8x7b-Instruct-v0.1	1114	+0/-0	52676	Mistral	Apache 2.0	2023/12
24	Claude-Instant-1	1109	+5/-5	21699	Anthropic	Proprietary	Unknown
25	Yi-34B-Chat	1107	+4/-6	10800	01 AI	Yi License	2023/6
25	GPT-3.5-Turbo-0125	1104	+4/-3	39816	OpenAI	Proprietary	2021/9
25	Tulu-2-DPO-70B	1103	+8/-8	6967	AllenAI/UW	AI2 ImpACT Low-risk	2023/11

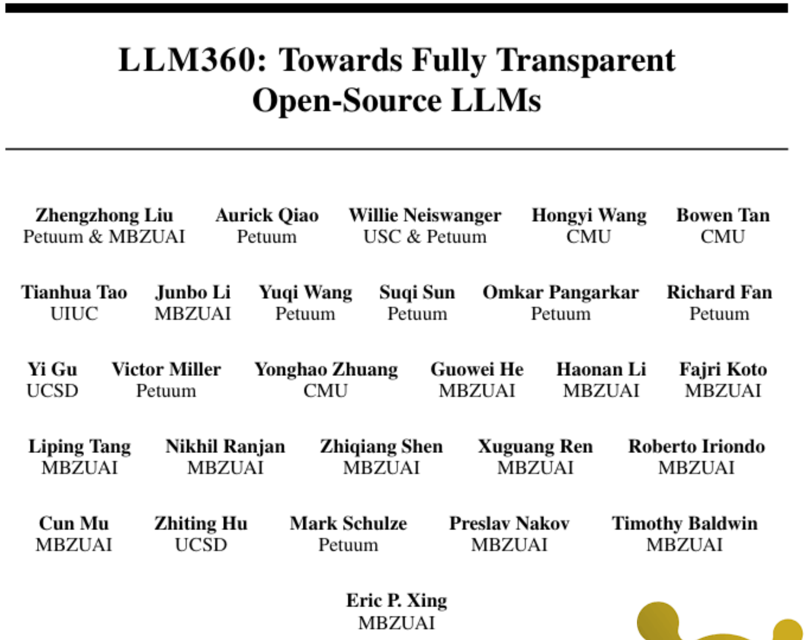
Open Source vs Open Weights

Open Language Model: OLMo

A State-of-the-Art, Truly Open LLM and Framework

Each model comes with the following:

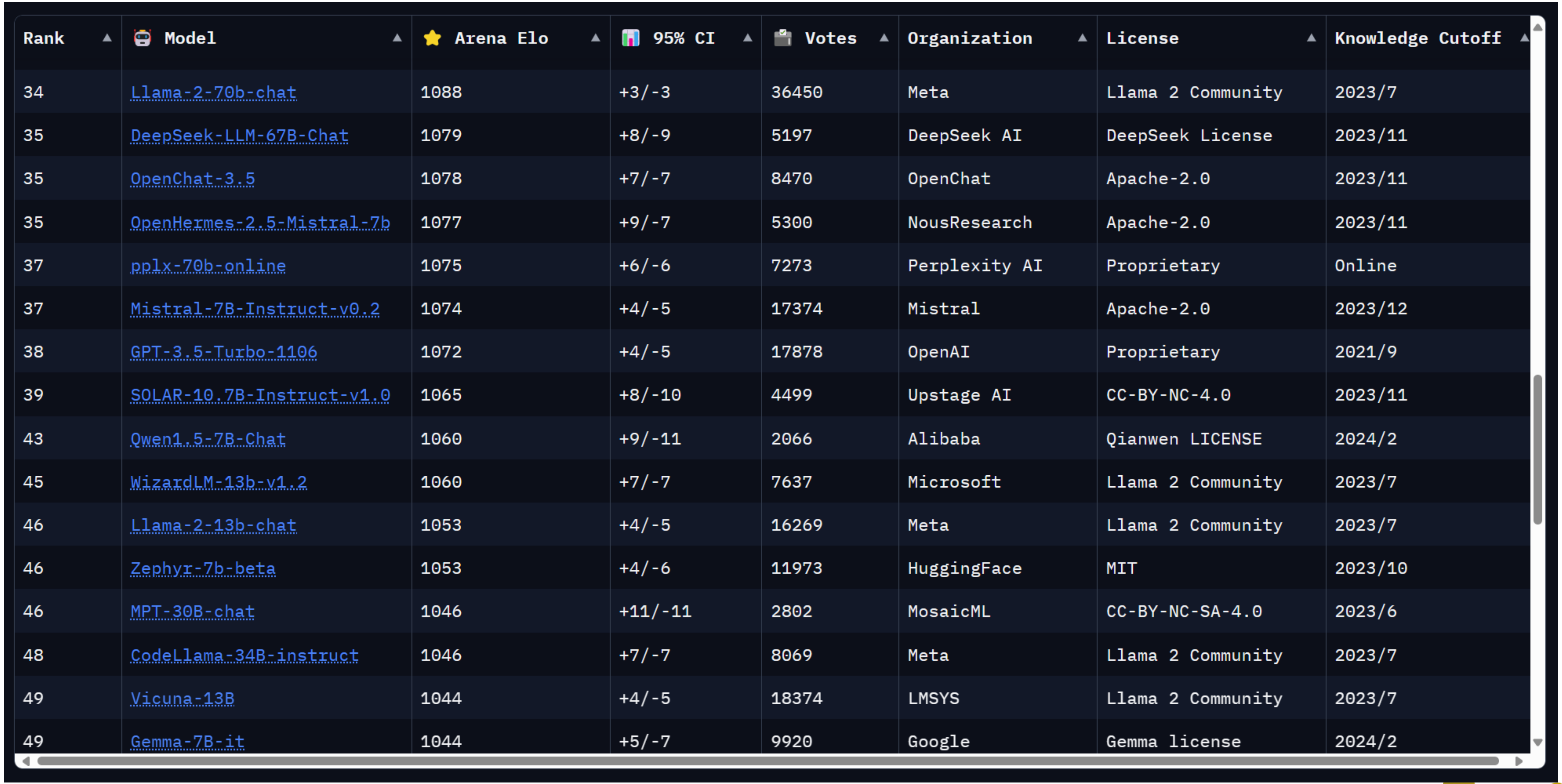
- Full training data used for these models, including code that produces the training data, from AI2's Dolma, and WIMBD for analyzing pretraining data.
- Full model weights, training code, training logs, training metrics in the form of Weights & Biases logs, and inference code.
- 500+ checkpoints per model, from every 1000 steps during the training process, available as revisions on HuggingFace.
- Evaluation code under the umbrella of AI2's Catwalk and Paloma.
- Fine-tuning code and adapted models (with Open Instruct)
- All code, weights, and intermediate checkpoints are released under the Apache 2.0 License.



LLM360: Towards Fully Transparent Open-Source LLMs

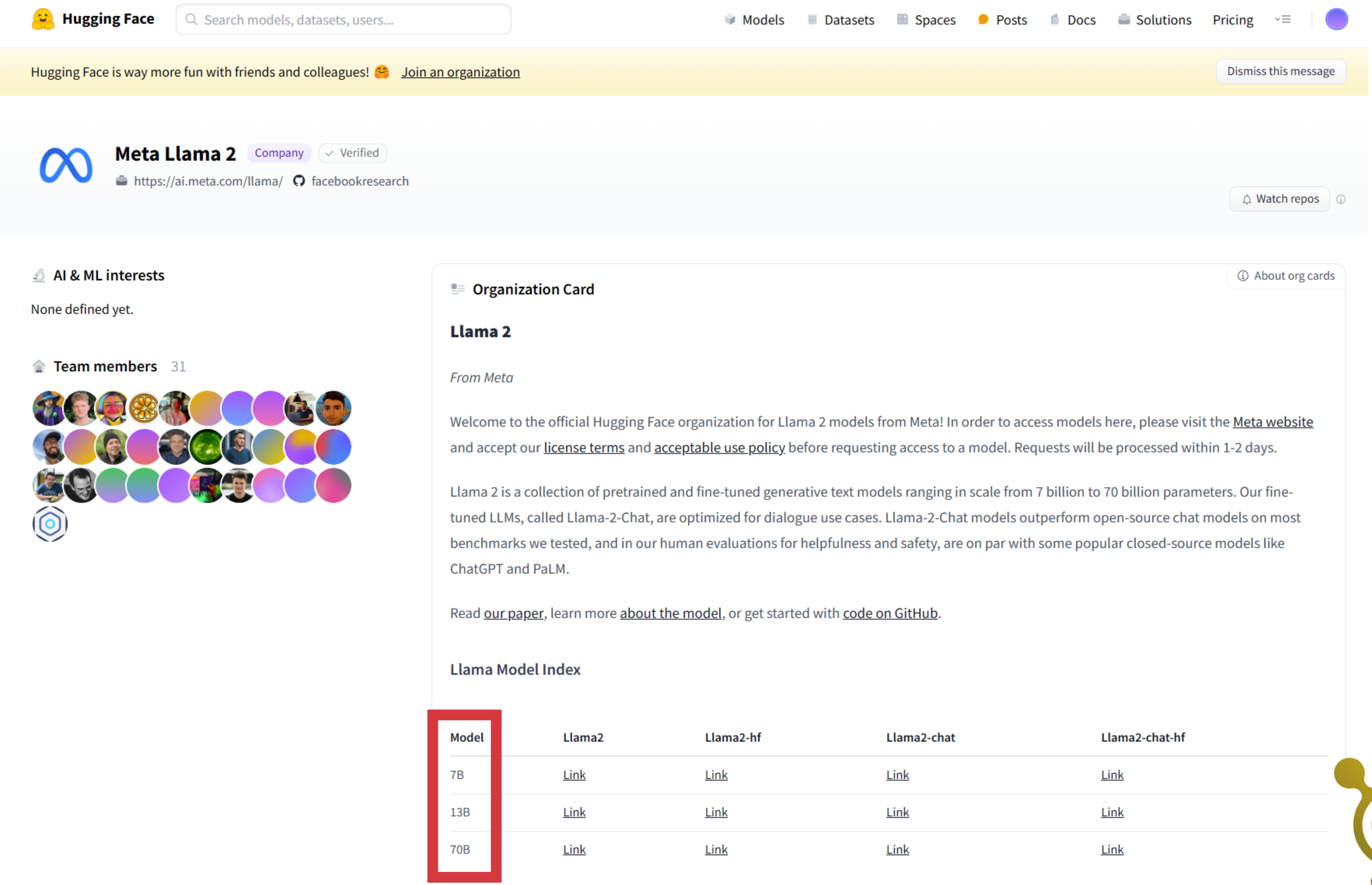
Zhengzhong Liu, Petuum & MBZUAI
Aurick Qiao, Petuum
Willie Neiswanger, USC & Petuum
Hongyi Wang, CMU
Bowen Tan, CMU
Tianhua Tao, UIUC
Junbo Li, MBZUAI
Yuqi Wang, Petuum
Suqi Sun, Petuum
Omkar Pangarkar, Petuum
Richard Fan, Petuum
Yi Gu, UCSD
Victor Miller, Petuum
Yonghao Zhuang, CMU
Guowei He, MBZUAI
Haonan Li, MBZUAI
Fajri Koto, MBZUAI
Liping Tang, MBZUAI
Nikhil Ranjan, MBZUAI
Zhiqiang Shen, MBZUAI
Xuguang Ren, MBZUAI
Roberto Iriondo, MBZUAI
Cun Mu, MBZUAI
Zhiting Hu, UCSD
Mark Schulze, Petuum
Preslav Nakov, MBZUAI
Timothy Baldwin, MBZUAI
Eric P. Xing, MBZUAI

Meta LLaMa2



Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
34	Llama-2-70b-chat	1088	+3/-3	36450	Meta	Llama 2 Community	2023/7
35	DeepSeek-LLM-67B-Chat	1079	+8/-9	5197	DeepSeek AI	DeepSeek License	2023/11
35	OpenChat-3.5	1078	+7/-7	8470	OpenChat	Apache-2.0	2023/11
35	OpenHermes-2.5-Mistral-7b	1077	+9/-7	5300	NousResearch	Apache-2.0	2023/11
37	pplx-70b-online	1075	+6/-6	7273	Perplexity AI	Proprietary	Online
37	Mistral-7B-Instruct-v0.2	1074	+4/-5	17374	Mistral	Apache-2.0	2023/12
38	GPT-3.5-Turbo-1106	1072	+4/-5	17878	OpenAI	Proprietary	2021/9
39	SOLAR-10.7B-Instruct-v1.0	1065	+8/-10	4499	Upstage AI	CC-BY-NC-4.0	2023/11
43	Qwen1.5-7B-Chat	1060	+9/-11	2066	Alibaba	Qianwen LICENSE	2024/2
45	WizardLM-13b-v1.2	1060	+7/-7	7637	Microsoft	Llama 2 Community	2023/7
46	Llama-2-13b-chat	1053	+4/-5	16269	Meta	Llama 2 Community	2023/7
46	Zephyr-7b-beta	1053	+4/-6	11973	HuggingFace	MIT	2023/10
46	MPT-30B-chat	1046	+11/-11	2802	MosaicML	CC-BY-NC-SA-4.0	2023/6
48	CodeLlama-34B-instruct	1046	+7/-7	8069	Meta	Llama 2 Community	2023/7
49	Vicuna-13B	1044	+4/-5	18374	LMSYS	Llama 2 Community	2023/7
49	Gemma-7B-it	1044	+5/-7	9920	Google	Gemma license	2024/2

Different model sizes for different computing capabilities



Hugging Face is way more fun with friends and colleagues! 🤗 Join an organization

Meta Llama 2

Company ✓ Verified

https://ai.meta.com/llama/ facebookresearch

AI & ML interests

None defined yet.

Team members 31

Organization Card

Llama 2

From Meta

Welcome to the official Hugging Face organization for Llama 2 models from Meta! In order to access models here, please visit the Meta website and accept our license terms and acceptable use policy before requesting access to a model. Requests will be processed within 1-2 days.

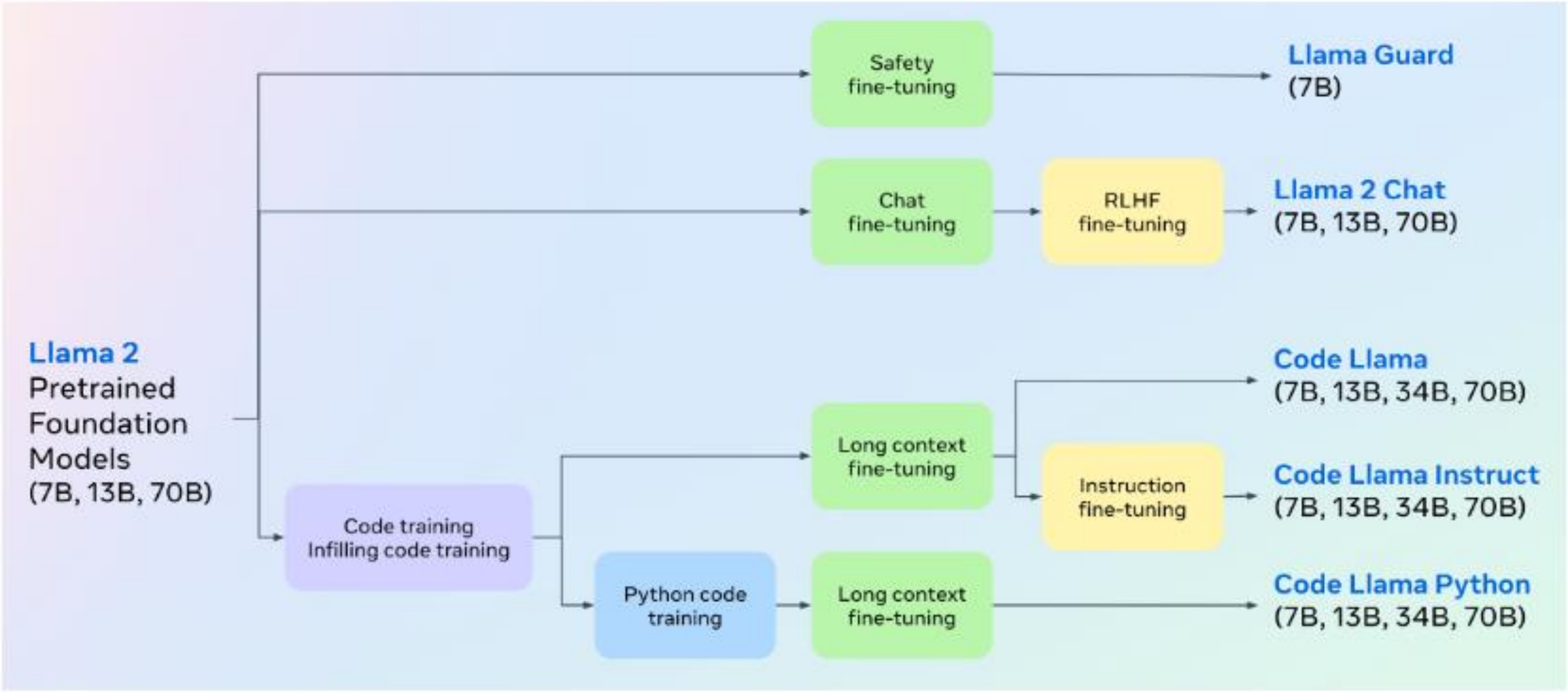
Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

Read our paper, learn more about the model, or get started with code on GitHub.

Llama Model Index

Model	Llama2	Llama2-hf	Llama2-chat	Llama2-chat-hf
7B	Link	Link	Link	Link
13B	Link	Link	Link	Link
70B	Link	Link	Link	Link

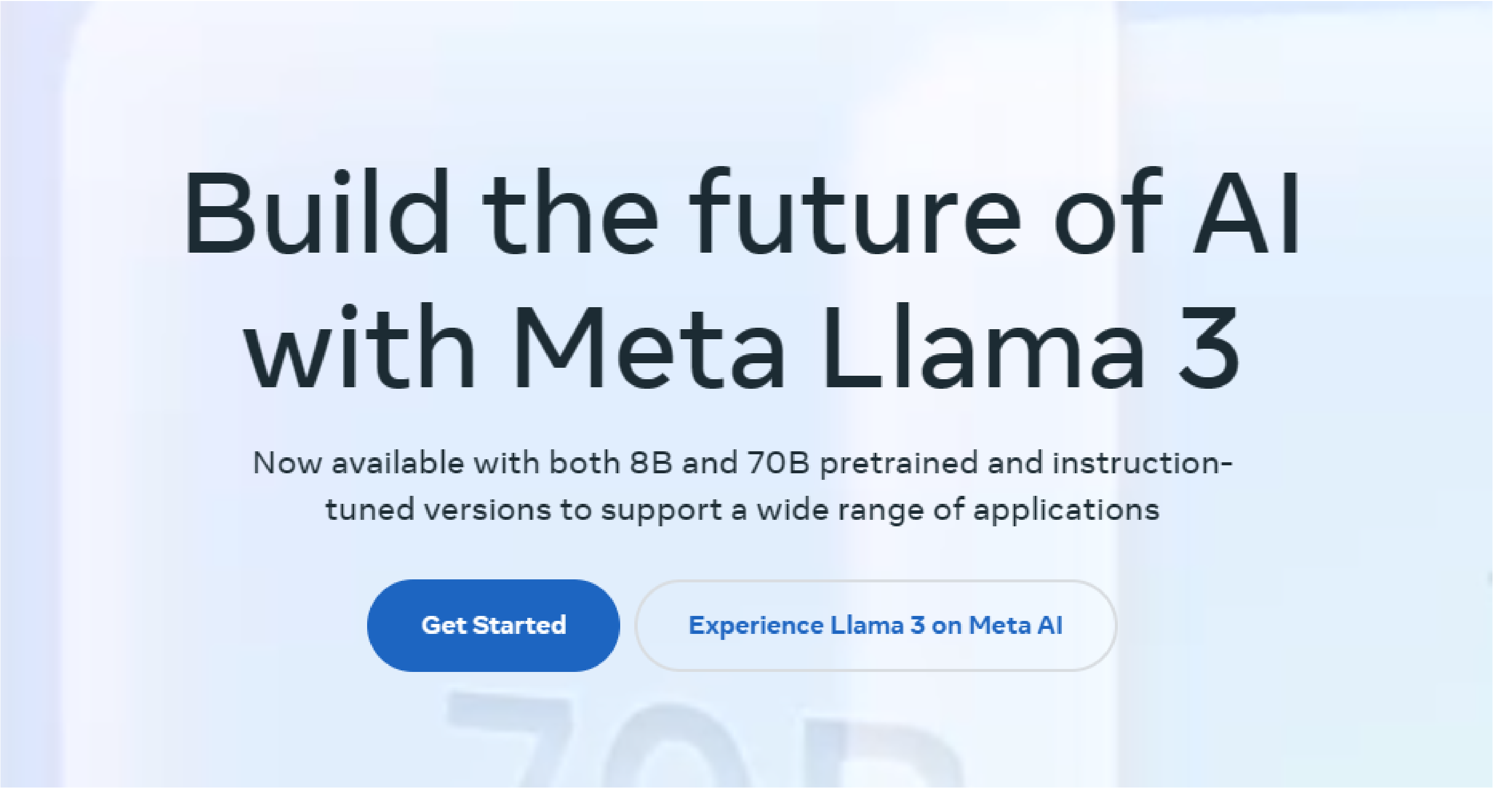
Different flavors for different use cases



Llama 2 Pretrained Foundation Models (7B, 13B, 70B)
Safety fine-tuning
Llama Guard (7B)
Chat fine-tuning
RLHF fine-tuning
Llama 2 Chat (7B, 13B, 70B)
Code training Infilling code training
Long context fine-tuning
Code Llama (7B, 13B, 34B, 70B)
Instruction fine-tuning
Code Llama Instruct (7B, 13B, 34B, 70B)
Python code training
Long context fine-tuning
Code Llama Python (7B, 13B, 34B, 70B)

- Llama 2 = Foundational model (good at text completion)
- Llama2 – chat = Good at question answering (like ChatGPT!)
- CodeLlama - Instruct: for instruction following and safer deployment
- LlamaGuard: input-output safeguard model.
- * – hf = Hugging Face format

Last minute news



Build the future of AI with Meta Llama 3

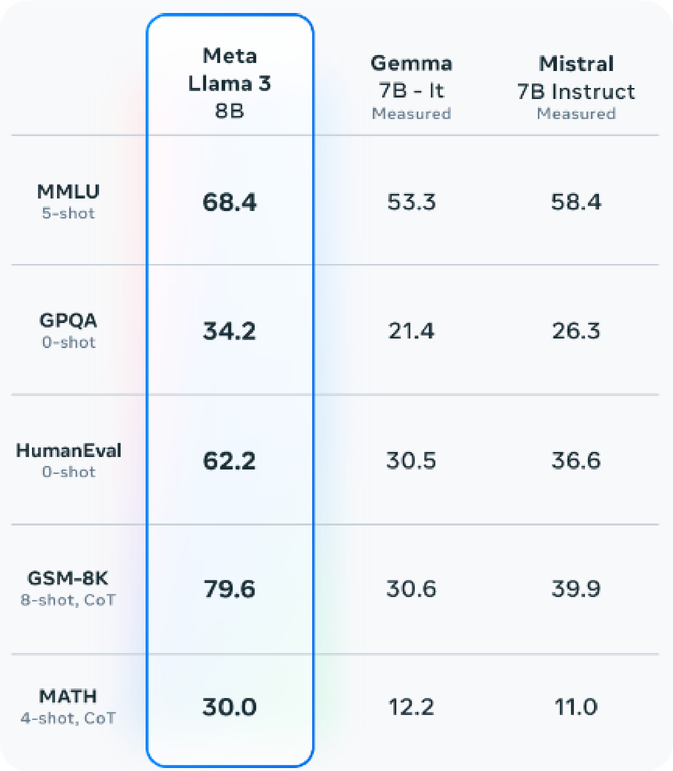
Now available with both 8B and 70B pretrained and instruction-tuned versions to support a wide range of applications

- Llama 3 released on April 18th
- 8B Model almost as good as Llama 2 70B
- 70B Model better than Claude 3 Sonnet
- **400B Model currently under training!**

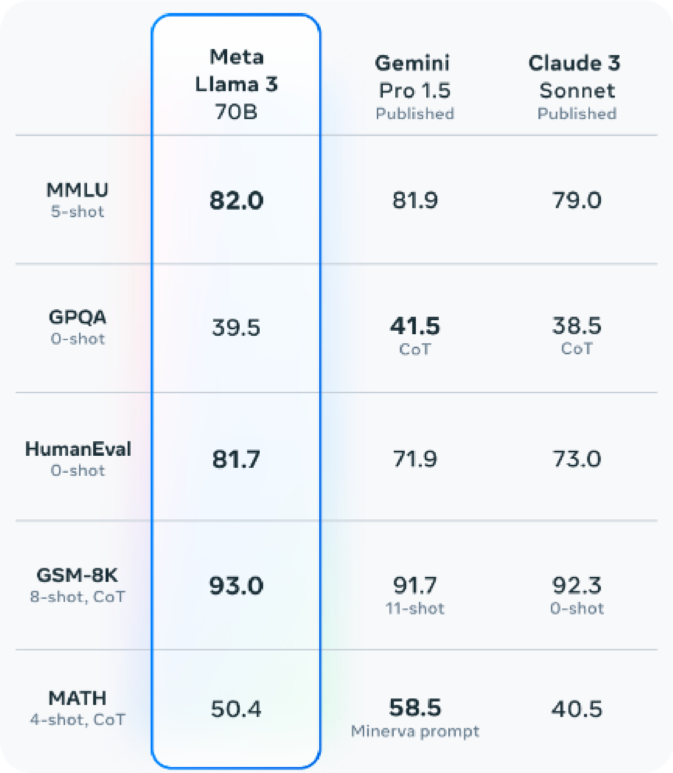
Llama 3 benchmark

- MMLU = Multiple Choice on many subjects
- Humaneval = Generate Python code
- GSM8K = 8th grade level math questions

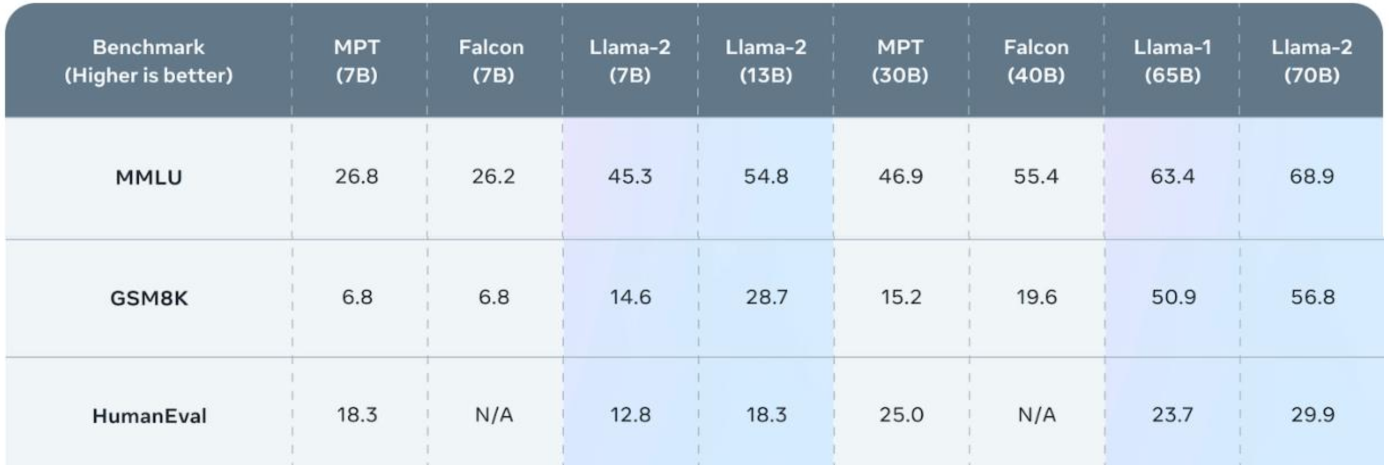
Meta Llama 3 Instruct model performance



	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0



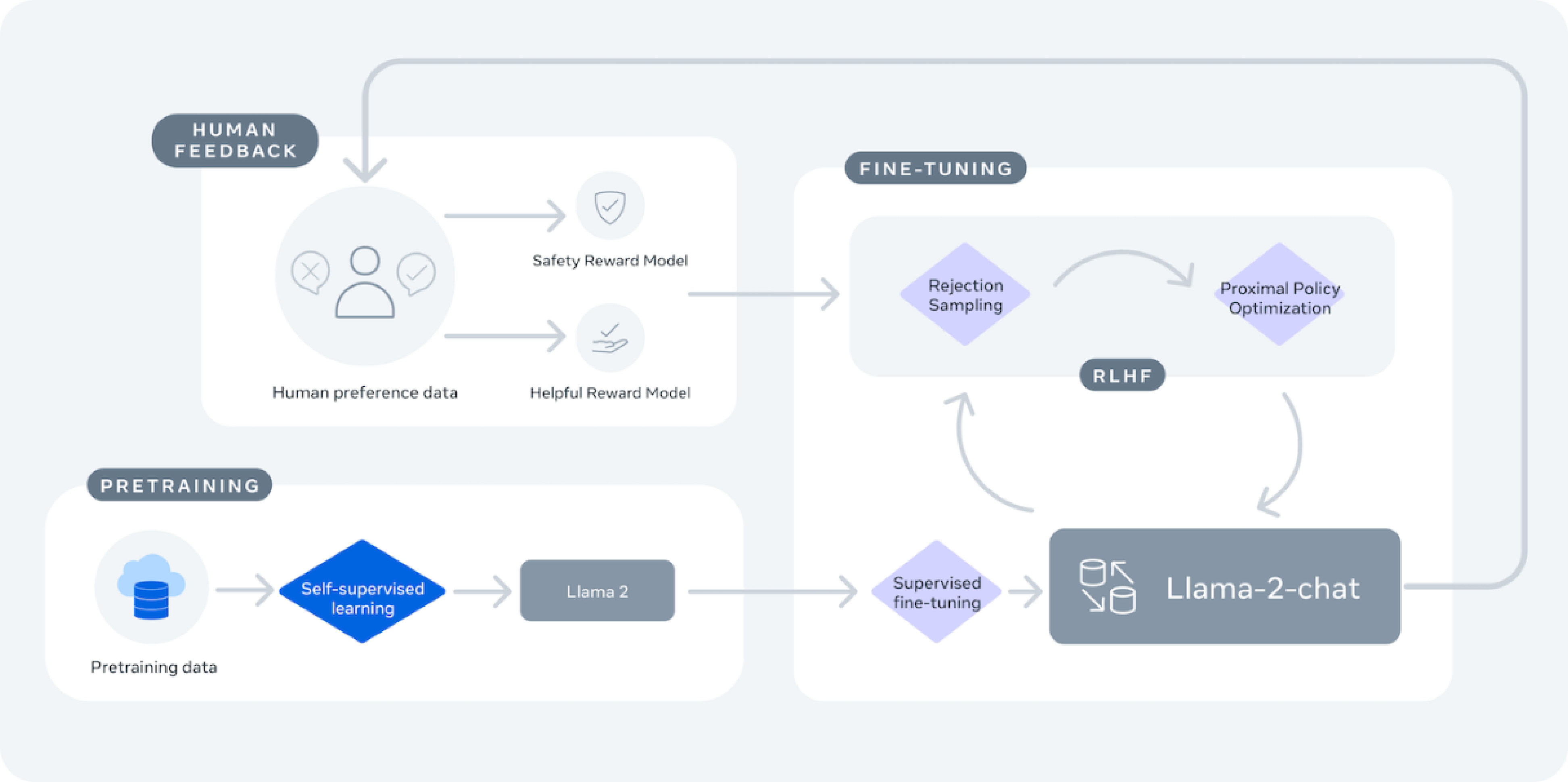
	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5



Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8
HumanEval	18.3	N/A	12.8	18.3	25.0	N/A	23.7	29.9

EUROCC ITALY
HPC Competence Center

LLaMa 2 Training



Questions?

Seguici sui canali EuroCC Italy

Sito web www.euroccitaly.it

LinkedIN

YouTube

Co-funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia under grant agreement No 101101903.

<<<<<<<<<<