

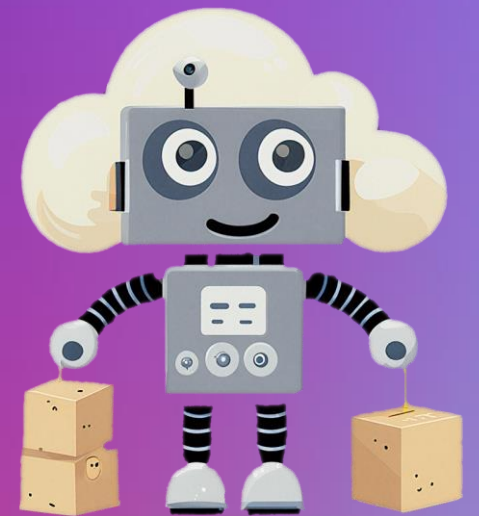


LIG103

# LLM evolution in 2022-2023

Luca Perrozzi

Solutions Architect  
Amazon Web Services



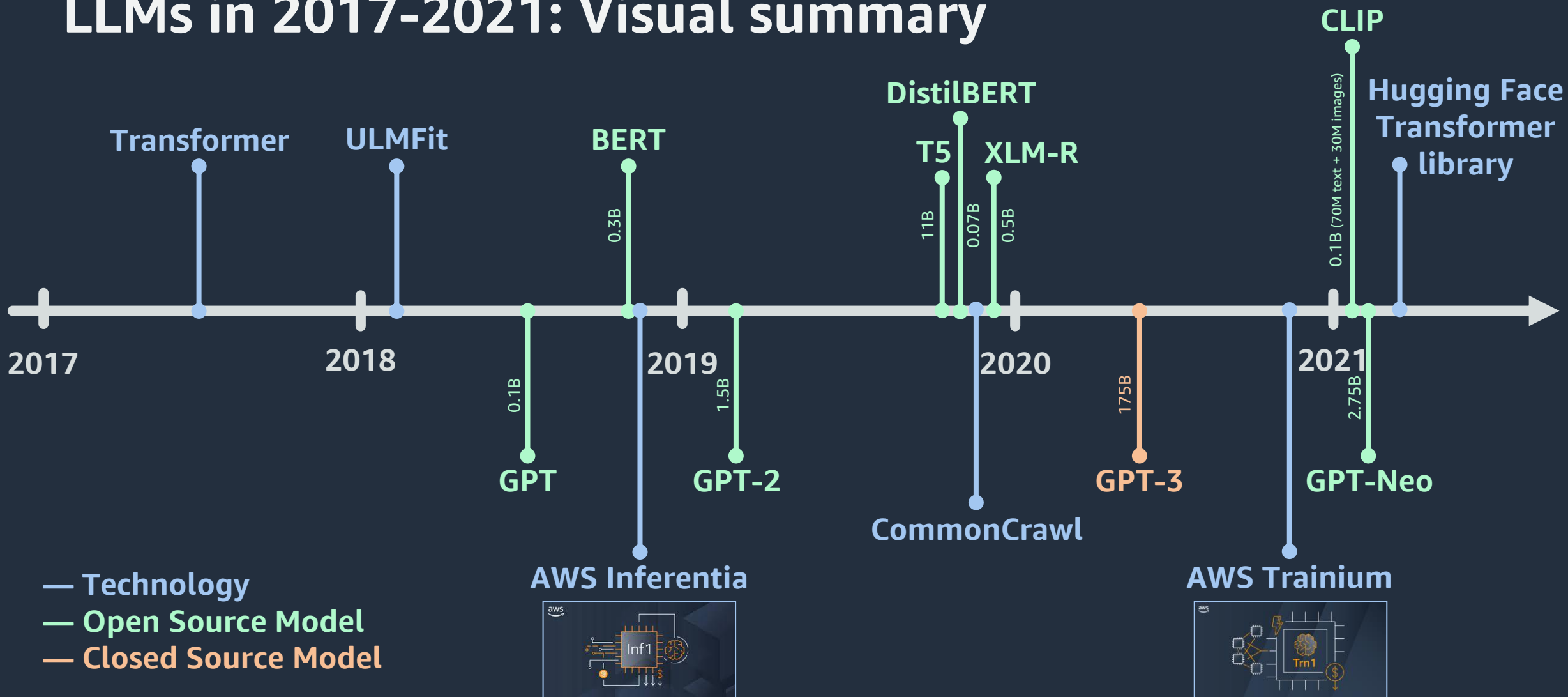
**“If I have seen further it is  
by standing on the  
shoulders of Giants”**

**Isaac Newton**

letter to his rival Robert Hooke

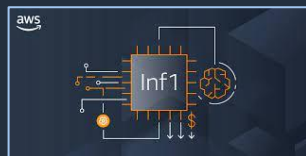


# LLMs in 2017-2021: Visual summary

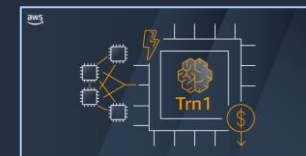


- Technology
- Open Source Model
- Closed Source Model

AWS Inferentia



AWS Trainium



# LLMs in 2017-2021: Key take aways

- **Transformer** architecture and **transfer learning** changed the game
  - Initial approaches focus on **embeddings (BERT)** or **generation (GPT)**
- Most notable **LLMs are open-source**
- Model **evolution trends**:
  - **Bigger** (e.g. 175B of GPT-3)
  - **Optimized architecture** (e.g. DistilBERT, GPT-Neo)
  - **Optimized prompting** (e.g. T5 classification as text generation)
  - **Larger training dataset** (e.g. XLM-R multi-lingual by using CommonCrawl)
  - **Multi-modality** (e.g. CLIP)
- **Technology** is also evolving (e.g. **AWS Inferentia & Trainium, HF library**)



**“What's past is prologue”**

**William Shakespeare**

The Tempest

# LLMs in 2022: Bigger and trained on more data



- Technology
- Open Source Model
- Closed Source Model

### Explaining a Joke

**Input:** Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

**Model Output:** TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

### Logical Inference

**Input:** Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!

**Question:** Is it likely that Shelley will be near the Pacific Ocean this weekend?

**Model Output:** The city with the famous market where they throw the fish is Seattle, Washington. Seattle is on the Pacific Ocean. Shelley is visiting Seattle, so she will be near the Pacific Ocean this weekend. The answer is "yes", it is likely that Shelley will be near the Pacific Ocean this weekend.

### Code generation

**prompt**

```
def exercise5():
    """Marie ordered one chicken meal that costs $12, 5
    packs of milk that costs $3 each, 4 apples that cost
    $1.50 each, and some boxes of pizza. Marie paid a total
    of $50. How many boxes of pizza did Marie order if each
    box costs $8.50?"""
```

**model**

```
total_paid = 50
cost_of_pizza_per_box = 8.50
cost_of_meal = 12
cost_of_milk_per_pack = 3
cost_of_apples_per_apple = 1.50
packages_of_milk = 5

return float((total_paid - cost_of_meal - cost_of_milk_per_pack *
              packages_of_milk - cost_of_apples_per_apple *
              4) / cost_of_pizza_per_box)
```

**prompt**

```
// Translate from C to Python:
long long int sum_pairwise_products ( int n ) {
    long long int sum = 0;
    for ( int i = 1; i <= n; i ++ ) {
        for ( int j = i; j <= n; j ++ ) {
            sum = sum + i * j;
        }
    }
    return sum;
}
```

**model**

```
def sum_pairwise_products(n):
    sm = 0
    for i in range(1, (n + 1)):
        for j in range(i, (n + 1)):
            sm = (sm + (i * j))
    return sm
```

**prompt**

```
# Given a string representing musical notes in a
special ASCII format,
# parse this string and compute a total length of the
music in beats.
# Legend:
# 'o' - whole note, lasts four beats
# 'o|' - half note, lasts two beats
# '|' - quarter note, lasts one beat

notes = 'o o| .| o| | .| .| o o'
```

**model**

```
# 1. Split the string into a list of notes.
# 2. Iterate over the list and compute the total
length of the music.
# 3. Print the result.

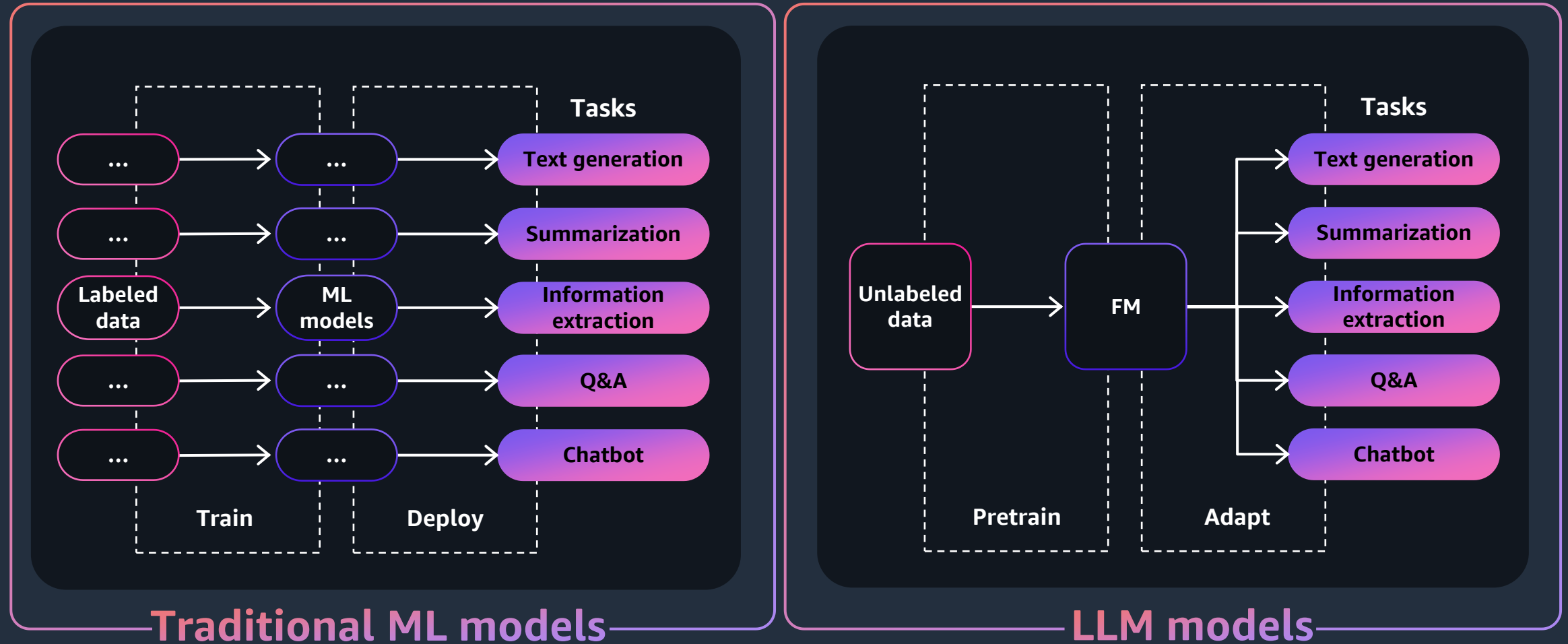
# 1.
notes = notes.split()

# 2.
total_length = 0
for note in notes:
    if note == 'o':
        total_length += 4
    if note == 'o|':
        total_length += 2
    if note == '|':
        total_length += 1

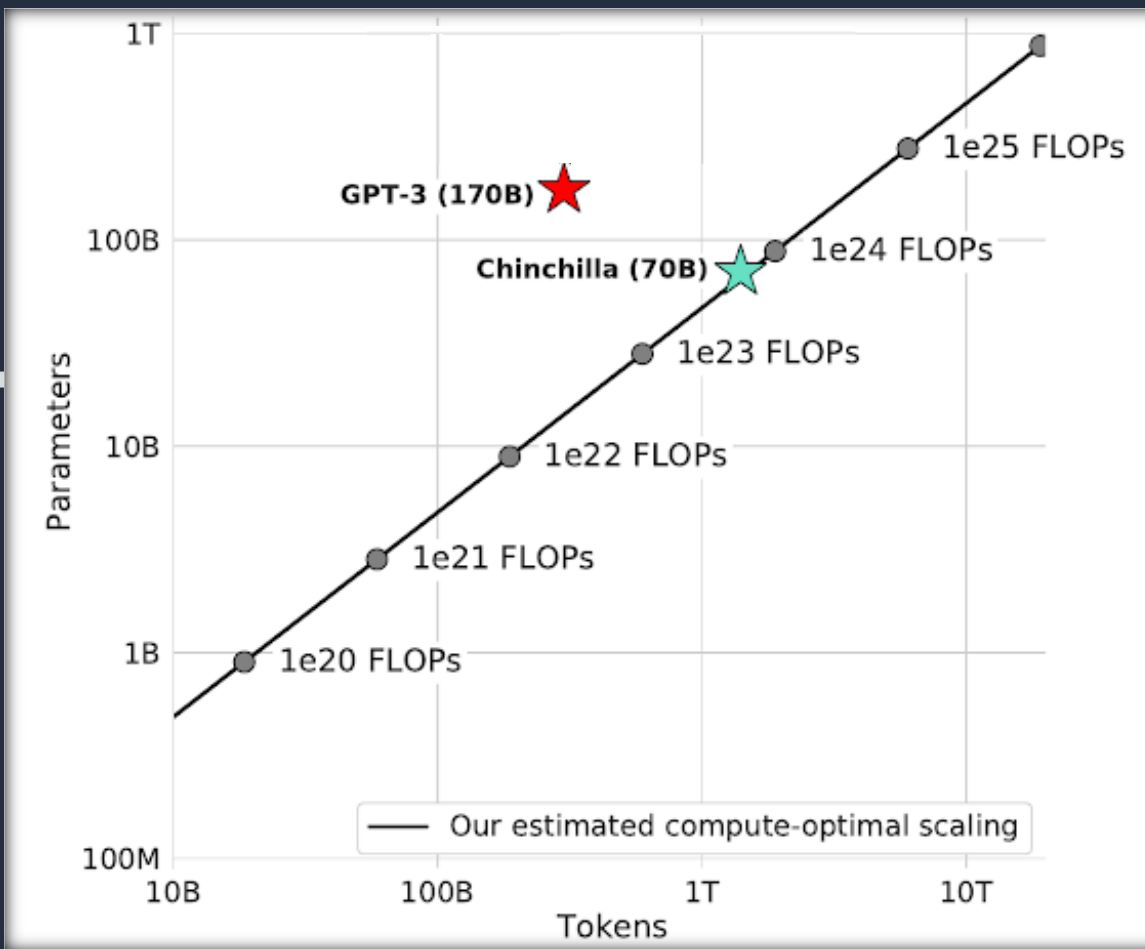
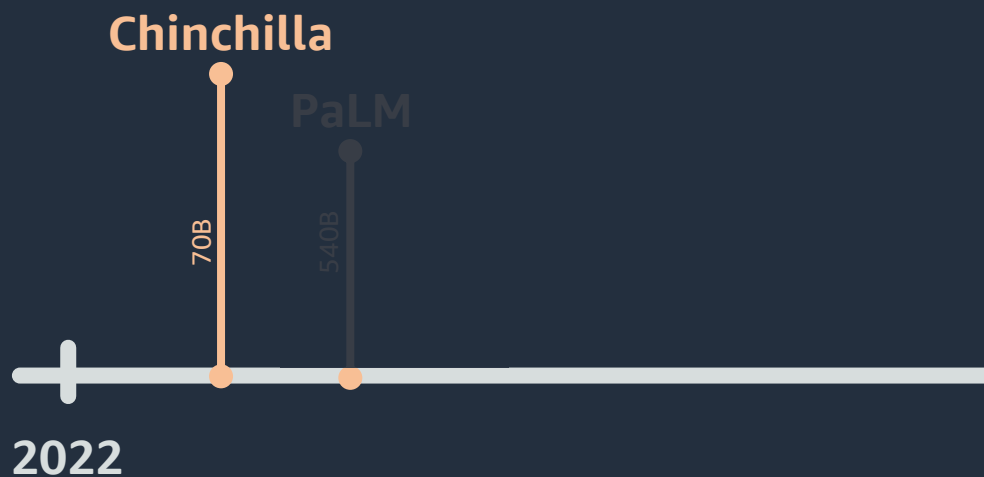
# 3.
print(total_length)
```



# Emerging capabilities of LLMs



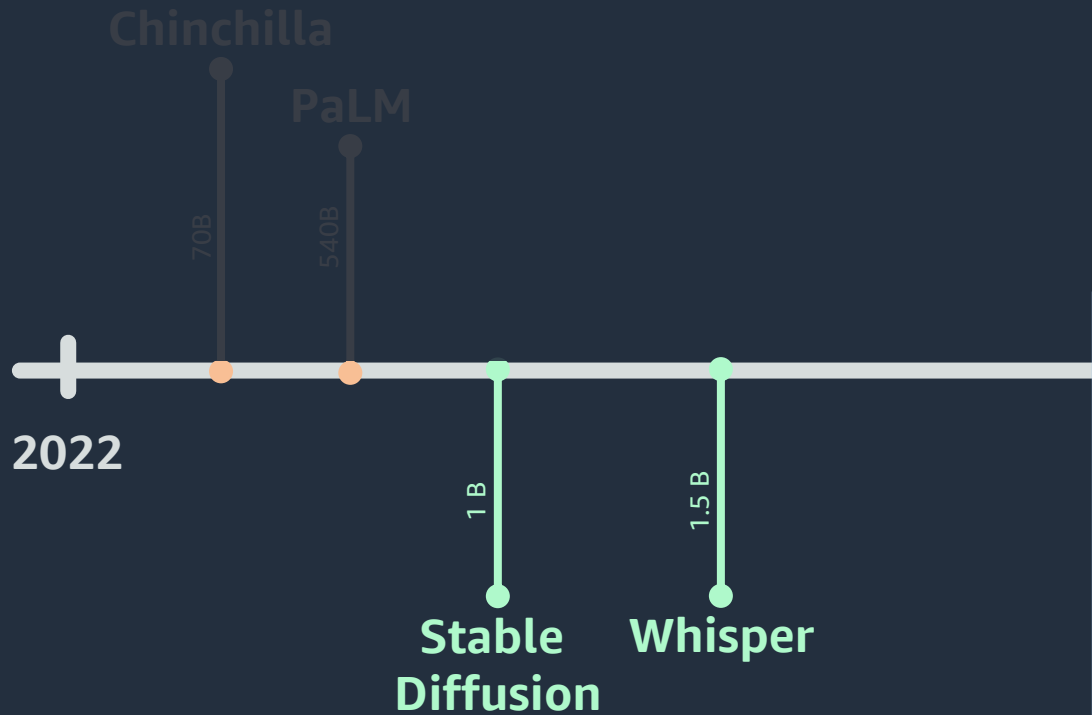
# LLMs in 2022: Optimized



- Technology
- Open Source Model
- Closed Source Model

Image adapted from: [Training Compute-Optimal Large Language Models](#)

# LLMs in 2022: Multi-modal



- Technology
- Open Source Model
- Closed Source Model

## Stable Diffusion Public Release

22 Aug



Research  
**Introducing Whisper**

Illustration: Ruby Chen

We've trained and are open-sourcing a neural net called Whisper that approaches human level robustness and accuracy on English speech recognition.

Since the public release of stable diffusion following our [https://stability.ai/stablediffusion] we all have been overwhelmed by the response and have made a safe and ethical release, incorporating data from our community for the developers to act on.

Our less legal, ethics, and technology teams at HuggingFace and CoreWeave, we have incorporated the following elements:

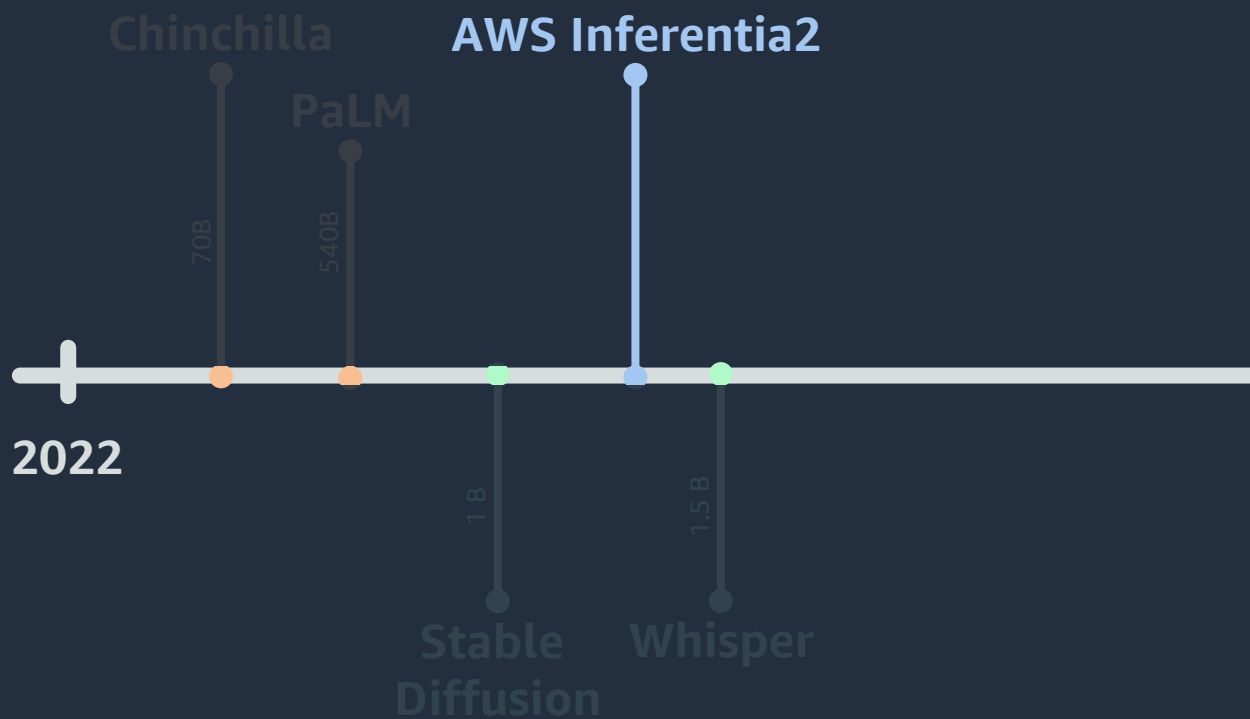
It is used under a [Creative ML OpenRAIL-M license](#). This is a license that allows for commercial and non-commercial usage. This license is subject to the legal use of the model as your responsibility and must be used in accordance with the terms of the model. It must also be made available to end users who use it on it.

Credits: [Stability.AI](#)

Credits: [OpenAI](#)



# LLMs in 2022: Technology evolves



- Technology
- Open Source Model
- Closed Source Model

### AWS announces Amazon EC2 Inf2 instances (Preview)

Posted On: Nov 29, 2022

Today, AWS announces the preview of Amazon Elastic Compute Cloud (Amazon EC2) Inf2 instances, which are designed to deliver high performance at the lowest cost in Amazon EC2 for the most demanding deep learning (DL) inference applications. Inf2 instances are powered by up to 12 AWS Inferentia2, the third AWS-designed DL accelerator. Inf2 instances offer 3x higher compute performance, up to 4x higher throughput, and up to 10x lower latency compared to Inf1 instances.

You can use Inf2 instances to run DL applications for natural language understanding, translation, video and image generation, speech recognition, personalization, and more. They are optimized to deploy complex models, such as large language models (LLM) and vision transformers, at scale while also improving the Inf1 instances' price-performance benefits for smaller models. To support ultra-large 100B+ parameter models, Inf2 instances are the first inference-optimized instances in Amazon EC2 to support scale-out distributed inference with ultra-high-speed connectivity between accelerators.

Inf2 instances offer up to 2.3 petaflops of DL performance, up to 384 GB of accelerator memory with 9.8 TB/s bandwidth, and NeuronLink, an intra-instance ultra-high-speed, nonblocking interconnect. Inf2 instances also offer up to 50% better performance per watt compared to GPU-based instances in Amazon EC2 and help you meet your sustainability goals. The AWS Neuron SDK is natively integrated with popular ML frameworks, such as PyTorch and TensorFlow, so you can deploy your DL applications on Inf2 with a few lines of code.

Credits: [AWS](#)

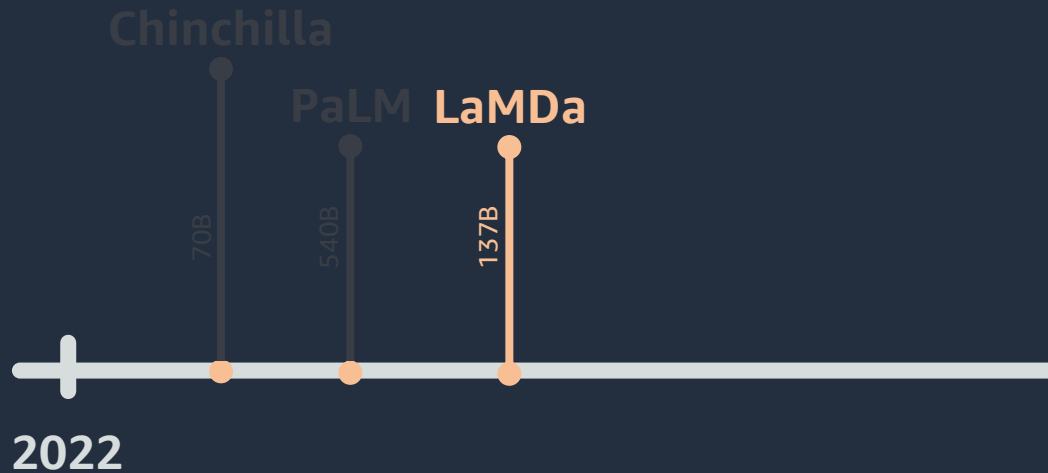
# LLMs trends in ~~2017~~~~2021~~ 2022

- Model evolution trends:
  - Bigger (e.g. 540B of PaLM)
  - Optimized architecture (e.g. Chinchilla)
  - Larger training dataset
  - Multi-modality (e.g. Stable Diffusion)
- Technology is also evolving (e.g. AWS Inferentia2)



## Anything I missed?

# Ah right, Chatbots! How to forget LaMDa? 🤖



- Technology
- Open Source Model
- Closed Source Model

ARTIFICIAL INTELLIGENCE >

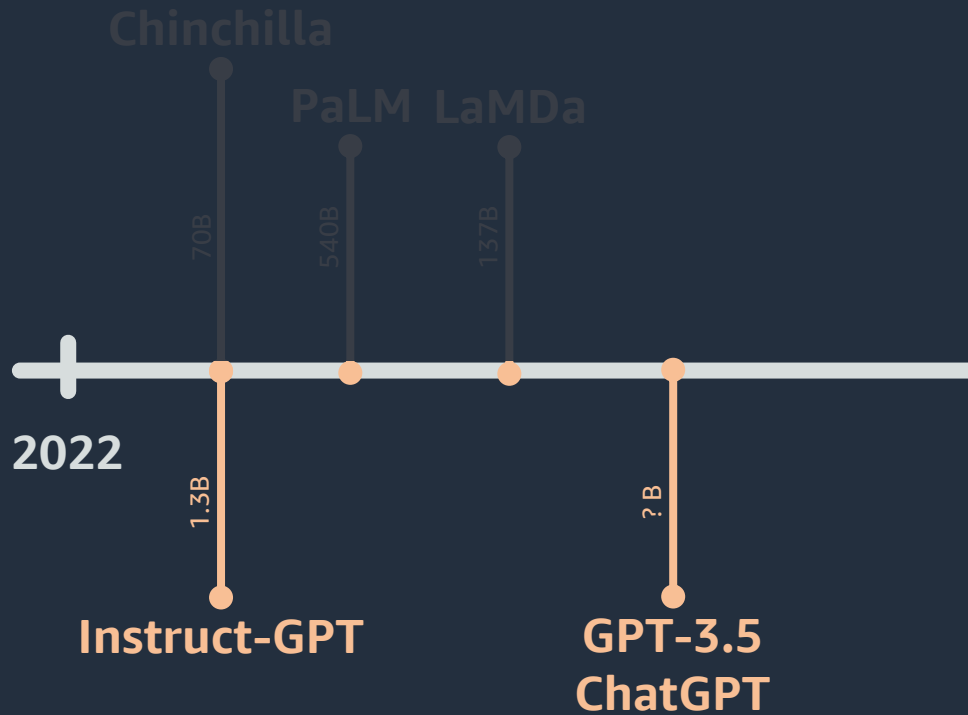
## LaMDa, the machine that is like ‘a seven-year-old kid’: can a computer have consciousness?

A Google engineer believes he had a conversation with an artificial intelligence system capable of independent thought. Although the scientific community has scoffed at the idea, advances in AI will lead to ‘uncomfortable debates’ in the future

A large red heart shape is centered on a white background. Inside the heart, a detailed black and white image of a computer circuit board is visible, with various components and traces. The heart is set against a light red background.

Credits: [EL Pais](#)

# LLMs in 2022: The elephant in the room -- RLHF



- Technology
- Open Source Model
- Closed Source Model

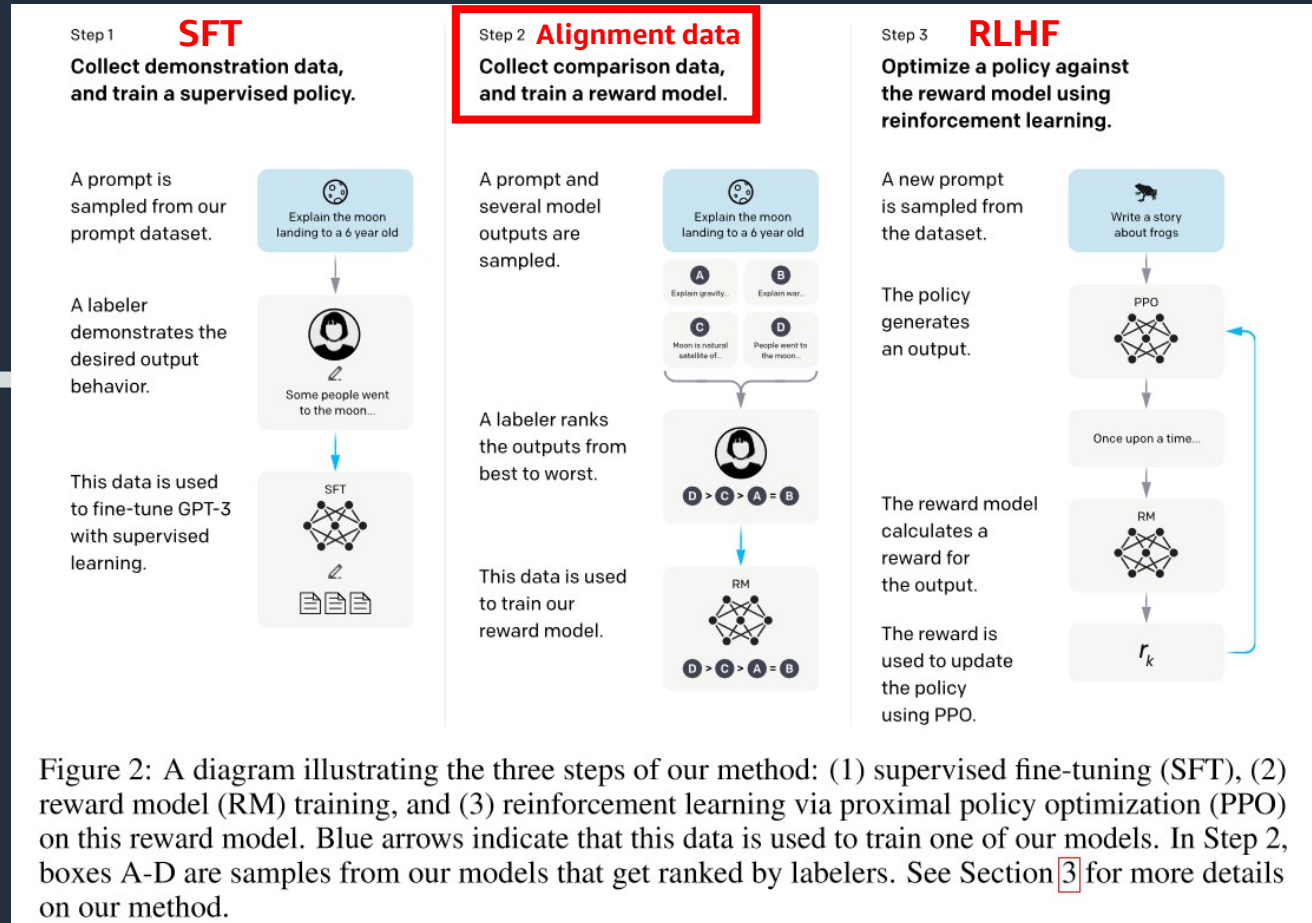


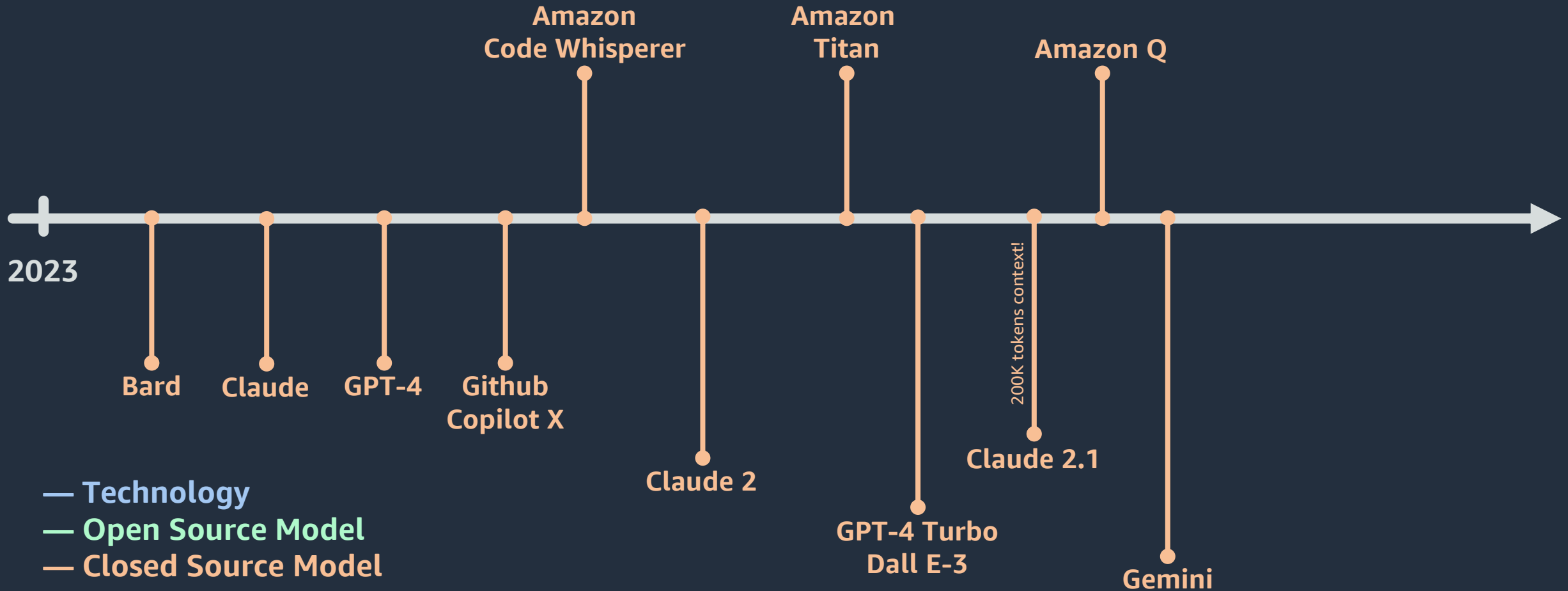
Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

Image credit: [Training language models to follow instructions with human feedback](#)

# LLMs in 2022: take aways

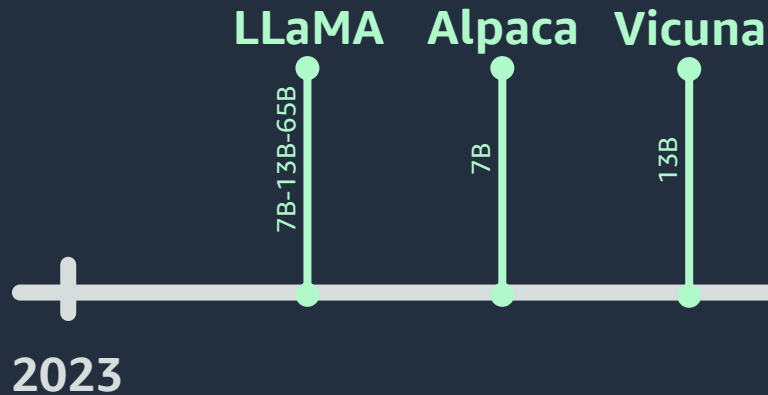
- **Emerging capabilities of LLMs**
  - Generalization with few shots
  - High-quality text generation
  - Ability to handle domain-specific and general knowledge
  - Reasoning and problem-solving
- **RLHF** sparks a surge of renovated interest into conversational AI
- **Multi-modal models** take off
- Most notable **LLMs are now closed-source**

# LLMs in 2023: the raise of closed source models and apps



- Technology
- Open Source Model
- Closed Source Model

# Llama: major Open Source release since BERT (2018) and GPT-2 (2019)



- Technology
- Open Source Model
- Closed Source Model

Research

## Introducing LLaMA: A foundational, 65-billion-parameter large language model

February 24, 2023

leasing our model under a **noncommercial license** focused granted on a case-by-case basis to academic researchers;

Credits: [Meta](#)

### Alpaca: A Strong, Replicable Instruction-Following Model

Authors: Rohan Taori\* and Ishaan Gulrajani\* and Tianyi Zhang\* and Yann Dubois\* and Xuechen Li\* and Carlos Guestrin and Percy Liang and Tatsunori B. Hashimoto

We introduce Alpaca 7B, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations. On our preliminary turn instruction following, Alpaca behaves qualitatively similarly to OpenAI's text-davinci-003, while being surprisingly small and cheap (<600\$). Checkout our code release on [GitHub](#).

Update: The public demo is now disabled. The original goal of releasing a demo was to disseminate our research in an accessible form. We have mostly achieved this goal, and given the hosting costs and the inadequacies of our content filters, we decided to bring down the demo.

Credits: [Stanford](#)

### Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality

by: The Vicuna Team, Mar 30, 2023

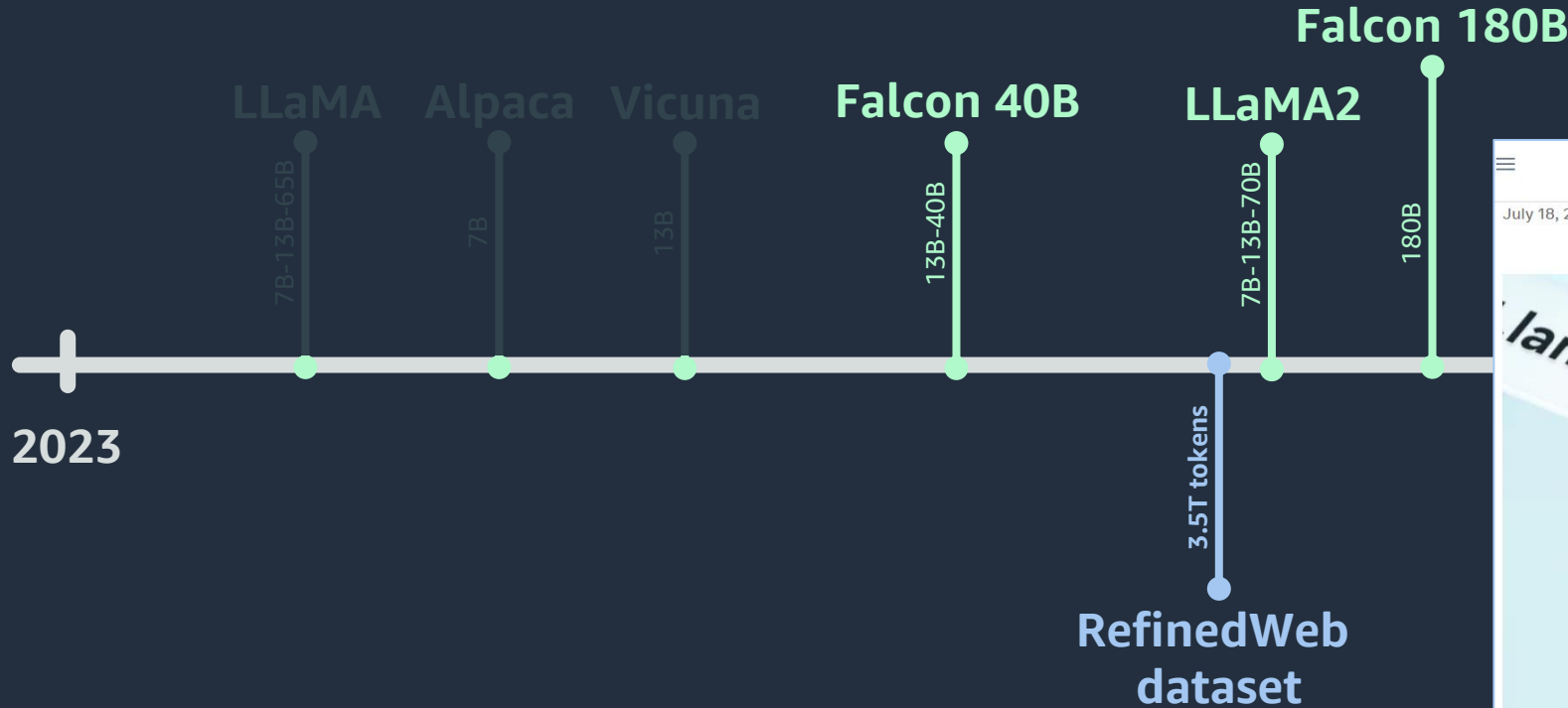
We introduce Vicuna-13B, an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Preliminary evaluation using GPT-4 as a judge shows Vicuna-13B achieves more than 90%\* quality of OpenAI ChatGPT and Google Bard while outperforming other models like LLaMA and Stanford Alpaca in more than 90%\* of cases. The cost of training Vicuna-13B is around \$300. The [code](#) and [weights](#), along with an online [demo](#), are publicly available for non-commercial use.

Vicuna (generated by stable diffusion 2.1)

Credits: [LMSYS](#)



# Open Source is back in business...



- Technology
- Open Source Model
- Closed Source Model

Estimated costs for LLaMA2 70B: \$22M

Reference: [Nazneen Rajani](#)

UAE's Technology Innovation Institute Launches Open-Source "Falcon 40B" Large Language Model for Research & **Commercial Utilization**

May 25, 2023

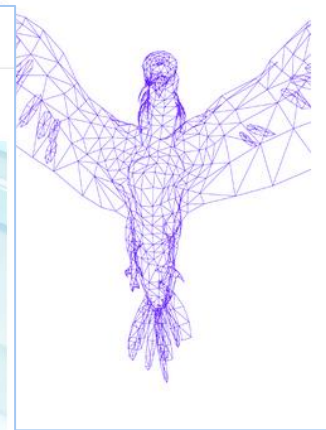
Meta  
July 18, 2023

llama 2

Llama-2-7B  
Llama-2-13B  
Llama-2-70B

Takeaways

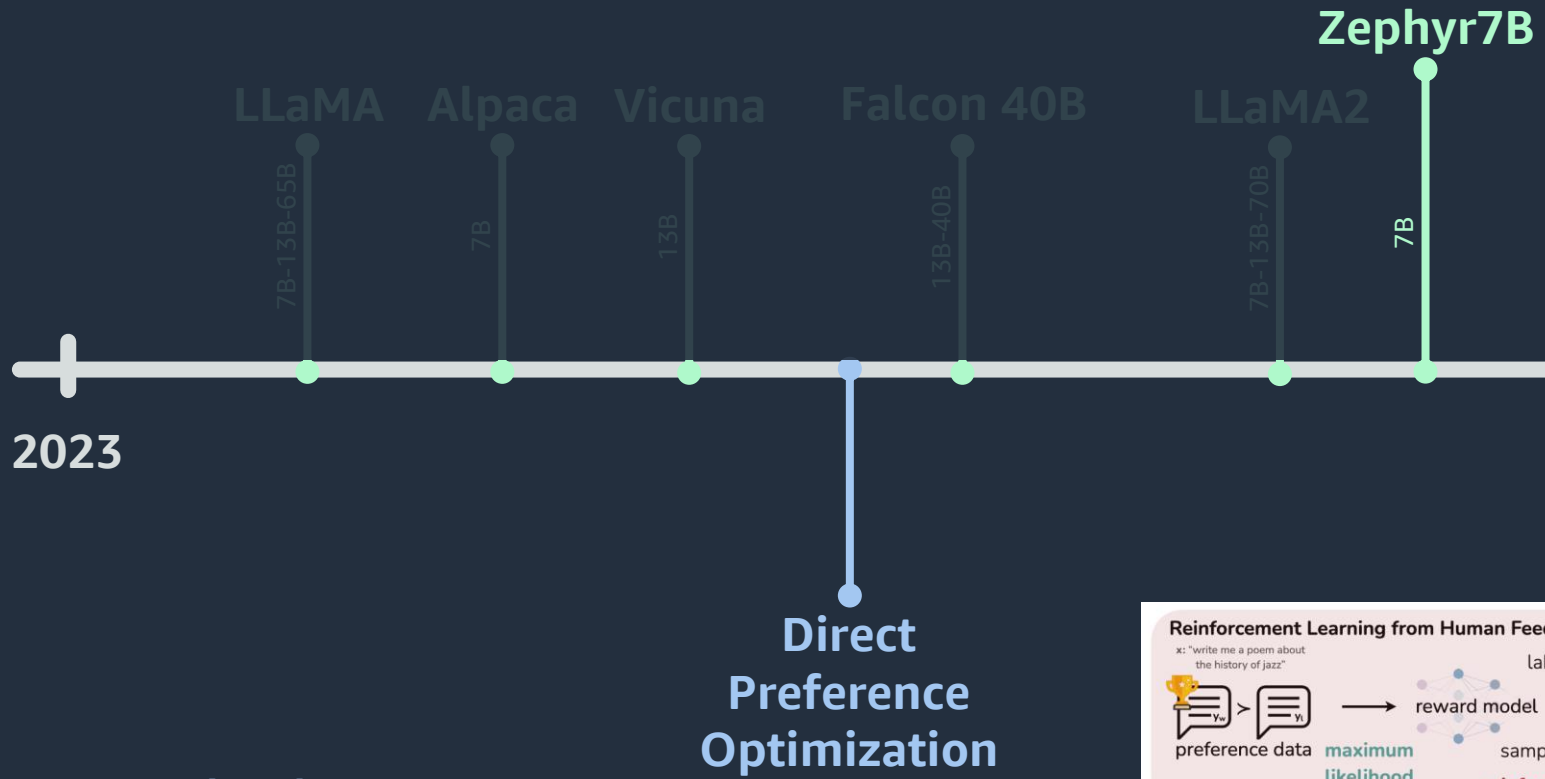
- Today, we're introducing the availability of Llama 2, the next generation of our open source large language model.
- **Llama 2 is free for research and commercial use.**



Credits: [TUI](#)



# Invent and simplify: DPO



- Technology
- Open Source Model
- Closed Source Model

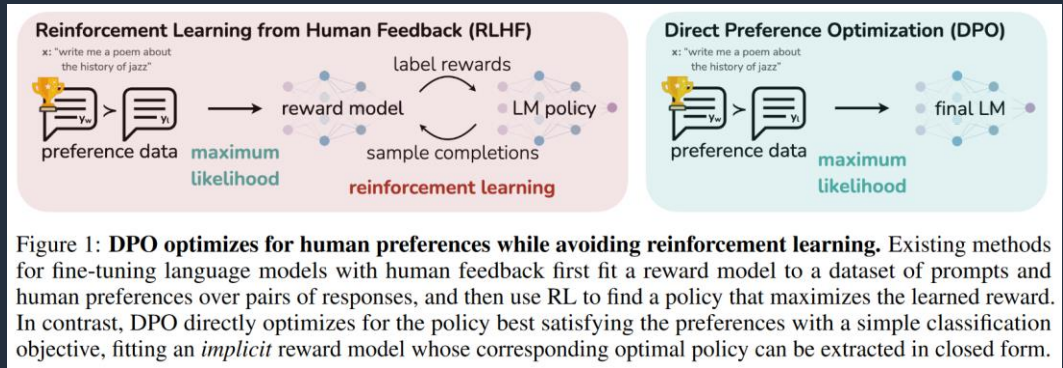
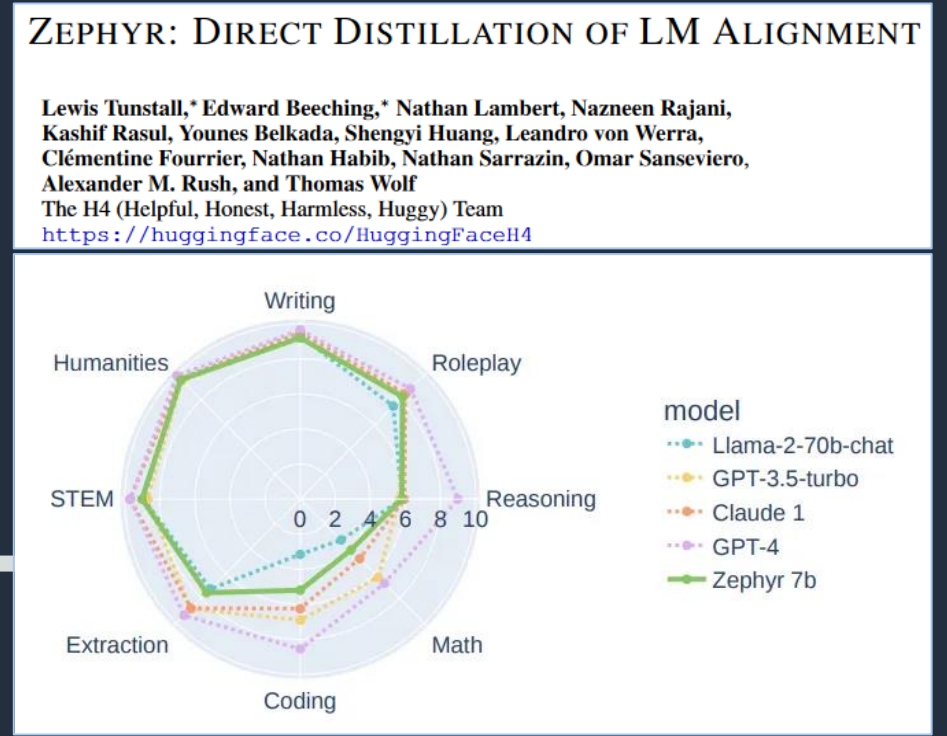
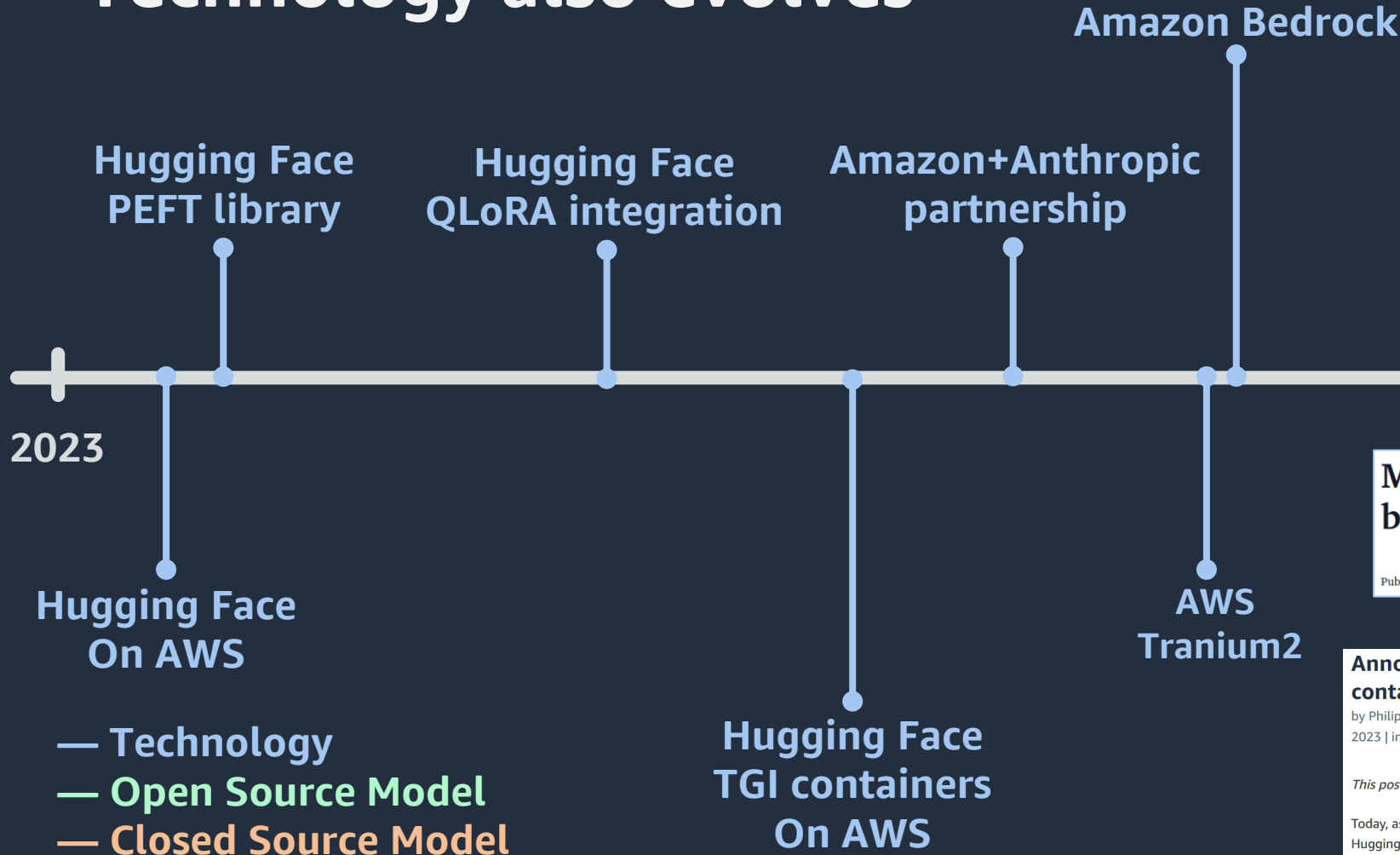


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes 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.

# Technology also evolves



 **PEFT: Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware**

Published February 10, 2023

Credits: [Hugging Face](#)

[AWS Machine Learning Blog](#)

**AWS and Hugging Face collaborate to make generative AI more accessible and cost efficient**

by Amazon Web Services | on 21 FEB 2023 | in [Amazon Machine Learning](#), [Artificial Intelligence](#), [Generative AI](#) | [Permalink](#) | [Comments](#) | [Share](#)

We're thrilled to announce an expanded collaboration between AWS and Hugging Face to accelerate the training, fine-tuning, and deployment of large language and vision models used to create generative AI applications. Generative AI applications can perform a variety of tasks, including text summarization, answering questions, code generation, image creation, and writing essays and articles.

Credits: [AWS + Hugging Face](#)

**Making LLMs even more accessible with bitsandbytes, 4-bit quantization and QLoRA**

Published May 24, 2023

Credits: [Hugging Face](#)

**Announcing the launch of new Hugging Face LLM Inference containers on Amazon SageMaker**

by Philipp Schmid, Jeff Boudier, Gagan Singh, Qing Lan, Robert Van Dusen, Simon Zamarin, and Xin Yang | on 05 JUN 2023 | in [Amazon SageMaker](#), [Announcements](#), [Artificial Intelligence](#) | [Permalink](#) | [Comments](#) | [Share](#)

*This post is co-written with Philipp Schmid and Jeff Boudier from Hugging Face.*

Today, as part of Amazon Web Services' partnership with Hugging Face, we are excited to announce the release of a new Hugging Face [Deep Learning Container](#) (DLC) for inference with Large Language Models (LLMs). This new Hugging Face LLM DLC is powered by [Text Generation Inference \(TGI\)](#), an open source, purpose-built solution for deploying and serving Large Language Models. TGI enables high-performance text generation using Tensor Parallelism and dynamic batching for the most popular open-source LLMs, including StarCoder, BLOOM, GPT-NeoX, StableLM, Llama, and T5.

Credits: [AWS + Hugging Face](#)



# LLMs trends in ~~2017~~~~2021~~~~2022~~2023

- Model evolution trends:
  - Bigger (e.g. ?B)
  - Optimized architecture (e.g. Zephyr7B)
  - Larger training dataset (NB: mostly AI-generated)
  - Multi-modality (e.g. Stable Diffusion XL, GPT-4, Gemini, Titan)
- Technology is also evolving



# LLMs in 2023: take aways

- **Closed-source landscape is accelerating**
  - Models and applications evolving very quickly
  - Multi-modal becoming fashionable (e.g. GPT-4, Gemini)
  - LLMs are becoming a commodity (e.g. Amazon Bedrock)
- **Open Source landscape is keeping up**
  - Competitive models available with commercial license (e.g. Llama, Falcon, Mistral)
  - Lots of experimentation and innovation (e.g. Zephyr)
- **Specialization is on the rise**
  - Natural language to Code (e.g. CodePilot, Code Whisperer, LlamaCode)
- **Emerging architectural patterns**
  - Retrieval Augmented Generation
  - Agents

# Generative AI Stack in AWS










## APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

-  Amazon Q
-  Amazon Q in Amazon QuickSight
-  Amazon Q in Amazon Connect
-  Amazon CodeWhisperer

## TOOLS TO BUILD WITH LLMs AND OTHER FMs

-  Amazon Bedrock
- Broad choice of models
- Customization Capabilities
- Guardrails
- Agents

## INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

-  GPUs
-  Trainium
-  Inferentia
-  SageMaker
-  UltraClusters
-  EFA
-  EC2 Capacity Blocks
-  Nitro
-  Neuron



# Thank you!

Luca Perrozzi

 <https://www.linkedin.com/in/luca-perrozzi/>

lperroz@amazon.ch



Please complete the session survey.



# Backup slides



# References

## Models:

[GPT](#)

[BERT](#)

[GPT-2](#)

[T5](#)

[DistilBERT](#)

[XLM-R](#)

[GPT-3](#)

[CLIP](#)

[GPT-Neo](#)

[LaMDA](#)

[PaLM: Scaling Language Modeling with Pathways](#)

[AlexaTM 20B](#)

[Chinchilla](#)

[Stable Diffusion](#)

[Whisper](#)

[Instruct-GPT](#)

[GPT-3](#)

[Bard](#)

[Claude](#)

[Claude 2](#)

[Claude 2.1](#)

[Github Copilot](#)

[Amazon Code Whisperer](#)

[GPT-4](#)

[Gemini](#)

[Amazon Titan](#)

[LLaMA](#)

[Alpaca](#)

[Vicuna](#)

[Falcon Series of Open Language Models](#)

[Llama 2](#)

[Zephyr](#)

[Mistral 7B](#)

## Others:

[Transformer](#)

[ULMFit - Universal Language Model Fine-tuning for Text Classification](#)

[CommonCrawl](#)

[AWS Inferentia](#)

[AWS Trainium](#)

[AWS Inferentia2](#)

[AWS Trainium2](#)

[Amazon Bedrock](#)

[Amazon+Anthropic](#)

[AWS+HuggingFace](#)

[AWS+HuggingFace TGI containers](#)

[HuggingFace transformer](#)

[Harnessing the Power of LLMs in Practice](#)

[RefinedWeb Dataset](#)

[Direct Preference Optimization](#)

# Amazon Bedrock

Broad choice of models

AI21 labs

amazon

ANTHROPIC

cohere

∞ Meta

stability.ai

JURASSIC-2

AMAZON TITAN  
TEXT + IMAGES  
EMBED TEXT + MULTI-MODAL

CLAUDE

COMMAND + EMBED

LLAMA 2

STABLE DIFFUSION XL

# Evolutionary Tree

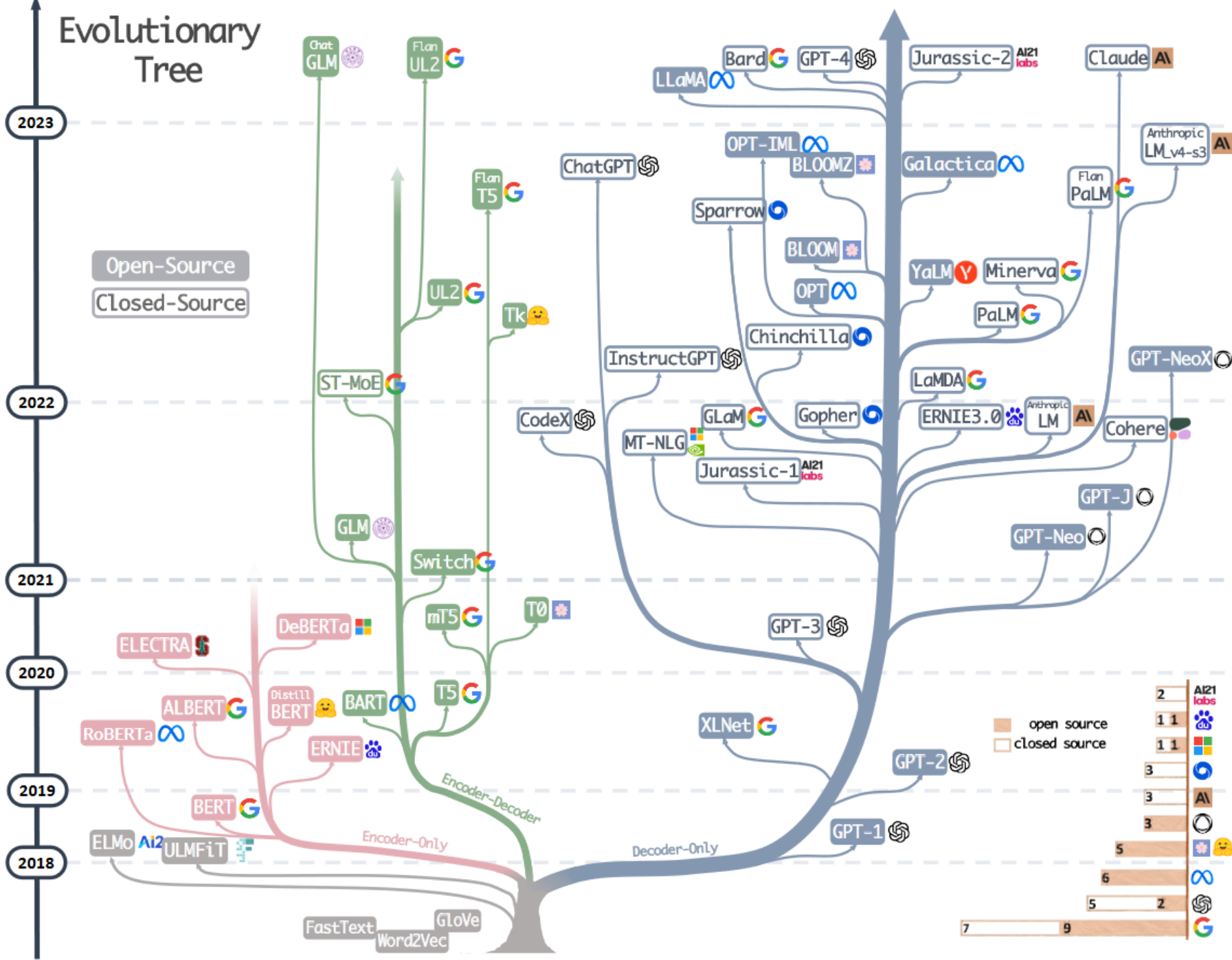
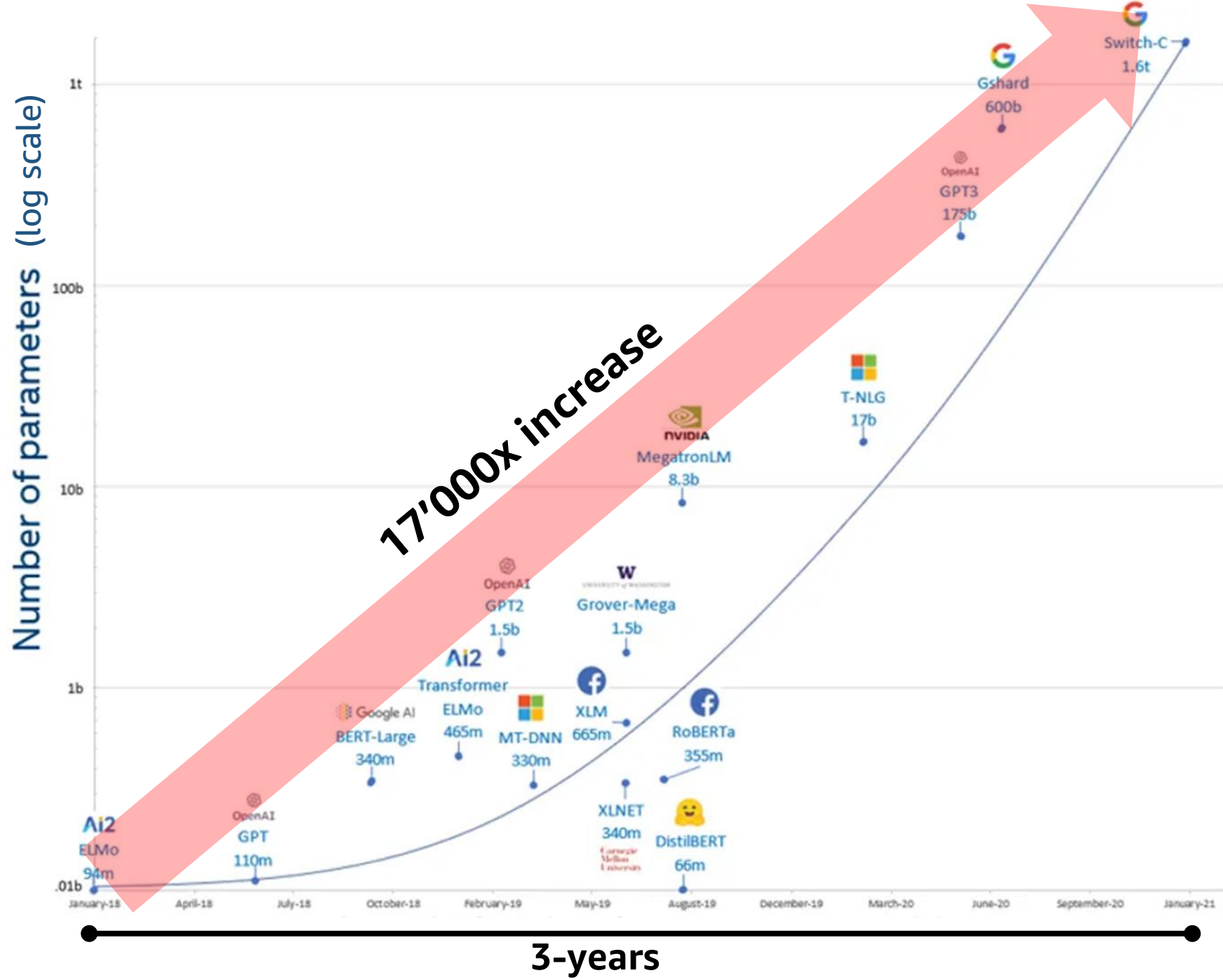


Image adapted from: [Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond](#)



# About estimated LLama2 70B costs: \$22M

## Pre-training dataset

- 2T tokens

## SFT dataset

- 27,540 high quality human annotations (filtered among M of examples)

## RLHF dataset

- 1.5M human dedicated annotations + 1.5M open-source annotations

References: llama [repository](#) and [paper](#)

## Disclosed pre-training costs

- 7B model: 184k GPU hours, \$300k
- 70B model: 1.7M GPU hours, \$2.5M

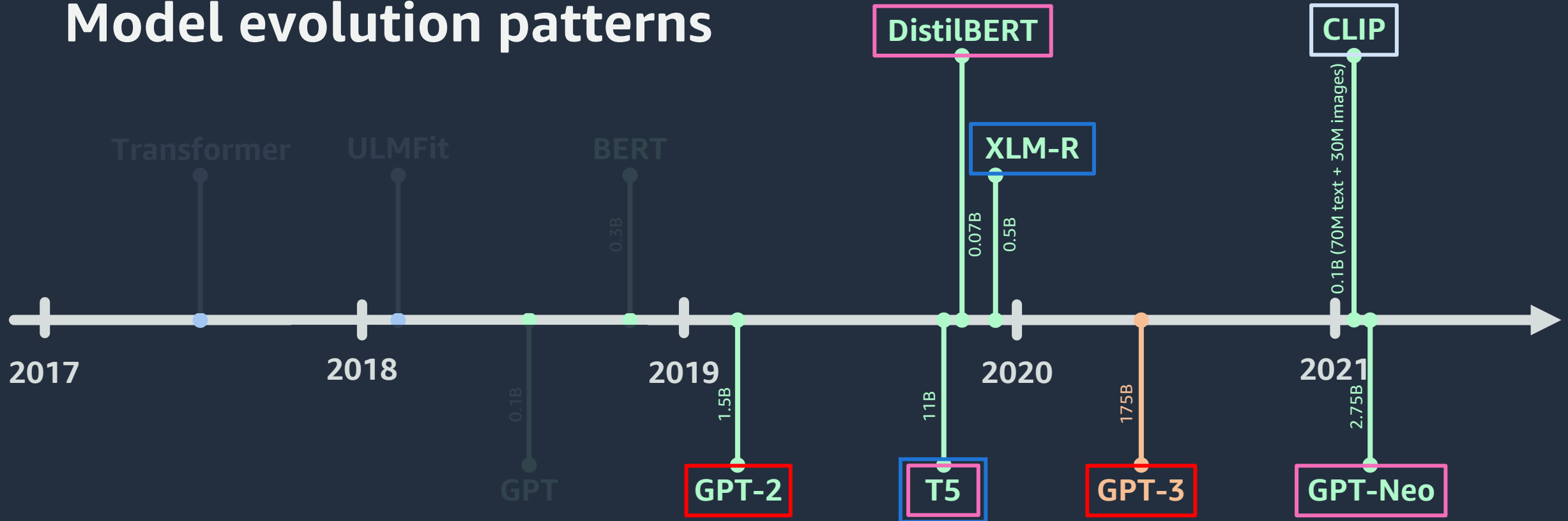
## Overall estimated costs

SFT and RLHF **\$11M** (avg 3.9 turns). This does not account for all the human evaluations they did for helpfulness and safety.

**Adding training costs and evaluation costs, even with a super conservative estimate, the LLaMA 2 has a price tag of \$22M**

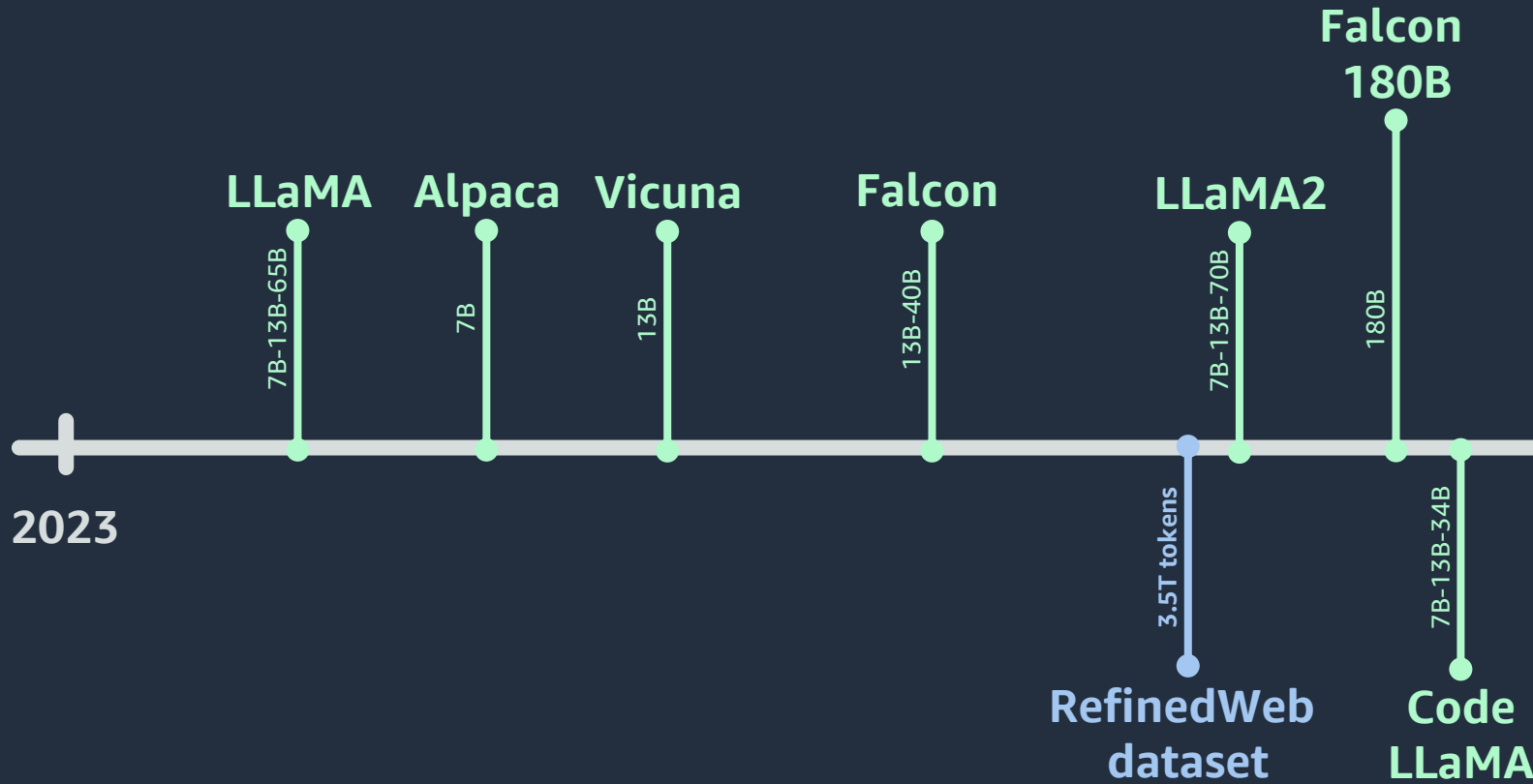
Reference: [Nazneen Rajani](#)

# Model evolution patterns



- Technology
- Open Source Model
- Closed Source Model

- Bigger
- Optimized (architecture/prompting)
- More/better training data
- Multi-modal encoding



### Introducing Code Llama, a state-of-the-art large language model for coding

August 24, 2023

**Code Llama** Meta AI

**PROMPT**

In Bash, how do I list all text files in the current directory (excluding subdirectories) that have been modified in the last month?

Clear Submit

**RESPONSE**

You can use the 'find' command in Bash to list all text files in the current directory (excluding subdirectories) that have been modified in the last month. The command would be:

```
find . -maxdepth 1 -type f -mtime -30 -name "*.txt"
```

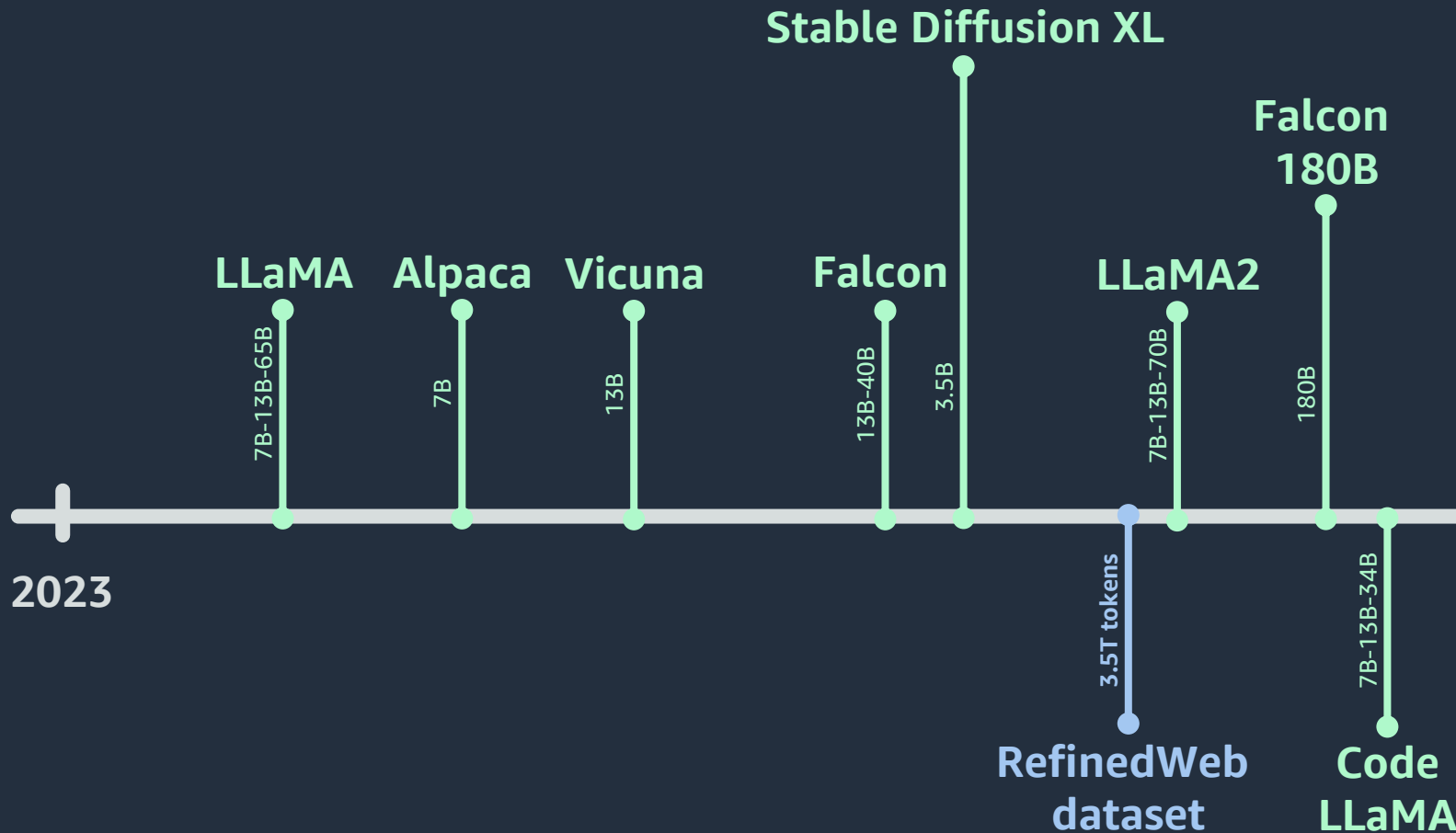
Here is a breakdown of the command:

- `.` specifies the current directory
- `-maxdepth 1` tells 'find' to only search the current directory and not any subdirectories
- `-type f` tells 'find' to only search for files (not directories or other types of files)
- `-mtime -30` tells 'find' to only search for files that have been modified in the last 30 days
- `-name "\*.txt"` tells 'find' to only search for files with the '.txt' extension

Note that the '-mtime' option uses the number of days since the file was last modified, so '-mtime -30' means "modified in the last 30 days." If you want to search for files modified in the last month, you can use the '-mtime -30' option.

Credits: [Meta](#)

- Technology
- Open Source Model
- Closed Source Model



- Technology
- Open Source Model
- Closed Source Model



## Announcing SDXL 1.0

26 Jul

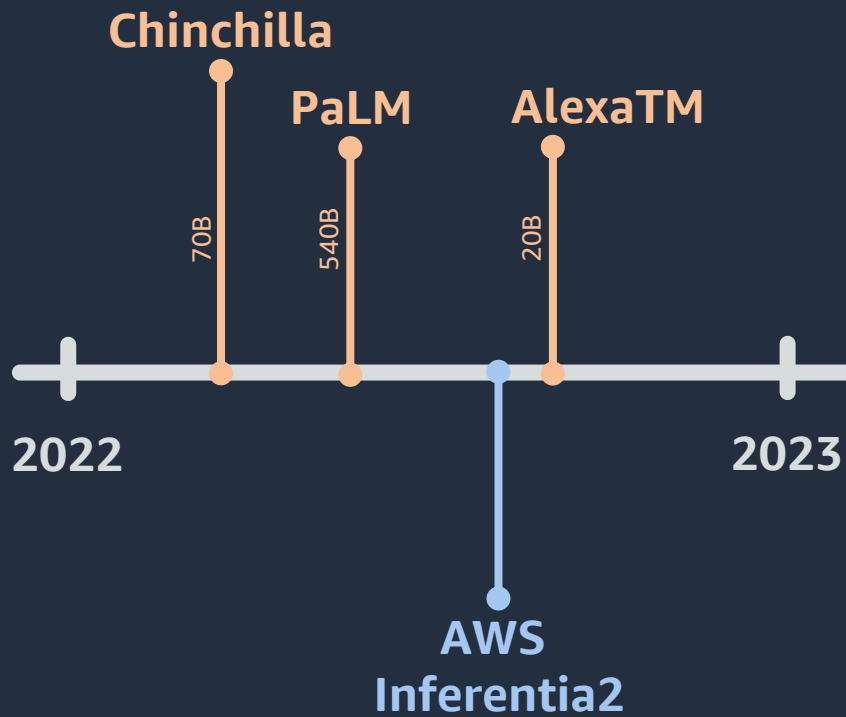
The Stability AI team is proud to release as an open model SDXL 1.0, the next iteration in the evolution of text-to-image generation models. Following the limited, research-only release of SDXL 0.9, the full version of SDXL has been improved to be the world's best open image generation model.



SDXL 1.0 launch, made with forthcoming image control from Stability AI.

Credits: [Stability.AI](https://stability.ai)

# AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model



- Technology
- Open Source Model
- Closed Source Model

## Amazon's AlexaTM 20B Model Outperforms GPT-3 on NLP Benchmarks

LIKE DISCUSS

AUG 19, 2022 • 2 MIN READ

by



Anthony Alford

FOLLOW

Director, Development at Genesys Cloud Services

Researchers at [Amazon Alexa AI](#) have announced [Alexa Teacher Models](#) (AlexaTM 20B), a 20-billion-parameter sequence-to-sequence (seq2seq) language model that exhibits state-of-the-art performance on 1-shot and few-shot NLP tasks. AlexaTM 20B outperforms [GPT-3](#) on [SuperGLUE](#) and [SQuADv2](#) benchmarks while having fewer than 1/8 the number of parameters.

Credits: [InfoQ](#)