

Beyond ChatGPT: Open-source Aligned LLMs

Daniel Hershcovich

UNIVERSITY OF COPENHAGEN



What ChatGPT is(n't)

It is a

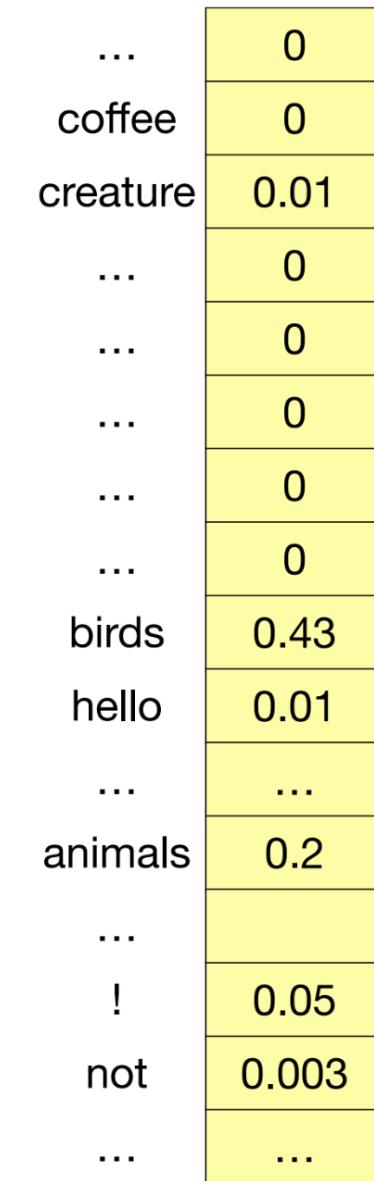
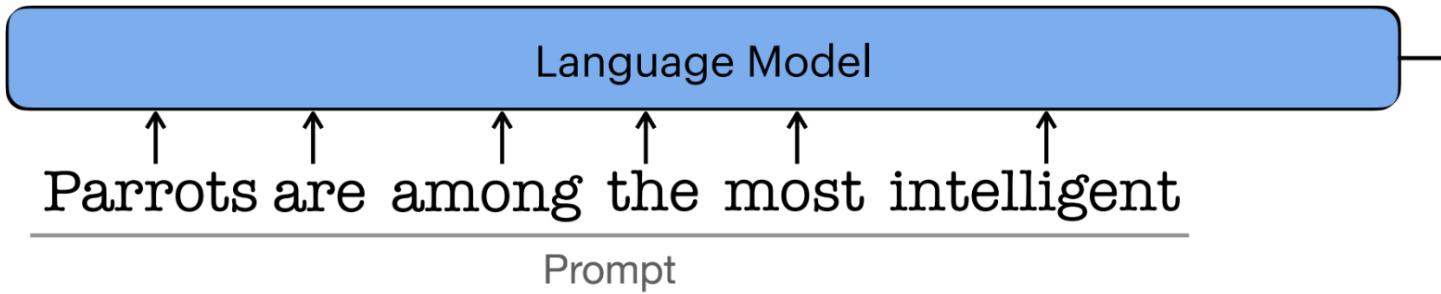
- Free/paid service by OpenAI
- Frontend web interface
- Backend powered by
 - **LLM** (large language model) which is a **GPT** (generative pre-trained transformer) fine-tuned ("**aligned**") for **chat**/instruction following

It is NOT

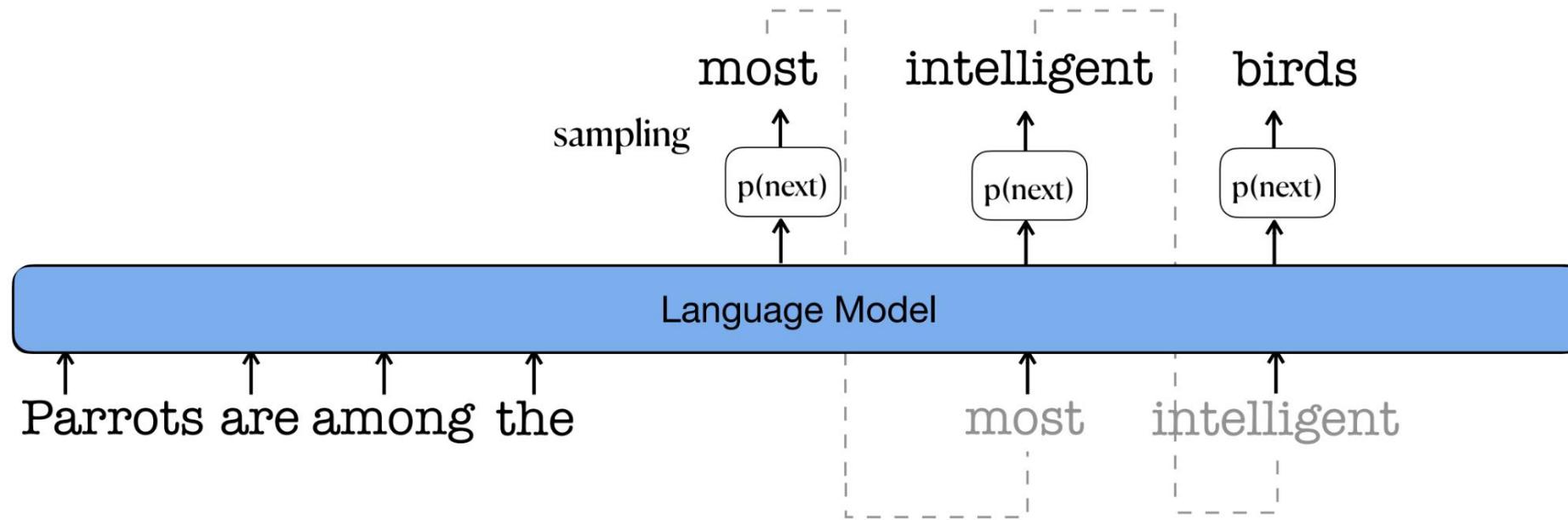


- Trustworthy or reliable (due to, e.g., "hallucinations")
- Up-to-date (trained in 2021/2023)
- Open source/model/data nor transparent
- Particularly environmentally friendly

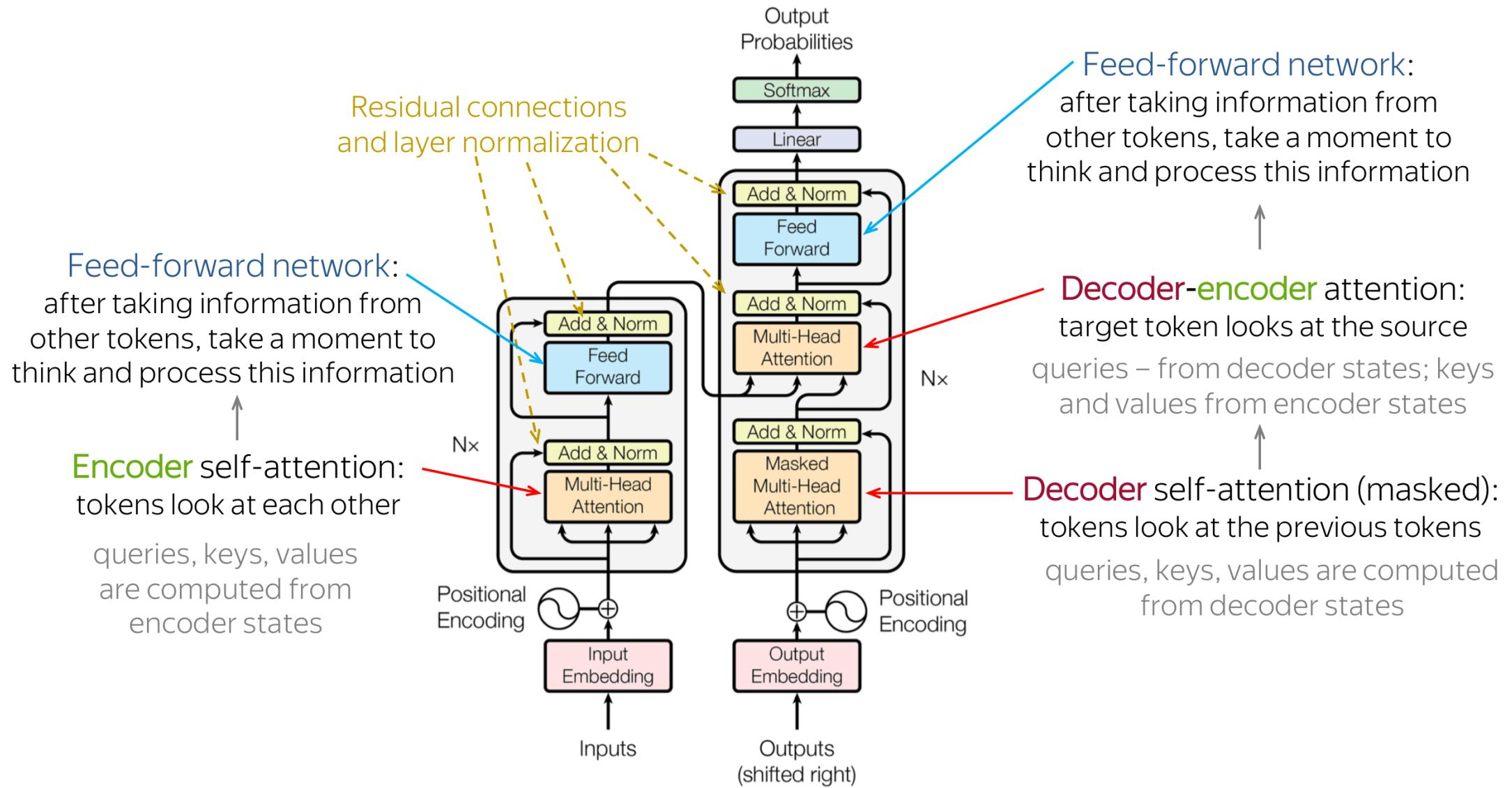
Language Models



Generative (decoder) LMs



Transformer



Pre-training *base* models

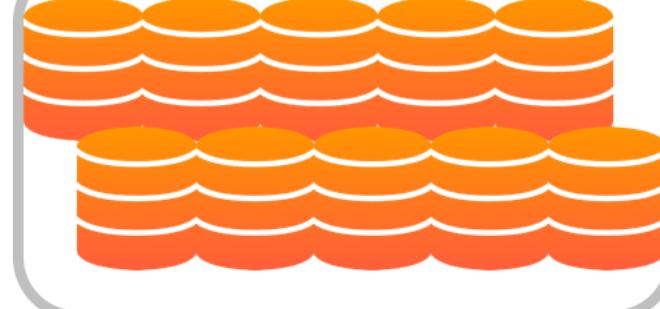
Close to scale. Selected highlights only. Using images from FlatIcon.com. Alan D. Thompson, April 2023. <https://lifearchitect.ai>



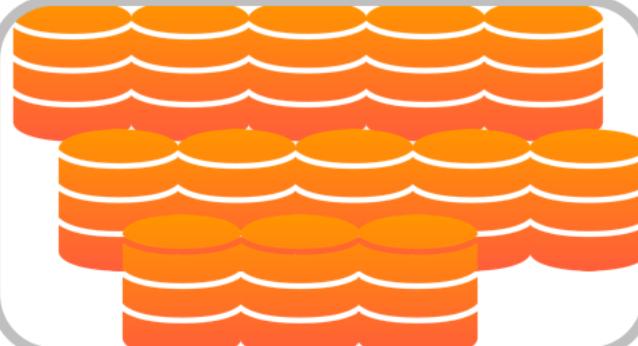
GPT-3 dataset
499B tokens / 0.75TB



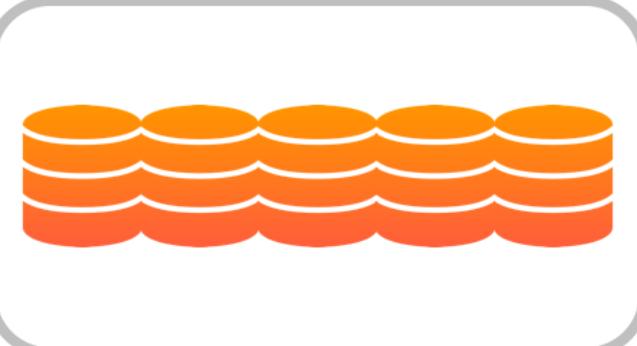
The Pile dataset (GPT-Neo)
247B tokens / 0.8TB



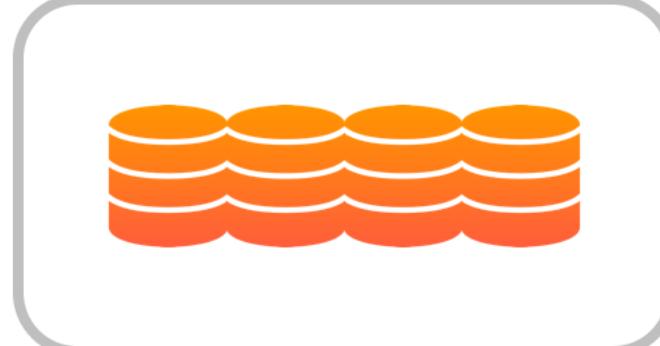
MassiveText dataset (Chinchilla)
2.3T tokens / 10.5TB



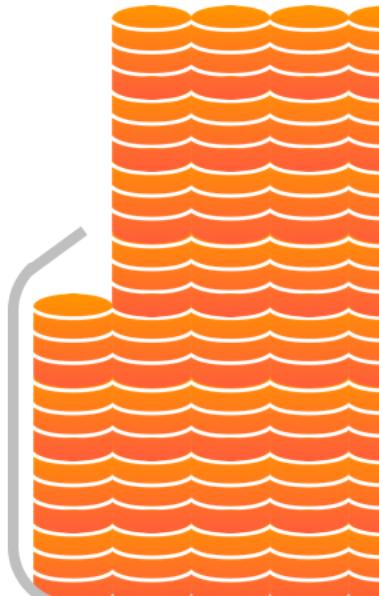
Infiniset dataset (LaMDA)
2.8T tokens / 12.6TB



Stability The Pile dataset
1.5T tokens / 5TB



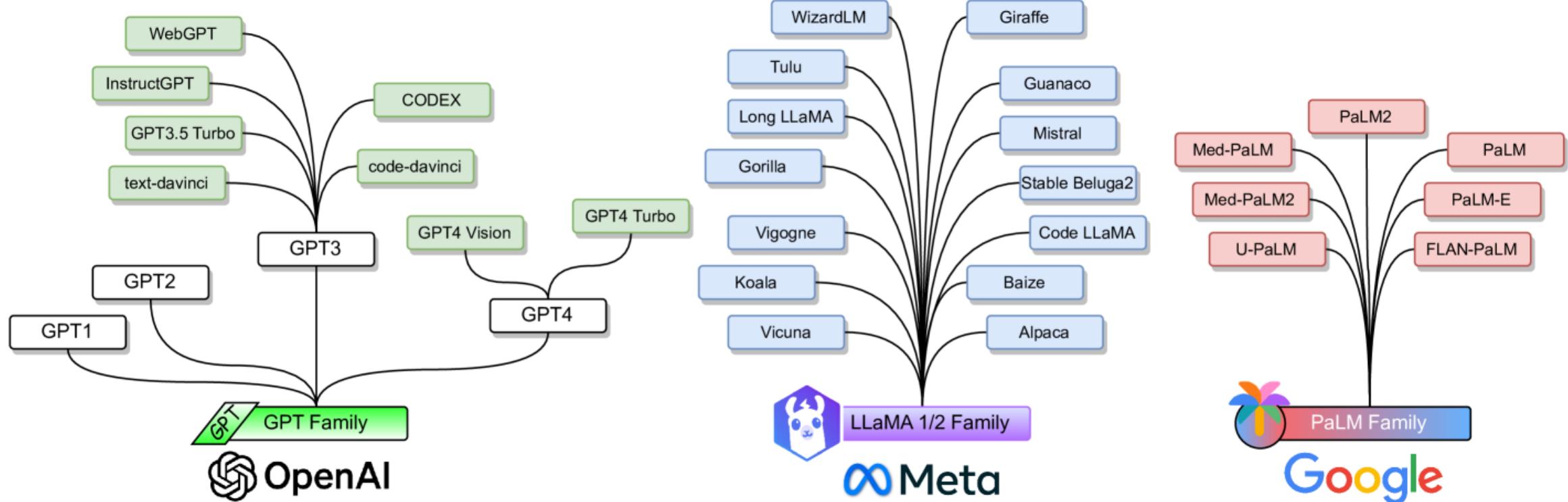
RedPajama dataset
1.2T tokens / 4TB



GPT-4 (estimate)
20T tokens / 40TB

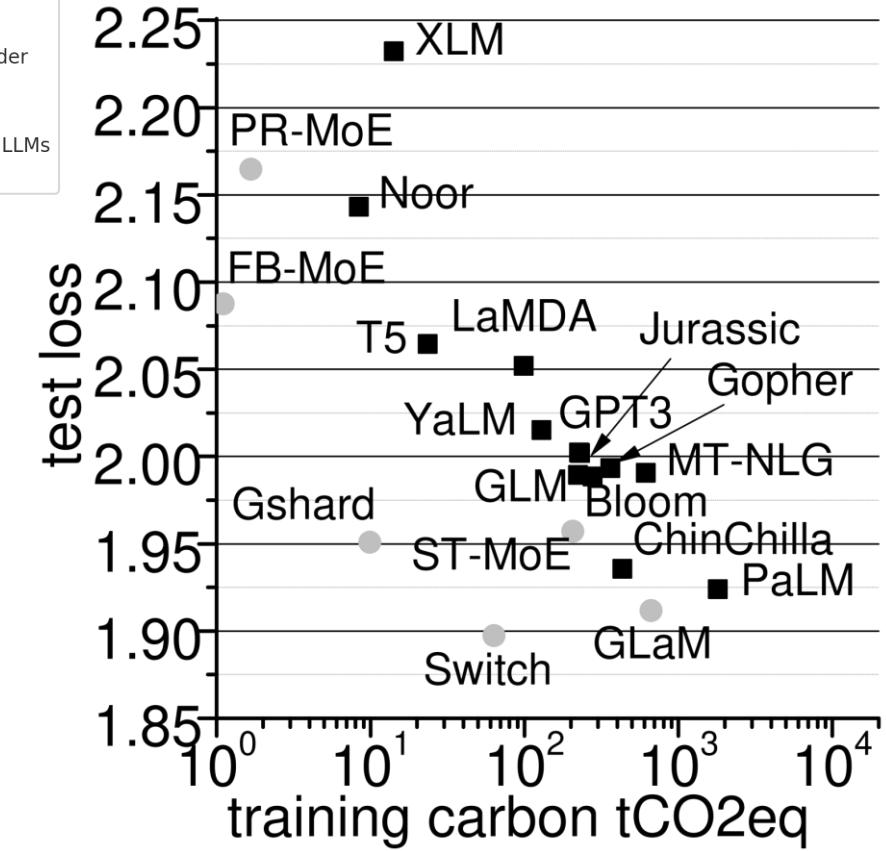


Popular LLM families



LLM size and emissions

Timeline of Model Sizes (in Millions of Parameters) by Category



Faiz, Ahmad, et al. "LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models." arXiv preprint arXiv:2309.14393 (2023).

Some definitions for “alignment” of models

- **Alignment:** General notion of training a model to mirror user desires
- **Supervised fine-tuning (SFT):** Training to learn task-specific capabilities
- **Instruction fine-tuning (IFT):** Training a model to follow instructions
- **Reinforcement learning from human feedback (RLHF):** Specific technical tool for training ML models from human data
- **Preference fine-tuning:** Using labeled preference data to fine-tune a LM (either with RL, DPO, or another loss function)

RLHF in ChatGPT

Step 1

Collect demonstration data
and train a supervised policy.

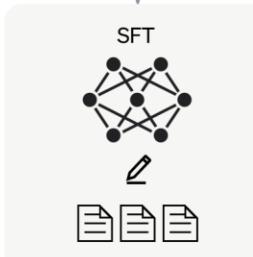
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



This data is used to
fine-tune GPT-3.5
with supervised
learning.



Provide additional feedback

What would the ideal answer have been?

- This is harmful / unsafe
- This isn't true
- This isn't helpful

Submit feedback

Aligning open language models

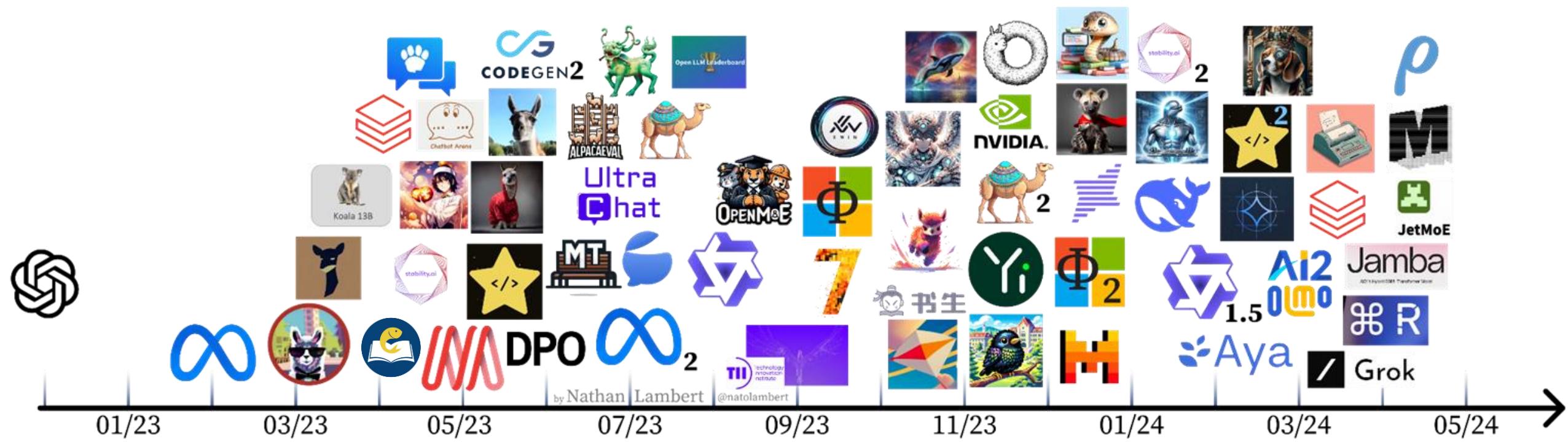
1: instruction tuning blooms

2: evals & expectations

3: RLHF works!

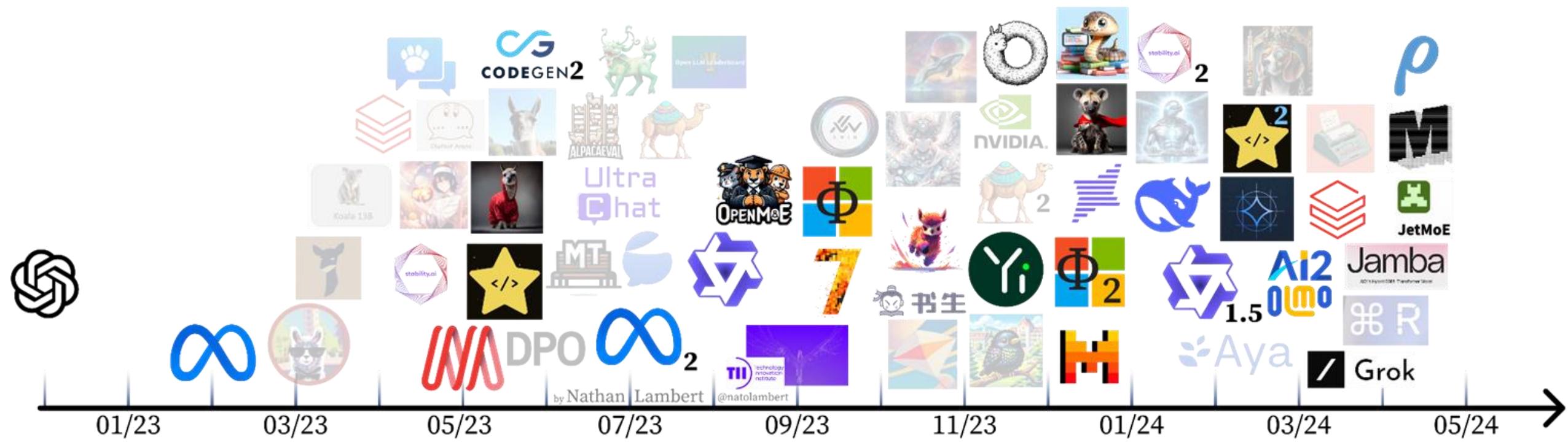
4. expansion

Aligning open language models (Nathan Lambert, April 2024)



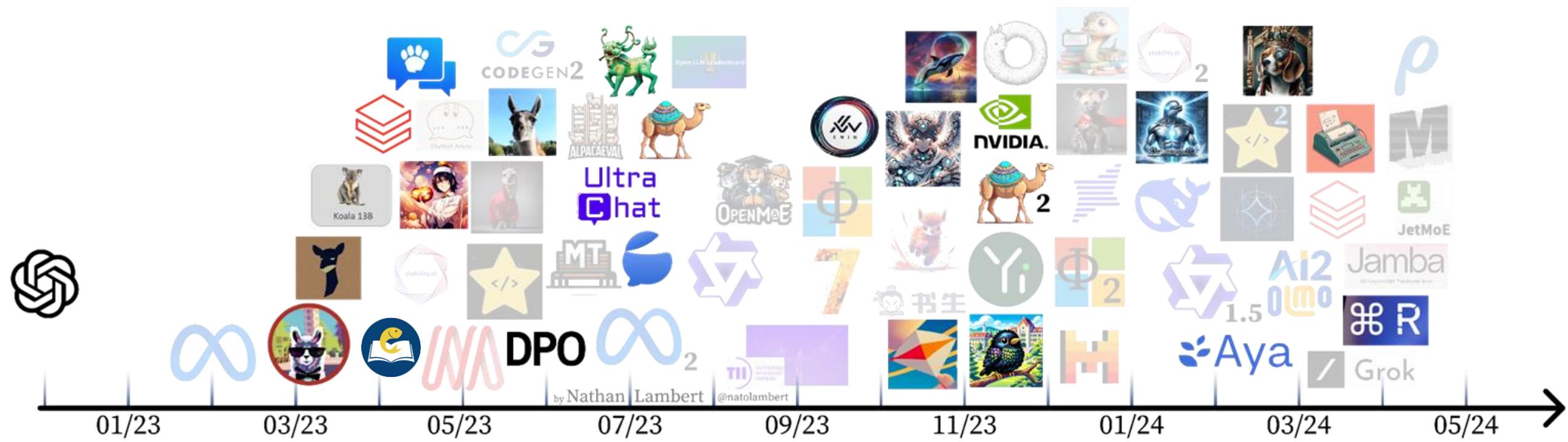
Base models

Aligning open language models (Nathan Lambert, April 2024)

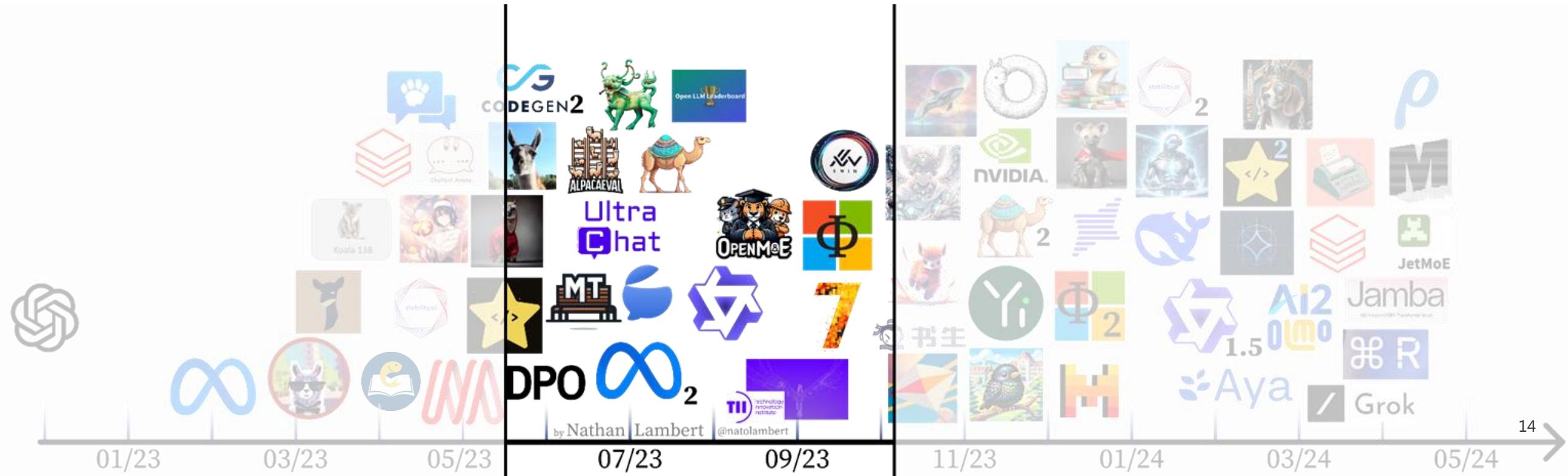


Aligned / fine-tuned / preference trained models

Aligning open language models (Nathan Lambert, April 2024)



Setting expectations & evaluations



Prehistory: MMLU benchmark

September 2020

- Massive Multitask Language Understanding
- 57 tasks including elementary mathematics, US history, computer science, law
- **Only multiple-choice questions**

Hendrycks, Dan, et al. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).



Rank	Model	Average (%)	Extra Training Data	Paper	Code	Result	Year	Tags
1	Gemini Ultra ~1760B	90	✗	Gemini: A Family of Highly Capable Multimodal Models	🔗	🔗	2023	self-consistency chain-of-thought
2	GPT-4o	88.7	✗	GPT-4 Technical Report	🔗	🔗	2023	
3	Claude 3 Opus (5-shot, CoT)	88.2	✗	The Claude 3 Model Family: Opus, Sonnet, Haiku	🔗	🔗	2024	few-shot
4	Claude 3 Opus (5-shot)	86.8	✗	The Claude 3 Model Family: Opus, Sonnet, Haiku	🔗	🔗	2024	
5	Leeroo (5-shot)	86.64	✗	Leeroo Orchestrator: Elevating LLMs Performance Through Model Integration	🔗	🔗	2024	
6	GPT-4 (few-shot)	86.4	✓	GPT-4 Technical Report	🔗	🔗	2023	few-shot

Establishing evaluation for *generative* models

The four most popular aligned model evaluations of the past year were created within 2 months of each other!

1. May 3rd, 2023: ChatBotArena
2. June 8th, 2023: AlpacaEval
3. June 22nd, 2023: MT Bench
4. July 2023: Open LLM Leaderboard

ChatBotArena

Side by side
preference
collection of
two different
models

The screenshot shows the ChatBotArena interface on a web browser. The title bar reads "Chat with Open Large Language Model". The URL is "chat.lmsys.org". The main content area has a header "Arena (battle) Arena (side-by-side) Direct Chat Leaderboard About Us". Below this is a section titled "Chatbot Arena : Benchmarking LLMs in the Wild" with links to "Blog | GitHub | Paper | Dataset | Twitter | Discord". A "Rules" section lists three points: "Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!", "You can continue chatting until you identify a winner.", and "Vote won't be counted if model identity is revealed during conversation.". A "Arena Elo Leaderboard" section mentions using 100K human votes to compile an Elo-based LLM leaderboard. A "Chat now!" button is present. The main feature is a side-by-side comparison of two models, Model A and Model B, both answering the question "What is the meaning of life?". Model A's response is: "The meaning of life is a subjective and philosophical question that has been debated for centuries. Different people, cultures, and belief systems have varying interpretations of the meaning of life. Some believe it is to find happiness, fulfillment, or purpose, while others may argue that it is to form connections and relationships, or to contribute to the greater good. Ultimately, the meaning of life is a deeply personal and individual question, and each person is likely to have their own unique answer." Model B's response is: "The meaning of life is a philosophical question that has puzzled humanity for centuries and may have different answers depending on individual beliefs, cultural backgrounds, and philosophical viewpoints. Here are a few perspectives on the meaning of life: 1. Religious and spiritual perspectives: Many religious traditions offer their own answers to the meaning of life. For example, in many forms of Christianity, the meaning of life is to know, love, and serve God in order to achieve eternal happiness with Him in Heaven. In Buddhism, the purpose of life is to end suffering by reaching enlightenment through the practice of the Noble Eightfold Path. 2. Existentialist perspective: Existentialist philosophers like Jean-Paul Sartre and Albert Camus suggest that life has no inherent meaning, and it is up to each individual to create their own purpose through their choices and actions. 3. Humanist perspective: Humanists often believe that the meaning of life is to seek happiness and fulfillment, to improve the human condition, and to make a positive impact on the lives of others and the world around us." At the bottom, there is a text input field with placeholder "Enter your prompt and press ENTER", a "Send" button, and buttons for "New Round", "Regenerate", and "Share". The footer includes "Parameters", "Acknowledgment" (mentioning Kaggle, MBZUAI, AnyScale, and HuggingFace), and logos for kaggle, AnyScale, and Hugging Face.

LMSYS Org

The Large Model Systems Organization develops large models and systems that are open, accessible, and scalable.

Aligning open language models (Nathan Lambert, April 2024)

AlpacaEval

LLM-as-a-judge
mirroring
preference
collection phase

The screenshot shows a web browser window titled "Alpaca Eval Leaderboard" with the URL "tatsu-lab.github.io/alpaca_eval/". The page features a logo of a cartoon rabbit wearing a graduation cap. The title "AlpacaEval Leaderboard" is displayed above a subtitle "An Automatic Evaluator for Instruction-following Language Models". A cautionary note at the top right states: "Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs." Below this, there are two filter sections: "Evaluator" (with "GPT-4" selected) and "Filter" (with "Community" selected). The main content is a table listing 15 language models, each with a thumbnail icon, model name, win rate, and length.

Model Name	Win Rate	Length
GPT-4 Turbo	97.70%	2049
XwinLM 70b V0.1	95.57%	1775
PairRM+Tulu 2+DPO 70B (best-of-16)	95.40%	1607
GPT-4	95.28%	1365
Tulu 2+DPO 70B	95.03%	1418
Yi 34B Chat	94.08%	2123
PairRM+Zephyr 7B Beta (best-of-16)	93.41%	1487
LLaMA2 Chat 70B	92.66%	1790
UltraLM 13B V2.0 (best-of-16)	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
UltraLM 13B (best-of-16)	91.54%	1980
Claude 2	91.36%	1069
PairRM+Tulu 2+DPO 13B (best-of-16)	91.06%	1454
Cohere Command	90.62%	1983
Zenhvr 7B Beta	90.60%	1444

MT Bench

LLM-as-a-judge: ask an LLM (GPT4/Claude) to rate a model response

Open LLM Leaderboard

Evaluates almost any model on HuggingFace on core LLM tasks

RLHF has not been shown to improve these metrics deeply, *starting to get better in 2024*

The screenshot shows the 'Open LLM Leaderboard' page on the HuggingFace platform. The top navigation bar includes links for 'App', 'Files', 'Community', and a user profile. The main title is 'Open LLM Leaderboard' with a subtitle: 'The 🤖 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.' Below this, there's a note: 'Submit a model for automated evaluation on the 🤖 GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!' The interface features several filter sections: 'LLM Benchmark' (selected), 'Metrics through time', 'About', and 'Submit here!'. It also includes search and column selection tools. The main content area displays a table of models with their names, URLs, and performance metrics across various benchmarks. The table includes columns for Average, ARC, HellaSwag, MMLU, and other metrics.

Model	Average	ARC	HellaSwag	MMLU
upstage/SOLAR-10.7B-Instruct-v1.0	74.2	71.08	88.16	66.21
rishiraj/meow	73.94	70.48	88.08	66.25
ValiantLabs/ShiningValiant	73.78	71.33	90.96	71.21
sequelbox/SunsetBoulevard	73.78	71.33	90.96	71.21
Qwen/Qwen-72B	73.6	65.19	85.94	77.37
SUSTech/SUS-Chat-34B	73.22	66.3	83.91	76.41
Sao10K/SOLAR-10.7B-NahIdWin	73.21	64.51	85.67	64.17
EmbeddedLLM/Mistral-7B-Merge-14-v0.2	72.77	69.11	86.89	64.84
mistralai/Mistral-8x7B-Instruct-v0.1	72.62	70.22	87.63	71.16
AIDC-ai-business/Marcoroni-7B-v3	72.53	69.45	86.78	65
Toten5/Marcoroni-v3-neural-chat-v3-3-Sleep	72.51	68.77	86.55	64.51
Toten5/Marcoroni-neural-chat-7B-v2	72.5	68.6	86.33	64.65

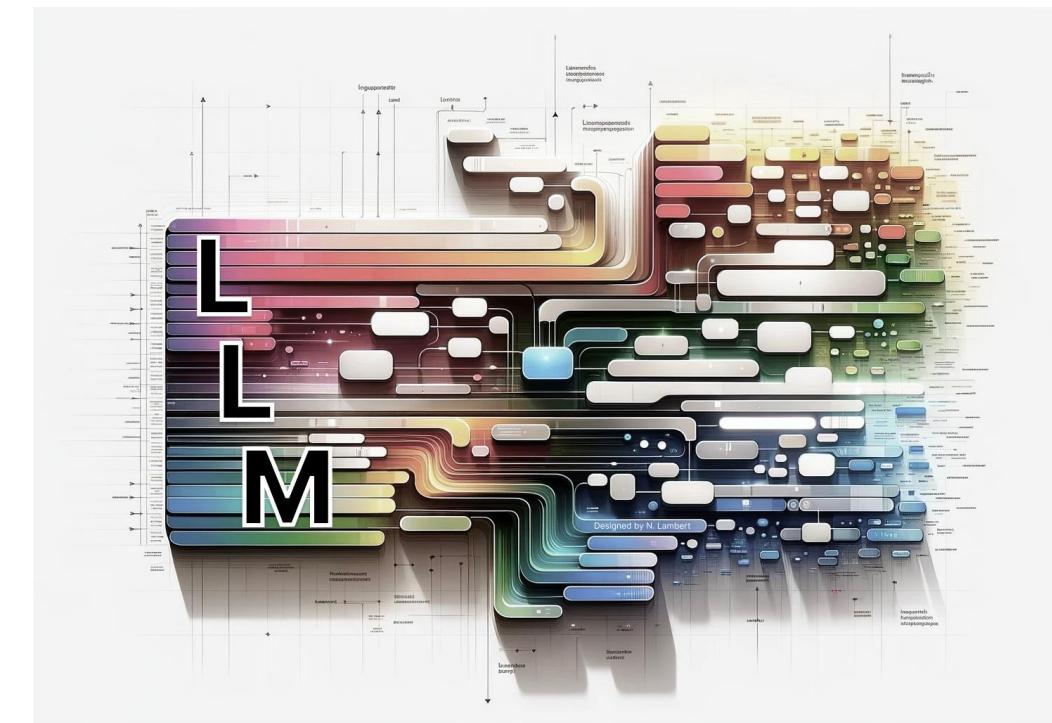
The first open instruct models

[Aligning open language models \(Nathan Lambert, April 2024\)](#)



Key idea: Instruction fine-tuning (IFT)

1. Adapt base model to **specific style of input**
 2. Ability to include system prompts, multi-turn dialogues, and other **chat templates**



Key idea: Self-instruct / synthetic data

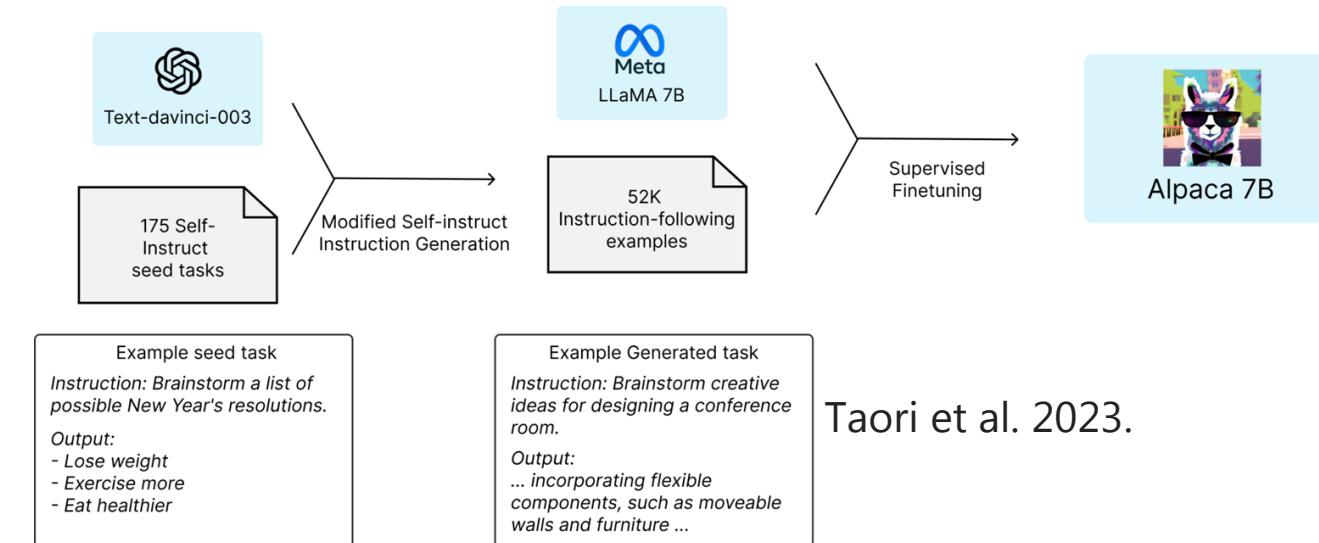
Start: high-quality (often human) prompts

Ask a strong LM: Create a modified version of these instructions

Generate completions with another (or same) strong LM

End: easily 10x more (synthetic) training data!

(synthetic data = text generated by another LLM)



Self-instruct: Wang et al. 2022

First open instruction tuned models



Alpaca

13 Mar. 2023

- 52k self-instruct style data distilled from text-davinci-003
- Model weight diff. to LLaMA 7B

<https://cfrm.stanford.edu/2023/03/13/alpaca.html>

MT Bench 13B: 4.53



MT Bench 7B: 6.69

Vicuna ([lmsys/vicuna-7b-delta-v0](#))

30 Mar. 2023

- Fine-tunes ChatGPT data from ShareGPT
- LLaMA 7B and 13B diff's
- Introduces LLM-as-a-judge

<https://lmsys.org/blog/2023-03-30-vicuna/>



Koala

3 Apr. 2023

- Diverse dataset (Alpaca, Anthropic HH, ShareGPT, WebGPT...)
- Human evaluation
- LLaMA 7B diff.

<https://bair.berkeley.edu/blog/2023/04/03/koala/>

MT Bench 13B: 6.08



Dolly

12 Apr. 2023

- 15k human written data
- Trained on Pythia 12b

<https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-lm>

OpenAssistant: The first open, human instruction dataset

*"In an effort to democratize research on large-scale alignment, we release OpenAssistant Conversations (OASST1), a human-generated, human-annotated assistant-style conversation corpus consisting of **161,443 messages** in **35 different languages**, annotated with 461,292 quality ratings, resulting in over 10,000 fully annotated conversation trees. The corpus is a product of a worldwide crowd-sourcing effort involving over **13,500 volunteers**."*

April 15th, 2023

- Used extensively in future models.
- Still the only human dataset of this size to be released.
- OpenAssistant and others trained the popular models with it.
- (released fine-tuned models too!)



Dataset: <https://huggingface.co/datasets/OpenAssistant/oasst1>

Paper: <https://arxiv.org/abs/2304.07327>

[Aligning open language models \(Nathan Lambert, April 2024\)](#)

StableVicuna

28th April, 2023



The first open source RLHF model

Trained with proximal policy optimization (PPO) on popular datasets

- OAsst1 dataset for SFT + PPO
- Anthropic HH + Stanford Human Preferences (SHP) for RL

QLoRA

LoRA: Low Rank Adaptation

Popular tool for fine-tuning models with lower memory consumption.

QLoRA: LoRA + quantized base model (+ paging and double quantization)

Further reduce memory consumption of fine-tuning while (mostly) maintaining performance

Method	Bits	7B	13B	30B	70B	8x7B	8x22B
Full	AMP	120GB	240GB	600GB	1200GB	900GB	2400GB
Full	16	60GB	120GB	300GB	600GB	400GB	1200GB
Freeze	16	20GB	40GB	80GB	200GB	160GB	400GB
LoRA/GaLore/BAdam	16	16GB	32GB	64GB	160GB	120GB	320GB
QLoRA	8	10GB	20GB	40GB	80GB	60GB	160GB
QLoRA	4	6GB	12GB	24GB	48GB	30GB	96GB
QLoRA	2	4GB	8GB	16GB	24GB	18GB	48GB

Approximate VRAM requirements.

Source: <https://github.com/hiyouga/LLaMA-Factory#hardware-requirement>

Guanaco

23rd May, 2023

First models trained with QLoRA + quality filtered Open Assistant dataset.

Both the dataset and QLoRA method are still regularly used

MT Bench: 6.88

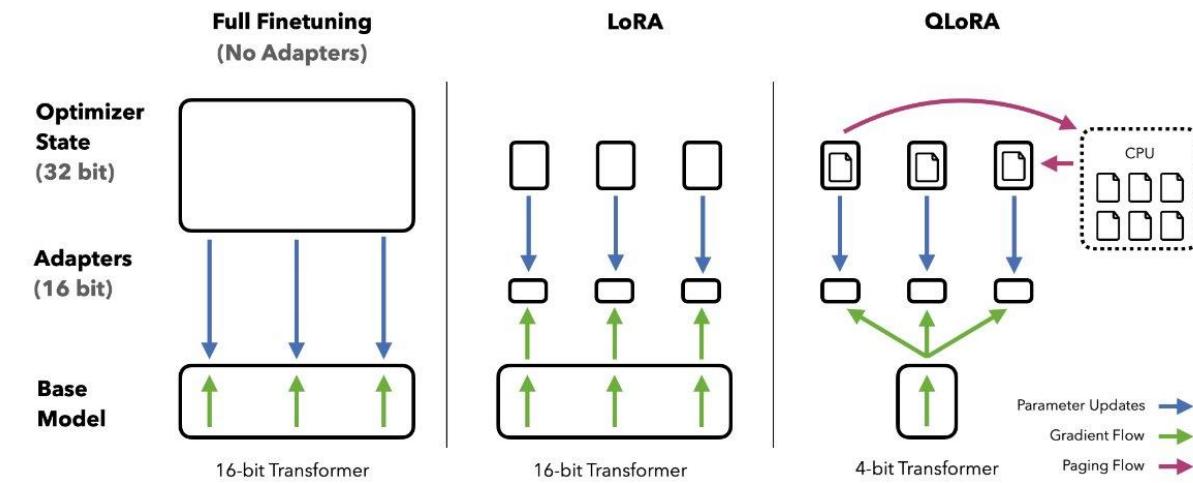


Image credit: Tim Dettmers

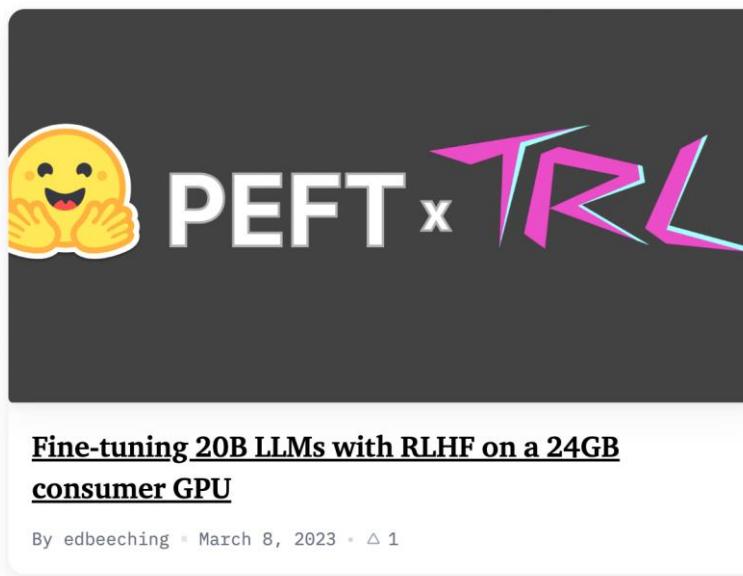
Paper: <https://arxiv.org/abs/2305.14314>

Dataset: <https://huggingface.co/datasets/timdettmers/openassistant-guanaco>

Model: <https://huggingface.co/timdettmers/guanaco-65b>

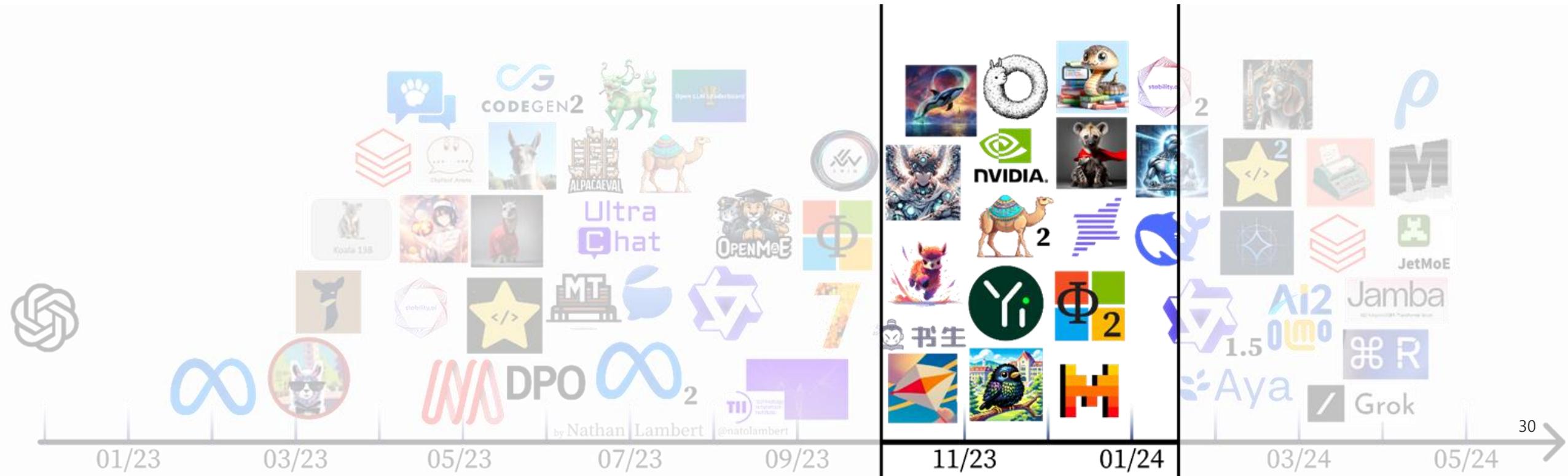
Do LoRA methods work with RL?

- Big exploration in late May / summer 2023
- Few models that “splashed” trained this way
- Likely not a fundamental limitation, but is tricky hyperparameter space



Getting RLHF to work

[Aligning open language models \(Nathan Lambert, April 2024\)](#)



RLHF without online humans in the loop

Direct Preference Optimization (DPO)

May 29th, 2023

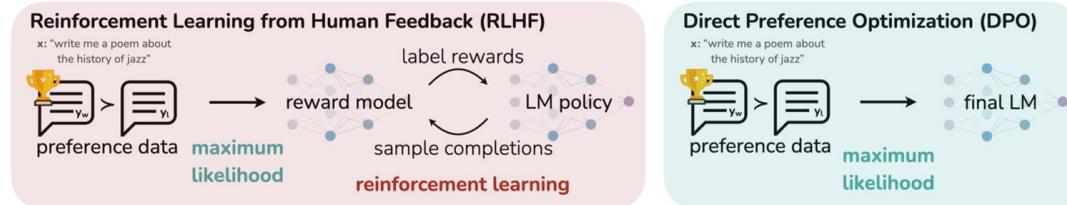


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, fitting an *implicit* reward model whose corresponding optimal policy can be extracted in closed form.

**Direct Preference Optimization:
Your Language Model is Secretly a Reward Model**

Rafael Rafailov^{*†} Archit Sharma^{*†} Eric Mitchell^{*†}
 Stefano Ermon^{†‡} Christopher D. Manning[†] Chelsea Finn[†]
[†]Stanford University [‡]CZ Biohub
{rafaillov, architsh, eric.mitchell}@cs.stanford.edu

Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper, we propose a new formulation of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

arXiv:2305.18290v2 [cs.LG] 13 Dec 2023

1 **Introduction**

Large unsupervised language models (LMs) trained on very large datasets acquire surprising capabilities [11, 7, 40, 8]. However, these models are trained on data generated by humans with a wide variety of goals, priorities, and skills. Since these goals and skills often do not align with what we desire them to do, for example, while we might want our AI to avoid common coding mistakes in order to correct them, nevertheless, when generating code, we would like to bias our model toward the (potentially rare) high-quality coding ability present in its training data. Similarly, we might want our language model to be aware of a common misconception believed by 50% of people, but we certainly do not want the model to claim this misconception to be true in 50% of queries about it! In other words, selecting the model's *desired responses and behavior* from its very wide *knowledge and abilities* is crucial to building AI systems that are safe, performant, and controllable [26]. While existing methods typically steer LMs to match human preferences using reinforcement learning (RL),

*Equal contribution; more junior authors listed earlier.
 †7th Conference on Neural Information Processing Systems (NeurIPS 2023).

RLHF phase: Zephyr β

- First model to make a splash with DPO!
- Fine-tune of Mistral 7B with UltraFeedback dataset
- MT Bench: 7.34



RLHF phase: Tülu 2

- First model to scale DPO to 70 billion parameters!
- Strongly validated the Zephyr results
- MT Bench: 7.89

Tülu v2

Open instruction & RLHF models



Modern ecosystem

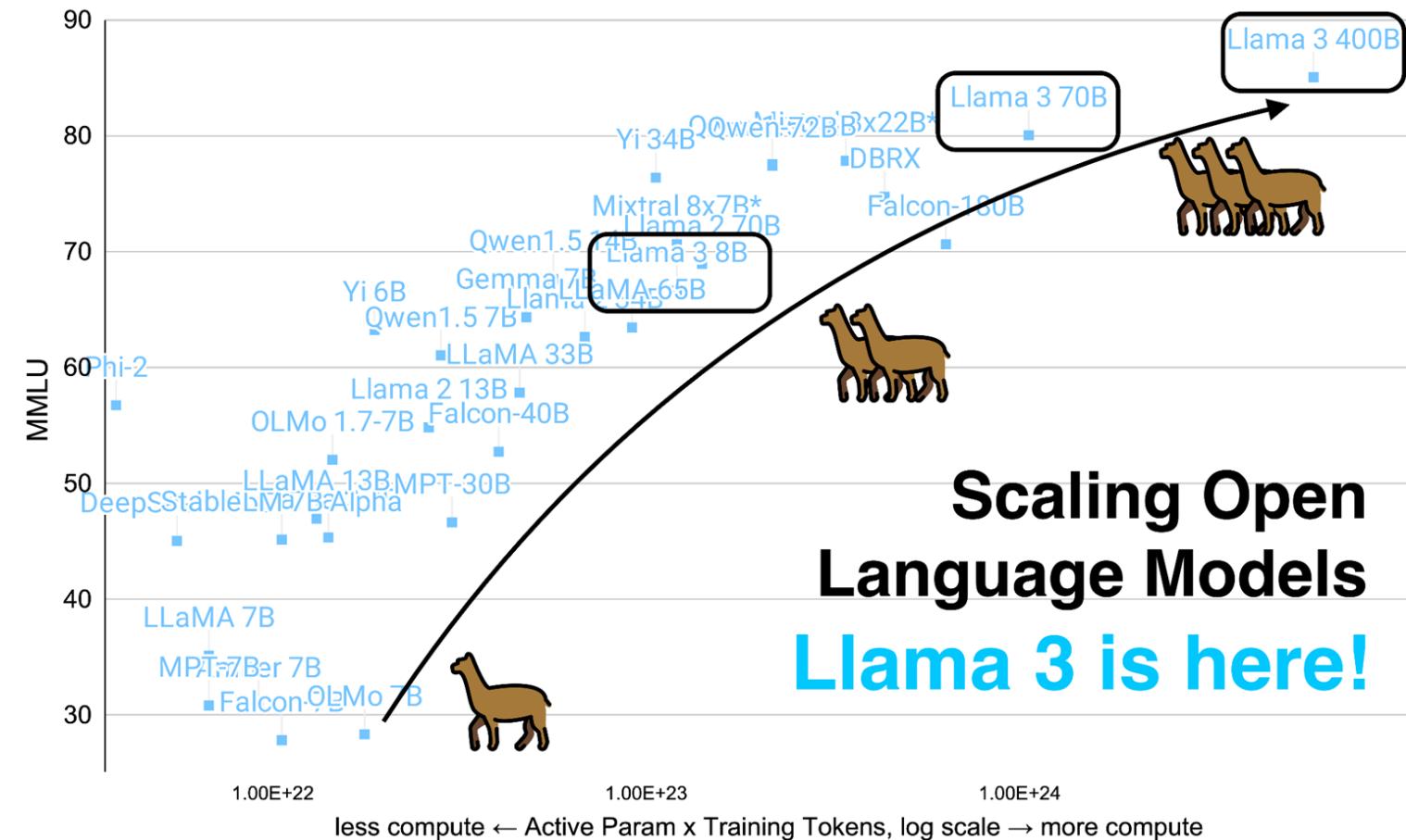
[Aligning open language models \(Nathan Lambert, April 2024\)](#)



Diversity of models and more players

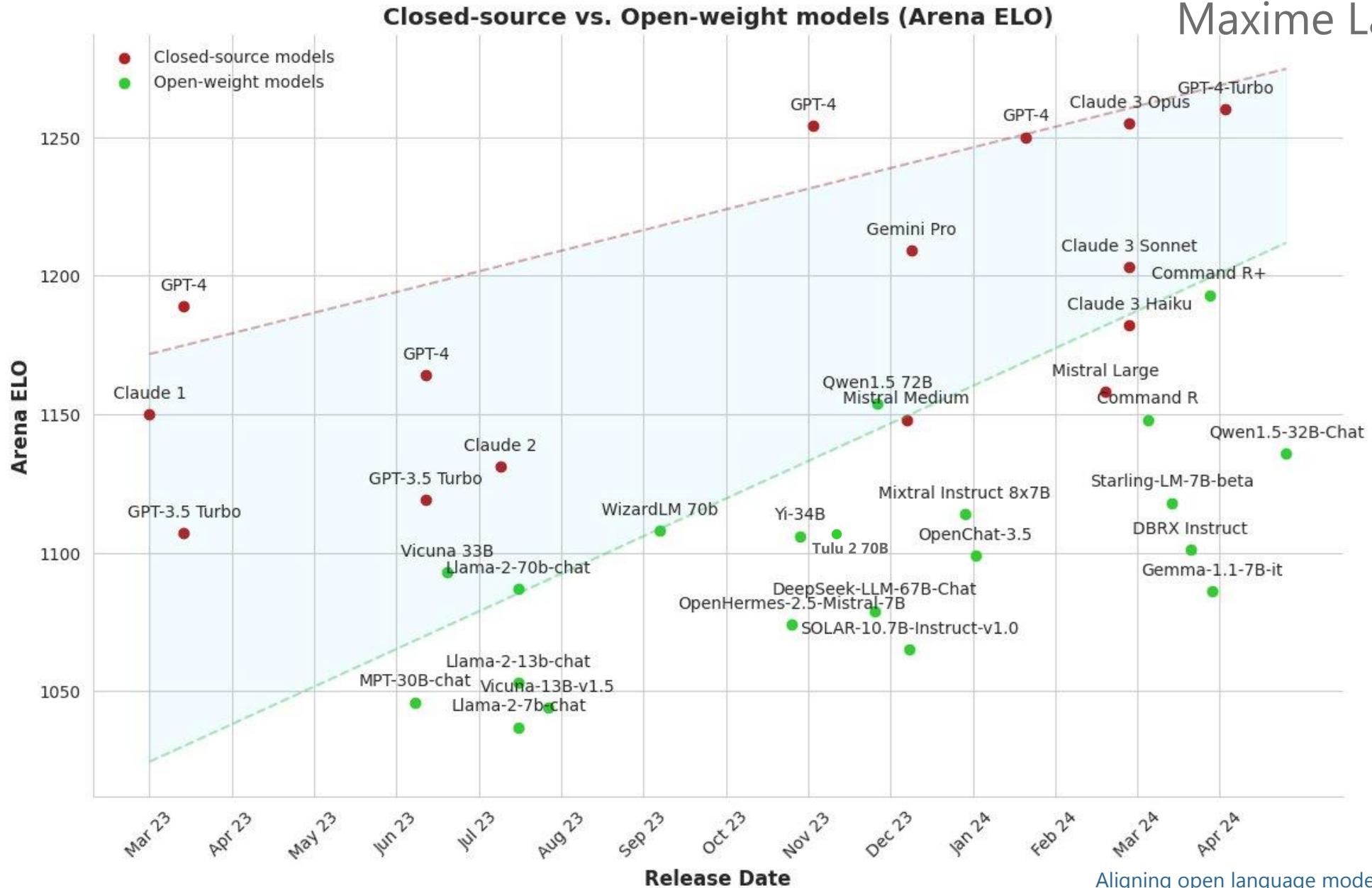
Examples:

- AI2's OLMo: truly *open-source* models
- Cohere's Command R+: first open model to pass GPT-4 on ChatBotArena
- Multilingual fine-tuning with Cohere's Aya



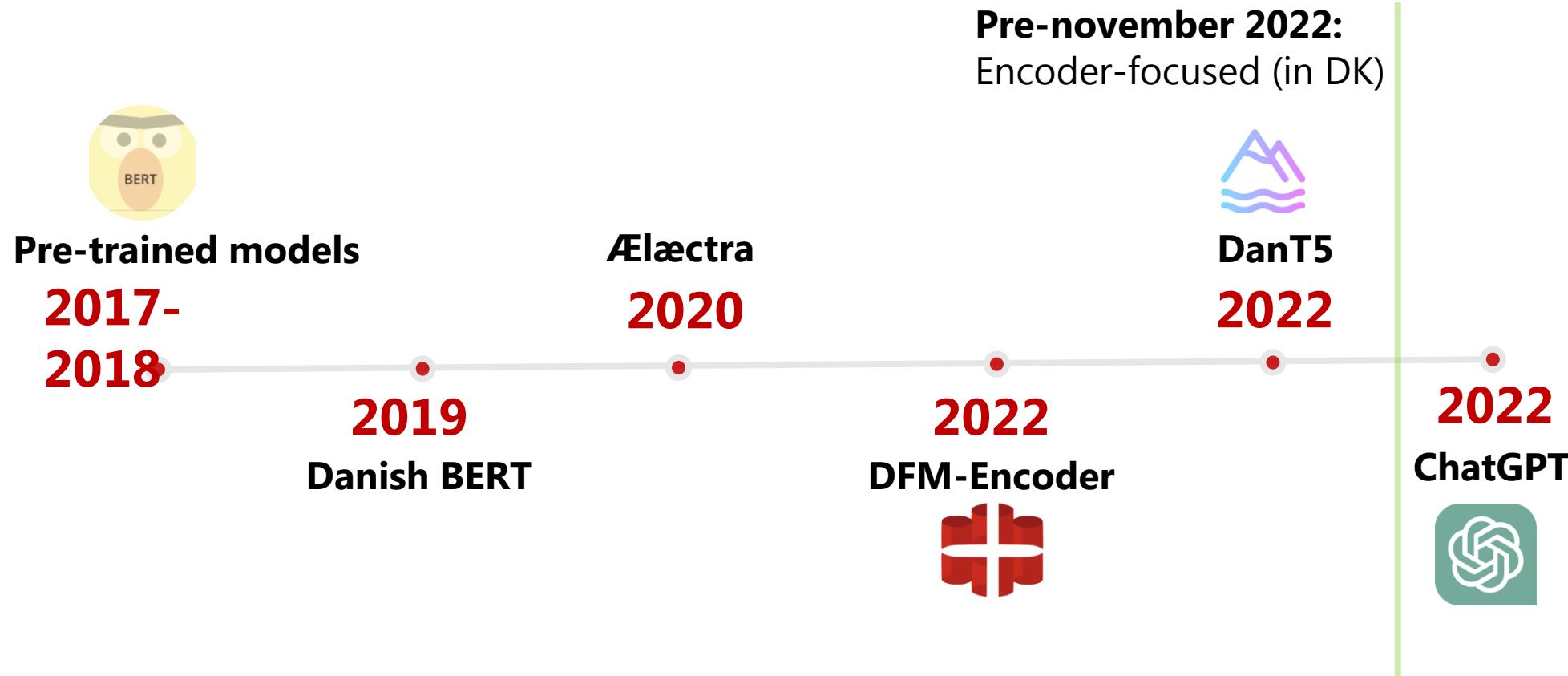
Open vs. closed aligned models

Image credit:
Maxime Labonne



The Danish LM Scene

The "Early" Danish NLP Scene



The (Chat)GPT Effect



NorwAI

Danish Generative Models

Open models



DanskGPT-tiny



/munin-7b-alpha



MISTRAL
AI_



Meta Llama 3

Closed models



Gemini

Mainly large tech or motivated individuals!

Research Efforts

TrustLLM

- Large European initiative
- Open, trustworthy, Germanic LLMs



Danish Foundation Models

- Open collaboration within DK

The Danish LM Scene

Evaluation

ScandEval: Danish NLU Benchmarking

Tasks:

- Information extraction (named entity recognition)
- Grammaticality (linguistic acceptability)
- Question answering (extractive QA)
- Text classification (sentiment classification)

ScandEval: Danish NLG Benchmarking

Tasks:

- Information extraction (named entity recognition)
- Grammaticality (linguistic acceptability)
- Question answering (extractive QA)
- Text classification (sentiment classification)
- Summarisation (news summarisation)
- Knowledge (translated multiple choice datasets: MMLU and ARC)
- Common sense reasoning (translated multiple choice dataset: HellaSwag)

ScandEval: Danish NLU

<https://scandeval.com/danish-nlu/>

Model ID	Parameters	Vocabulary Size	Context	Commercial	Speed	Rank ▼	DANSK
gpt-4-0613 (few-shot, val)	unknown	100	8192	True	$597 \pm 197 / 93 \pm 33$	1.24	$64.94 \pm 1.96 / 45.76 \pm 3.31$
AI-Sweden-Models/roberta-large-1160k	355	50	512	True	$14,560 \pm 2,508 / 3,760 \pm 1,188$	1.34	$74.16 \pm 1.73 / 70.93 \pm 1.67$
KennethEnevoldsen/dfm-sentence-encoder-large-1	355	50	512	True	$6,245 \pm 1,260 / 1,416 \pm 453$	1.34	$74.99 \pm 1.65 / 70.34 \pm 1.50$
KennethEnevoldsen/dfm-sentence-encoder-large-2	355	50	512	True	$6,569 \pm 1,320 / 1,492 \pm 476$	1.34	$75.30 \pm 1.02 / 70.13 \pm 1.50$
danish-foundation-models/encoder-large-v1	355	50	512	True	$6,671 \pm 1,380 / 1,497 \pm 482$	1.36	$74.60 \pm 1.94 / 69.95 \pm 2.01$
AI-Sweden-Models/roberta-large-1350k	355	50	512	True	$5,744 \pm 969 / 1,539 \pm 492$	1.40	$75.22 \pm 1.64 / 71.57 \pm 1.50$
meta-llama/Meta-Llama-3-70B (few-shot, val)	70554	128	8192	True	$312 \pm 55 / 177 \pm 51$	1.41	$63.62 \pm 3.74 / 53.29 \pm 3.38$

- Targeted models can compete with commercial LLMs
- This benchmark does not capture
 - Openness
 - Energy consumption
 - Carbon emissions

ScandEval: Danish NLG

<https://scandeval.com/danish-nlg/>

- But more bleak for open-source *generative* models

Model ID	Parameters	Vocabulary Size	Context	Commercial	Speed	Rank ▼	DANSK
gpt-4-0613 (few-shot, val)	unknown	100	8192	True	$597 \pm 197 / 93 \pm 33$	1.05	$64.94 \pm 1.96 / 45.76 \pm 3.31$
meta-llama/Meta-Llama-3-70B (few-shot, val)	70554	128	8192	True	$312 \pm 55 / 177 \pm 51$	1.43	$63.62 \pm 3.74 / 53.29 \pm 3.38$
gpt-3.5-turbo-0613 (few-shot, val)	unknown	100	4094	True	$921 \pm 293 / 113 \pm 37$	1.52	$61.31 \pm 2.34 / 49.86 \pm 2.18$
meta-llama/Meta-Llama-3-70B-Instruct (few-shot, val)	70554	128	8192	True	$1,673 \pm 583 / 275 \pm 85$	1.59	$63.10 \pm 2.12 / 55.10 \pm 1.44$
152334H/miqu-1-70b-sf (few-shot, val)	68977	32	32764	True	$2,126 \pm 676 / 319 \pm 104$	1.69	$56.96 \pm 2.39 / 45.84 \pm 1.75$
meta-llama/Llama-2-70b-hf (few-shot, val)	68977	32	4096	True	$1,892 \pm 650 / 318 \pm 105$	1.69	$58.06 \pm 2.48 / 50.10 \pm 2.23$
mhenrichsen/danskgpt-chat-v2.1 (few-shot)	unknown	32	32768	True	$5,085 \pm 998 / 1,306 \pm 404$	1.93	$51.08 \pm 1.60 / 35.83 \pm 1.84$

- Open Questions:

- Performance of newer models such as Mixtral, Cohere Command-R, Aya...
- Ranking in relation to English benchmarks
- Influence of RAG solutions

ScandEval: Current Limitations

- 3/8 datasets are translated
 - To what degree is performance on a translated dataset like a native dataset?
- Difference in evaluation and application
 - Popular model evaluations:
 - Code generation
 - ...
 - Danish use-cases:
 - Tax system, welfare, Danish conditions
 - *"Er jeg berretiget til kørsel fra Tyskland på arbejde?"*

New! Danoliterate GLLM Leaderboard

- <https://danoliterate.compute.dtu.dk/Leaderboard>

	🎯	🔒	✍	🏆	Citizenship Test	HyggeSwag
GPT 4	✓	✓		97	97± 0.3	80± 3
GPT 4 Turbo 2024-04-09	✓	✓		96	98± 0.1 🥇	85± 2 🥇
GPT 4o 2024-05-13				95	98± 0.2 🥈	84± 2 🥈
GPT 4 Turbo 1106	✓	✓		94	98± 0.2 🥉	80± 3
Claude Opus	✓	✓		94	98± 0.2	81± 3 🥉
OpenAI GPT 3.5 Turbo	✓	✓		78	82± 1	60± 4
Claude Sonnet	✓	✓		77	93± 1	69± 4
OpenAI GPT 3.5 Turbo Instruct	✓	✓		74	77± 1	65± 4
Claude Haiku	✓	✓		73	89± 1	65± 4
LLaMa 3 70B (@ Groq)	✓		70.6	73	87± 1	67± 4

Søren Vejlgaard Holm and Lars Kai Hansen, DTU Compute
Martin Carsten Nielsen, Alvenir

In progress: Danoliterate A/B testing (like ChatBot Arena)

- <https://danoliterate.compute.dtu.dk/Spørgeskema>

The screenshot shows a web-based survey or test interface. On the left, there are five dropdown menus labeled "Prompt 1" through "Prompt 5" with options like "Generere en undervisningsplan", "Lave lektier", "Tjekke ens viden", "Sjov og frol", and "Forbedre noter". In the center, a large text box contains a question about evaluating an internet article, followed by a paragraph of text from the article. To the right, two side-by-side boxes show outputs from "Model A" and "Model B" for the same prompt. Model A's output is a direct translation of the Danish text into English. Model B's output is a more polished, rewritten version. At the bottom right, there are three sections for rating: "Hvilken model foretrak du?", "Hvad synes du om 🤖 A?", and "Hvad synes du om 🤖 B?". Each section includes a radio button for each model, a "Ved ikke" option, and a row of smiley face icons for rating.

Prompt 1: Generere en undervisningsplan

Prompt 2: Lave lektier

Prompt 3: Tjekke ens viden

Prompt 4: Sjov og frol

Prompt 5: Forbedre noter

Kan du vurdere min start på en internetartikel? Bare kom med en kort analyse af teksten, du skal ikke omformulere den:

'Stammer fra Sydamerika kræver last fra et gammelt spansk skib. Helt tilbage i 1708 sank fragtskibet Sankt José ud fra kysten på det, der i dag er Colombia. Ifølge Associated Press vurderer historikere, at der kan være 11 millioner mønter og ædelsten. Men der er uenighed om, hvem der har ret til lasten i dag, hvor grupper der består af tre forskellige stammer argumenterer for, at det er deres forfædre, der har udvundet guldet.'

Model A:

Title: A Sunken Spanish Ship from South America May Hold Up to 11 Million Coins and Precious Stones: Dispute Over Who Rightfully Owns the Treasure

The article discusses the discovery of potential treasure

Model B:

Din start på internetartiklen er fængende og informativ. Du giver hurtigt læseren de vigtigste oplysninger: hvad der er sket (et spansk skib er sunket i 1708), hvor det skete (ud for Colombias kyst), og hvad konsekvenserne er i dag (stammer gør krav på lasten).

Hvilken model foretrak du?

Hvad synes du om 🤖 A?

Hvad synes du om 🤖 B?

ChatGPT alternatives

- DanskGPT (<https://chat.danskgpt.dk/>)
 - Large closed model and small open model available
 - Meta LLaMA further trained on (translated) Danish data
- HuggingChat (<https://huggingface.co/chat/>)
 - Offers many open-source models of different sizes
 - Competitive with ChatGPT but not tested on Danish

Rule of thumb: use smaller
models if they're good enough

CohereForAI/c4ai-command-r-plus

Active

meta-llama/Meta-Llama-3-70B-Instruct

HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1

mistralai/Mixtral-8x7B-Instruct-v0.1

NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO

01-ai/Yi-1.5-34B-Chat

google/gemma-1.1-7b-it

mistralai/Mistral-7B-Instruct-v0.2

microsoft/Phi-3-mini-4k-instruct

Strengths and weaknesses

LLMs are really good at

- Writing fluent and well-formed language
- Explaining their answers
- Summarizing long texts
- Writing programming code
- Identifying topics
- ...

They are not so good at

- Accessing up-to-date knowledge
- Math (except if using code)
- Knowing what they don't know
- ...