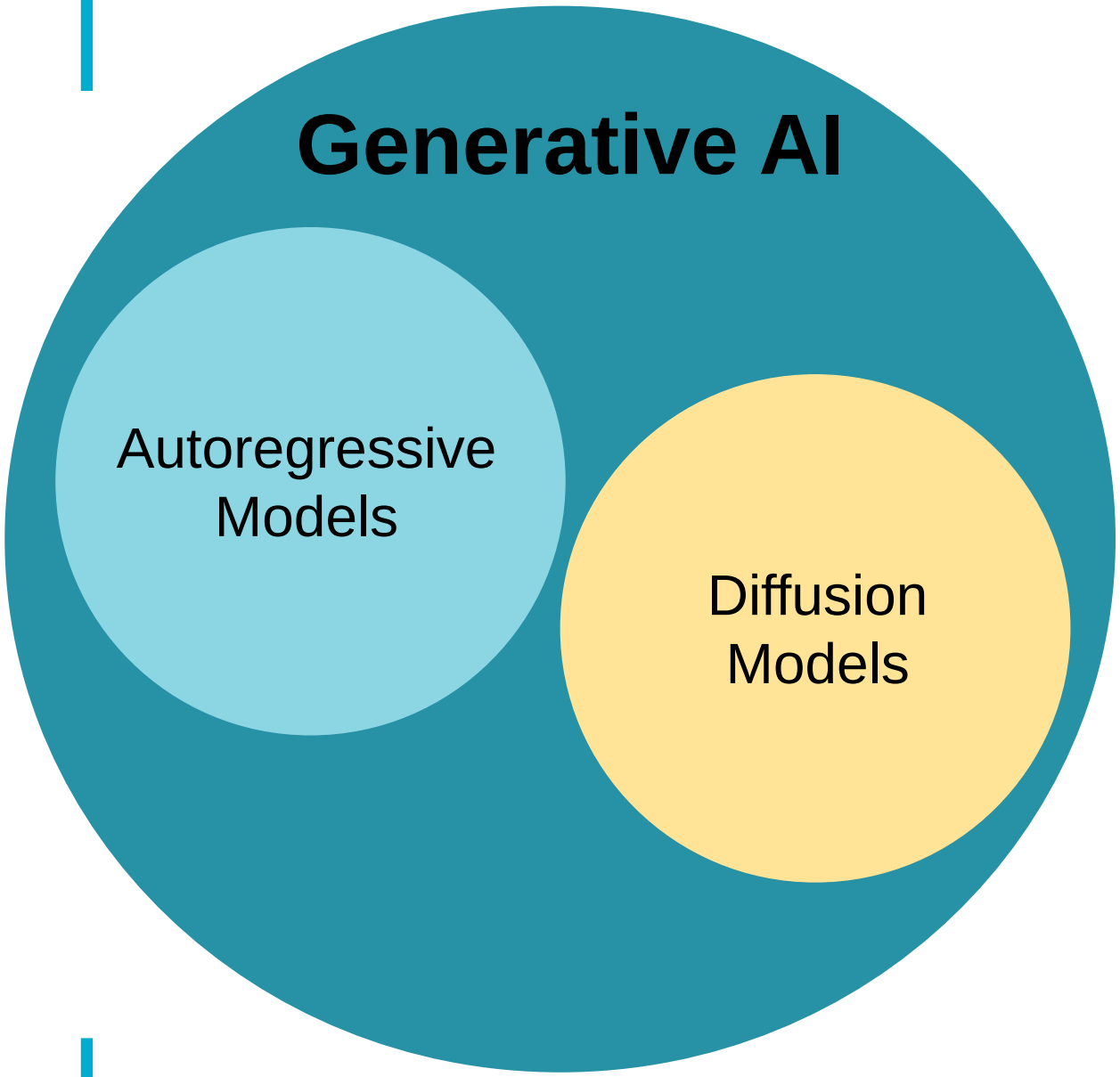
# A Gentle introduction to LLMs and LLaMa

**Ivan Gentile**
**IFAB**
**EuroCC Italy**

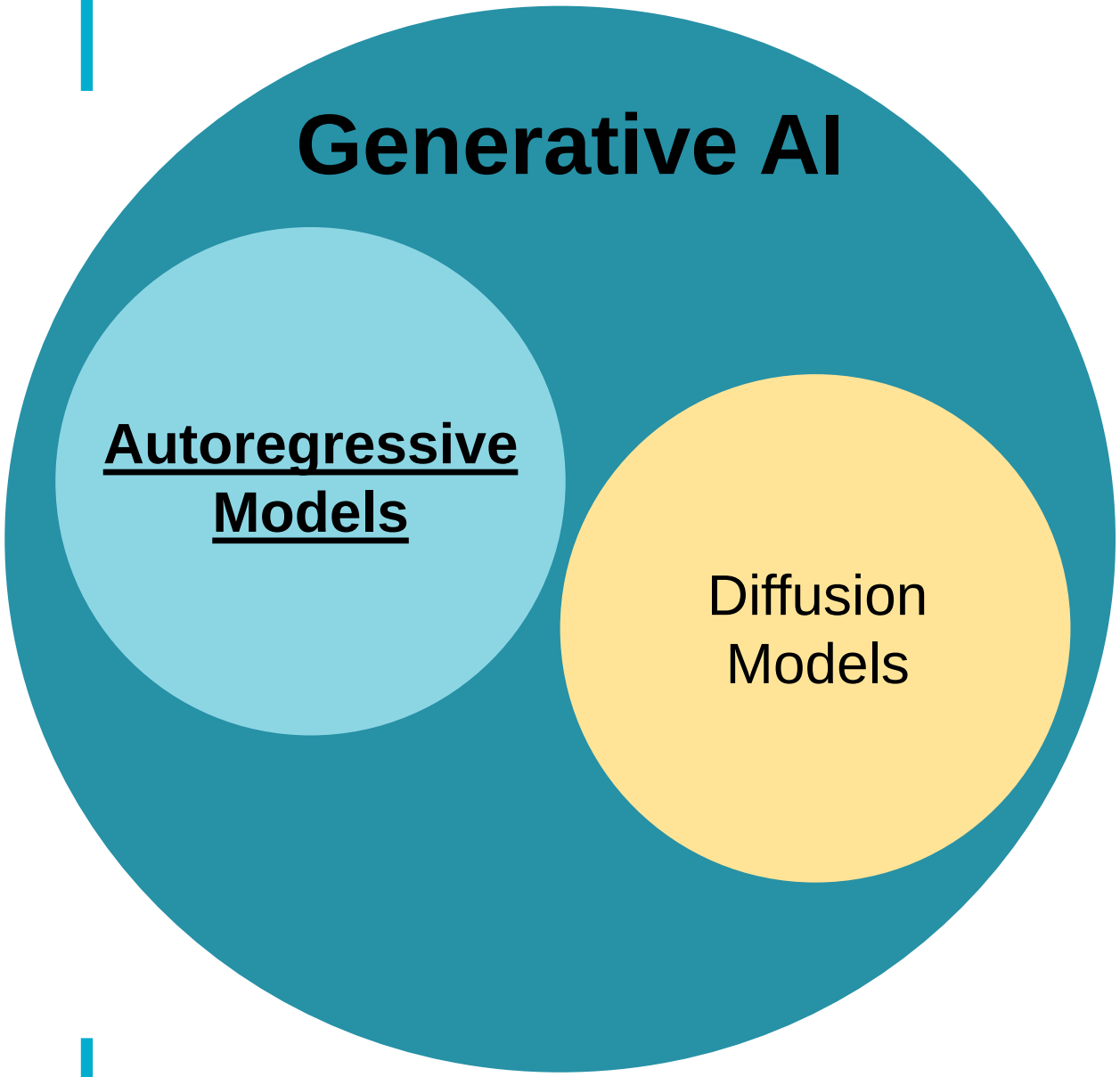
# Gen AI




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

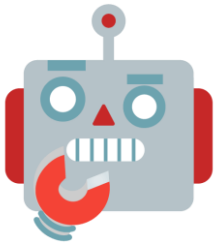



With **Gen AI** we refer to applications of AI where **new data is created** (generated) upon user interaction

- **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.

# Llama Generation Flow

1

**Input Processing**

- **Tokenization**: Breaking down input text into tokens (using BPE)
- **Embedding**: Converting tokens into dense semantic vectors (learned)
- **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**: top-k sampling for token selections
- **Iterative process**: Selected token is fed back to generate the next one

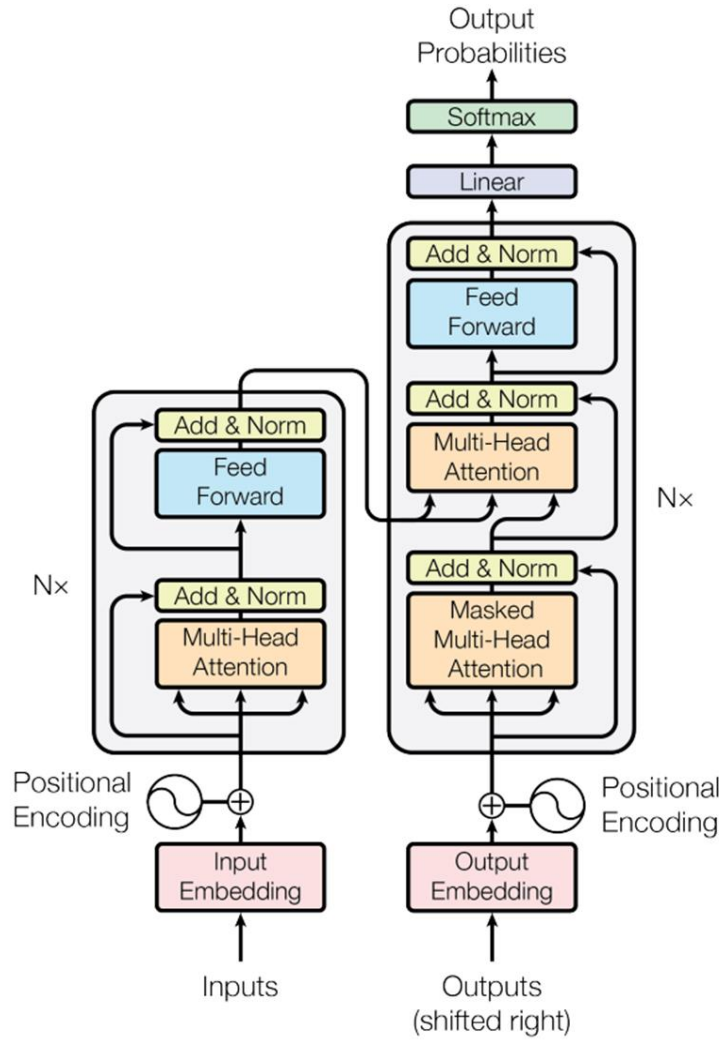



Image from: Attention Is All You Need, Ashish Vaswani et al.

# State of the Art Models

## LMSYS Chatbot Arena Leaderboard

Category: Overall

Overall Questions

#models: 92 (100%) #votes: 910,122 (100%)

| Rank* (UB) | Model | Arena Elo | 95% CI | Votes | Organization | License | Knowledge Cutoff |
|---|---|---|---|---|---|---|---|
| 1 | GPT-4-Turbo-2024-04-09 | 1259 | +4/-3 | 35931 | OpenAI | Proprietary | 2023/12 |
| 2 | GPT-4-1106-preview | 1253 | +2/-3 | 73547 | OpenAI | Proprietary | 2023/4 |
| 2 | Claude 3 Opus | 1251 | +3/-3 | 80997 | Anthropic | Proprietary | 2023/8 |
| 2 | Gemini 1.5 Pro API-0409-Preview | 1250 | +3/-3 | 39482 | Google | Proprietary | 2023/11 |
| 2 | GPT-4-0125-preview | 1247 | +3/-2 | 67354 | OpenAI | Proprietary | 2023/12 |
| 6 | Llama-3-70b-Instruct | 1210 | +3/-4 | 53404 | Meta | Llama 3 Community | 2023/12 |
| 6 | Bard (Gemini Pro) | 1209 | +5/-6 | 12387 | Google | Proprietary | Online |
| 7 | Claude 3 Sonnet | 1201 | +2/-3 | 78956 | Anthropic | Proprietary | 2023/8 |
| 9 | Command R+ | 1191 | +3/-3 | 44988 | Cohere | CC-BY-NC-4.0 | 2024/3 |
| 9 | GPT-4-0314 | 1190 | +3/-4 | 52079 | OpenAI | Proprietary | 2021/9 |
| 11 | Claude 3 Haiku | 1181 | +2/-3 | 69660 | Anthropic | Proprietary | 2023/8 |

Image from: arena.lmsys.org

# Open Source vs Open Weights

## : Accelerating the Science of Language Models


Dirk Groeneveld[α] Iz Beltagy[α]

Pete Walsh[α] Akshita Bhagia[α] Rodney Kinney[α] Oyvind Tafjord[α]

Ananya Harsh Jha[α] Hamish Ivison[αβ] Ian Magnusson[α] Yizhong Wang[αβ]

Shane Arora[α] David Atkinson[α] Russell Authur[α] Khyathi Raghavi Chandu[α]
Arman Cohan[γα] Jennifer Dumas[α] Yanai Elazar[αβ] Yuling Gu[α]
Jack Hessel[α] Tushar Khot[α] William Merrill[δ] Jacob Morrison[α]
Niklas Muennighoff Aakanksha Naik[α] Crystal Nam[α] Matthew E. Peters[α]
Valentina Pyatkin[αβ] Abhilasha Ravichander[α] Dustin Schwenk[α] Saurabh Shah[α]
Will Smith[α] Emma Strubell[αμ] Nishant Subramani[α] Mitchell Wortsman[β]

Pradeep Dasigi[α] Nathan Lambert[α] Kyle Richardson[α]
Luke Zettlemoyer[β] Jesse Dodge[α] Kyle Lo[α] Luca Soldaini[α]

Noah A. Smith[αβ] Hannaneh Hajishirzi[αβ]

[α]Allen Institute for Artificial Intelligence
[β]University of Washington [γ]Yale University
[δ]New York University [μ]Carnegie Mellon University

olmo@allenai.org

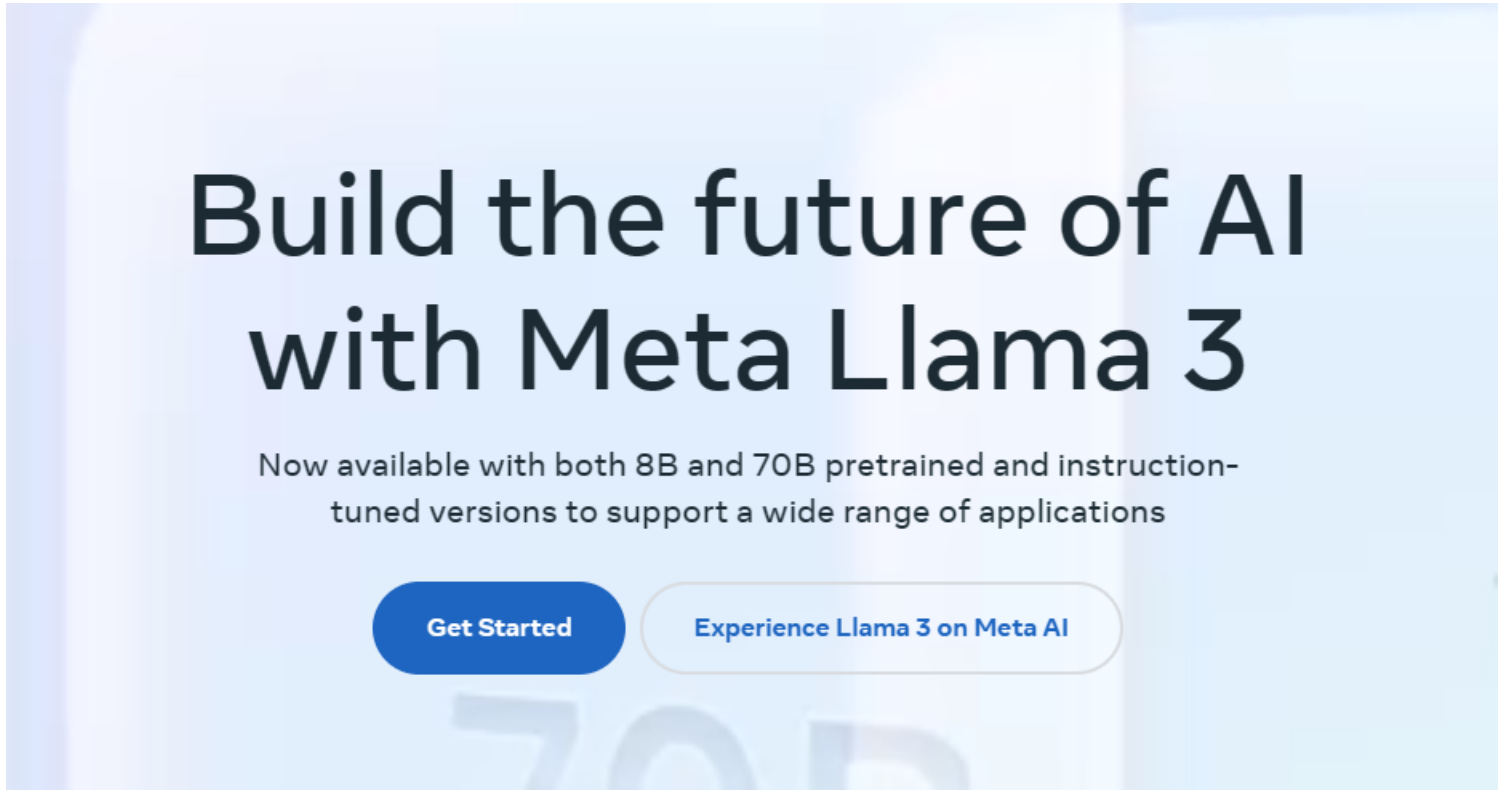


### Abstract

Language models (LMs) have become ubiquitous in both NLP research and in commercial product offerings. As their commercial importance has surged, the most powerful models have become closed off, gated behind proprietary interfaces, with important details of their training data, architectures, and development undisclosed. Given the importance of these details in scientifically studying these models, including their biases and potential risks, we believe it is essential for the research community to have access to powerful, truly open LMs. To this end, this technical report details the first release of OLMo, a state-of-the-art, truly **O**pen **L**anguage **Mo**del and its framework to build and study the science of language modeling. Unlike most prior efforts that have only released model weights and inference code, we release OLMo and the whole framework, including training data and training and evaluation code. We hope this release will empower and strengthen the open research community and inspire a new wave of innovation.


| | | |
|---|---|---|
| 🤗 | **Weights** | https://huggingface.co/allenai/OLMo-7B |
| | **Code** | https://github.com/allenai/OLMo |
| 🤗 | **Data** | https://huggingface.co/datasets/allenai/dolma |
| | **Evaluation** | https://github.com/allenai/OLMo-Eval |
| | **Adaptation** | https://github.com/allenai/open-instruct |
| | **W&B Logs** | https://wandb.ai/ai2-llm/OLMo-7B/reports/OLMo-7B--Vml |

# New entry within Open Weights LLMs by Meta




- Llama 3 released on April 18th
- 8B Model 70B
- LLaMa 3 license: royalty free if Monthly active users < 700M and not to be used to improve other LLMs
- **400B Model currently under training, planned weight release**



Image from: llama.meta.com/llama3/

# Impressive in English

## LMSYS Chatbot Arena Leaderboard

Category: English

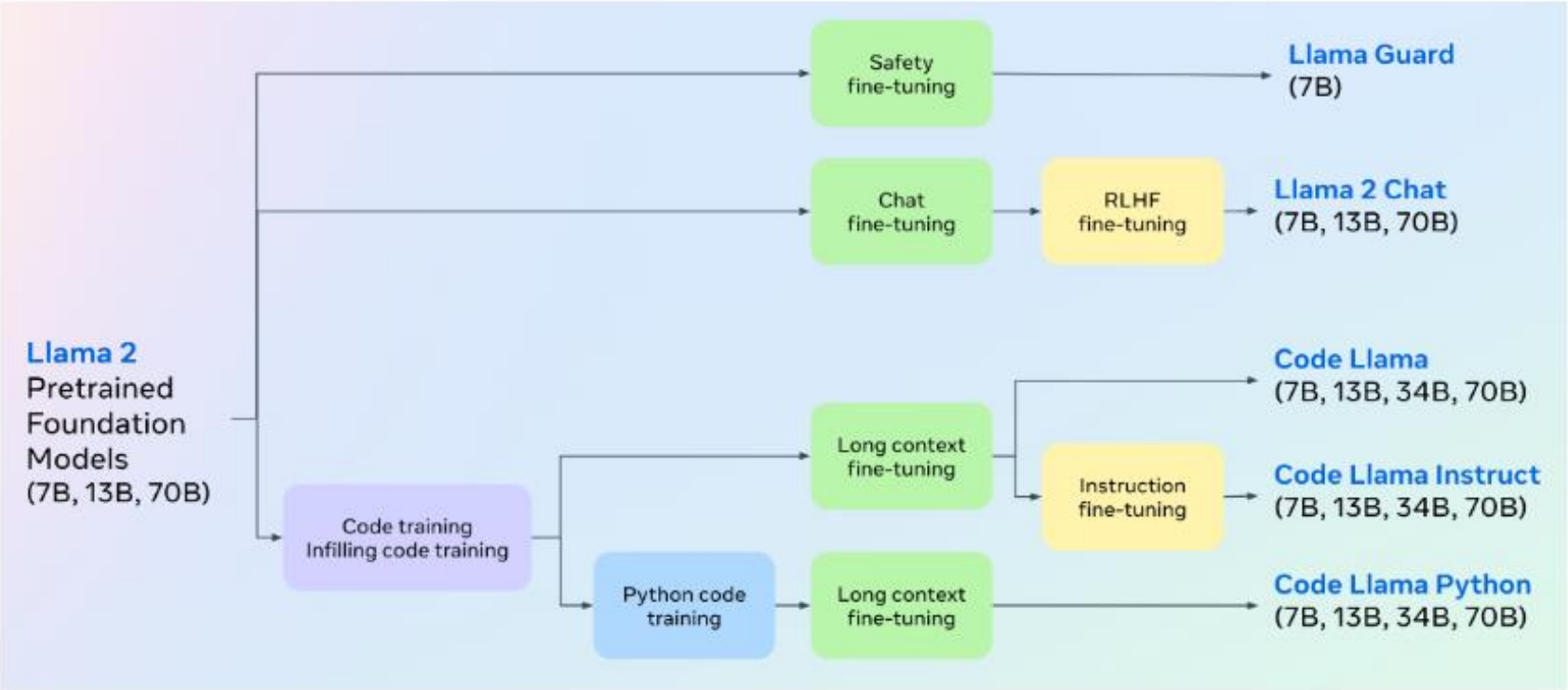
English Prompts

#models: 92 (100%) #votes: 579,014 (64%)

| Rank* (UB) | Delta | Model | Arena Elo | 95% CI | Votes | Organization | License | Knowledge Cutoff |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | GPT-4-Turbo-2024-04-09 | 1246 | +5/-4 | 19999 | OpenAI | Proprietary | 2023/12 |
| 2 | 0 | GPT-4-1106-preview | 1235 | +4/-3 | 49485 | OpenAI | Proprietary | 2023/4 |
| 2 | 0 | Gemini 1.5 Pro API-0409-Preview | 1232 | +5/-4 | 22566 | Google | Proprietary | 2023/11 |
| 2 ↑ | 4 | Llama-3-70b-Instruct | 1231 | +4/-5 | 32039 | Meta | Llama 3 Community | 2023/12 |
| 2 | 0 | GPT-4-0125-preview | 1230 | +4/-3 | 39443 | OpenAI | Proprietary | 2023/12 |
| 6 ↓ | -4 | Claude 3 Opus | 1218 | +4/-4 | 45494 | Anthropic | Proprietary | 2023/8 |
| 7 ↓ | -1 | Bard (Gemini Pro) | 1183 | +7/-7 | 9601 | Google | Proprietary | Online |
| 7 | 0 | Claude 3 Sonnet | 1176 | +4/-4 | 43302 | Anthropic | Proprietary | 2023/8 |
| 9 | 0 | GPT-4-0314 | 1166 | +4/-4 | 33659 | OpenAI | Proprietary | 2021/9 |
| 9 | 0 | Command R+ | 1164 | +4/-4 | 23641 | Cohere | CC-BY-NC-4.0 | 2024/3 |
| 9 ↑ | 4 | Llama-3-8b-Instruct | 1163 | +4/-4 | 31389 | Meta | Llama 3 Community | 2023/3 |
| 10 ↑ | 1 | Claude 3 Haiku | 1159 | +4/-4 | 37016 | Anthropic | Proprietary | 2023/8 |

Image from: arena.lmsys.org

# Which Llama2?




- 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

image from llama.meta.com/llama2/

# Llama 2 Training

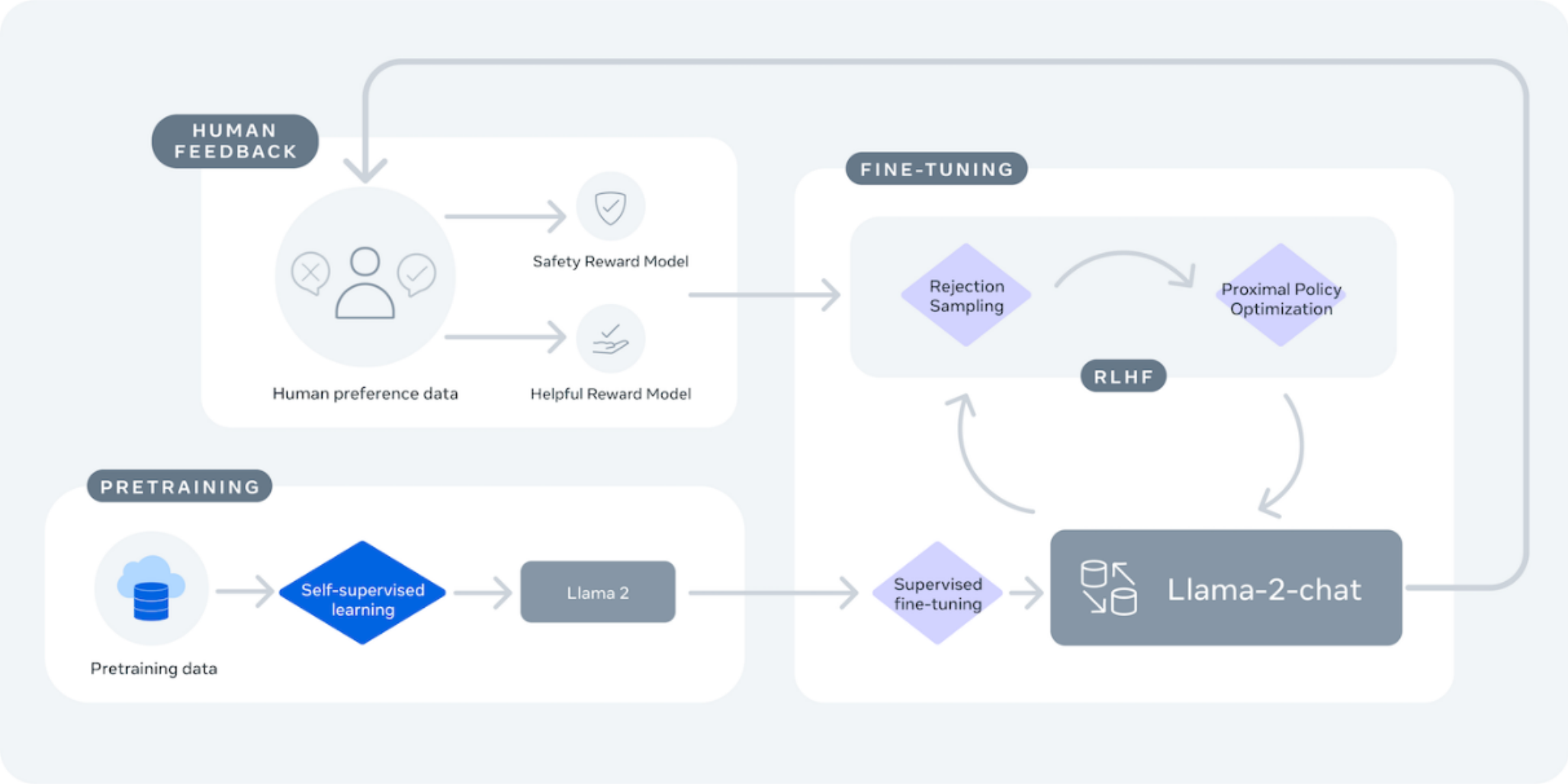



image from llama.meta.com/llama2/

# Optimizing Llama 2 Inference

image from llama.meta.com/llama2/

# Let’s continue the discussion!

If you want to talk more about LLMs feel free to add me on the event website or Linkedin or just come say hi in the next days :)

REINVENTING HPC

ISC High Performance

MAY 12 – 16, 2024 | HAMBURG, GERMANY

**Ivan Gentile**

Data Scientist at IFAB | Machine Learning & Big Data | Harnessing AI & HPC for Societal Impact

EUROCC ITALY HPC Competence Center

iFAB

# Acknowledgements

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.

Co-funded by the European Union


**Disclaimer**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European High Performance Computing Joint Undertaking (JU) and the participating countries in the project . Neither the European Union nor the granting authority can be held responsible for them.



EUROCC ITALY HPC Competence Center iFAB

# Questions?

**Follow IFAB**
and EuroCC Italy

Sito web www.euroccitaly.it

IFAB International Foundation (ifabfoundation.org)

International Foundation Big Data and Artificial Intelligence for Human Development

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.

<<<<<<<<<<