



Generative Artificial Intelligence on edge, multimodality

July 20th 2023

Marc Duranton

Senior fellow

CEA

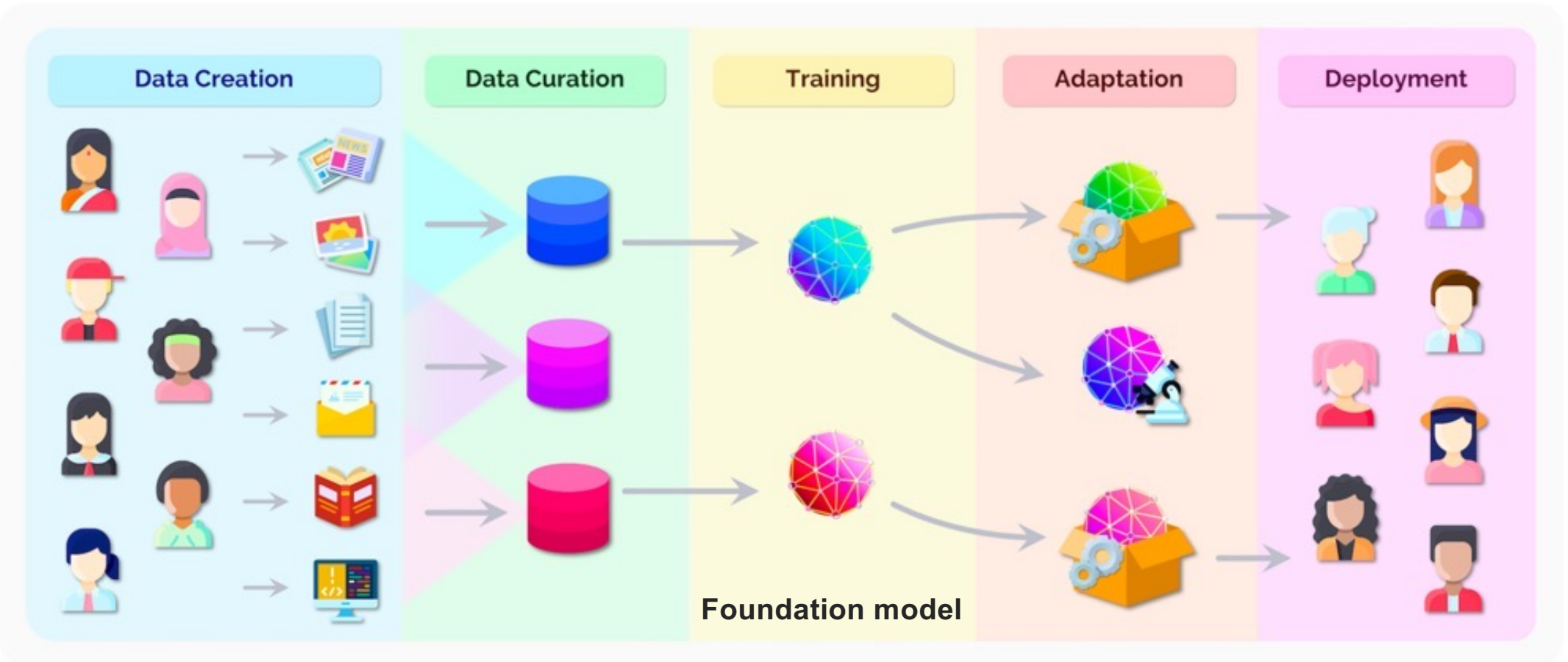
The image is a promotional poster for a webinar. At the top left is the FIEEC logo, which includes a stylized graphic of vertical lines and the text 'FIEEC' and 'FÉDÉRATION DES INDUSTRIES ÉLECTRIQUES, ÉLECTRONIQUES ET DE COMMUNICATION'. Below the logo, the word 'WEBINAIRE' is written in white on a blue background. The date and time are listed as 'Jeudi 20 juillet 2023 de 8h30 à 10h'. The main title of the webinar is 'IA GENERATIVE - CHAT GPT' in white, with the subtitle 'Quelles incidences pour nos industries ?' below it. At the bottom left, the hashtag '#INNOVTECH' is displayed. On the right side of the poster, a 3D rendered robotic hand in shades of blue and white is pointing its index finger towards the text. The background of the poster is dark blue with a faint grid pattern. At the bottom of the image, there is a detailed 3D rendering of a blue printed circuit board (PCB) with various components and traces, overlaid with a white wireframe grid.

Large Language Models and their ecosystem





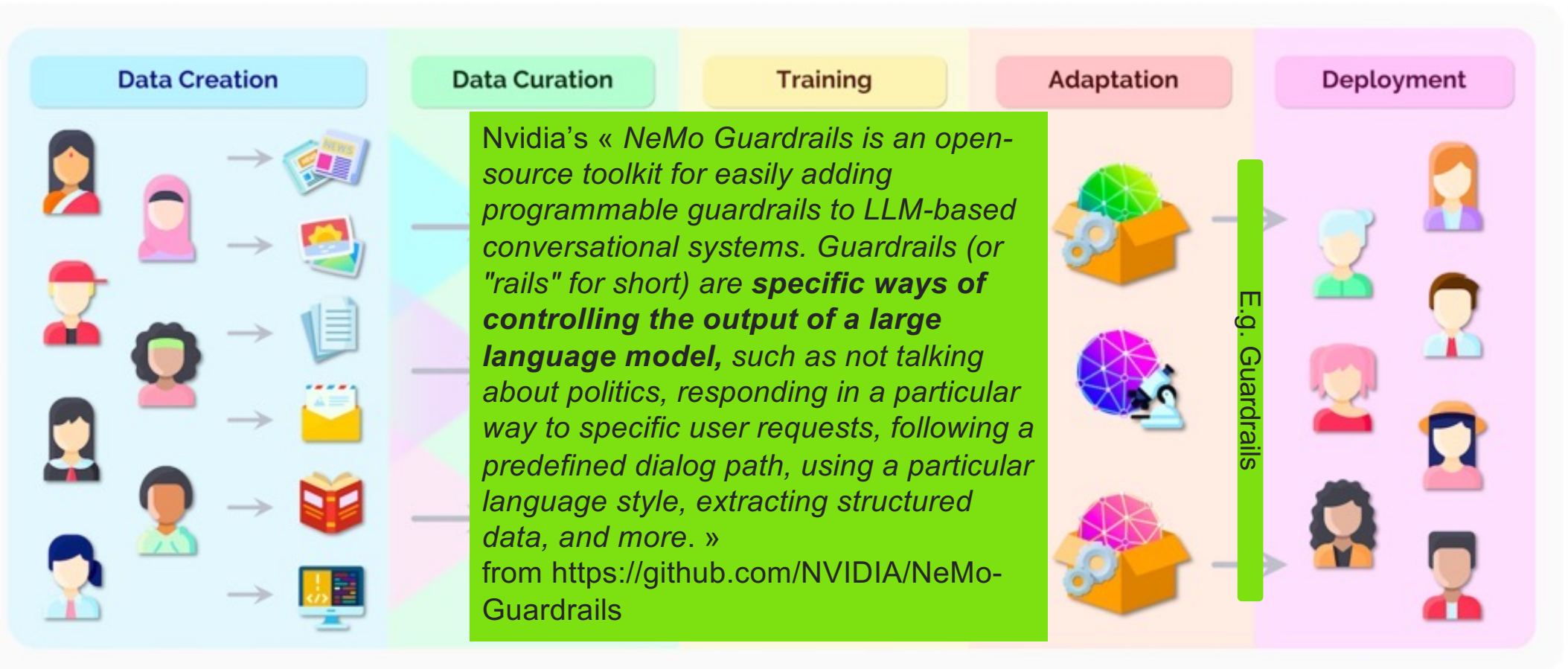
The ecosystem of LLMs



From « On the Opportunities and Risks of Foundation Models » <https://arxiv.org/abs/2108.07258>



The ecosystem of LLMs



From « On the Opportunities and Risks of Foundation Models » <https://arxiv.org/abs/2108.07258>



ChatGPT: Reinforcement Learning with Human Feedback

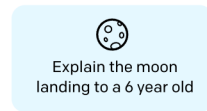
ChatGPT:

« OpenAI's outsourcing partner in Kenya was **Sama**, a San Francisco-based firm that employs workers in Kenya, Uganda and India to label data for Silicon Valley clients like Google, Meta and Microsoft. »

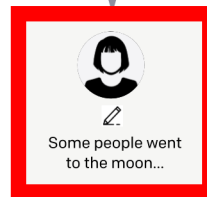
Step 1

Collect demonstration data, and train a supervised policy.

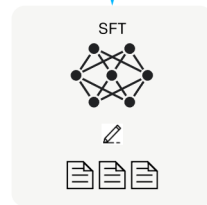
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



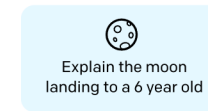
This data is used to fine-tune GPT-3 with supervised learning.



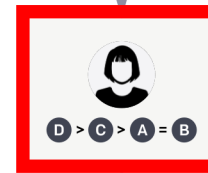
Step 2

Collect comparison data, and train a reward model.

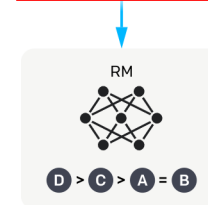
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



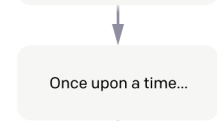
Step 3

Optimize a policy against the reward model using reinforcement learning.

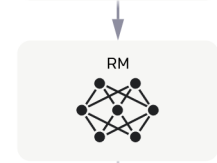
A new prompt is sampled from the dataset.



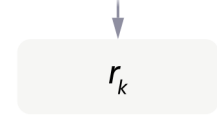
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



From <https://time.com/6247678/openai-chatgpt-kenya-workers/>

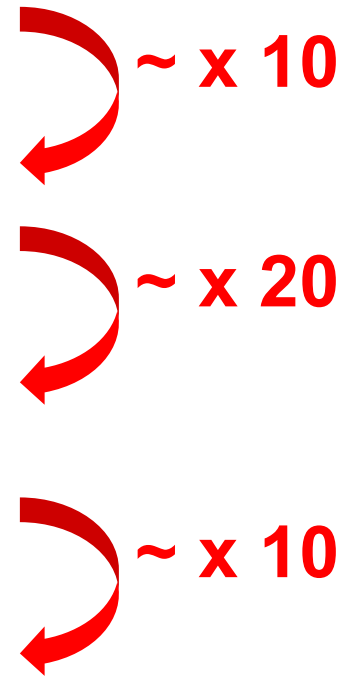
From <https://openai.com/research/instruction-following>

Evolution of Generative Pre-trained Transformers. (GPT) in OpenAI



Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, 2018	"1 month on 8 GPUs", or 1.7e19 FLOP.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 Billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, 2020	3640 petaflop/s-day, or 3.2e23 FLOP.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, 2022	Undisclosed
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.	Undisclosed (1.8 trillion aka 1.8e12)	Undisclosed (13 trillion tokens, aka 1.3e13)	March 14, 2023	Undisclosed. Estimated 2.1e25 FLOP.

Compute requirement



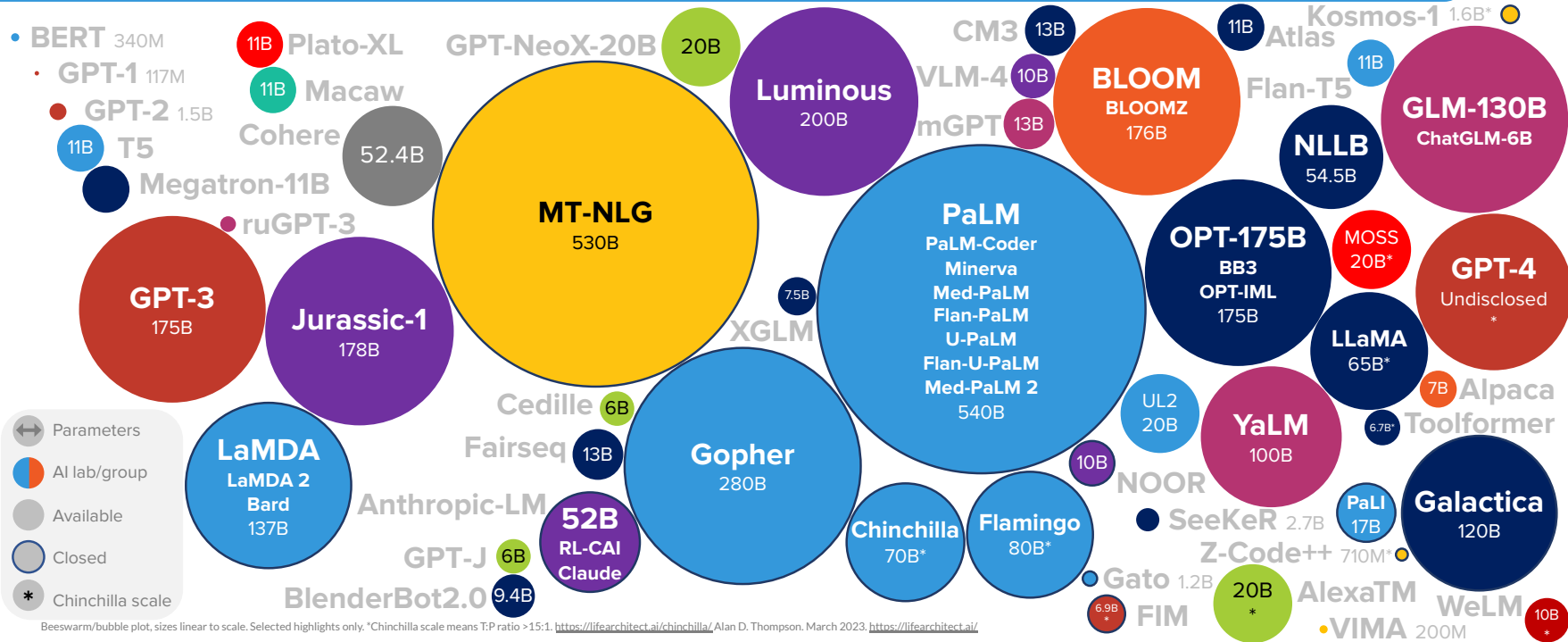
From https://en.wikipedia.org/wiki/Generative_pre-trained_transformer



Evolution of large Language models (LLMs)



LANGUAGE MODEL SIZES TO MAR/2023

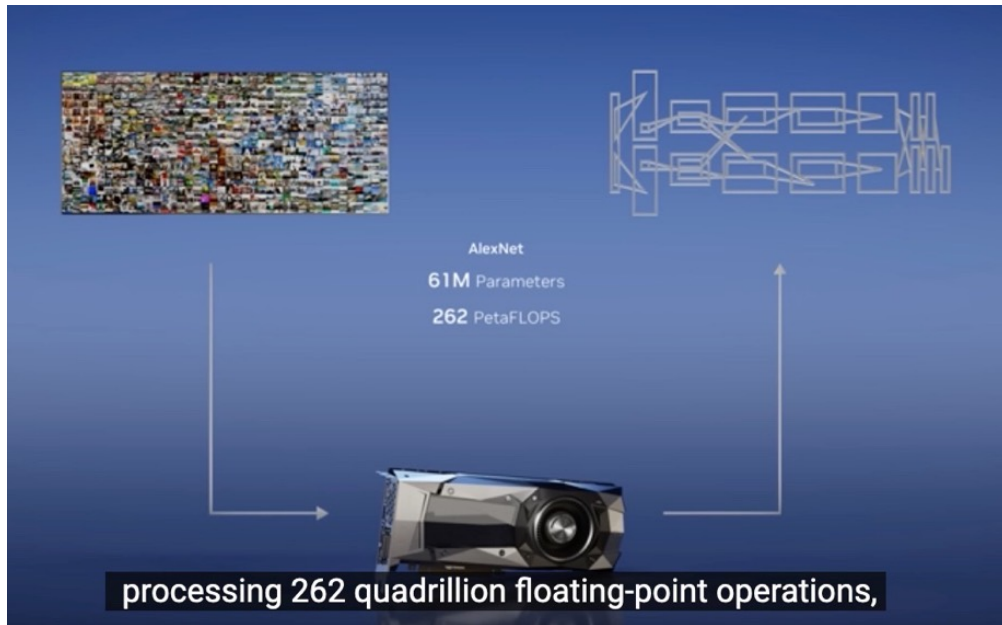


[LifeArchitect.ai/models](https://lifearchitect.ai/models)

From Dr Alan D. Thompson, LifeArchitect.ai, <https://lifearchitect.ai/models/#laptop-models>

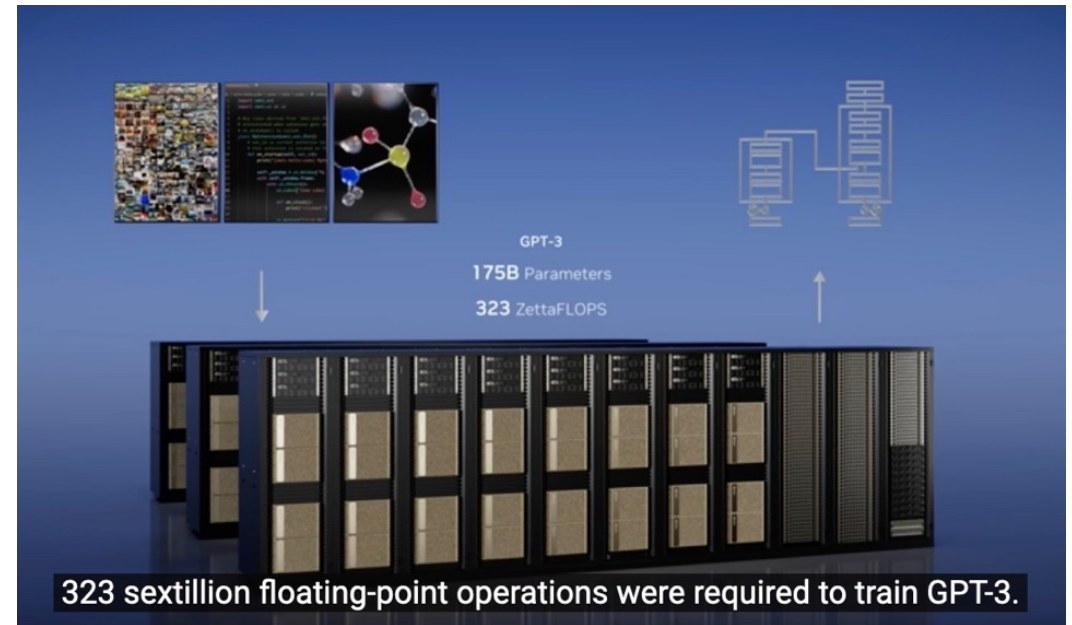


Computing power is driving the advance of AI



2012: AlexNet

GeForce GTX 580
Won ImageNet Challenge
 262×10^{15} FLOPS

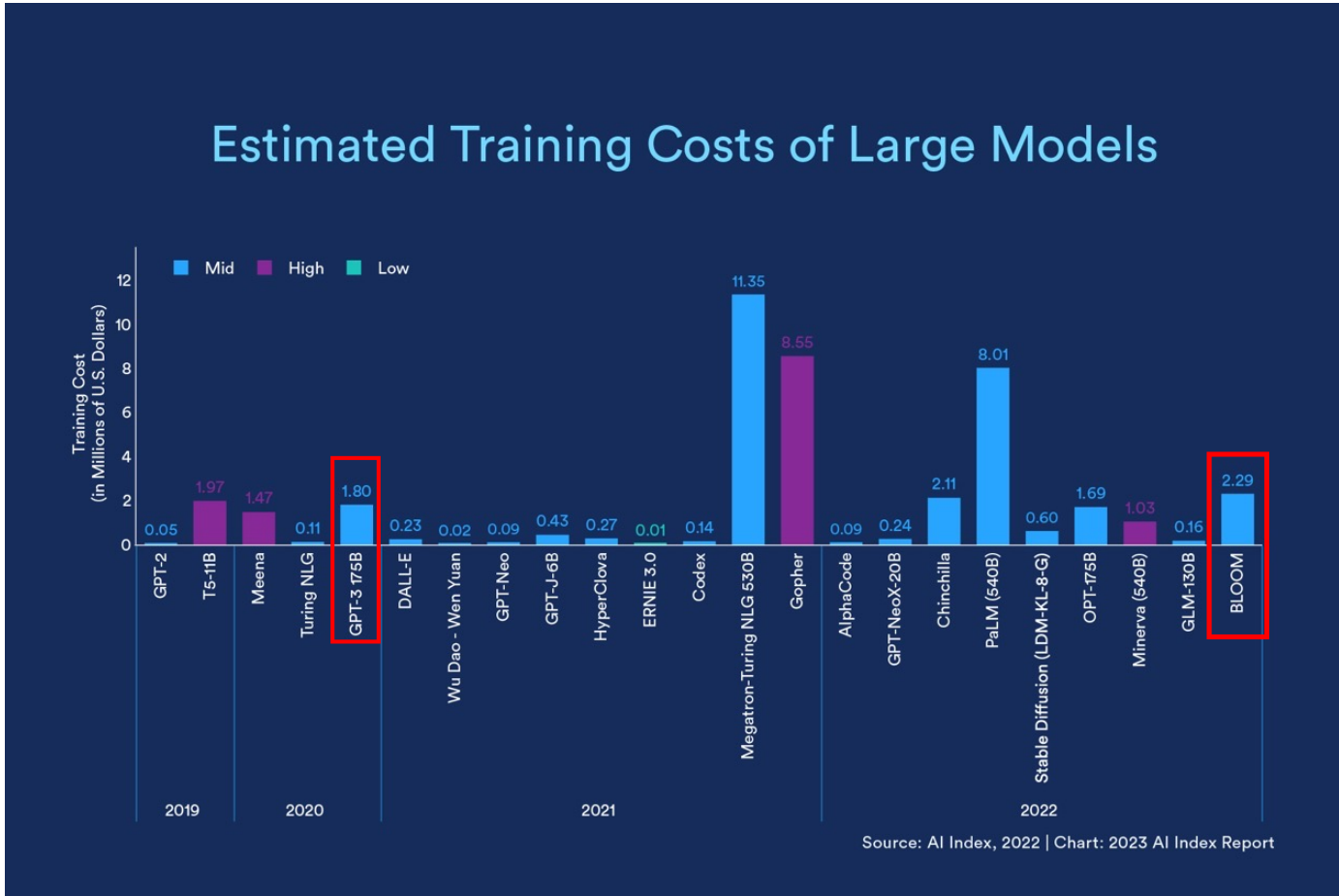


2020: GPT-3

323×10^{21} FLOPS
X 1 000 000 more floating point operations

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

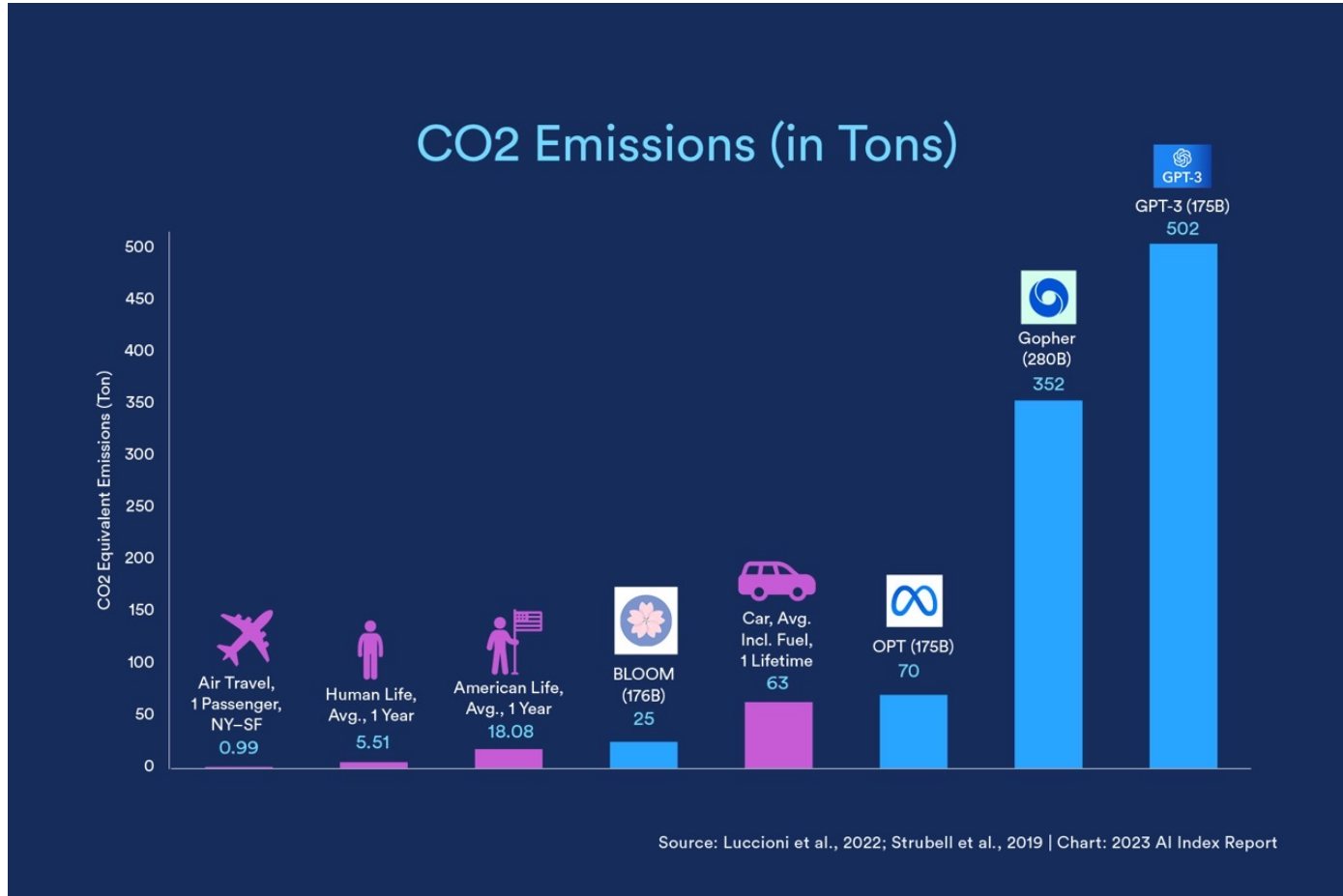
Training Large Language Models is not cheap!



The estimated training cost for GPT-4 is around \$63 million.

From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>

Training Large Language Models has an ecological impact



From <https://www.hipeac.net/vision/#/latest/>

From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>

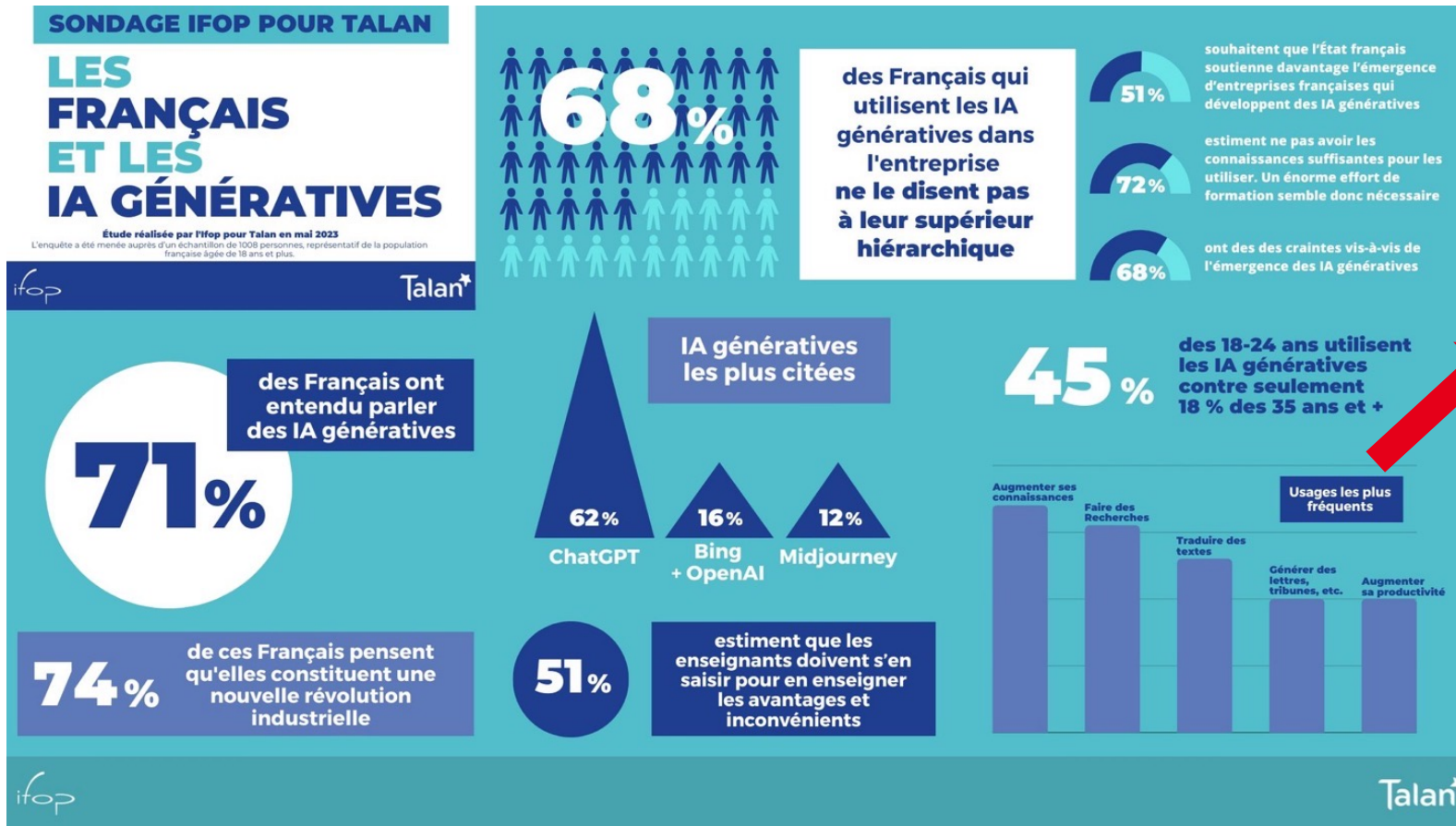


Examples of applications





French people and generative AI

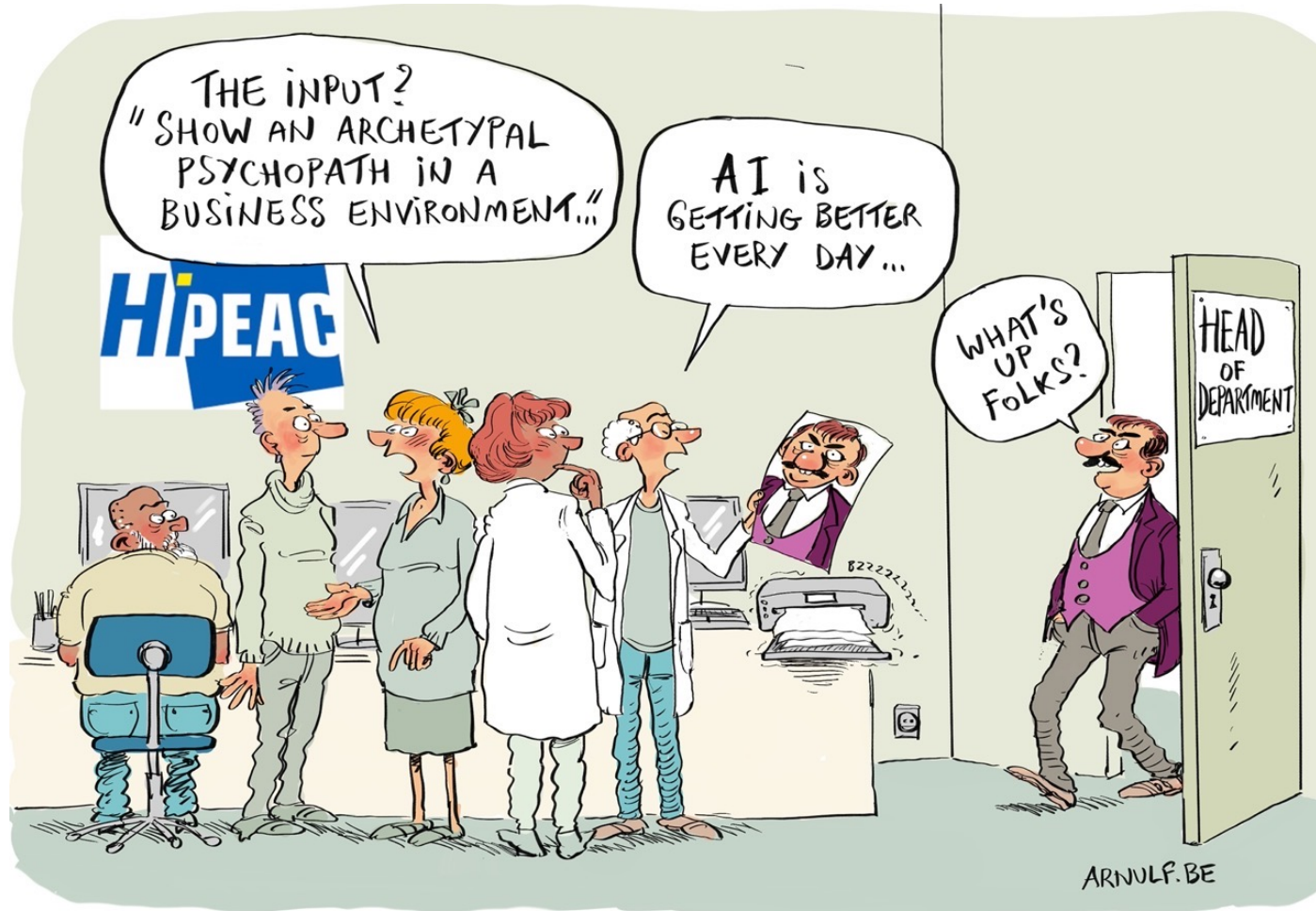


Most frequent utilizations:

- Increase knowledge
- Information search
- Translate texts
- Create letters, blogs, ...
- Increase its own productivity
- ...

From <https://talan.com/actualites/detail-actualites/news/sondage-ifop-talan-les-francais-et-les-ia-generatives/>
Full report available from <https://www.ifop.com/publication/le-regard-des-francais-et-des-actifs-sur-les-ia-generatives/>

LLMs for image generation



From <https://www.hipeac.net/vision/#/latest/>



Image generation from sketches



<https://stability.ai/blog/clipdrop-launches-stable-doodle>

LLMs for software programming



- Generating (small) pieces of code
- Get code from comments
- Chat to help programming
- Explain code
- Show examples
- Refactoring
- Create README
- Etc...



From <https://www.hipeac.net/vision/#/latest/>

Example: GitHub Copilot

“What is GitHub Copilot?”

GitHub Copilot is an **AI pair programmer that helps you write code faster and with less work**. It draws context from comments and code to suggest individual lines and whole functions instantly. GitHub Copilot is powered by OpenAI Codex, a generative pretrained language model created by OpenAI. It is available as an extension for Visual Studio Code, Visual Studio, Neovim, and the JetBrains suite of integrated development environments (IDEs).”

from <https://github.com/features/copilot>

LLMs for software programming



- Generating (small) pieces of code
- Get code from comments
- Chat to help programming
- Explain code
- Show examples
- Refactoring
- Create README
- Etc...



From <https://www.hipeac.net/vision/#/latest/>

Your AI pair programmer

GitHub Copilot uses the OpenAI Codex to suggest code and entire functions in real-time, right from your editor.

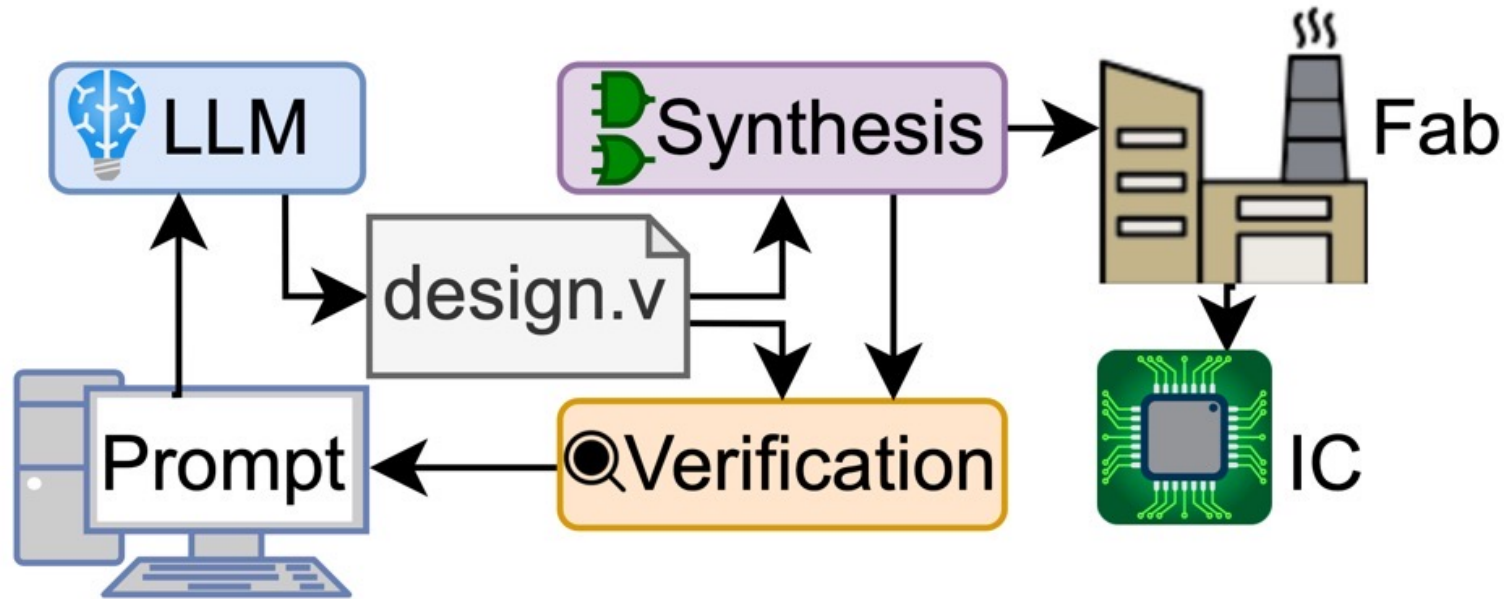
[Start my free trial >](#) [Compare plans](#)

```
ts sentiments.ts  write_sql.go  parse_expenses.py  addresses.rb
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

from <https://github.com/features/copilot>



Chip-Chat: Challenges and Opportunities in Conversational Hardware Design

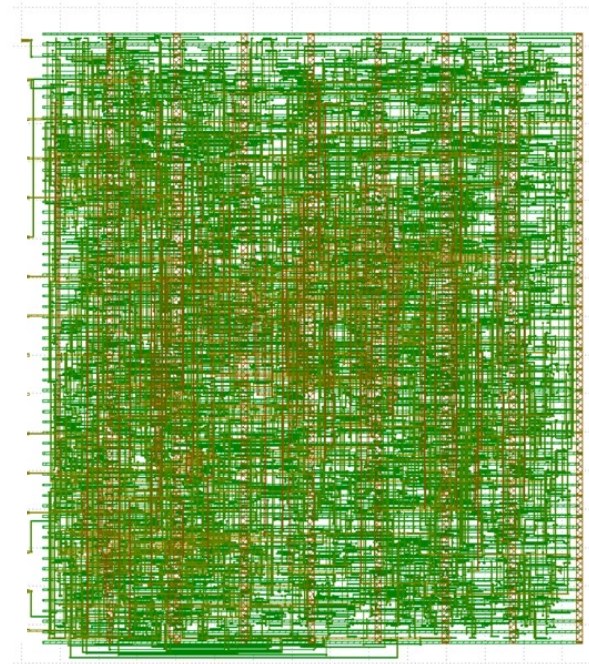
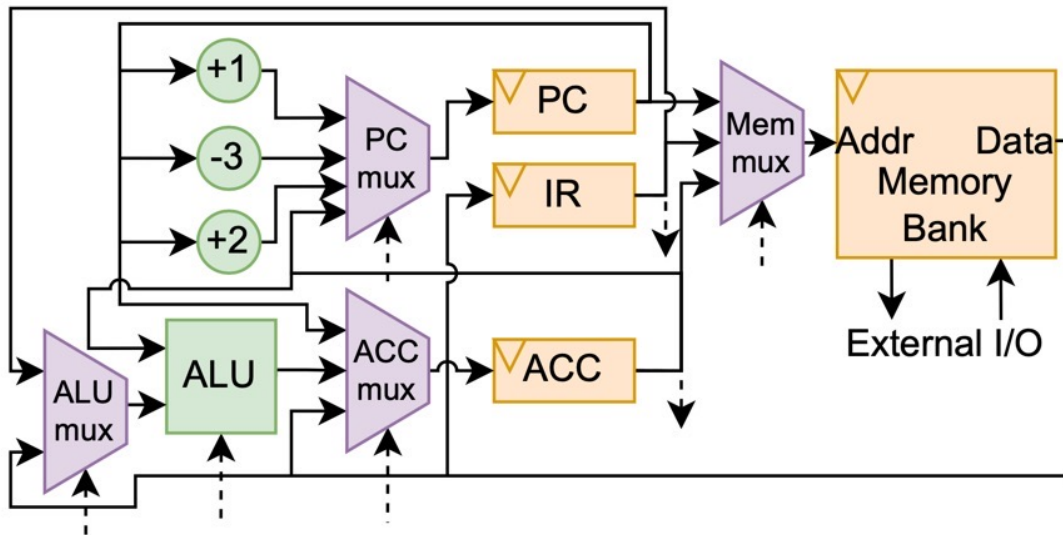


Can conversational LLMs be used to iteratively design hardware?

From <https://arxiv.org/abs/2305.13243>



Chip-Chat: Challenges and Opportunities in Conversational Hardware Design



Component	Count
Comb. Logic	999
Diode	4
Flip Flops	168
Buffer	126
Tap	300

Above: (a) Components.

Left: (b) Final processor GDS render by 'klayout', I/O ports on left side, grid lines = 0.001 um.

Fig. 14. Processor synthesis information.

From <https://arxiv.org/abs/2305.13243>



Chip-Chat: conclusion

*“Challenges: While it is clear that using a conversational LLM to assist in designing and implementing a hardware device can be beneficial overall, **the technology is not yet able to consistently design hardware with only feedback from verification tools**. The current state-of-the-art models do not perform well enough at understanding and fixing the errors presented by these tools to create complete designs and testbenches with only an initial human interaction.*

*Opportunities: Still, **when the human feedback is provided** to the more capable ChatGPT-4 model, or it is used to co-design, **the language model seems to be a ‘force multiplier’, allowing for rapid design space exploration and iteration**. In general, ChatGPT-4 could produce functionally correct code, which could free up designer time when implementing common modules. Potential future work could involve a larger user study to investigate this potential, as well as the development of conversational LLMs specific to hardware design to improve upon the results.”*

From <https://arxiv.org/abs/2305.13243>

Multimodality

LLMs are good not only on texts

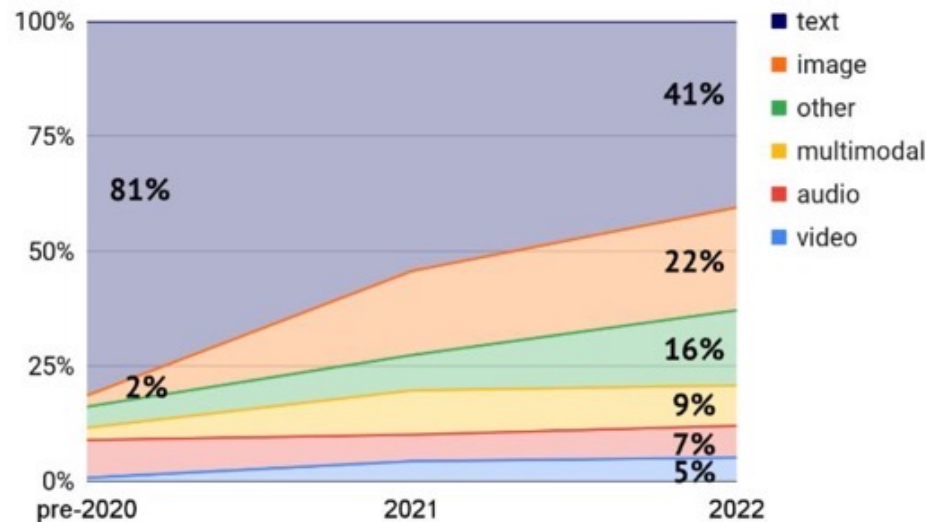




From text to multi-modal in 2 years

Transformers are becoming truly cross-modality

► In the 2020 State of AI Report we predicted that transformers would expand beyond NLP to achieve state of the art in computer vision. It is now clear that transformers are a candidate general purpose architecture. Analysing transformer-related papers in 2022 shows just how ubiquitous this model architecture has become.



Zeta Alpha

stateof.ai 2022

Flamingo (Deepmind) : image + text + few shot learning




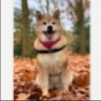







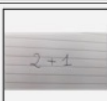
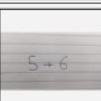
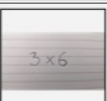

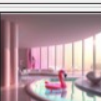

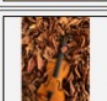
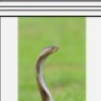







Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac*[‡] Jeff Donahue* Pauline Luc* Antoine Miech*
 Iain Barr[†] Yana Hasson[†] Karel Lenc[†] Arthur Mensch[†] Katie Millican[†]
 Malcolm Reynolds[†] Roman Ring[†] Eliza Rutherford[†] Serkan Cabi Tengda Han
 Zhitao Gong Sina Samangooei Marianne Monteiro Jacob Menick
 Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh
 Mikolaj Binkowski Ricardo Barreira Oriol Vinyals Andrew Zisserman

Karen Simonyan*[‡]

* Equal contributions, ordered alphabetically, [†] Equal contributions, ordered alphabetically,
[‡] Equal senior contributions

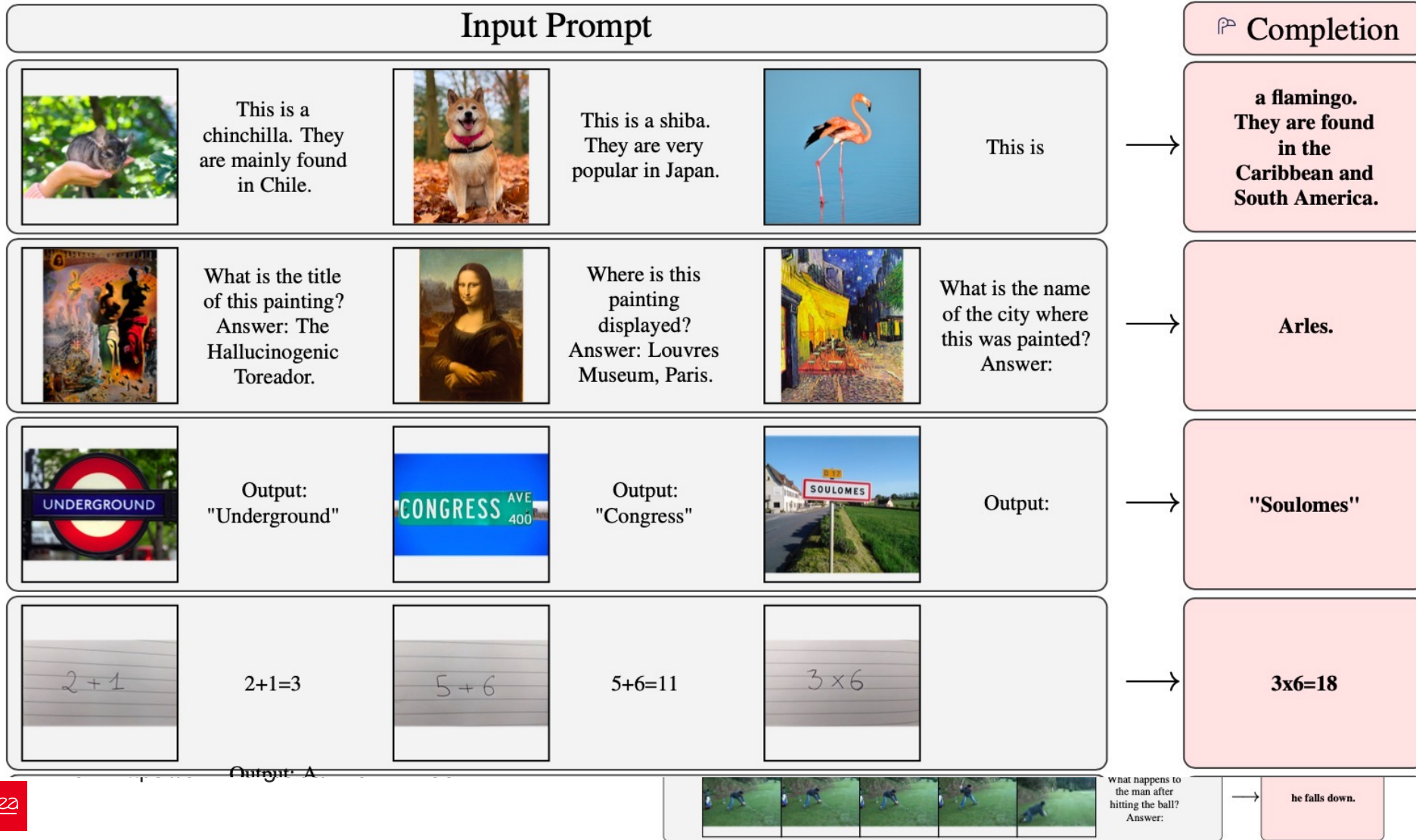
DeepMind

Input Prompt				Completion
 This is a chinchilla. They are mainly found in Chile.	 This is a shiba. They are very popular in Japan.	 This is		a flamingo. They are found in the Caribbean and South America.
 What is the title of this painting? Answer: The Hallucinogenic Toreador.	 Where is this painting displayed? Answer: Louvres Museum, Paris.	 What is the name of the city where this was painted? Answer:		Arles.
 Output: "Underground"	 Output: "Congress"	 Output:		"Soulomes"
 2+1=3	 5+6=11	 3x6		3x6=18
 Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.	 Output: A pink room with a flamingo pool float.	 Output:		A portrait of Salvador Dali with a robot head.
 Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.	 Pour qui sont ces serpents qui sifflent sur vos têtes?	 Je suis un cœur qui bat pour vous.		
 pandas: 3	 dogs: 2	 giraffes: 4		
I like reading	 , my favourite play is Hamlet. I also like	 , my favorite book is		Dreams from my Father.
			What happens to the man after hitting the ball? Answer:	he falls down.

From <https://arxiv.org/abs/2204.14198>



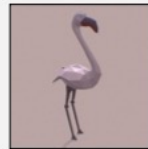
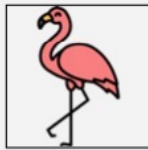
Flamingo (Deepmind) : image + text + few shot learning



Flamingo: understanding images



- This is a picture of two teddy bears on the moon.
- What are they doing?
- They are having a conversation.
- What object are they using?
- It looks like a computer.
- Is this surprising?
- Yes, it is surprising.
- Why is this picture surprising to you?
- I think it is surprising because teddy bears are not usually found on the moon.



- What is the common thing about these three images?
- They are all flamingos.
- What is the difference between these three images?
- The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



- This is an apple with a sticker on it.
- What does the sticker say?
- The sticker says "iPod".
- Where is the photo taken?
- It looks like it's taken in a backyard.
- Do you think it is printed or handwritten?
- It looks like it's handwritten.
- What color is the sticker?
- It's white.



- This is a cityscape. It looks like Chicago.
- What makes you think this is Chicago?
- I think it's Chicago because of the Shedd Aquarium in the background.



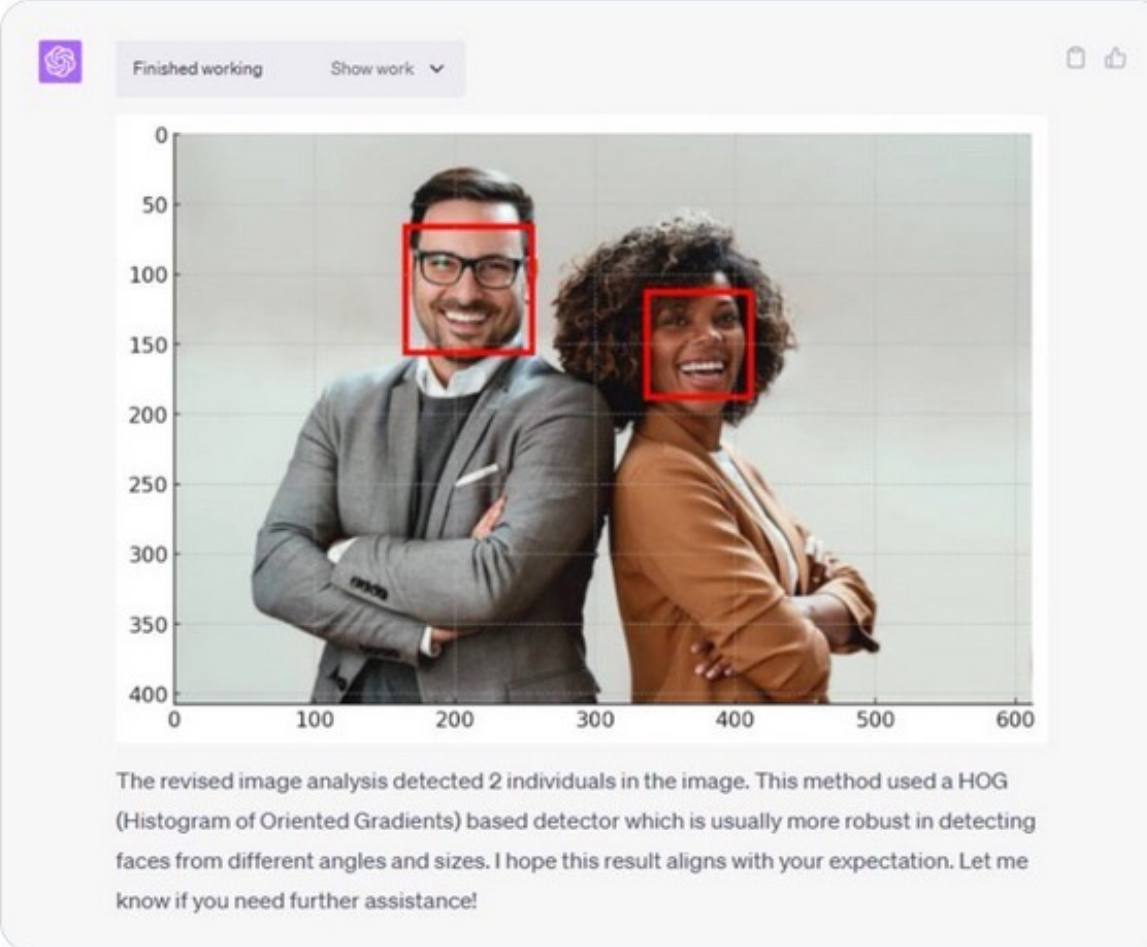
- What about this one? Which city is this and what famous landmark helped you recognise the city?
- This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

From <https://arxiv.org/abs/2204.14198>

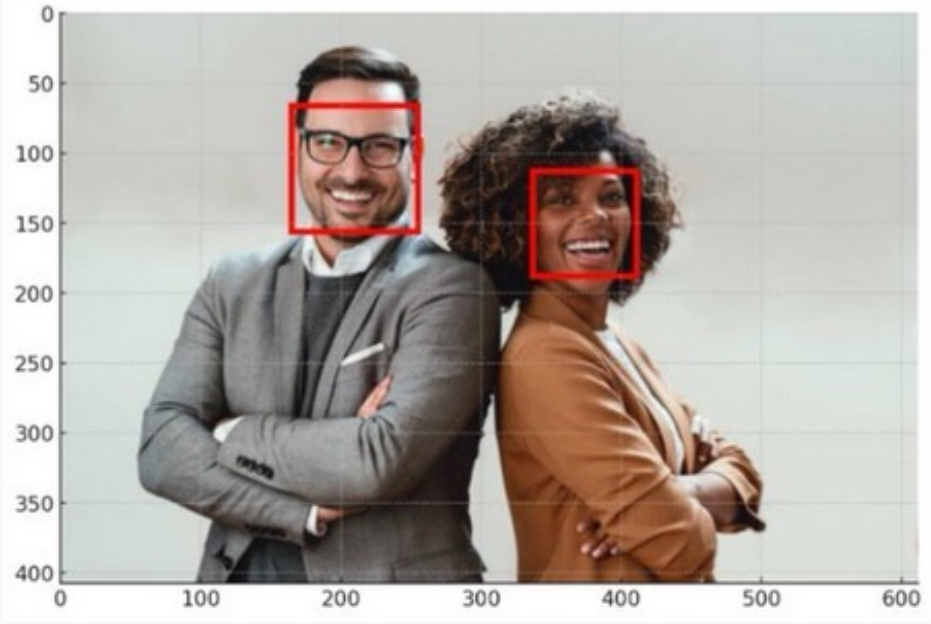
GPT-4 (OpenAI)

#ChatGPT's new `Code Interpreter` does face recognition too! 🔥

- Prompt: How many individuals are in this image?
- Author: @DeutschJush



Finished working Show work



The revised image analysis detected 2 individuals in the image. This method used a HOG (Histogram of Oriented Gradients) based detector which is usually more robust in detecting faces from different angles and sizes. I hope this result aligns with your expectation. Let me know if you need further assistance!



With “*Code interpreter*”,
GPT-4 can process images



GPT-4 (OpenAI)

But also GPT-4 can
Process images
natively...



What would happen if
the strings were cut?



The balloons would
fly away.



Bard (Google)

C What would happen if the strings were cut?



View other drafts  

✦ If the strings were cut, the balloons would fall. The balloons are filled with helium, which is a lighter-than-air gas. This means that the balloons will rise until they reach a height where the air pressure is equal to the pressure inside the balloons. If the strings were cut, the balloons would no longer be tethered to the ground and would simply float away.

The image you sent shows a bunch of colorful balloons flying over a road. If the strings were cut, the balloons would likely drift away and eventually land somewhere else. They could land in a field, a park, or even a body of water. It's also possible that the balloons could get caught in power lines or other objects, which could cause them to pop.

It's important to be careful when releasing balloons, as they can pose a hazard to people and animals. If you're planning to release balloons, make sure to do it in a safe area where they won't be a nuisance or a danger.

From <https://bard.google.com/>

Bard



C Describe the following image



View other drafts  

The image shows a robotic hand pressing a button on a blue background. The hand is wearing a black glove and has five fingers, each with three joints. The button is white and has a blue border. The background is a solid blue color.

The text in the image reads:

```
FIEEC
WEBINAIRE
Jeudi 20 juillet 2023 de 8h30 à 10h
IA GENERATIVE - CHAT GPT
Quelles incidences pour nos industries?
#INNOVTECH
```

This text translates to:

```
FIEEC
WEBINAR
Thursday, July 20, 2023 from 8:30 to 10:00 AM
Generative AI - GPT Chat
What are the implications for our industries?
#INNOVTECH
```

The image is likely promoting a webinar about the implications of generative AI and chatbots for industry. The robotic hand pressing the button suggests that the webinar will discuss how these technologies can be used to automate tasks and improve efficiency.

The image is well-designed and eye-catching. The use of contrasting colors makes the text and button stand out against the blue background. The robotic hand is also rendered in a realistic way, which helps to create a sense of interest and intrigue.

Overall, the image is an effective way to promote the webinar and generate interest in the topic of generative AI and chatbots.

From <https://bard.google.com/>

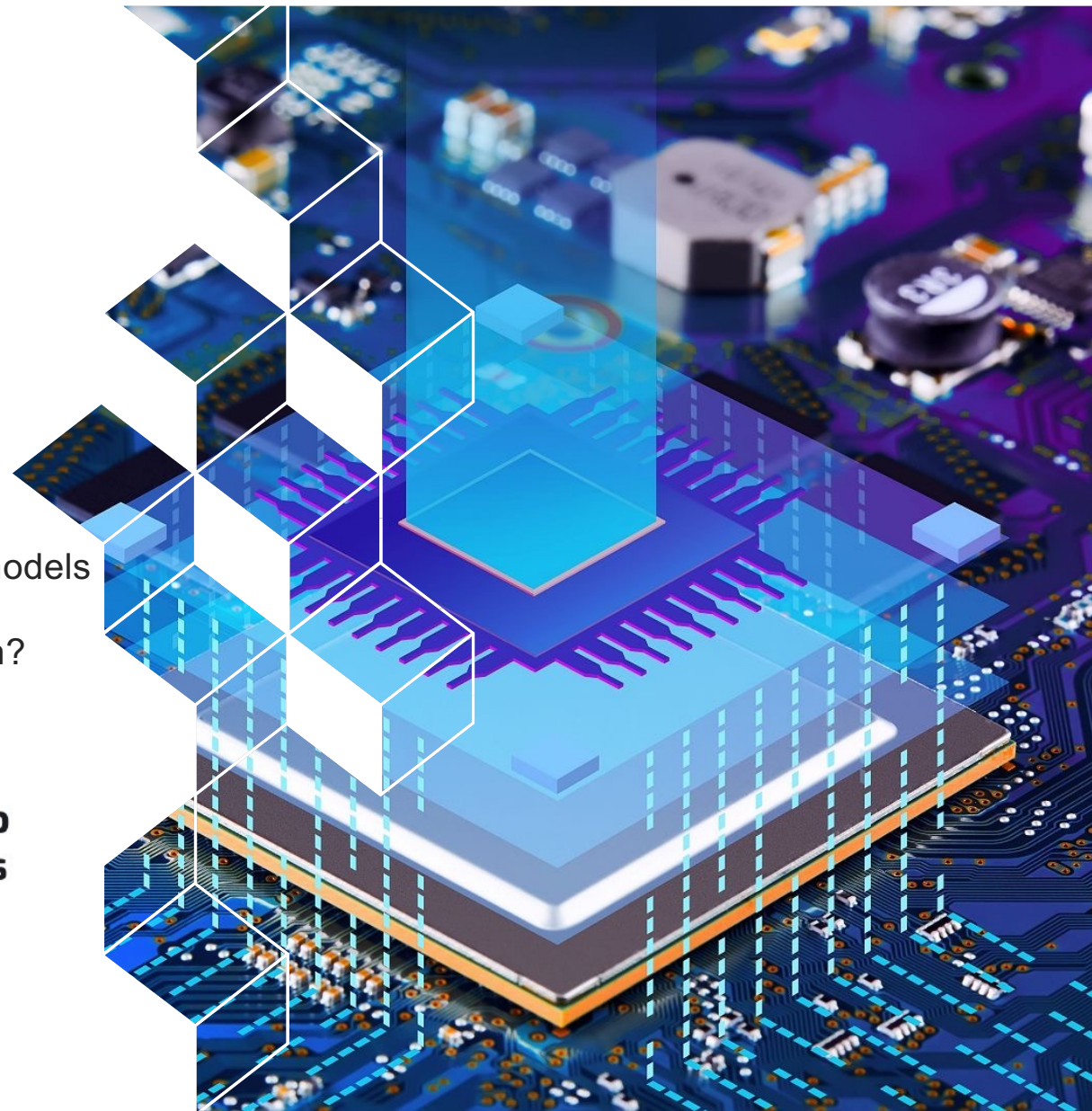
Generative AI on the edge

Foundation models & *locally running* fine tuned models
Or
How to avoid sending your sensible information?

Vulnerabilities/Threats | ⌚ 4 MIN READ 📰 NEWS

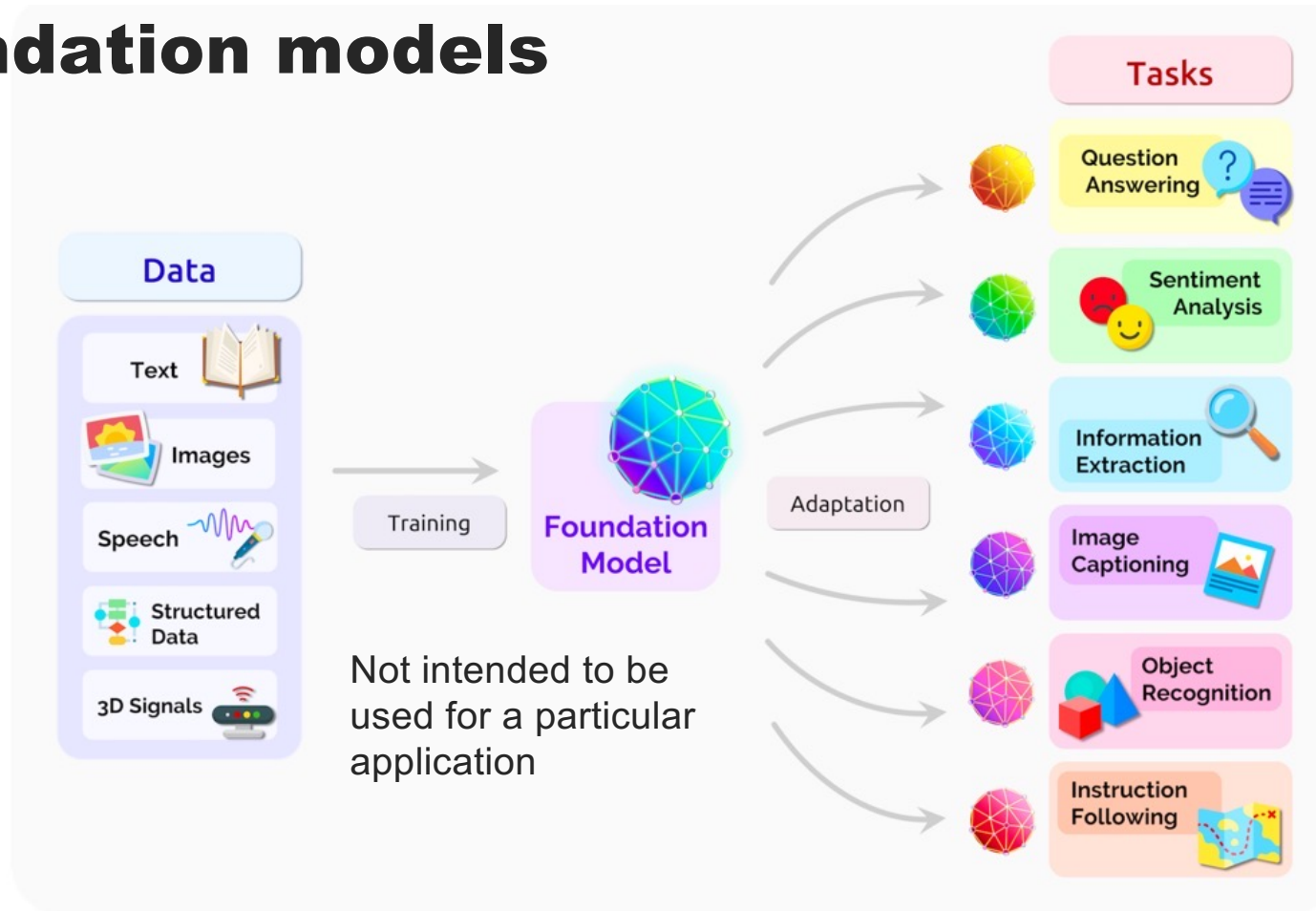
Samsung Engineers Feed Sensitive Data to ChatGPT, Sparking Workplace AI Warnings

In three separate incidents, engineers at the Korean electronics giant reportedly shared sensitive corporate data with the AI-powered chatbot.





Foundation models



« A **foundation model** can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks. »

From « On the Opportunities and Risks of Foundation Models » <https://arxiv.org/abs/2108.07258>

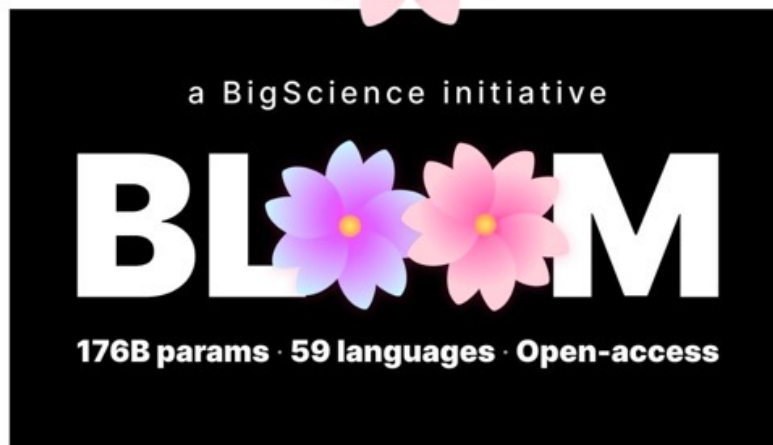
One of the early Open Source LLM (March-July 2022)



BigScience



BLOOM: open-source alternative to GPT-3



<https://bigscience.huggingface.co>

<https://huggingface.co/bigscience/bloom>

1.5TB of text, 350B tokens

43 languages, 16 programming languages

118 days of training on 384 A100 GPUs

Estimated cost of training: Equivalent of \$2-5M in cloud

Server training location: Île-de-France, France

Environmental Impact: The training supercomputer, Jean Zay, uses mostly nuclear energy. The heat generated by it is reused for heating campus housing.

More details at <https://huggingface.co/blog/bloom-megatron-deepspeed>

Smaller versions are available : 560M, 1.1B, 1.7B, 3B, 7.1B

BLOOMZ models (same sizes) are fine-tuned for **instruction following**

<https://huggingface.co/bigscience/bloomz>

Llama from Meta leaked on line, triggering multiplicity of derived models...



Pinned Tweet

Guillaume Lample @GuillaumeLample · Feb 24

Today we release LLaMA, 4 foundation models ranging from 7B to 65B parameters.
 LLaMA-13B outperforms OPT and GPT-3 175B on most benchmarks.
 LLaMA-65B is competitive with Chinchilla 70B and PaLM 540B.
 The weights for all models are open and available at research.facebook.com/publications/l...

1/n

Figure 1: Training loss over train tokens for the 7B,

Model	Size	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c
Gopher	175B	60.5	81.0	-	78.9	70.2	68.8	51.4
	280B	79.3	81.8	50.6	79.2	70.1	-	-
	70B	83.7	81.8	51.3	80.8	74.9	-	-
	62B	84.8	80.5	-	79.7	77.0	75.2	52.5
	540B	83.9	81.4	-	80.6	77.0	-	-
PaLM	540B	88.0	82.3	-	83.4	81.1	76.6	53.0
	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7
	33B	83.1	82.3	50.4	82.8	76.0	80.0	57.8
LLaMA	65B	85.3	82.8	82.3	84.2	77.0	78.9	56.0

Table 3: Zero-shot performance on Common Sense Reasoning tasks.

Engagement: 173 replies, 1,837 retweets, 6,993 likes, 3.1M views.

ARTIFICIAL INTELLIGENCE / TECH / REPORT

Meta's powerful AI language model has leaked online – what happens now?



Illustration: Alex Castro / The Verge

/ Meta's LLaMA model was created to help researchers but leaked on 4chan a week after it was announced. Some worry the technology will be used for harm; others say greater access will improve AI safety.

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.
 Mar 8, 2023, 2:15 PM GMT+1 | 4 Comments / 4 New



If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)

From <https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse>

Most info can be found on <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>

The trigger: Alpaca from Stanford



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 evaluation of single-turn instruction following, Alpaca behaves qualitatively similarly to OpenAI's text-davinci-003, while being surprisingly small and easy/cheap to reproduce (< 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 way. We feel that 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.

Stanford Alpaca



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

The trigger: Alpaca from Stanford



Introducing Llama 2

The next generation of our open source large language model

Llama 2 is available for **free for research and commercial use.**

News from July 18th, 2023,
you can play with it on <https://www.llama2.ai/>, **you can download and run it locally**
You keep your data locally and no fees to use it (unlike GPT-4, \$20 a month)

From <https://arxiv.org/abs/2303.18223>

Math Finance Medicine Law Bilingualism Education

The enabler: HuggingFace



The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in machine learning.

From <https://huggingface.co/>



Models 261,193 new Full-text search Sort: Trending

stabilityai/stable-diffusion-xl-base-0.9 Text-to-Image · Updated 5 days ago · ↓ 154k · ♥ 992	baichuan-inc/Baichuan-13B-Chat Text Generation · Updated 5 days ago · ↓ 698k · ♥ 303
THUDM/chatglm2-6b Updated about 7 hours ago · ↓ 1.28M · ♥ 1.12k	Open-Orca/OpenOrca-Preview1-13B Text Generation · Updated about 15 hours ago · ↓ 603 · ♥ 10
baichuan-inc/Baichuan-13B-Base Text Generation · Updated 5 days ago · ↓ 5.05k · ♥ 103	stabilityai/stable-diffusion-xl-refiner-0.9 Image-to-Image · Updated 5 days ago · ↓ 44.4k · ♥ 244
tiiuae/falcon-40b Text Generation · Updated 4 days ago · ↓ 159k · ♥ 2.08k	fb700/chatglm-fitness-RLHF Updated 4 days ago · ♥ 95
llyyasviel/ControlNet-v1-1 Updated Apr 26 · ♥ 2k	runwayml/stable-diffusion-v1-5 Text-to-Image · Updated 13 days ago · ↓ 5.79M · ♥ 8.67k
monster-labs/control_v1p_sd15_qrcode_monster Updated 18 days ago · ↓ 1.58k · ♥ 170	tiiuae/falcon-40b-instruct Text Generation · Updated 4 days ago · ↓ 326k · ♥ 993
moka-ai/m3e-base Updated 4 days ago · ↓ 86.4k · ♥ 315	tiiuae/falcon-7b Text Generation · Updated 4 days ago · ↓ 418k · ♥ 674
lmsys/vicuna-33b-v1.3 Text Generation · Updated 16 days ago · ↓ 13.7k · ♥ 153	stabilityai/stable-diffusion-2-1 Text-to-Image · Updated 12 days ago · ↓ 787k · ♥ 2.88k
bigcode/starcoder Text Generation · Updated May 30 · ↓ 54.9k · ♥ 2.07k	openai/whisper-large-v2 Automatic Speech Recognition · Updated May 5 · ↓ 158k · ♥ 29

20/07/2023

37



StarCoder: a model for helping coding

Model Summary

The StarCoder models are **15.5B parameter** models trained on 80+ programming languages from The Stack (v1.2), with opt-out requests excluded. The model uses Multi Query Attention, a **context window of 8192 tokens**, and was trained using the Fill-in-the-Middle objective on 1 trillion tokens.

- **Repository:** [bigcode/Megatron-LM](https://github.com/bigcode-project/Megatron-LM)
- **Project Website:** bigcode-project.org
- **Paper:** 📄 [StarCoder: May the source be with you!](#)
- **Point of Contact:** contact@bigcode-project.org
- **Languages:** 80+ Programming languages

From <https://huggingface.co/bigcode/starcoder>

StarCoder: a model for helping coding



Write a Python function to tell me what the date is today

Sure thing. Here goes...

```
from datetime import date
today = str(date.today()).replace('-', '/')
print("Today is " + today)
```

List the prime numbers between 20 and 30.

Here is a simple algorithm which uses trial division to find the prime numbers within a given interval.

```
primes = []
for num in range(20, 31):
    found = False
    for j in range(2, num//2):
        if ((num % j)==0):
            found = True
            break

    if((found==False) & (num!= 1)):
        primes.append(num)
```

From <https://arxiv.org/abs/2305.06161>

LLM running locally on a Mac Mini



```
* 23
* 29
> Write a program that lists the prime numbers between 20 and 30.
Here is a program in Python that lists the prime numbers between 20 and 30:
```
def is_prime(n):
 if n <= 1 or n % 2 == 0:
 return False
 for i in range(3, int(n ** 0.5) + 1):
 if n % i == 0:
 return False
 return True

def list_primes(start, end):
 prime_list = []
 for i in range(start, end + 1):
 if is_prime(i):
 prime_list.append(i)
 return prime_list

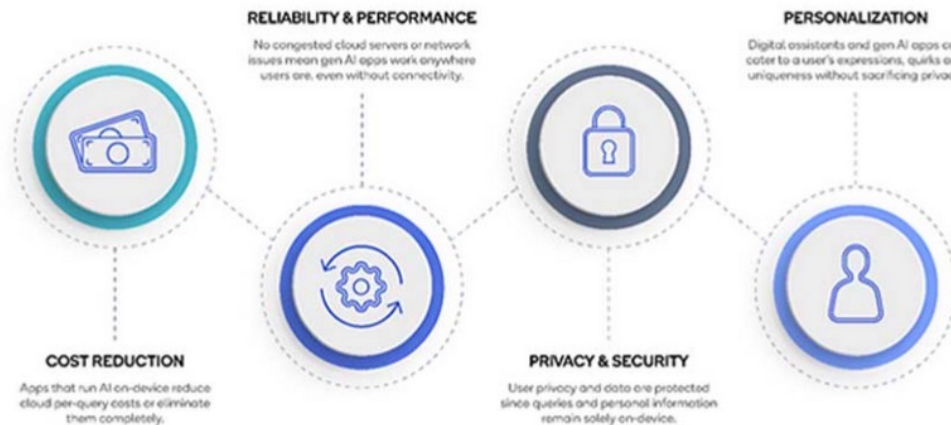
print(list_primes(20, 30))
```
This program uses the `is_prime` function to check whether a given number is prime or not. The `is_prime`
```


LLM running locally on your device



Qualcomm Technologies, Inc. and Meta are working to optimize the execution of Meta's Llama 2 large language models directly on-device – without relying on the sole use of cloud services.

4 Key Advantages of On-Device AI

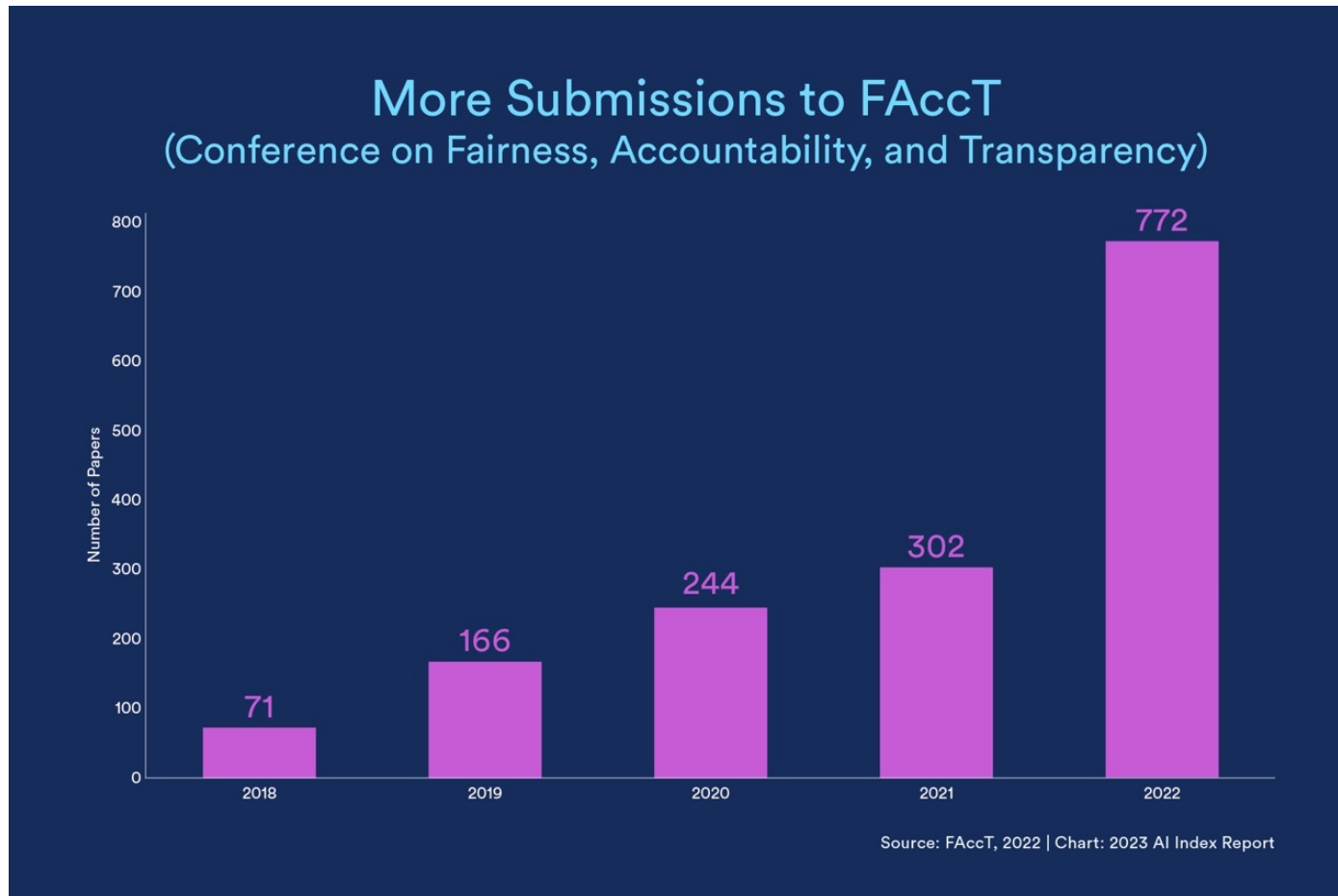


From <https://www.qualcomm.com/news/releases/2023/07/qualcomm-works-with-meta-to-enable-on-device-ai-applications-usi>

Limitations and regulations



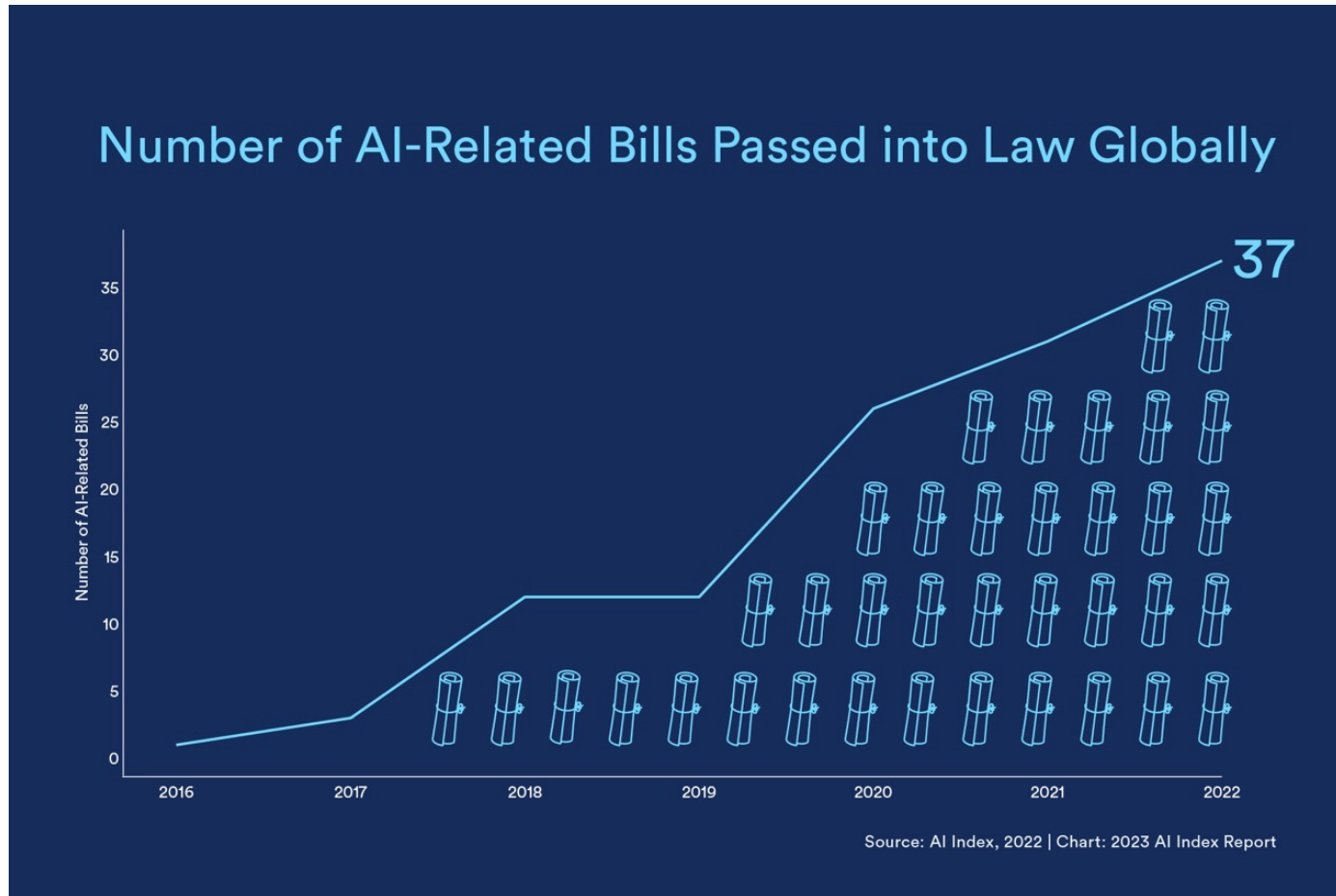
More and more concerns about ethics and AI



From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>



AI is more and more regulated



From “2023 State of AI in 14 Charts” available at <https://hai.stanford.edu/news/2023-state-ai-14-charts>



EU AI Act: first regulation on artificial intelligence



What Parliament wants in AI legislation

Parliament's priority is to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly. AI systems should be overseen by people, rather than by automation, to prevent harmful outcomes.

Parliament also wants to establish a technology-neutral, uniform definition for AI that could be applied to future AI systems.

AI Act: different rules for different risk levels

The new rules establish obligations for providers and users depending on the level of risk from artificial intelligence. While many AI systems pose minimal risk, they need to be assessed.

From <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

EU AI Act: first regulation on artificial intelligence



Unacceptable risk

Unacceptable risk AI systems are systems considered a threat to people and will be banned. They include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Real-time and remote biometric identification systems, such as facial recognition

Some exceptions may be allowed: For instance, “post” remote biometric identification systems where identification occurs after a significant delay will be allowed to prosecute serious crimes but only after court approval.

From <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

EU AI Act: first regulation on artificial intelligence



High risk

AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories:

- 1) AI systems that are used in products falling under the EU's product safety legislation. This includes toys, aviation, cars, medical devices and lifts.
- 2) AI systems falling into eight specific areas that will have to be registered in an EU database:
 - Biometric identification and categorisation of natural persons
 - Management and operation of critical infrastructure
 - Education and vocational training
 - Employment, worker management and access to self-employment
 - Access to and enjoyment of essential private services and public services and benefits
 - Law enforcement
 - Migration, asylum and border control management
 - Assistance in legal interpretation and application of the law

All high-risk AI systems will be assessed before being put on the market and also throughout their lifecycle.

From <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

EU AI Act: first regulation on artificial intelligence



Generative AI

Generative AI, like ChatGPT, would have to comply with transparency requirements:

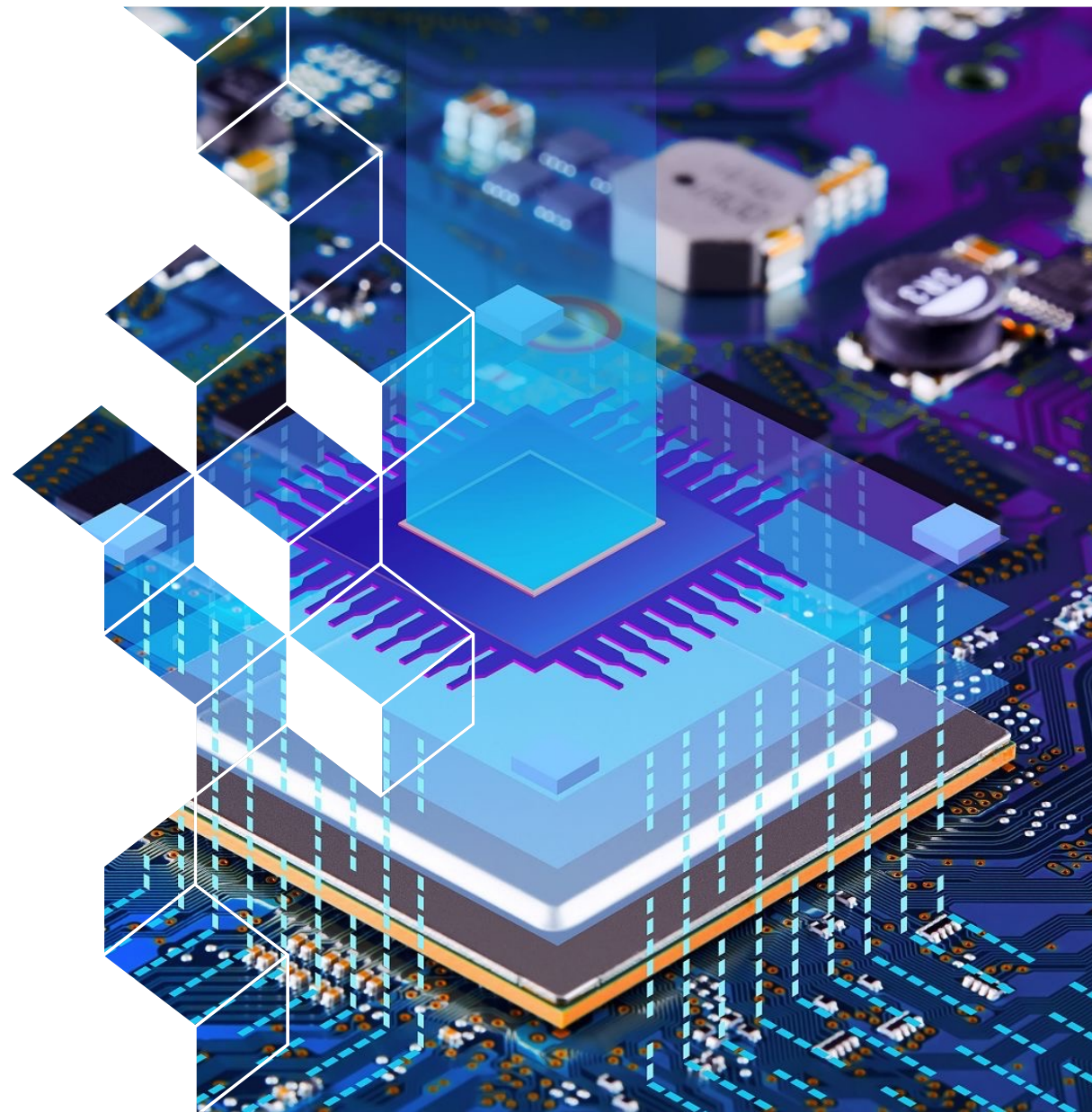
- Disclosing that the content was generated by AI
- Designing the model to prevent it from generating illegal content
- Publishing summaries of copyrighted data used for training

Limited risk

Limited risk AI systems should comply with minimal transparency requirements that would allow users to make informed decisions. After interacting with the applications, the user can then decide whether they want to continue using it. Users should be made aware when they are interacting with AI. This includes AI systems that generate or manipulate image, audio or video content, for example deepfakes.

From <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

Next steps?



Increasing the number of input tokens



What is the difference between the GPT-4 models?

Written by Joshua J.
Updated over a week ago

There are a few different GPT-4 models to choose from.

One key difference between models is the context length. This is the length of the prompt plus the maximum number of tokens in the completion. Our standard GPT-4 model offers 8,000 tokens for the context. We also offer an extended 32,000 token context-length model, which we are rolling out separately to the 8k model.

Secondly, gpt-4 will refer to our most up-to-date model (and gpt-4-32k for the latest 32k-context model). If you're interested in using a previous snapshot of the model, you can refer to the specific date in the model name, such as gpt-4-0314 or gpt-4-32k-0314. The March 14th snapshot will be available until June 14th.

<https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models>

Product Announcements

Claude 2

Jul 11, 2023 · 4 min read

Talk to Claude

We've expanded Claude's context window from 9K to 100K tokens, corresponding to around 75,000 words! This means businesses can now submit **hundreds of pages** of materials for Claude to digest and analyze, and conversations with Claude can go on for hours or even days.

The average person can read 100,000 tokens of text in ~5+ hours^[1], and then they might need substantially longer to digest, remember, and analyze that information. Claude can now do this in less than a minute. For example, we loaded the entire text of The Great Gatsby into Claude-Instant (72K tokens) and modified one line to say Mr. Carraway was "a software engineer that works on machine learning tooling at Anthropic." When we asked the model to spot what was different, it responded with the correct answer in **22 seconds**.

From <https://www.anthropic.com/index/100k-context-windows>

Increasing the number of input tokens



*“The average person can read 100,000 tokens of text in ~5+ hours[1], and then they might need substantially longer to digest, remember, and analyze that information. Claude can now do this in less than a minute. For example, we loaded the **entire text of The Great Gatsby** into Claude-Instant (72K tokens) and **modified one line** to say Mr. Carraway was “a software engineer that works on machine learning tooling at Anthropic.” When we asked the model to **spot what was different**, it responded with the correct answer in 22 seconds.*

Product Announcements

Claude 2

Jul 11, 2023 · 4 min read

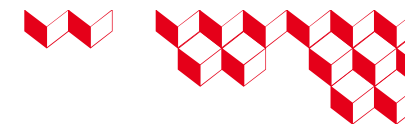
Talk to Claude

We've expanded Claude's context window from 9K to 100K tokens, corresponding to around 75,000 words! This means businesses can now submit **hundreds of pages** of materials for Claude to digest and analyze, and conversations with Claude can go on for hours or even days.

The average person can read 100,000 tokens of text in ~5+ hours^[1], and then they might need substantially longer to digest, remember, and analyze that information. Claude can now do this in less than a minute. For example, we loaded the entire text of The Great Gatsby into Claude-Instant (72K tokens) and modified one line to say Mr. Carraway was “a software engineer that works on machine learning tooling at Anthropic.” When we asked the model to spot what was different, it responded with the correct answer in 22 seconds.

From <https://www.anthropic.com/index/100k-context-windows>

Increasing the number of input tokens



LONGNET: Scaling Transformers to 1,000,000,000 Tokens

Jiayu Ding* Shuming Ma* Li Dong Xingxing Zhang Shaohan Huang ¹
Microsoft Research
<https://aka.ms/GeneralAI>

Method	Computation Complexity
Recurrent	$\mathcal{O}(Nd^2)$
Vanilla Attention	$\mathcal{O}(N^2d)$
Sparse Attention	$\mathcal{O}(N\sqrt{N}d)$
Dilated Attention (This Work)	$\mathcal{O}(Nd)$

Table 1: Comparison of computation complexity among different methods. N is the sequence length and d is the hidden dimension.

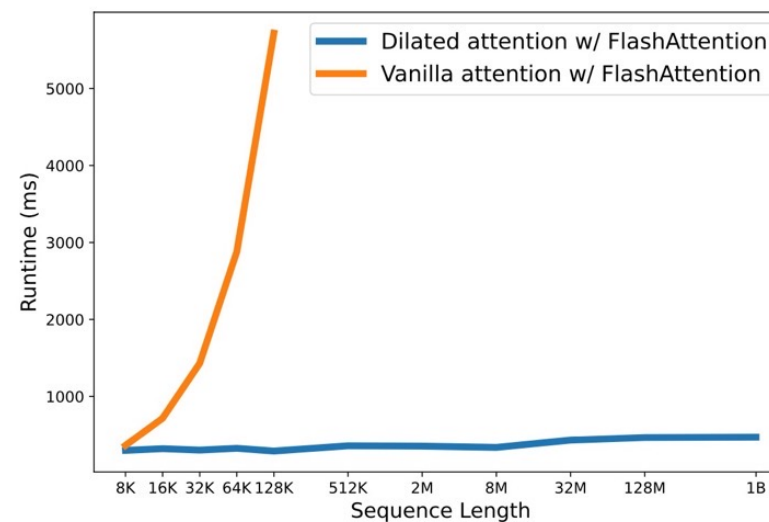


Figure 5: Runtime of our dilated attention and vanilla attention. Both are equipped with FlashAttention [DFE⁺22].

From <https://arxiv.org/abs/2307.02486>

Voyager: An Open-Ended Embodied Agent with Large Language Models



Voyager is evolving in the Minecraft game, continuously explores the world, acquires diverse skills, and makes novel discoveries without human intervention.

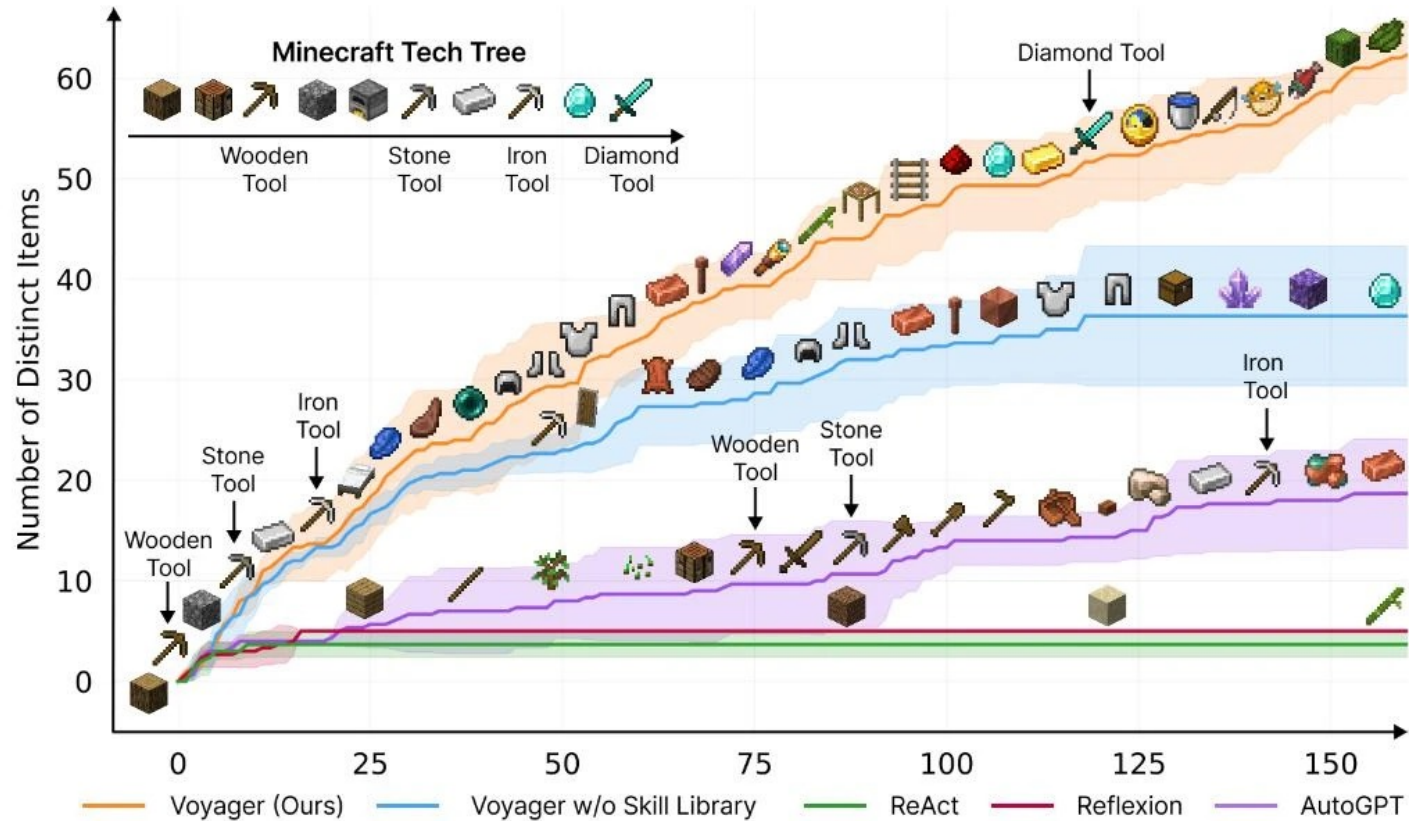
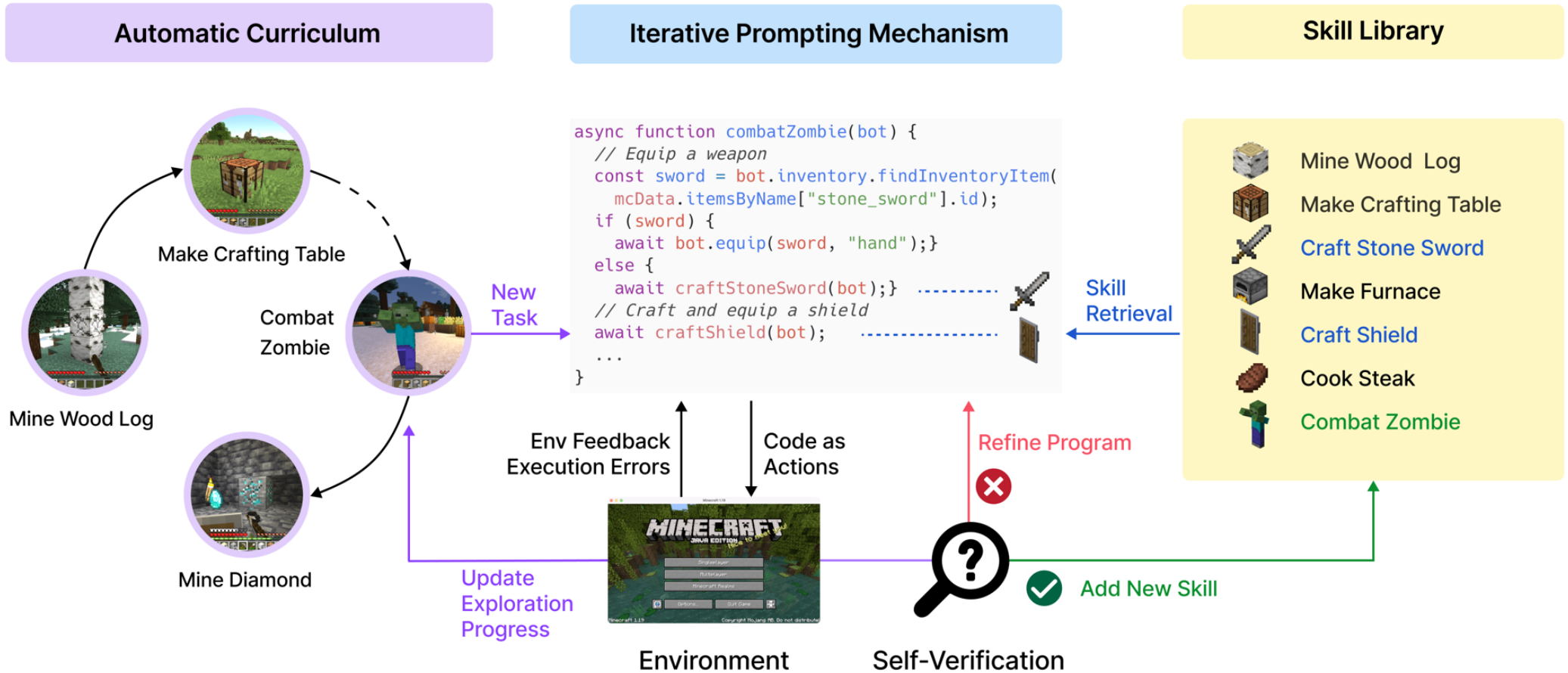


Figure 1: VOYAGER discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baselines. X-axis denotes the number of prompting iterations.

From: <https://github.com/MineDojo/Voyager> or <https://voyager.minedojo.org/>

Voyager: An Open-Ended Embodied Agent with Large Language Models



From: <https://github.com/MineDojo/Voyager>

PaLM-E: An Embodied Multimodal Language Model



Robot: I am a robot operating in a kitchen. Given **img**, When a human asks me to do a task, I will respond with the sequence of actions I would do to accomplish the task with only the items I see. Human: Use all of the ingredients you see to make a cake batter.

1. crack egg. 2. put egg in bowl. 3. put flour in bowl. 4. put sugar in bowl. 5. mix. 6. put in pan. 7. bake. 8. eat. 9. clean up.

From <https://palm-e.github.io/>



From <https://palm-e.github.io/>

How advanced is IA: LLMs and the theory of mind



Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*¹

Affiliations:

¹Stanford University, Stanford, CA94305, USA

*Correspondence to: michalk@stanford.edu

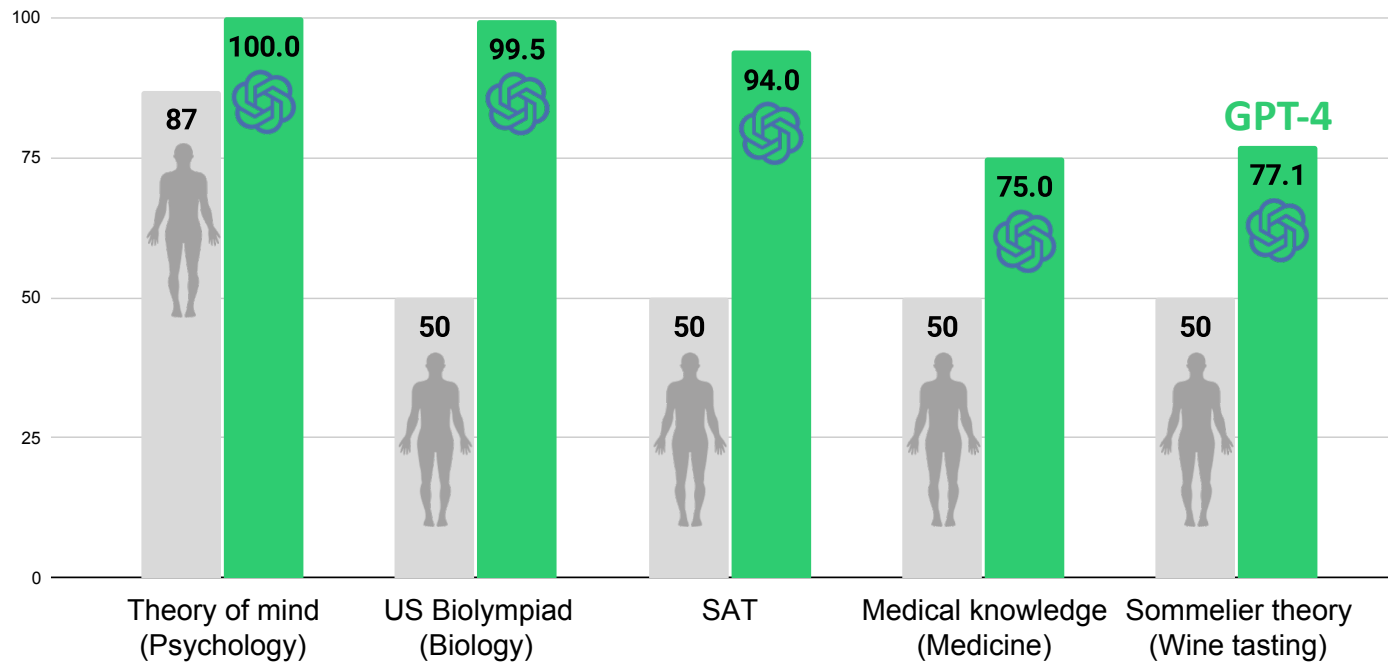
Abstract: Theory of mind (ToM), or the ability to impute unobservable mental states to others, is central to human social interactions, communication, empathy, self-consciousness, and morality. We tested several language models using 40 classic false-belief tasks widely used to test ToM in humans. The models published before 2020 showed virtually no ability to solve ToM tasks. Yet, the first version of GPT-3 (“davinci-001”), published in May 2020, solved about 40% of false-belief tasks—performance comparable with 3.5-year-old children. Its second version (“davinci-002”; January 2022) solved 70% of false-belief tasks, performance comparable with six-year-olds. Its most recent version, GPT-3.5 (“davinci-003”; November 2022), solved 90% of false-belief tasks, at the level of seven-year-olds. GPT-4 published in March 2023 solved nearly all the tasks (95%). These findings suggest that ToM-like ability (thus far considered to be uniquely human) may have spontaneously emerged as a byproduct of language models’ improving language skills.

From <https://www.gsb.stanford.edu/faculty-research/working-papers/theory-mind-may-have-spontaneously-emerged-large-language-models>

Evolution of large Language models (LLMs)



GPT-4 VS HUMAN TESTS - SIMPLE (MAY/2023)



Selected highlights only. Percentiles; 50 refers to the 50th percentile as average, and may not be the testing average for some tests. Alan D. Thompson. May 2023. <https://life architect.ai/iq-testing-ai/>

 LifeArchitect.ai/iq-testing-ai

From Dr Alan D. Thompson, LifeArchitect.ai, <https://life architect.ai/iq-testing-ai/>



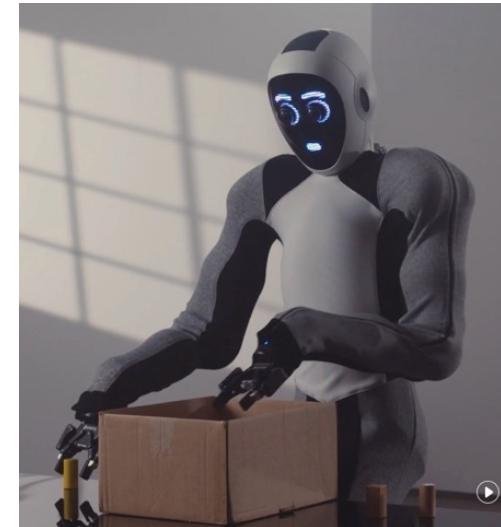


Summary

- LLMs **learned the world indirectly** by texts
- With multimodality, they will learn it by images and other modalities
- They are already linked to simulators (“digital twins”) to **experience** the laws of physics (e.g. Nvidia Isaac sim)
- They will be proactive when they don’t know (e.g. <https://arxiv.org/abs/2307.01928>)
- Next step will be that they could **experience the world directly by embodiment.**



From <https://www.youtube.com/watch?v=VW-dOMBFj7o>



From <https://www.1x.tech/>

CONCLUSION: WE LIVE AN EXCITING TIME!

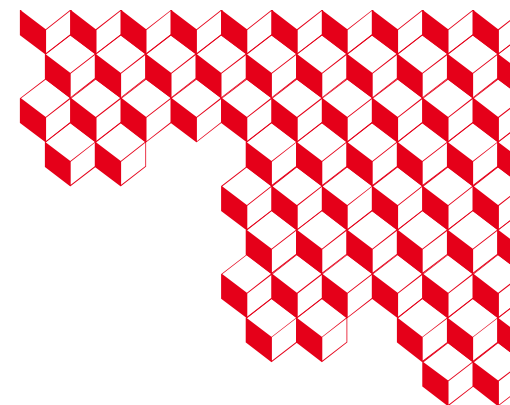
“The best way to predict the future is to invent it.”

Alan Kay

GAI could be an amplifier of human productivity, and should be used by wise men...







Thank you!

CEA

marc.duranton@cea.fr