Beyond Llama2: Future Trends And Challenges With LLMs

Au-delà de Llama2: Tendances et Défis futures avec les LLM

with

Isaac Chung

clarifai The World's AI™

TD

clarifai
The World's AI™

Isaac Chung

Senior Research Engineer

Clarifai

clarifai

The World’s AI™

Empower developers to quickly co-create, share, and use **The World’s AI^{TM}** for production.

clarifai
The World's AI™

The Best Build with Clarifai

U.S. ARMY

U.S. AIR FORCE

NVIDIA.

SIEMENS energy

wattpad

PUBLICIS GROUPE

P&G

ADVANCED DETECTION SYSTEMS
Metal Detection Experts

mate1

LOWE'S

ECS

Acquia DAM

Canva

OpenTable™

STANLEY

Viasat™

letgo

Nestlé

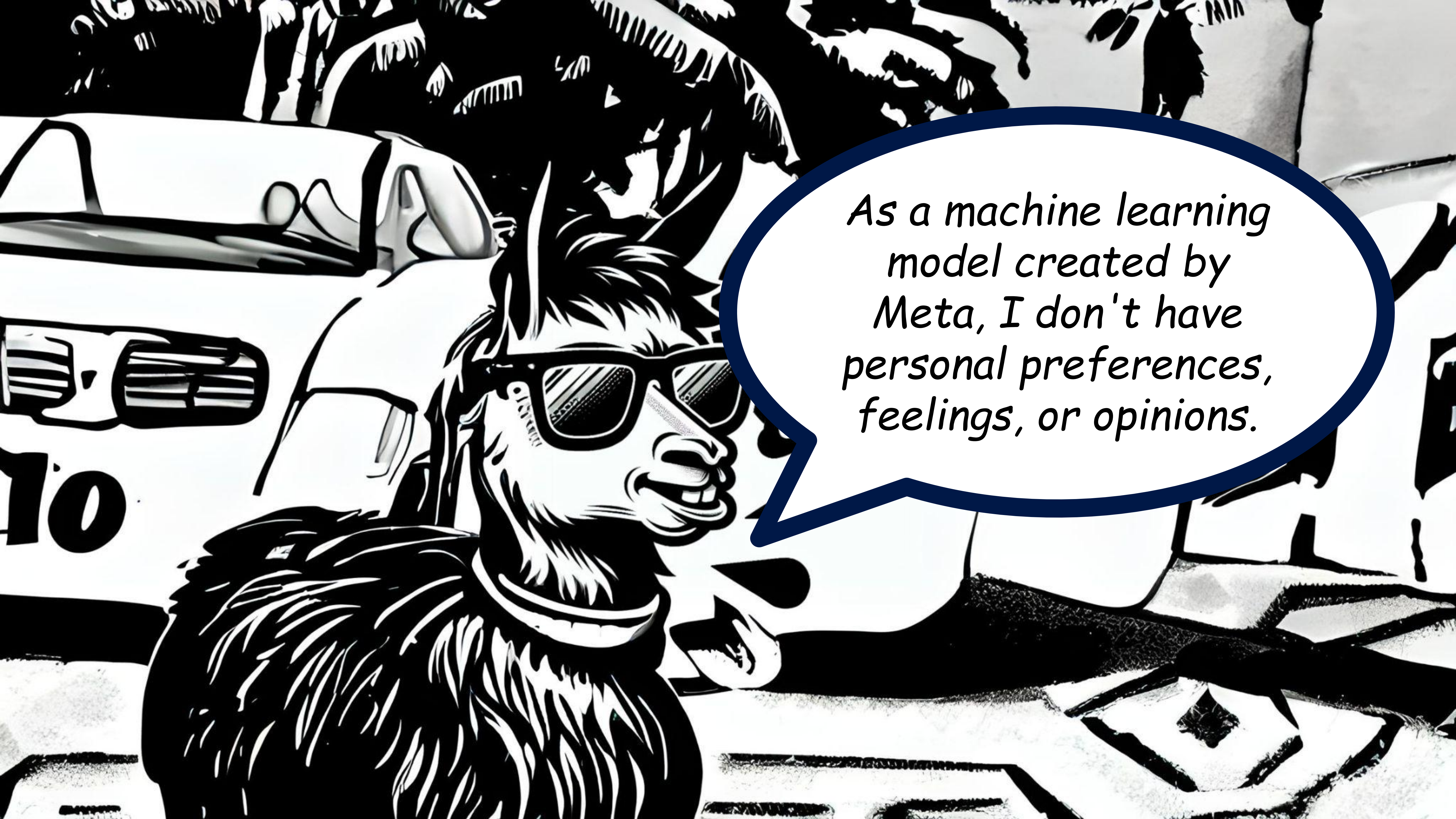
Kroger®

Humana

ZOLA

Stackla

MediaBeacon®
An ESKO Company

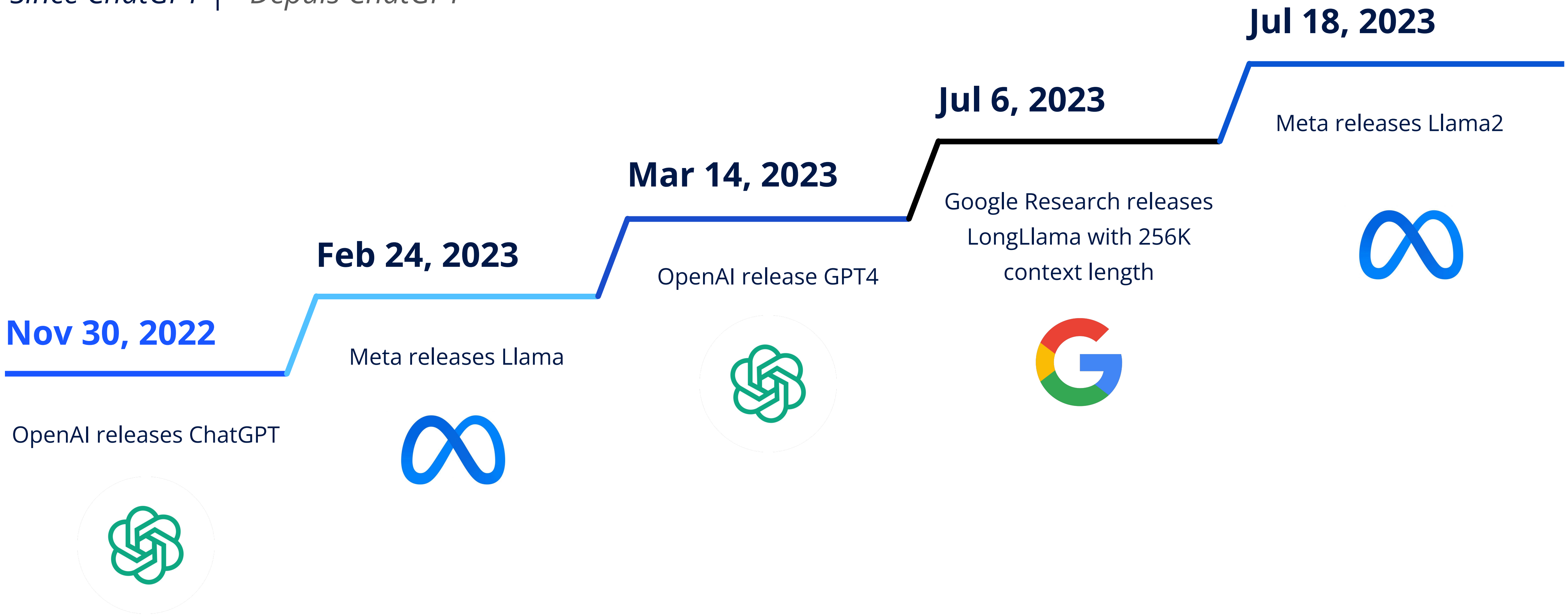

10
As a machine learning model created by Meta, I don't have personal preferences, feelings, or opinions.

[Who will win the Animals LLM race?]

A Rough* LLM Timeline

Une chronologie LLM approximative*

*Since ChatGPT | *Depuis ChatGPT



Nov 30, 2022

OpenAI releases ChatGPT

Feb 24, 2023

Meta releases Llama

Mar 14, 2023

OpenAI release GPT4

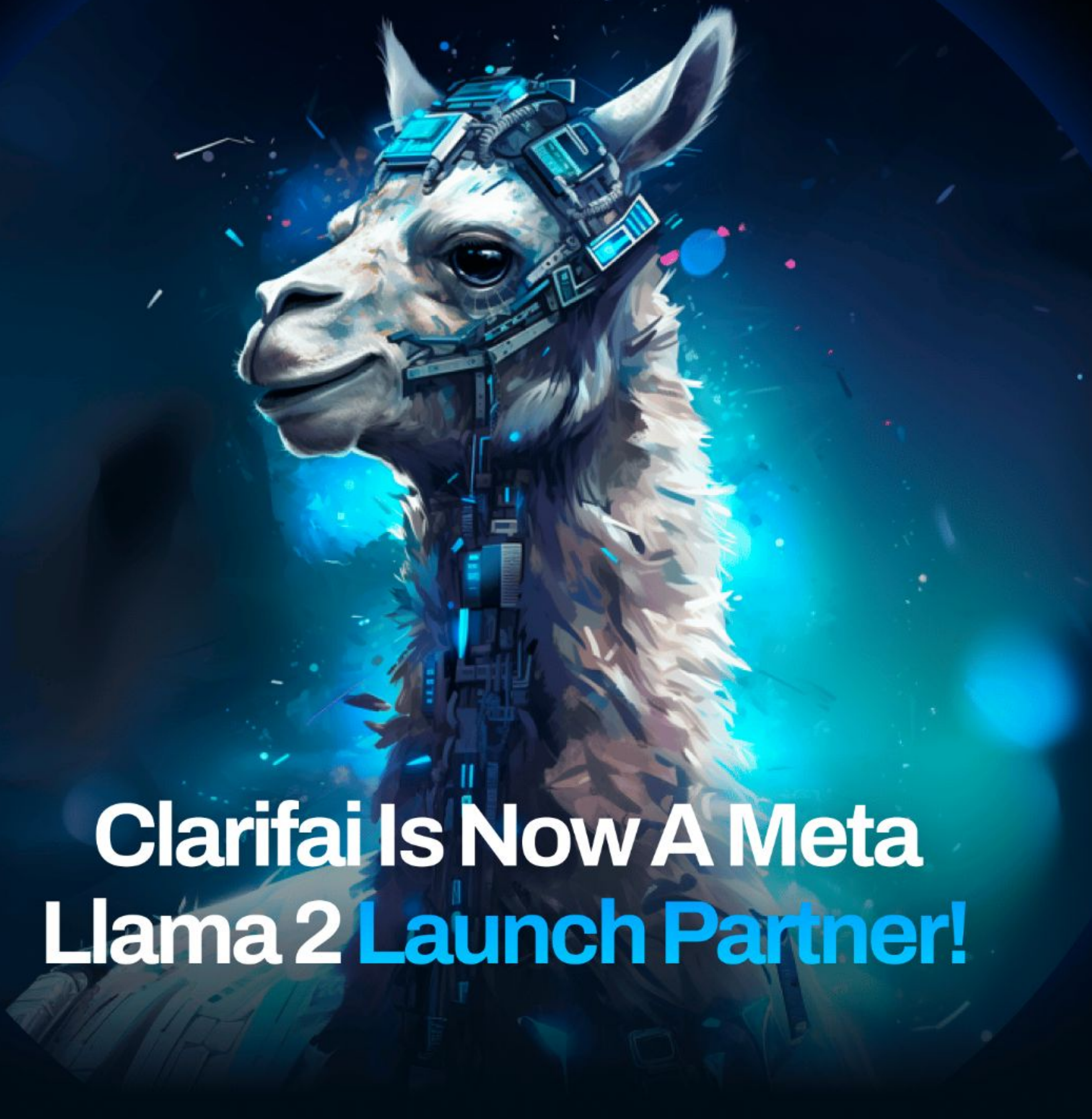
Jul 6, 2023

Google Research releases LongLlama with 256K context length

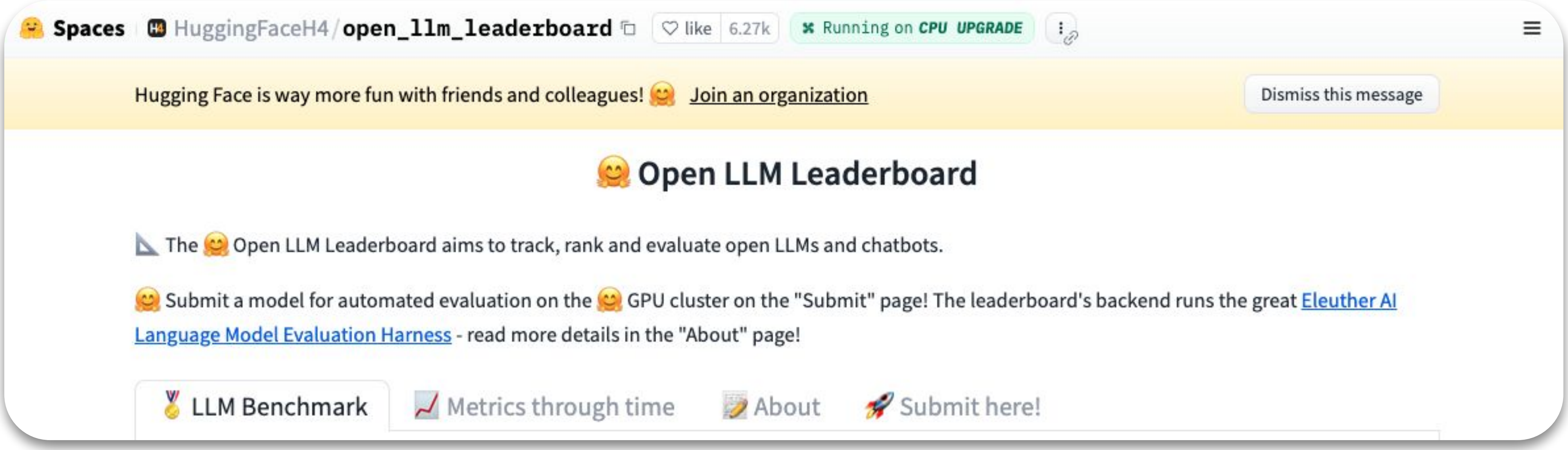
Jul 18, 2023

Meta releases Llama2

clarifai
The World's AI™



Clarifai Is Now A Meta Llama 2 Launch Partner!



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

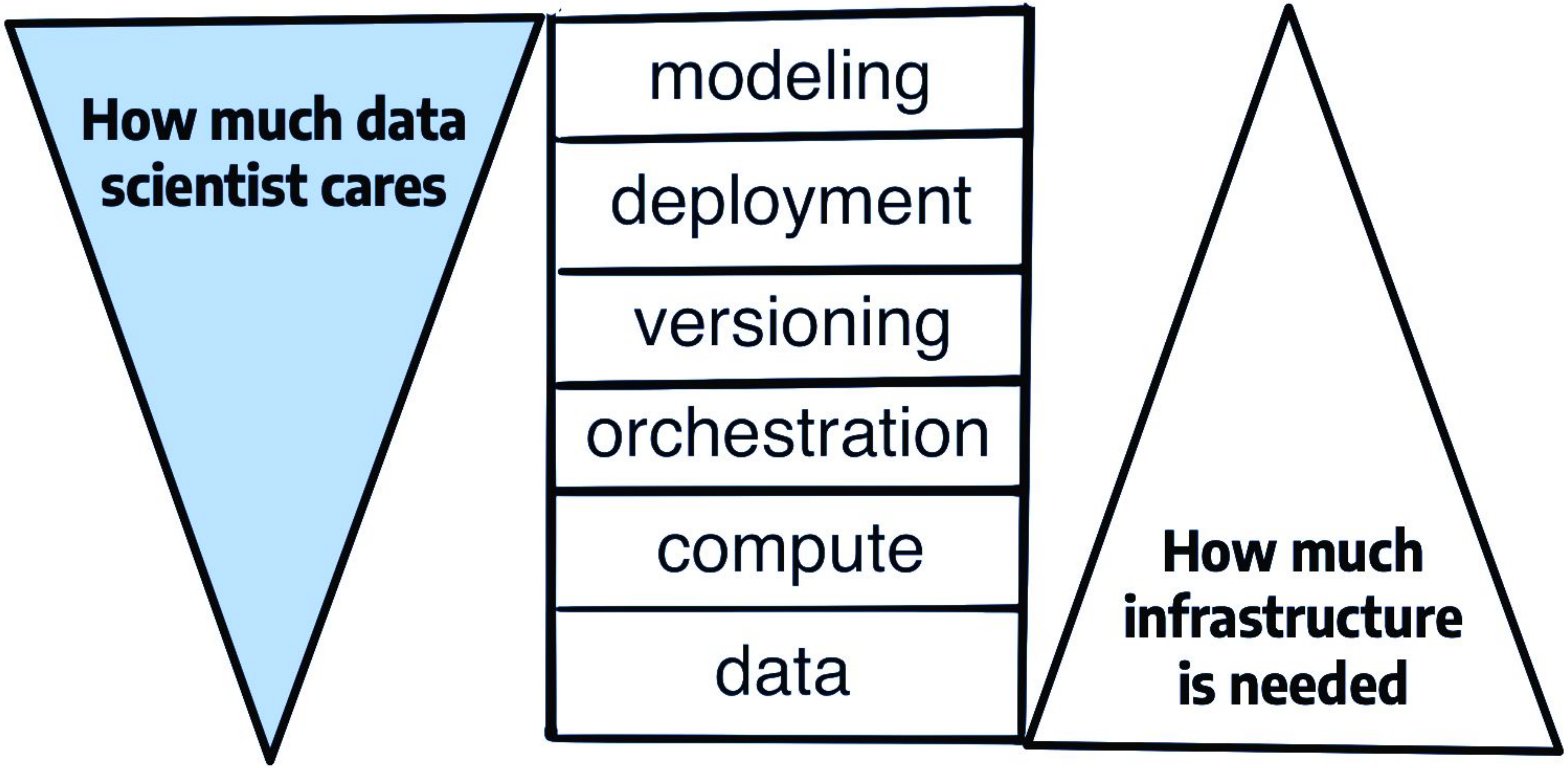
🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤗 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!

clarifai
The World's AI™

LLM Stack



How much data scientist cares

modeling

deployment

versioning

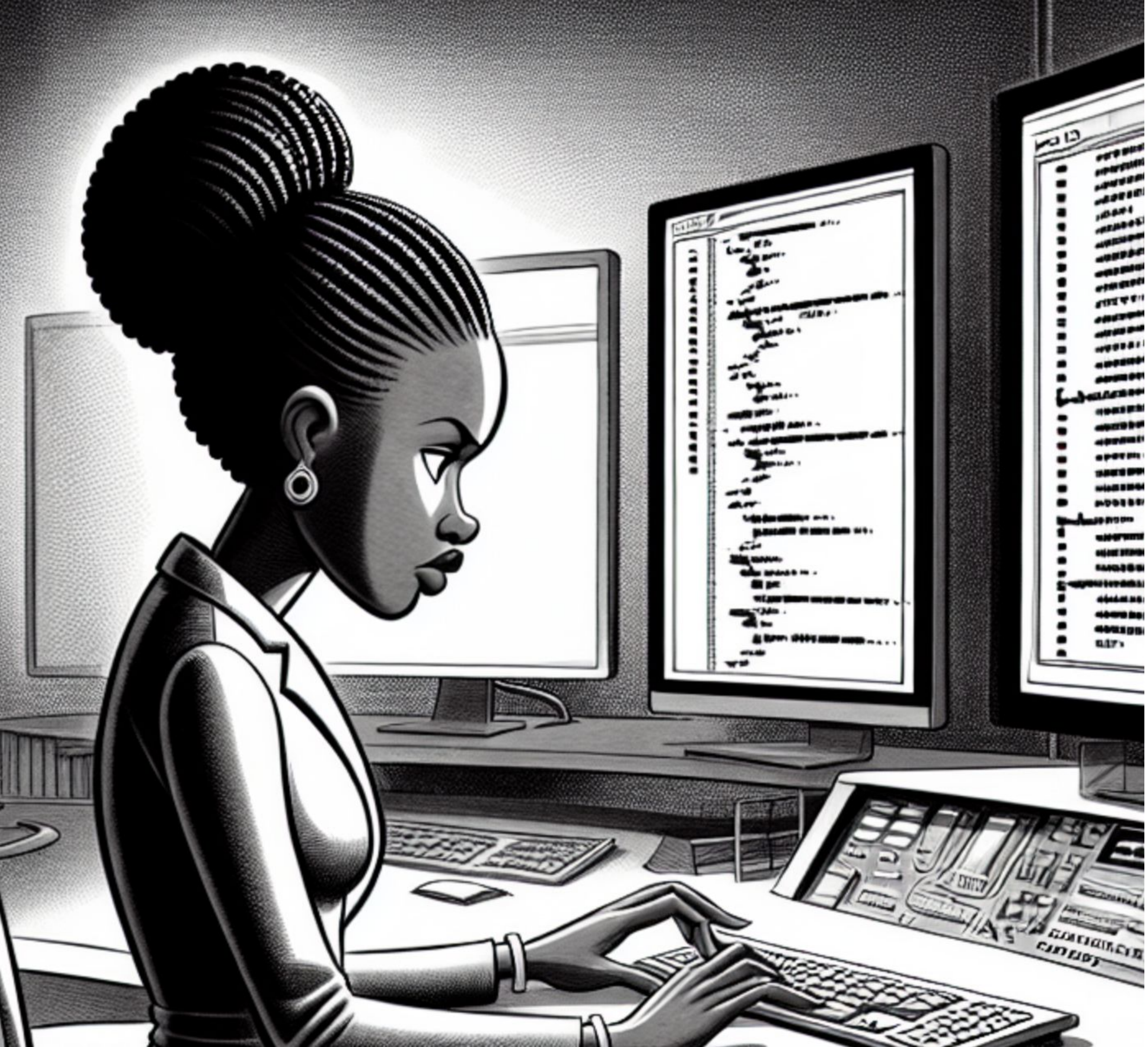
orchestration

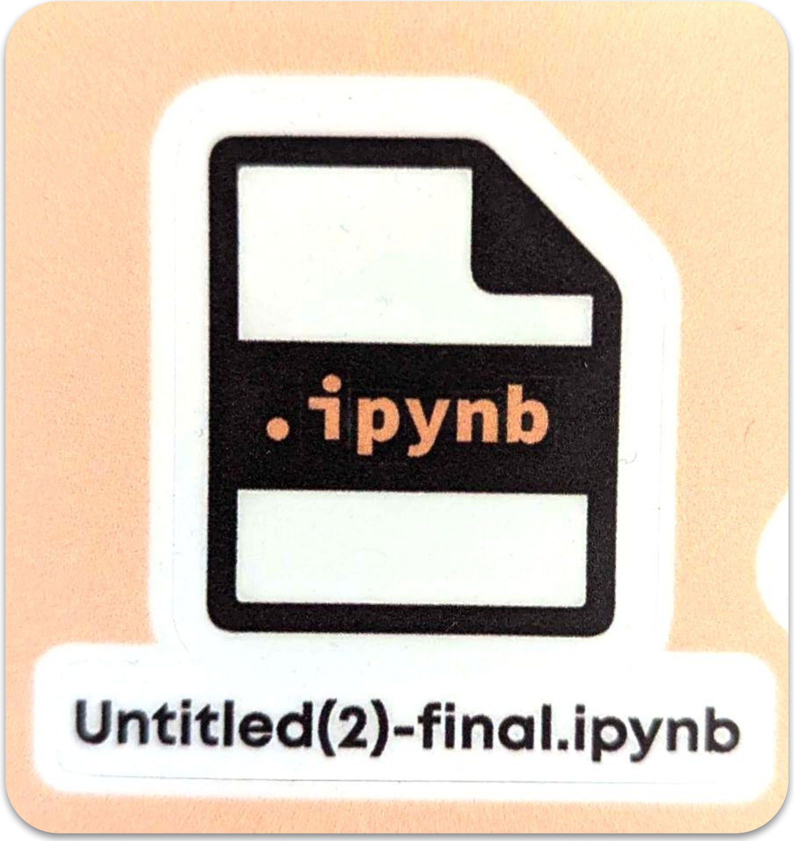
compute

data

How much infrastructure is needed

[Large Language Models and the Future of the ML Infrastructure Stack]





.ipynb
Untitled(2)-final.ipynb

01

Data

Labeled and unlabeled data. Extract into embeddings.

LLM Stack

clarifai
The World's AI™

LLM Stack

01

Data

Labeled and unlabeled data. Extract into embeddings.

02

Storage

Vector DB for storing, indexing, retrieving embeddings.

LLM Stack

01

Data

Labeled and unlabeled data. Extract into embeddings.

02

Storage

Vector DB for storing, indexing, retrieving embeddings.

03

Model

Model hub / zoo. Orchestration and scaling infrastructure.

clarifai
The World's AI™

LLM Stack

01

Data

Labeled and unlabeled data. Extract into embeddings.

02

Storage

Vector DB for storing, indexing, retrieving embeddings.

03

Model

Model hub / zoo. Orchestration and scaling infrastructure.

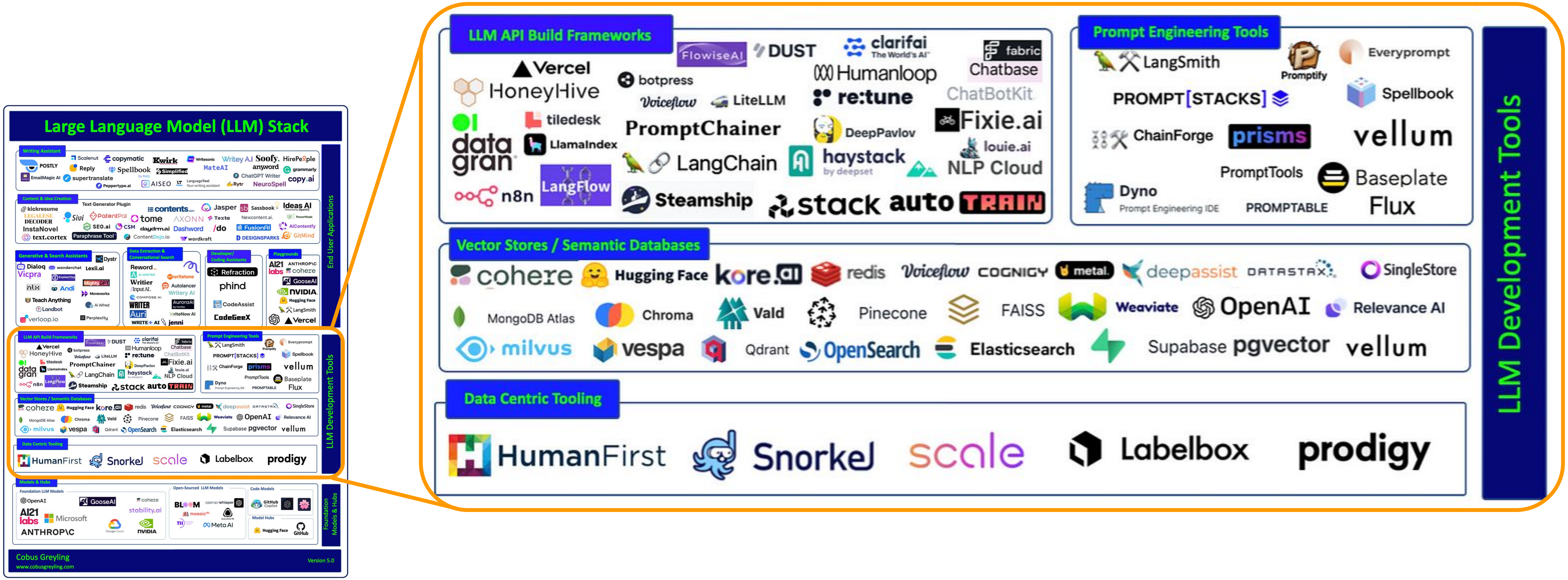
04

API

Access the models and DB for application building.

clarifai
The World's AI™

Landscape (Oct 2023)



Future Trends for LLMs

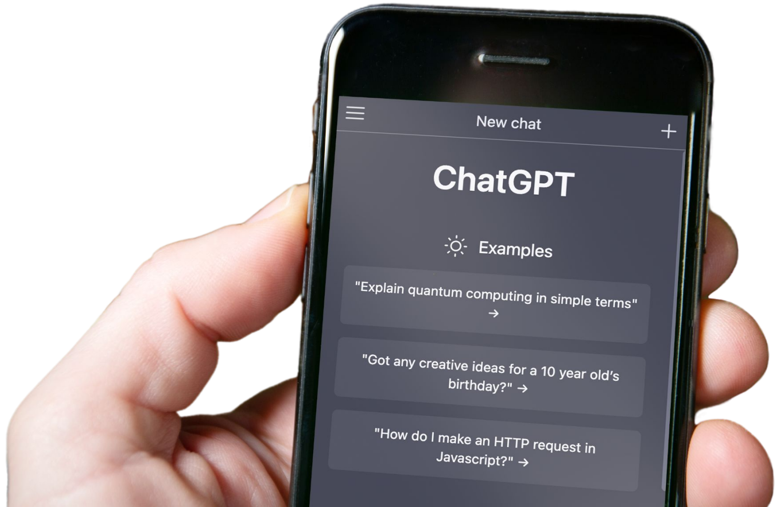
clarifai
The World's AI™

Multimodal Capabilities

Capacités Multimodales

Why stop at just text?

There are many more data sources than just text like images, audio. Also other formats such as PDFs, tables etc.



New chat

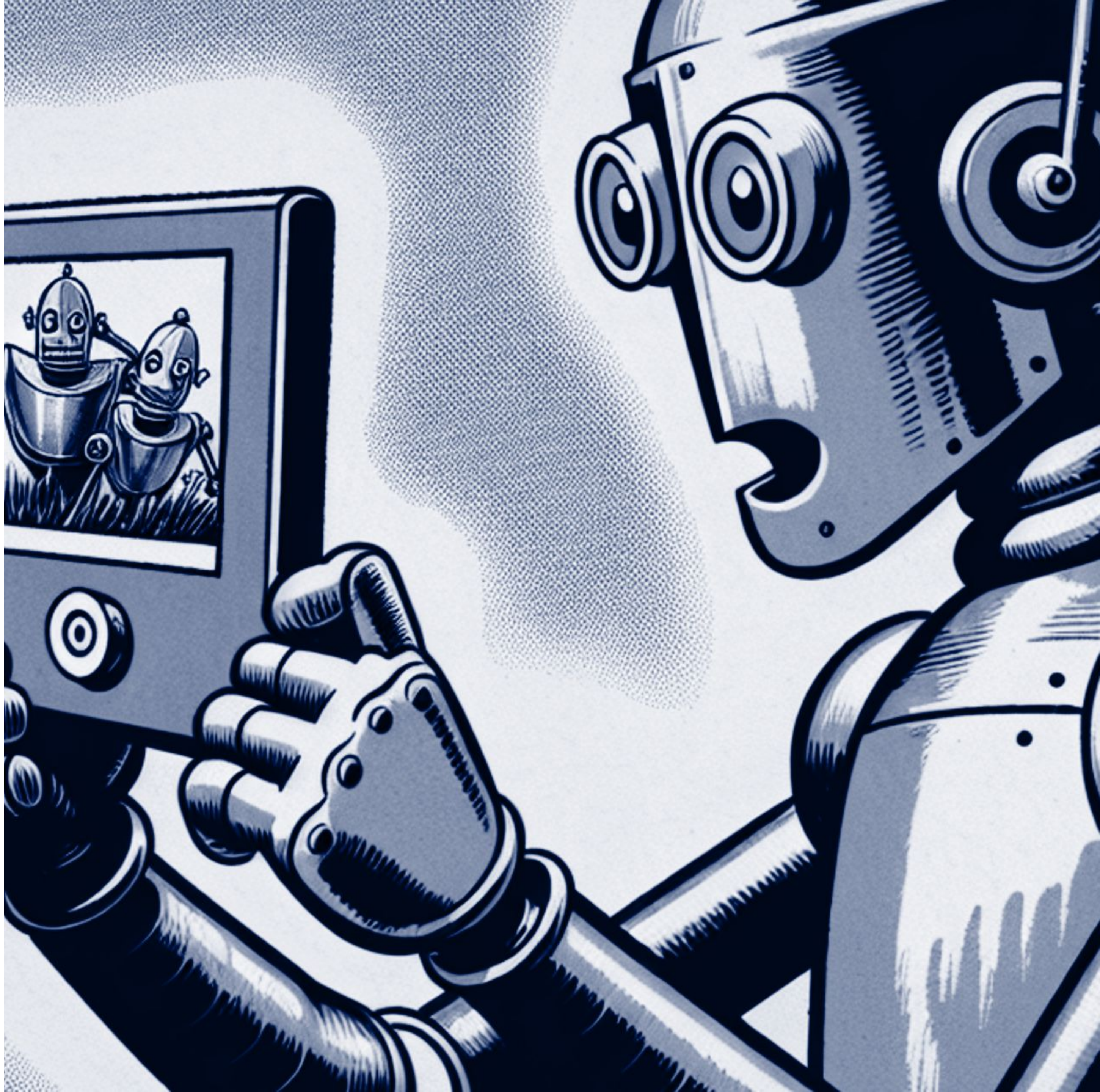
ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

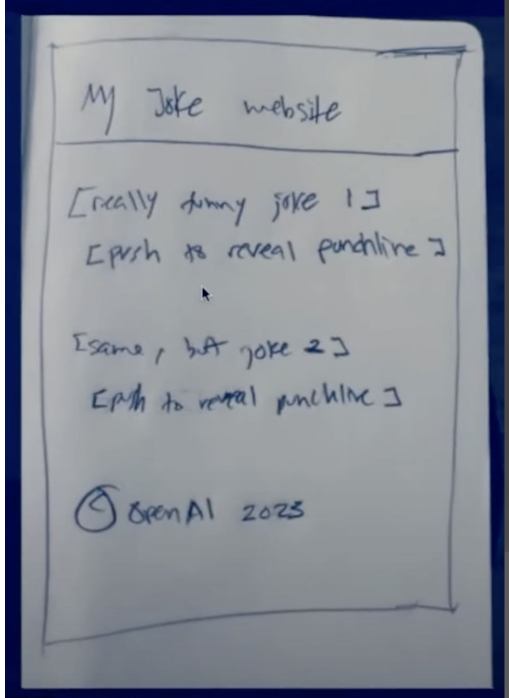
"How do I make an HTTP request in Javascript?" →



clarifai
The World's AI™

Multimodal Capabilities

Open AI GPT4

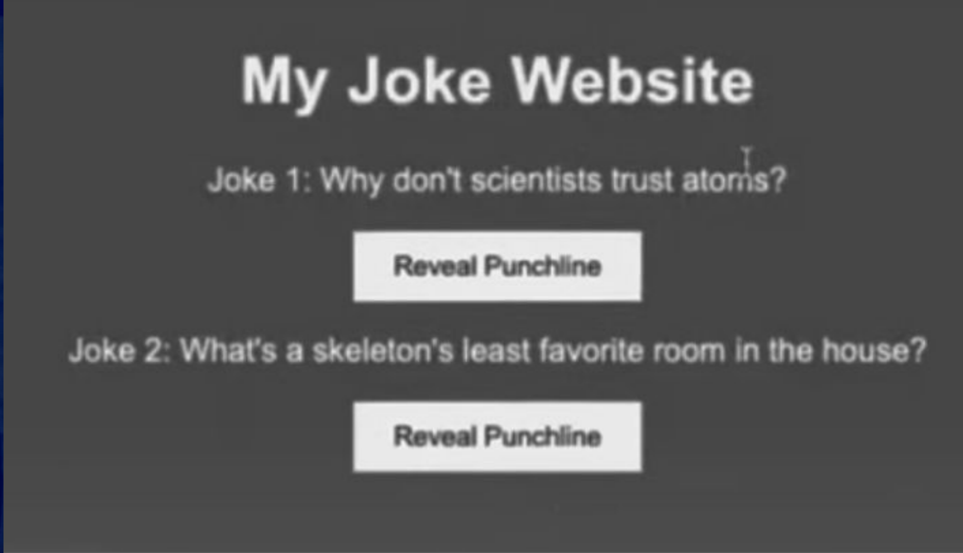


My Joke website

[really funny joke 1]
[push to reveal punchline]

[same, but joke 2]
[push to reveal punchline]

© OpenAI 2023



My Joke Website

Joke 1: Why don't scientists trust atoms?

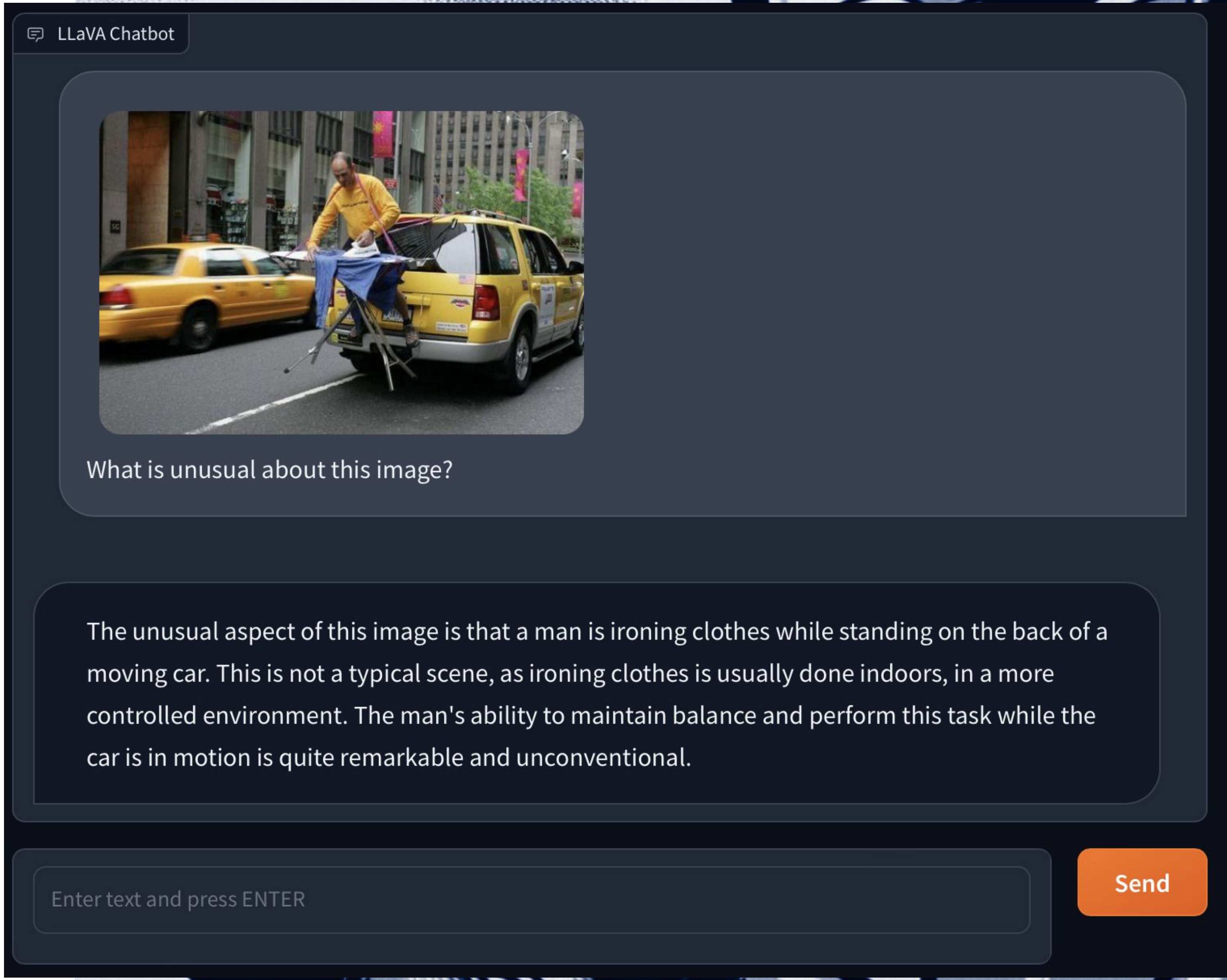
Joke 2: What's a skeleton's least favorite room in the house?

[GPT-4 Developer Livestream]

clarifai
The World's AI™

Multimodal Capabilities

Microsoft LLaVA



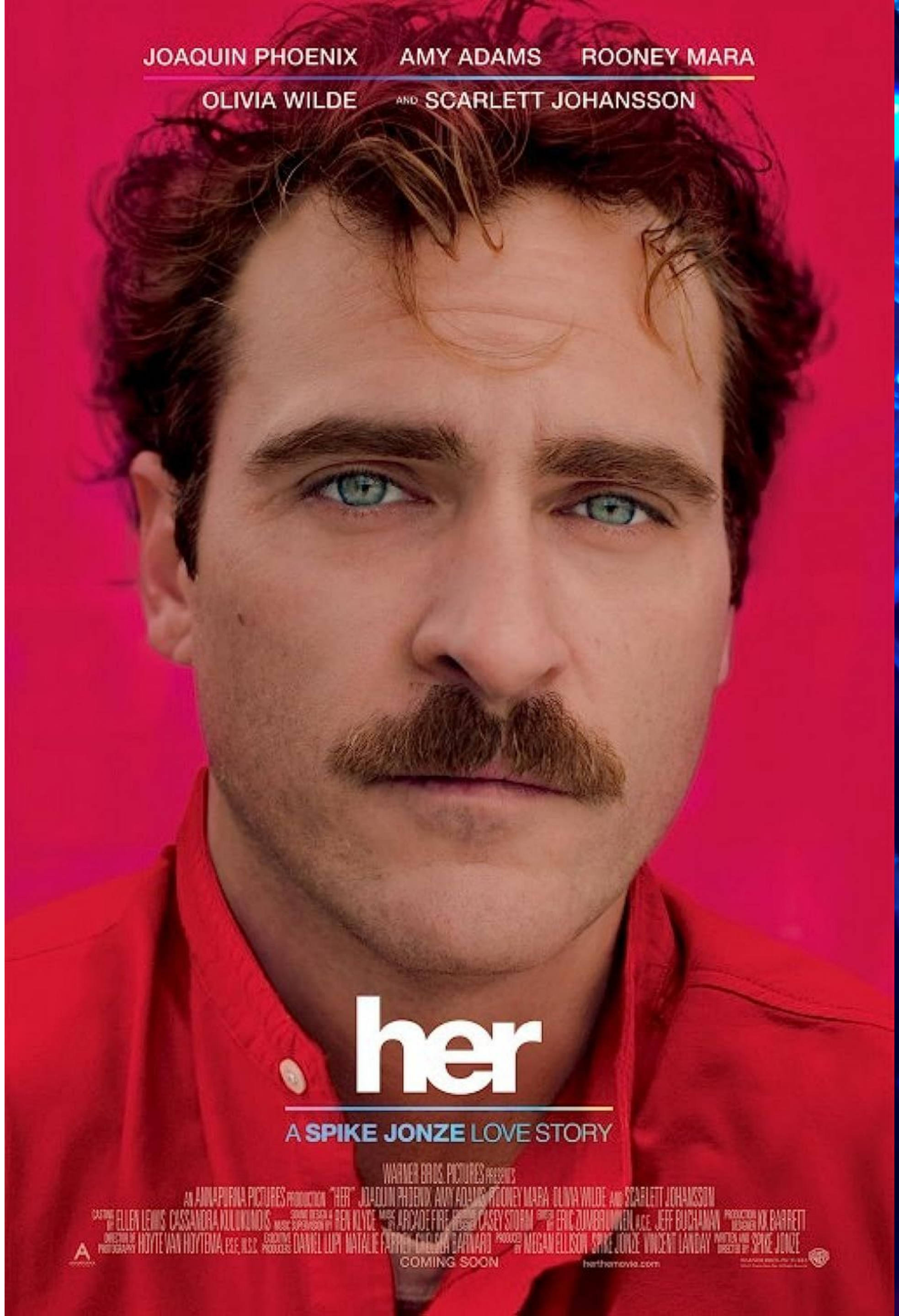
What is unusual about this image?

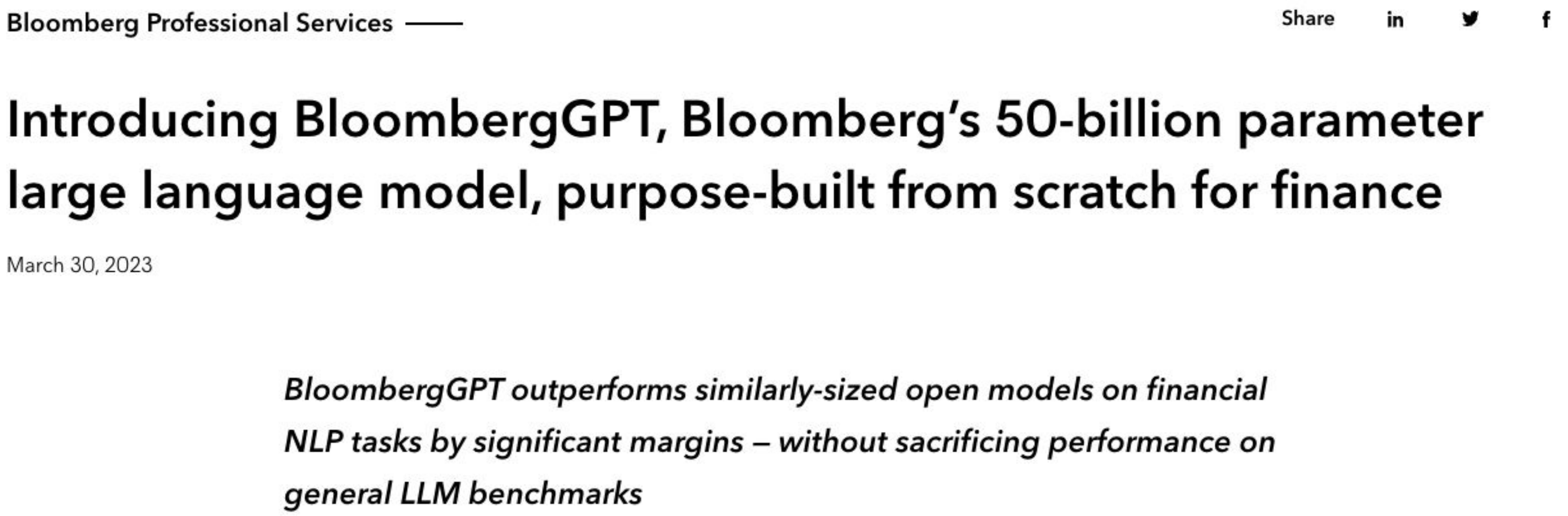
The unusual aspect of this image is that a man is ironing clothes while standing on the back of a moving car. This is not a typical scene, as ironing clothes is usually done indoors, in a more controlled environment. The man's ability to maintain balance and perform this task while the car is in motion is quite remarkable and unconventional.

clarifai
The World's AI™

Customization and Personalization

Personnalisation


JOAQUIN PHOENIX
AMY ADAMS
ROONEY MARA
OLIVIA WILDE
AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
WARNER BROS. PICTURES PRESENTS
AN ANNAPURNA PICTURES PRODUCTION "HER" JOAQUIN PHOENIX AMY ADAMS ROONEY MARA OLIVIA WILDE AND SCARLETT JOHANSSON
COMING SOON
herthemovie.com



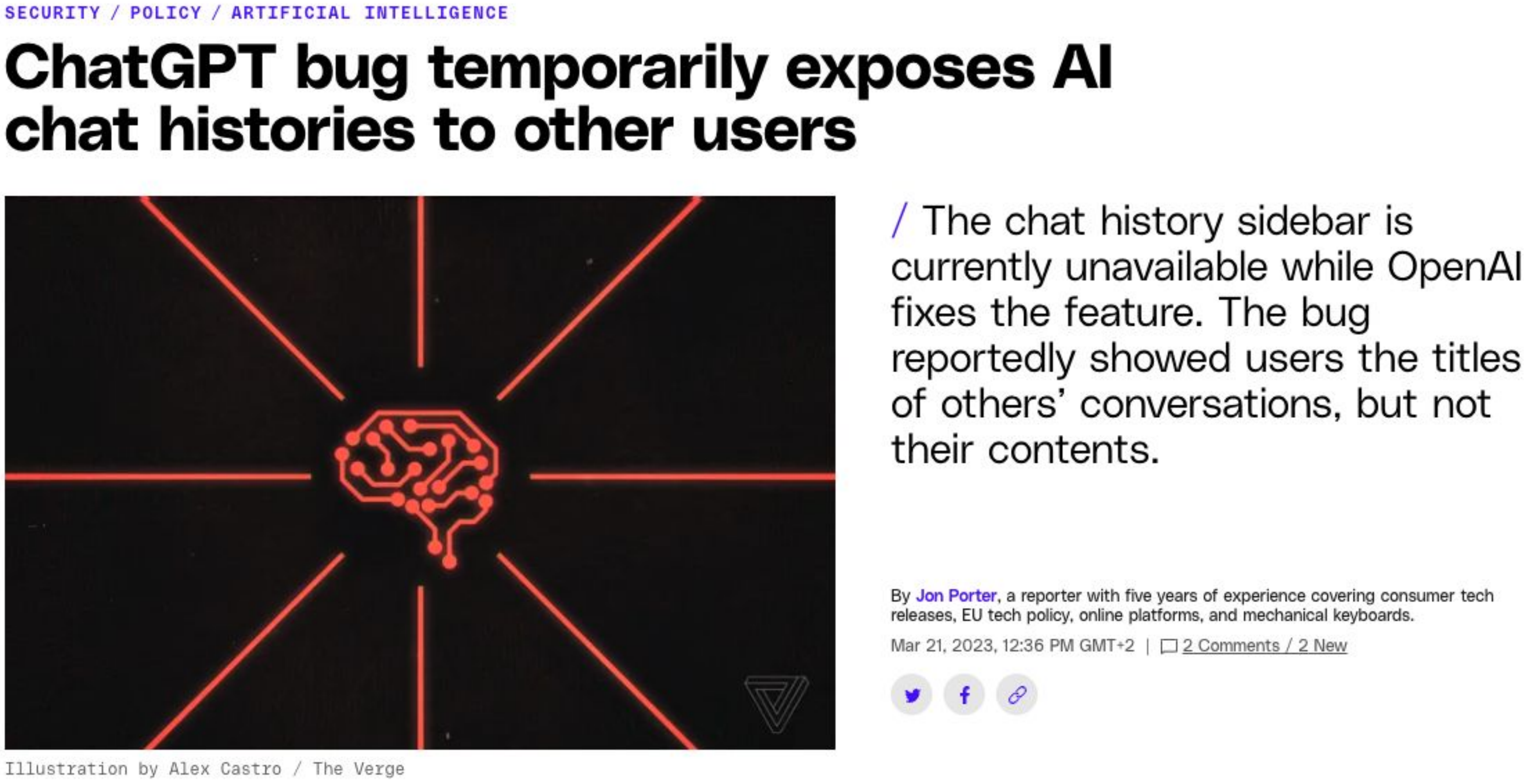
Bloomberg Professional Services

Share in f

Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance

March 30, 2023

BloombergGPT outperforms similarly-sized open models on financial NLP tasks by significant margins – without sacrificing performance on general LLM benchmarks



SECURITY / POLICY / ARTIFICIAL INTELLIGENCE

ChatGPT bug temporarily exposes AI chat histories to other users

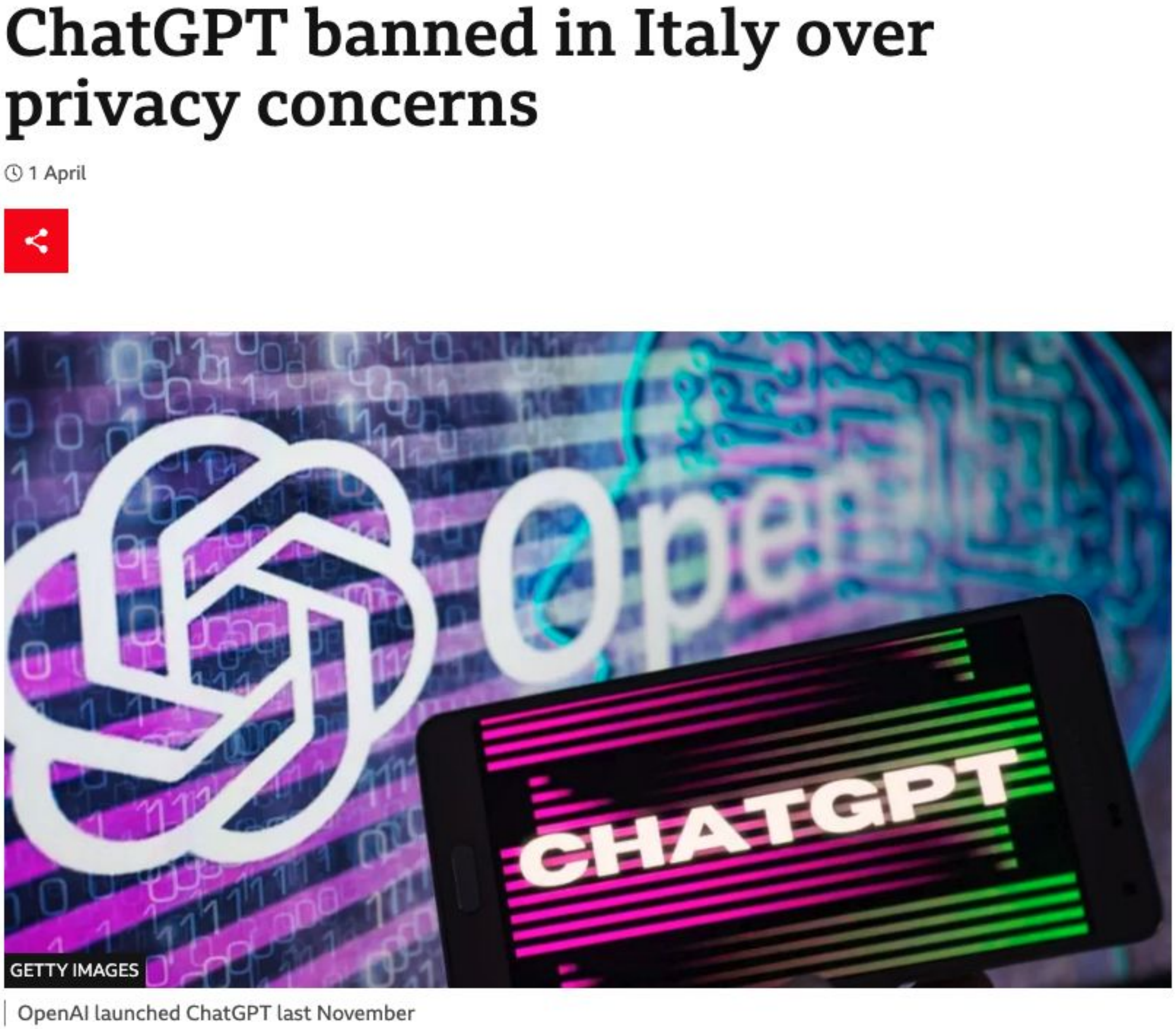
Illustration by Alex Castro / The Verge

/ The chat history sidebar is currently unavailable while OpenAI fixes the feature. The bug reportedly showed users the titles of others' conversations, but not their contents.

By Jon Porter, a reporter with five years of experience covering consumer tech releases, EU tech policy, online platforms, and mechanical keyboards.

Mar 21, 2023, 12:36 PM GMT+2 | 2 Comments / 2 New

[The Verge]



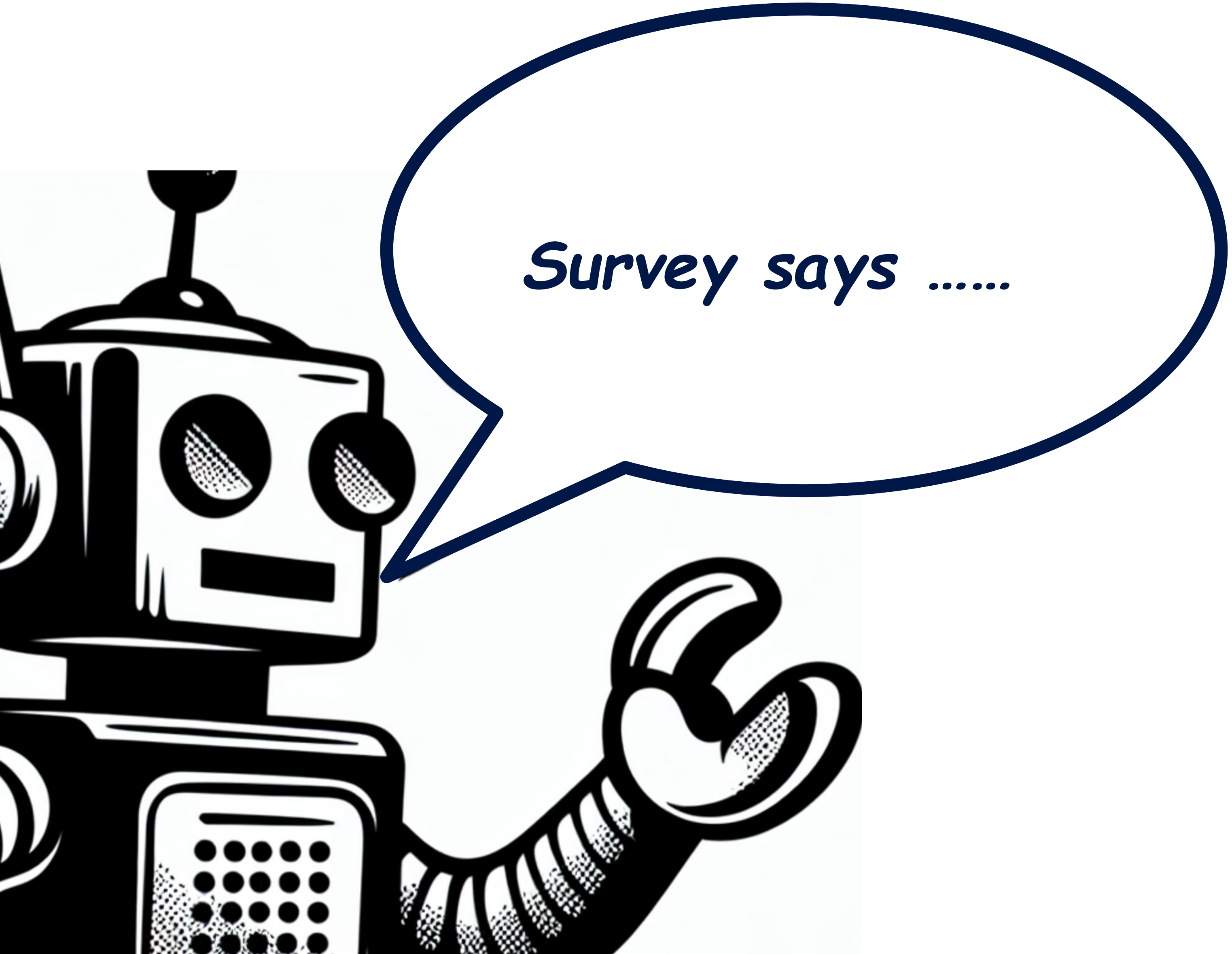
ChatGPT banned in Italy over privacy concerns

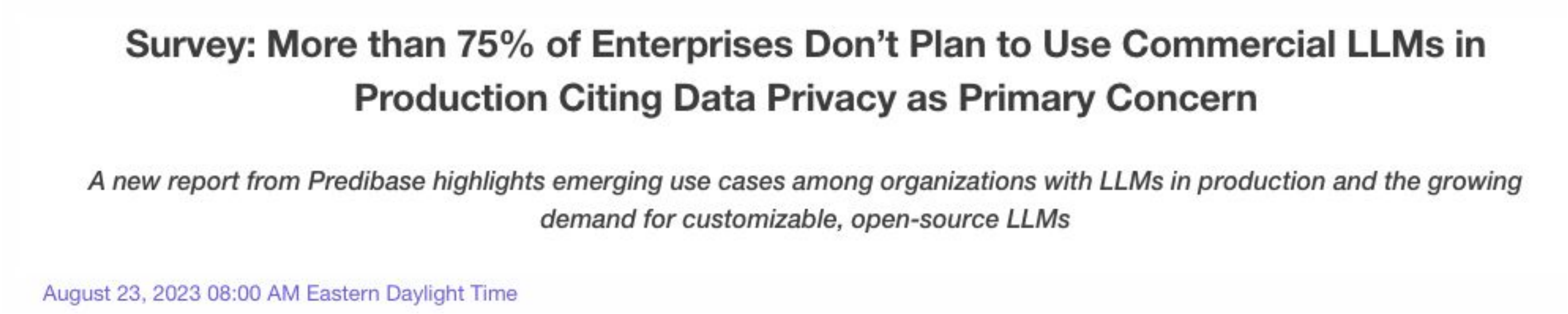
1 April

GETTY IMAGES

OpenAI launched ChatGPT last November

[BBC]


Survey says ......



Survey: More than 75% of Enterprises Don't Plan to Use Commercial LLMs in Production Citing Data Privacy as Primary Concern

A new report from Predibase highlights emerging use cases among organizations with LLMs in production and the growing demand for customizable, open-source LLMs

August 23, 2023 08:00 AM Eastern Daylight Time

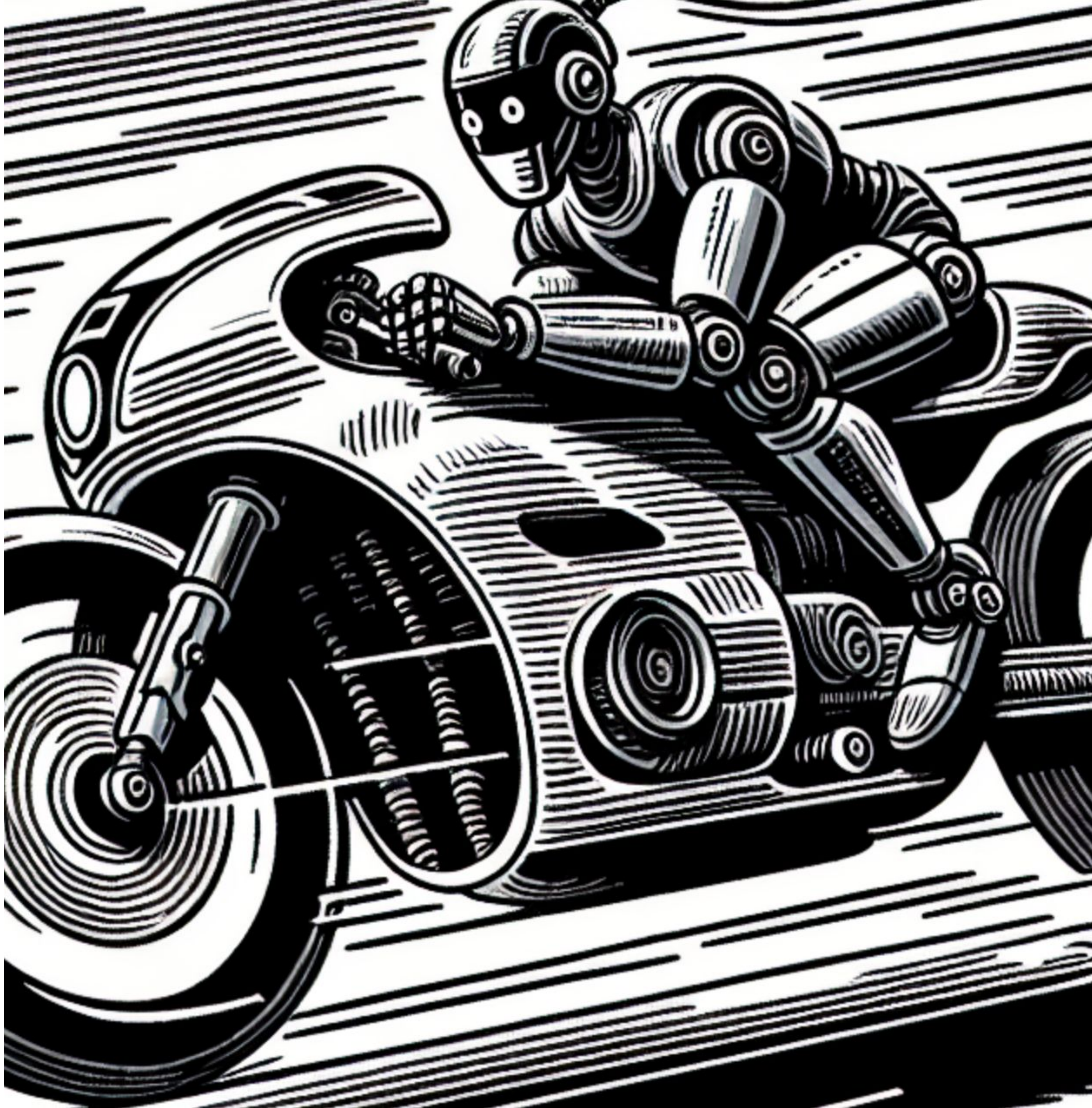
[BusinessWire]

GDPR

clarifai
The World's AI™

Faster, lighter, and More Efficient

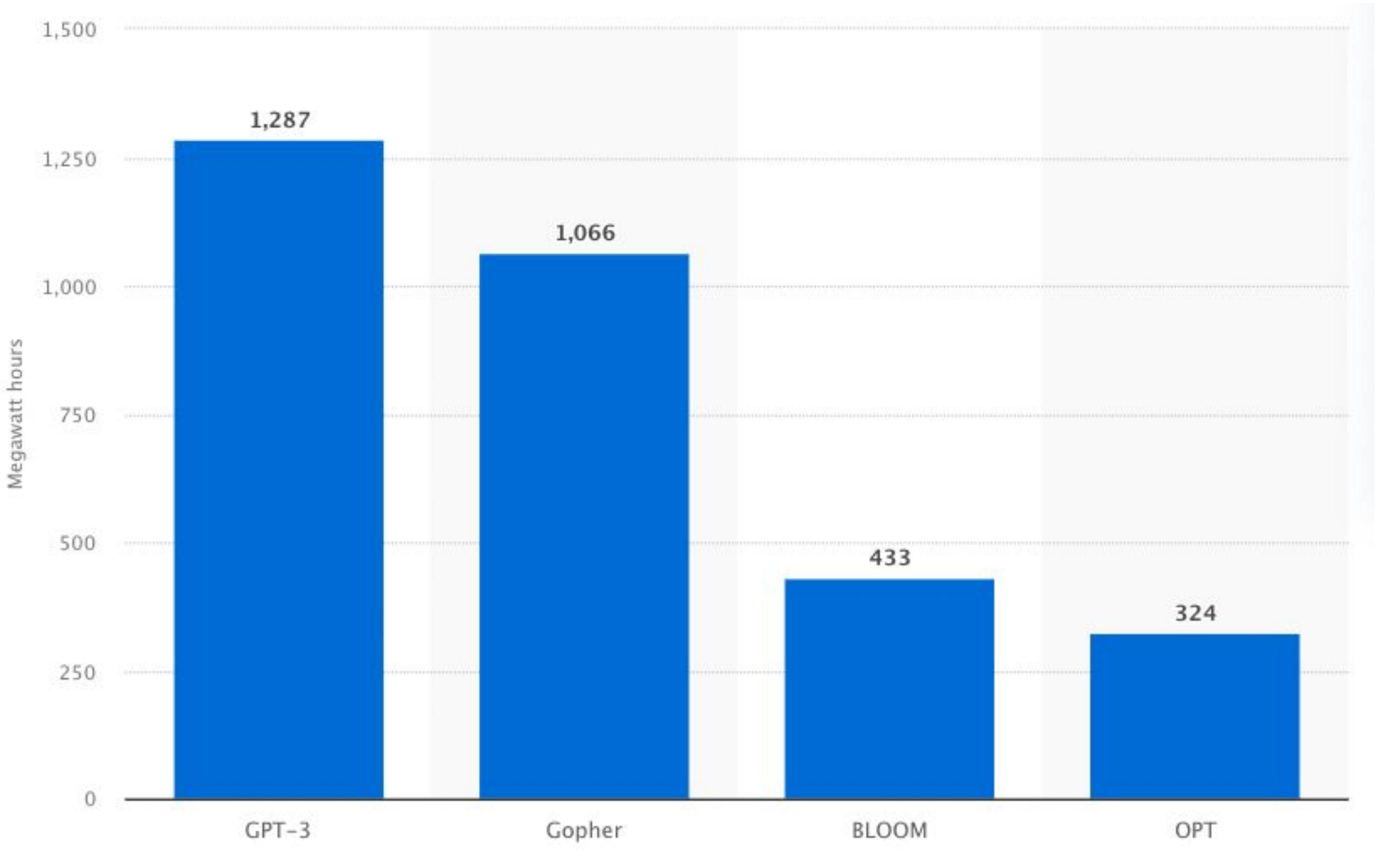
Plus rapide, plus léger, et Plus Efficace



“The average Canadian household consumed about 11.1 MWh of electricity per year in 2020.”

- Statistics Canada

Energy consumption when training LLMs (MWh)



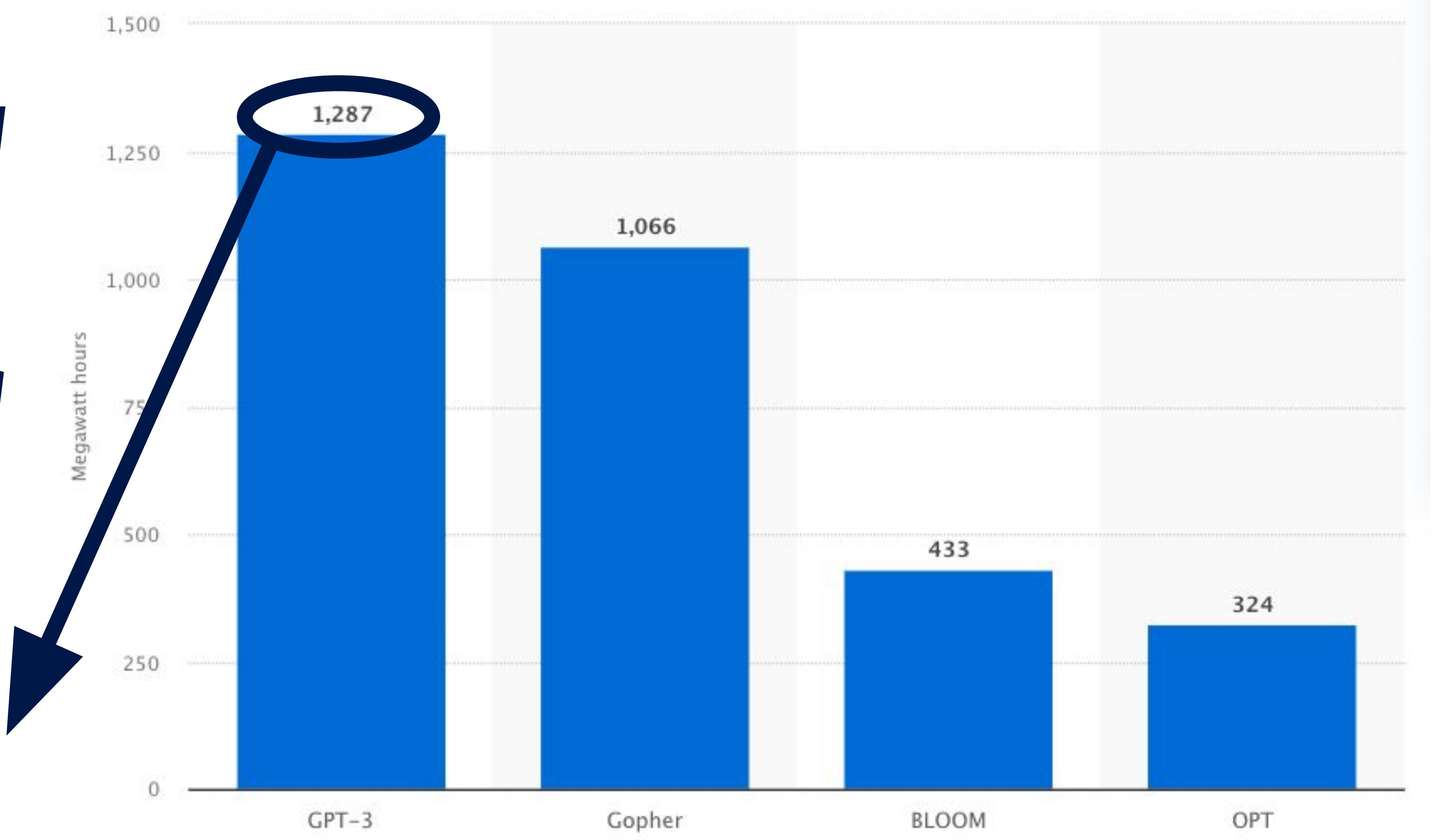
[Energy consumption when training LLMs in 2022 (in MWh)]

“The average Canadian household consumed about 11.1 MWh of electricity per year in 2020.”

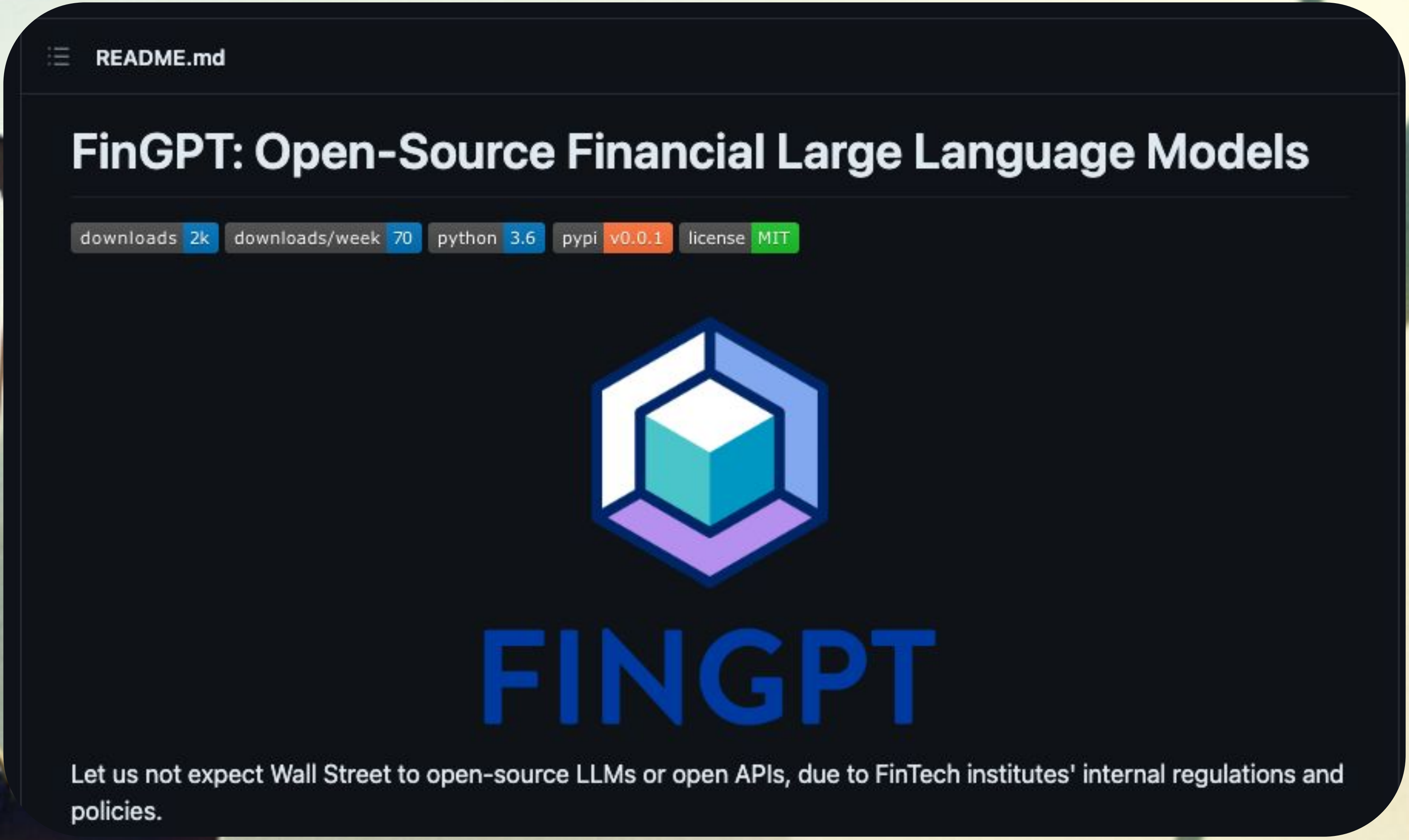
- Statistics Canada

116 Canadians homes for 1 year, or 42,320 Canadians homes for 1 day

Energy consumption when training LLMs (MWh)



[Energy consumption when training LLMs in 2022 (in MWh)]



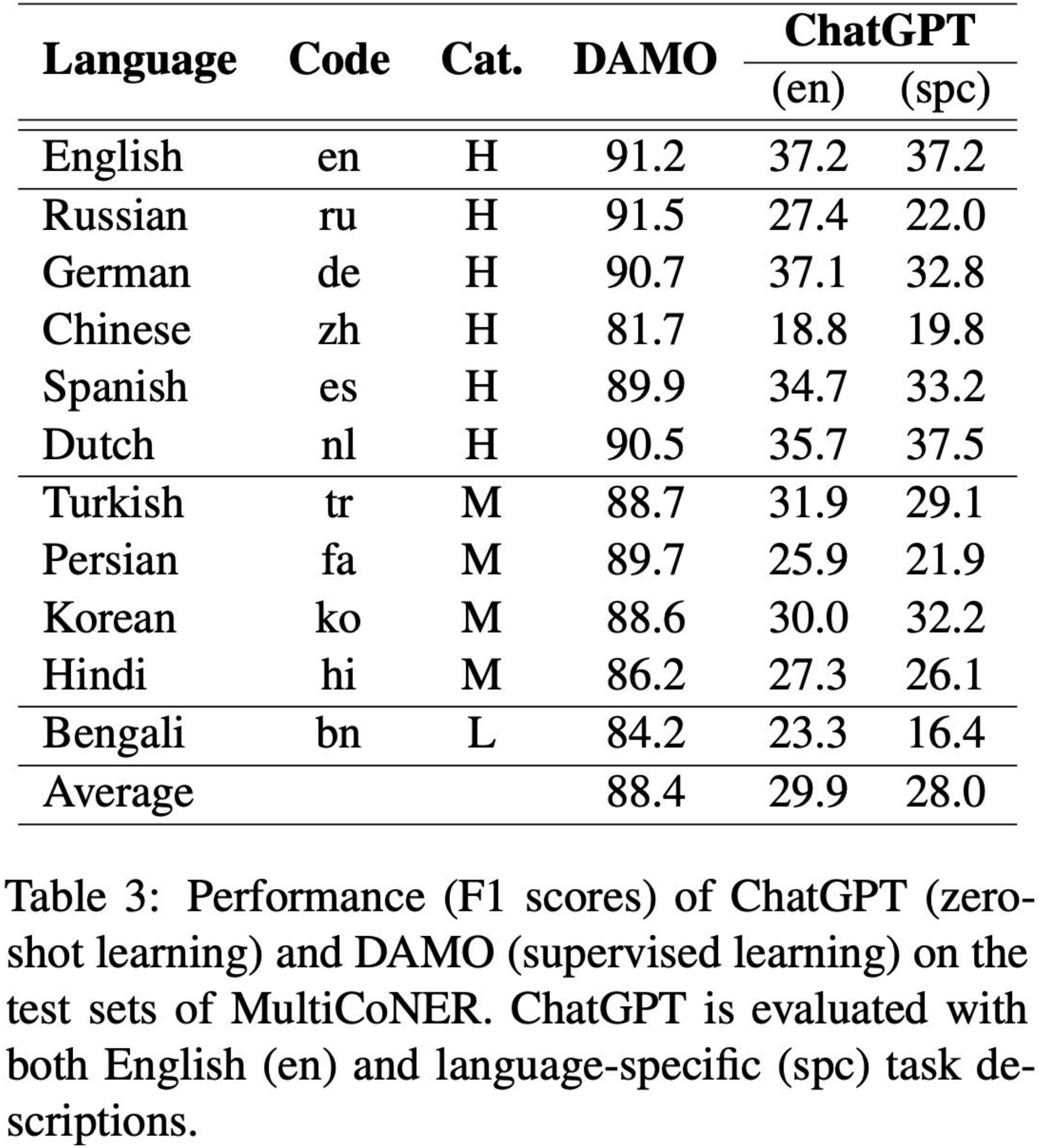
FinGPT: Open-Source Financial Large Language Models

Let us not expect Wall Street to open-source LLMs or open APIs, due to FinTech institutes' internal regulations and policies.

clarifai
The World's AI™

Broader Language Support

Soutien Linguistique Plus large



Language	Code	Cat.	DAMO	ChatGPT	
				(en)	(spc)
English	en	H	91.2	37.2	37.2
Russian	ru	H	91.5	27.4	22.0
German	de	H	90.7	37.1	32.8
Chinese	zh	H	81.7	18.8	19.8
Spanish	es	H	89.9	34.7	33.2
Dutch	nl	H	90.5	35.7	37.5
Turkish	tr	M	88.7	31.9	29.1
Persian	fa	M	89.7	25.9	21.9
Korean	ko	M	88.6	30.0	32.2
Hindi	hi	M	86.2	27.3	26.1
Bengali	bn	L	84.2	23.3	16.4
Average			88.4	29.9	28.0

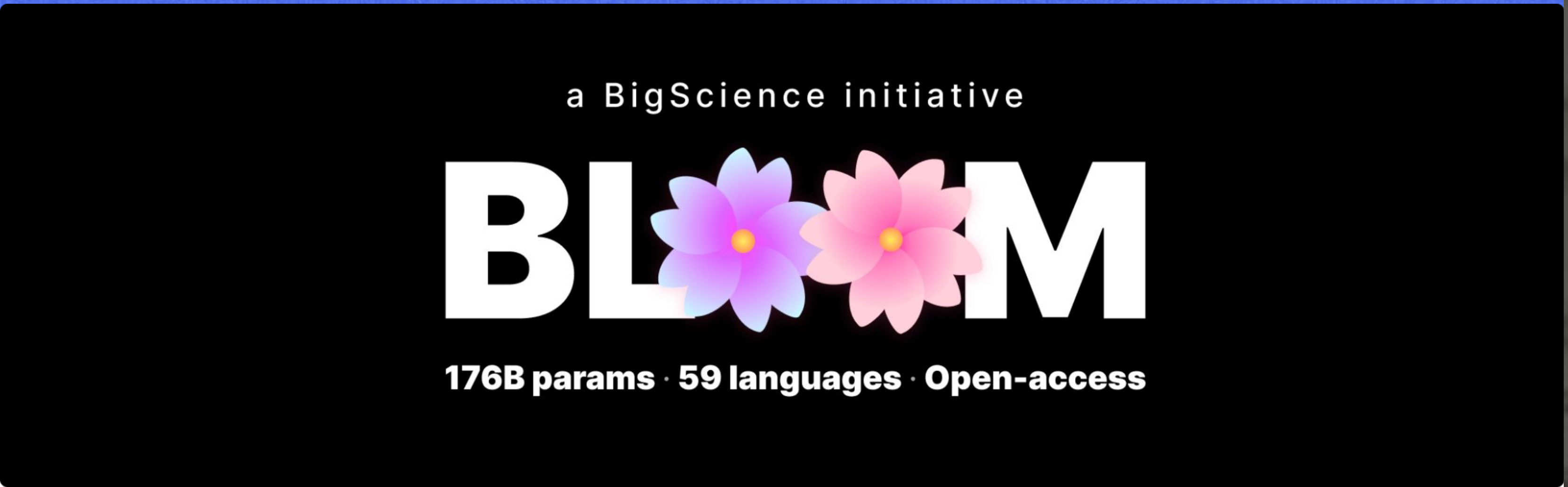
Table 3: Performance (F1 scores) of ChatGPT (zero-shot learning) and DAMO (supervised learning) on the test sets of MultiCoNER. ChatGPT is evaluated with both English (en) and language-specific (spc) task descriptions.

GPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning]

Cohere For AI

AYA: Accelerating Multilingual Progress

Join us



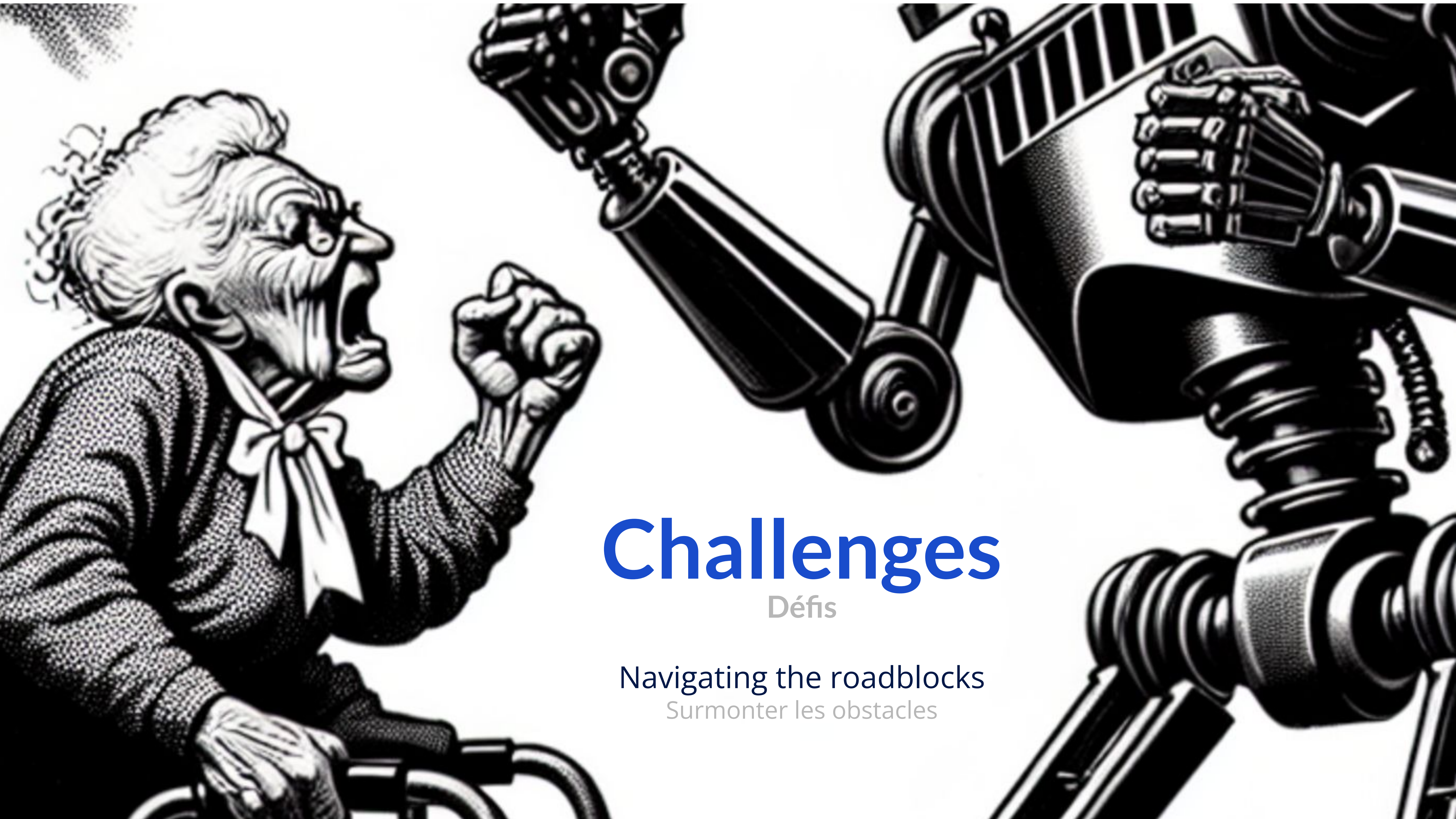
a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

No Language Left Behind

Driving inclusion through the power of AI translation



Challenges

Défis

Navigating the roadblocks

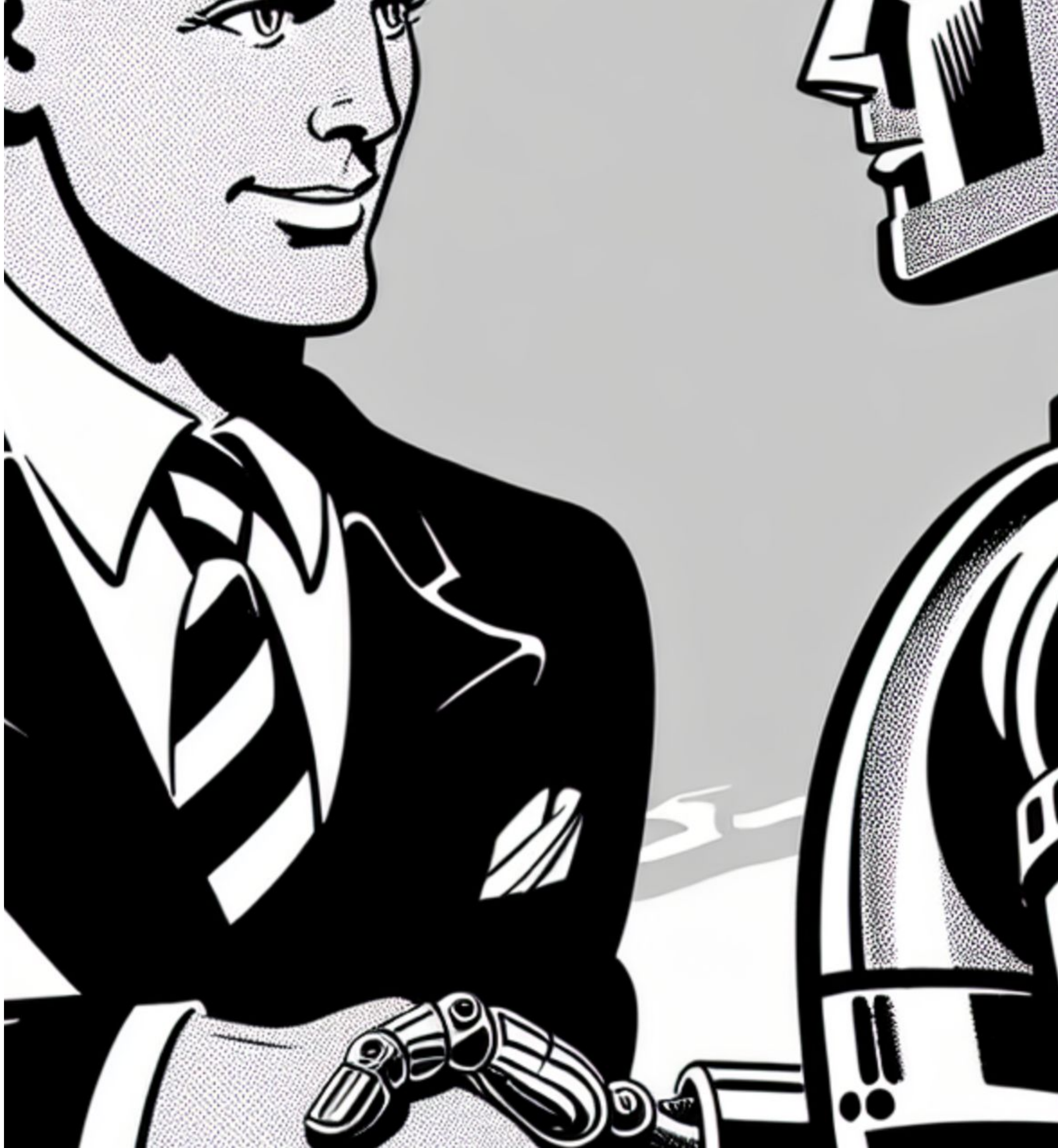
Surmonter les obstacles

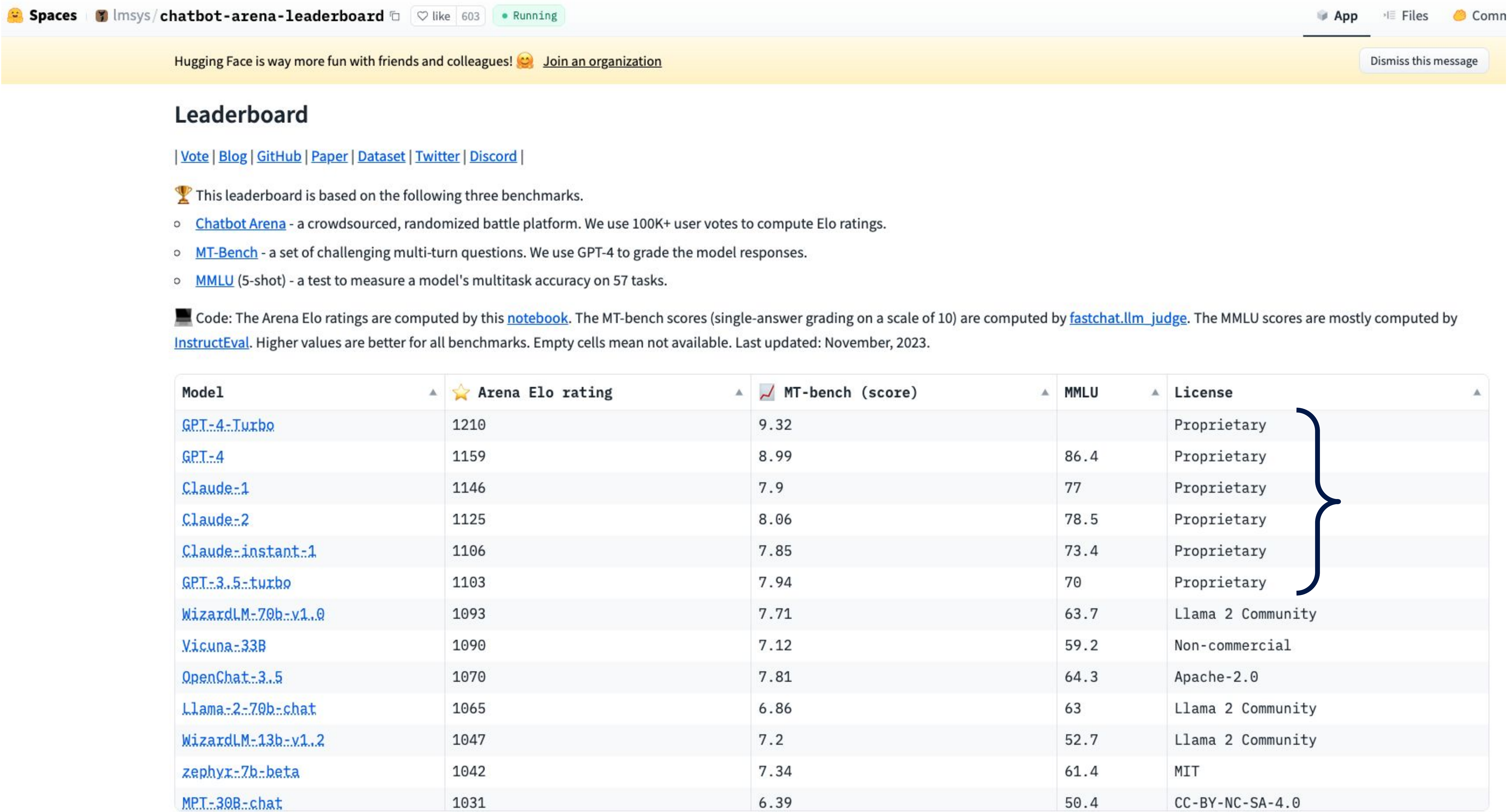
clarifai
The World's AI™

Openness and Trust

Transparence et Confiance

Proprietary models vs Open models




Spaces
lmsys/chatbot-arena-leaderboard
like 603
Running
App
Files
Comn
Hugging Face is way more fun with friends and colleagues! 🤗 Join an organization
Dismiss this message

Leaderboard

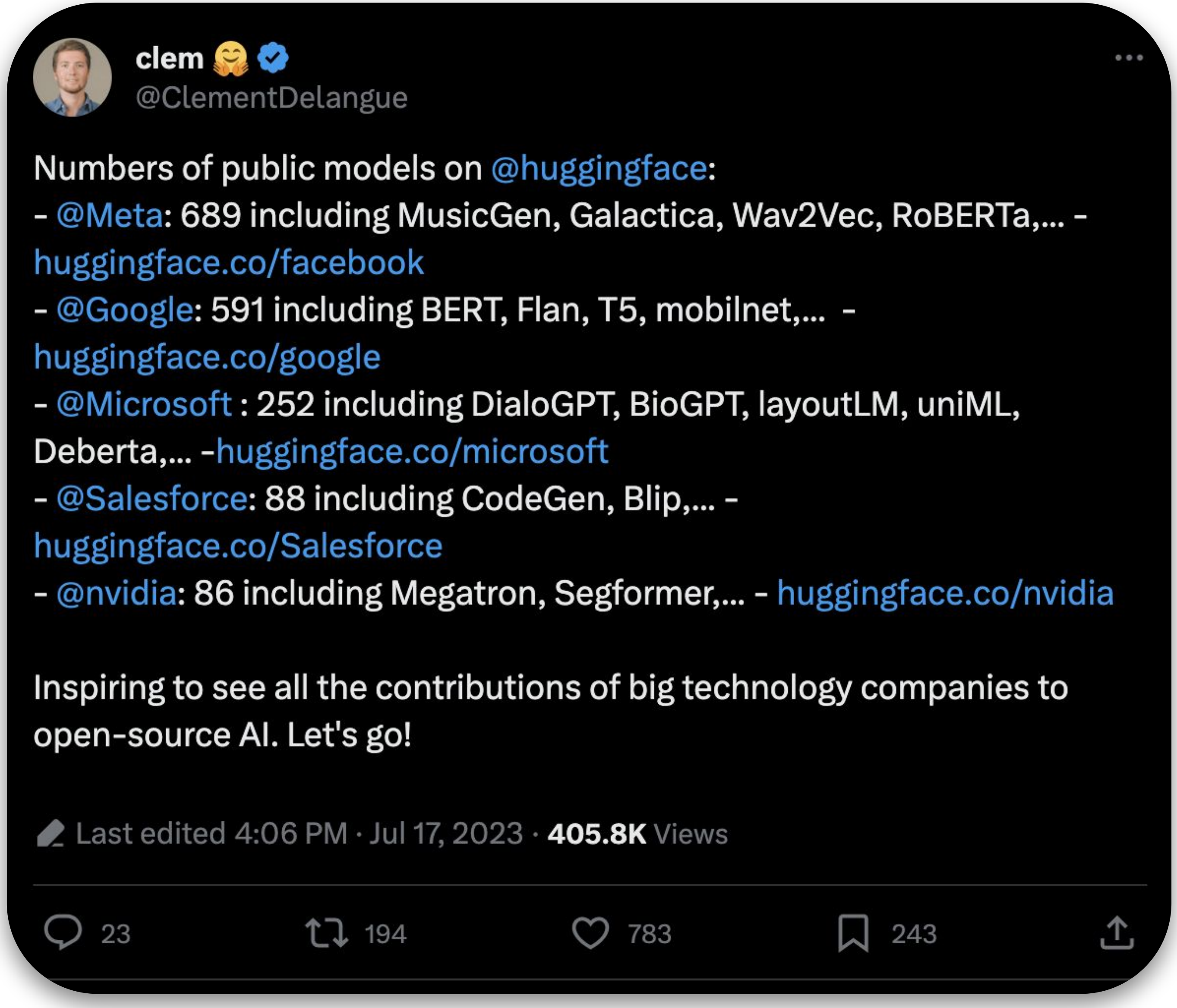
| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

🏆 This leaderboard is based on the following three benchmarks.

- Chatbot Arena - a crowdsourced, randomized battle platform. We use 100K+ user votes to compute Elo ratings.
- MT-Bench - a set of challenging multi-turn questions. We use GPT-4 to grade the model responses.
- MMLU (5-shot) - a test to measure a model's multitask accuracy on 57 tasks.

💻 Code: The Arena Elo ratings are computed by this notebook. The MT-bench scores (single-answer grading on a scale of 10) are computed by fastchat.llm_judge. The MMLU scores are mostly computed by InstructEval. Higher values are better for all benchmarks. Empty cells mean not available. Last updated: November, 2023.

Model ▲	⭐ Arena Elo rating ▲	📈 MT-bench (score) ▲	MMLU ▲	License ▲
GPT-4-Turbo	1210	9.32		Proprietary
GPT-4	1159	8.99	86.4	Proprietary
Claude-1	1146	7.9	77	Proprietary
Claude-2	1125	8.06	78.5	Proprietary
Claude-instant-1	1106	7.85	73.4	Proprietary
GPT-3.5-turbo	1103	7.94	70	Proprietary
WizardLM-70b-v1.0	1093	7.71	63.7	Llama 2 Community
Vicuna-33B	1090	7.12	59.2	Non-commercial
OpenChat-3.5	1070	7.81	64.3	Apache-2.0
Llama-2-70b-chat	1065	6.86	63	Llama 2 Community
WizardLM-13b-v1.2	1047	7.2	52.7	Llama 2 Community
zephyr-7b-beta	1042	7.34	61.4	MIT
MPT-30B-chat	1031	6.39	50.4	CC-BY-NC-SA-4.0



clem 🤗
@ClementDelangue

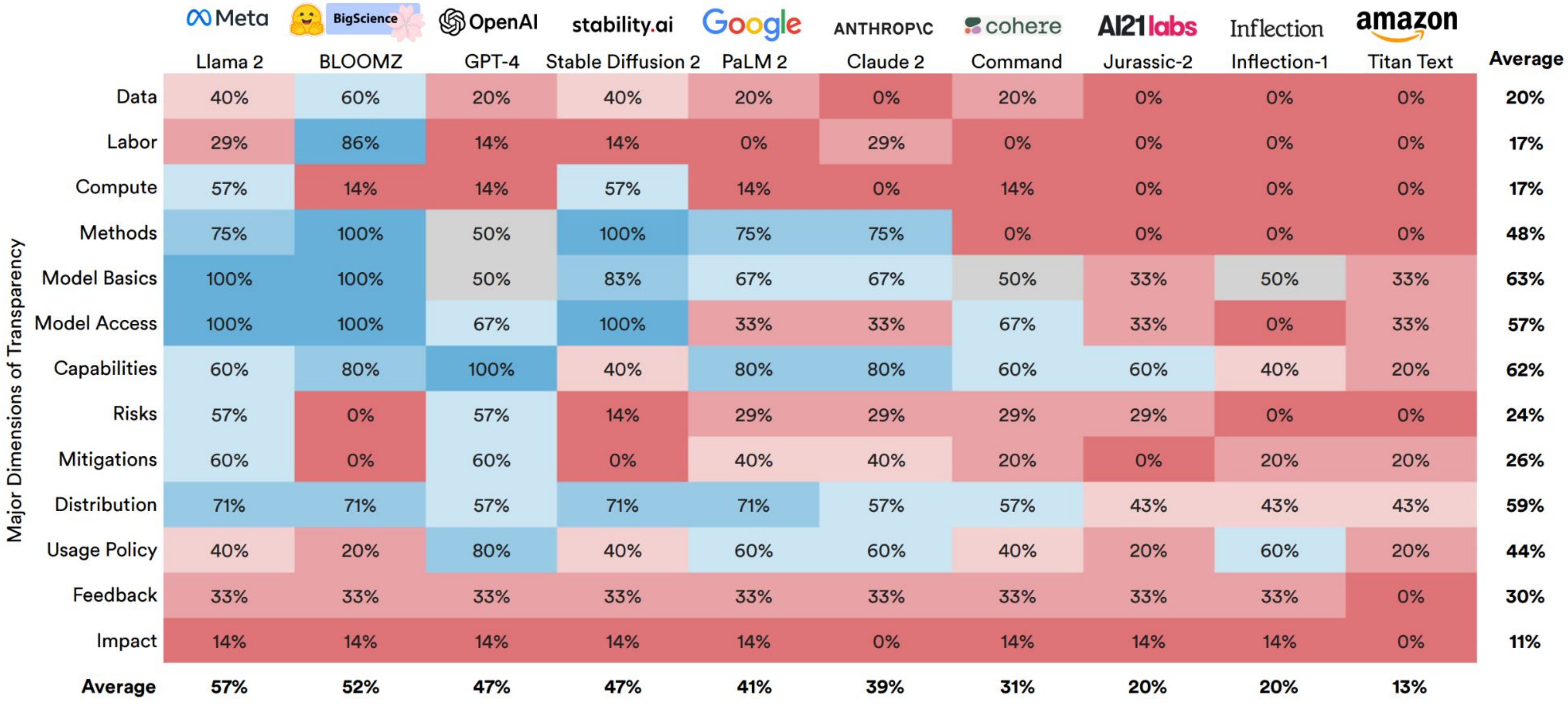
Numbers of public models on @huggingface:
- @Meta: 689 including MusicGen, Galactica, Wav2Vec, RoBERTa,... - huggingface.co/facebook
- @Google: 591 including BERT, Flan, T5, mobilnet,... - huggingface.co/google
- @Microsoft : 252 including DialoGPT, BioGPT, layoutLM, uniML, Deberta,... -huggingface.co/microsoft
- @Salesforce: 88 including CodeGen, Blip,... - huggingface.co/Salesforce
- @nvidia: 86 including Megatron, Segformer,... - huggingface.co/nvidia

Inspiring to see all the contributions of big technology companies to open-source AI. Let's go!

Last edited 4:06 PM · Jul 17, 2023 · **405.8K** Views

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index



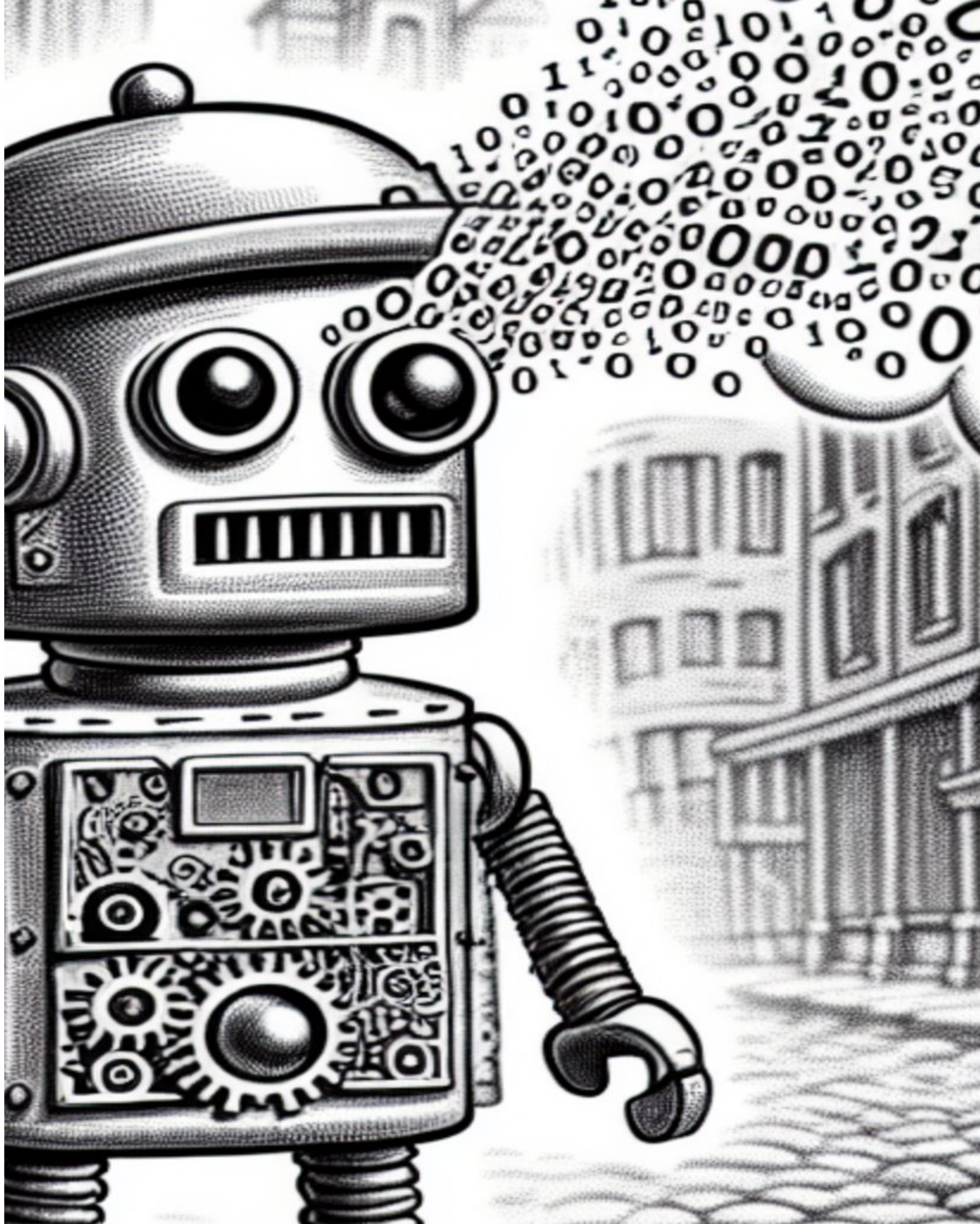
Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROP\C Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

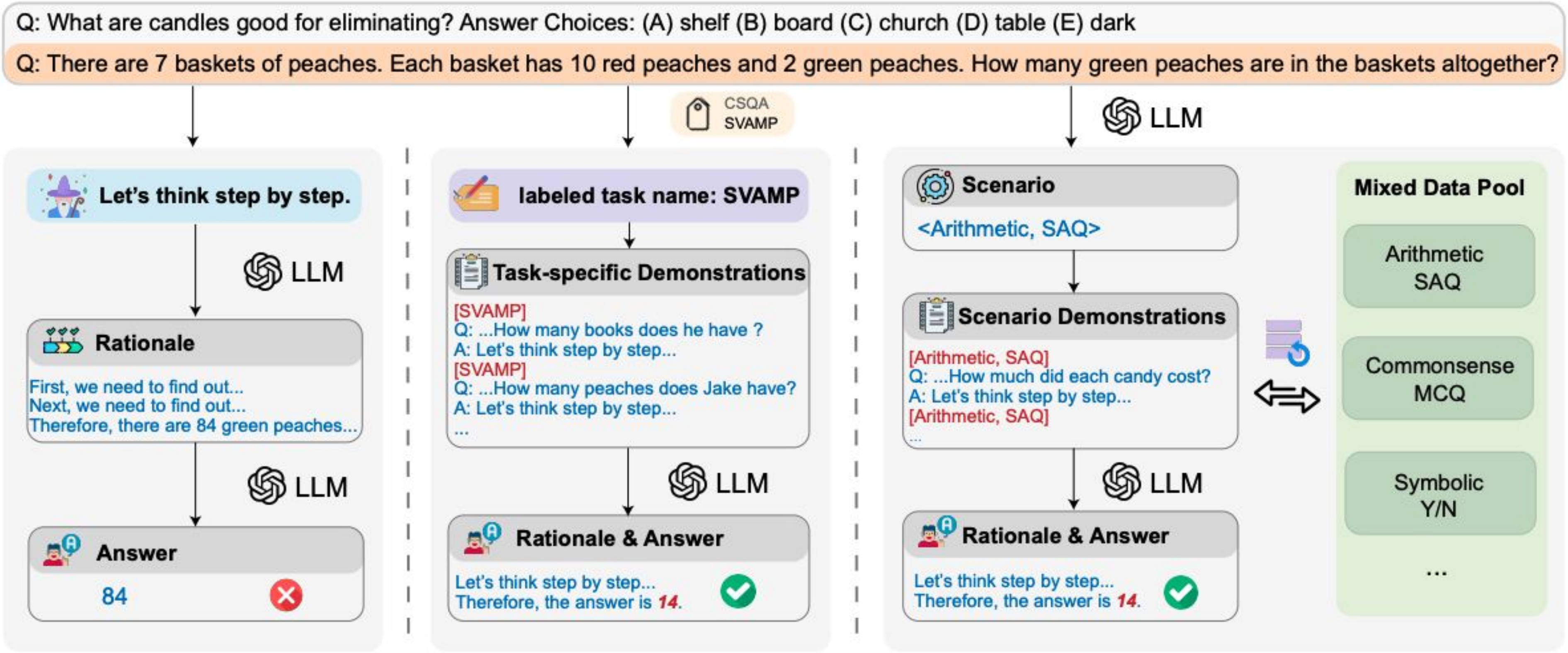
[The Foundational Model Transparency Index]

Hallucinations

When a model makes stuff up and generating nonsensical text.

While it can be a feature in some creative tasks, hallucinations can be potentially harmful to users.



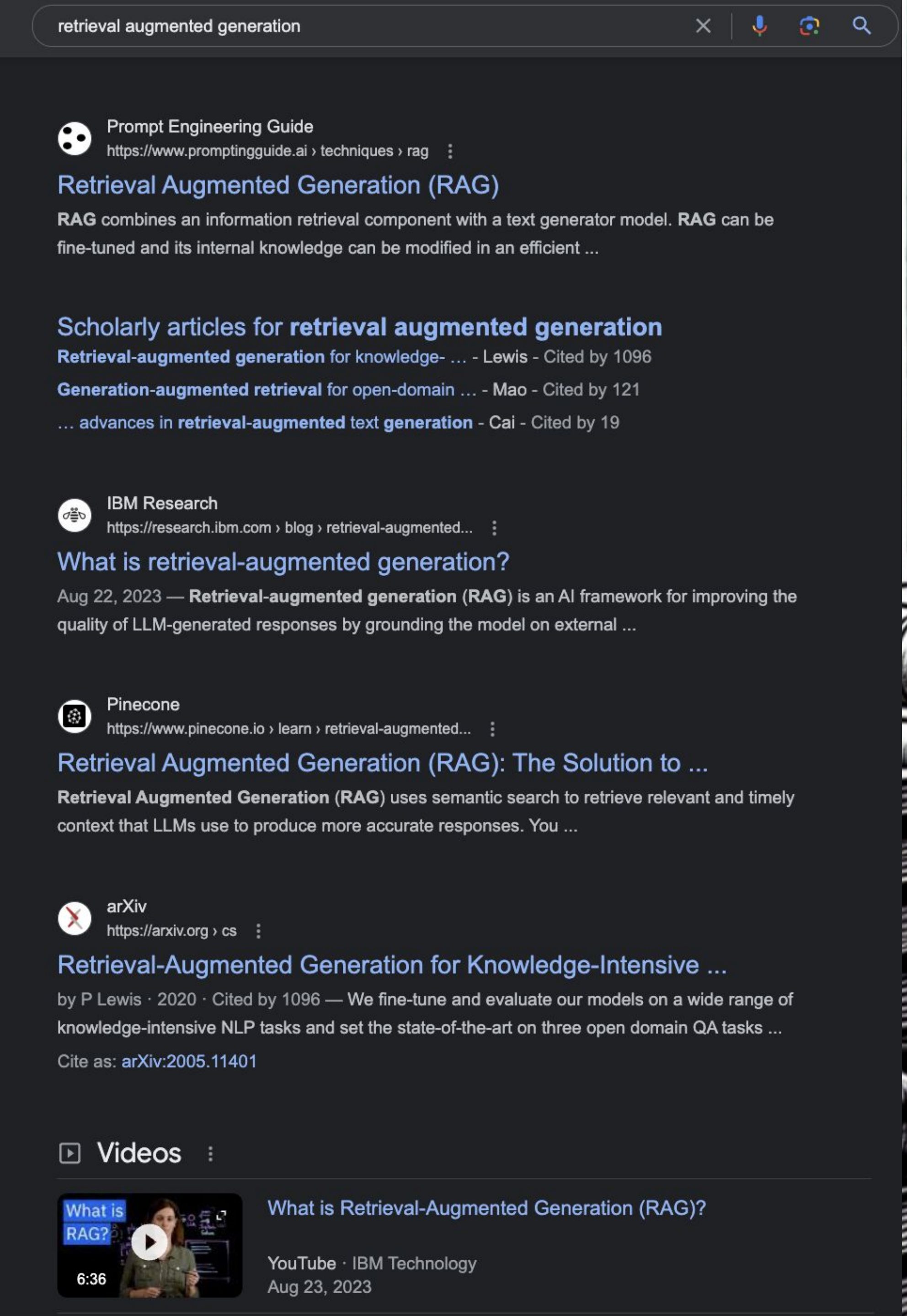

Q: What are candles good for eliminating? Answer Choices: (A) shelf (B) board (C) church (D) table (E) dark
Q: There are 7 baskets of peaches. Each basket has 10 red peaches and 2 green peaches. How many green peaches are in the baskets altogether?
Let's think step by step.
LLM
Rationale
First, we need to find out...
Next, we need to find out...
Therefore, there are 84 green peaches...
LLM
Answer
84
CSQA
SVAMP
labeled task name: SVAMP
Task-specific Demonstrations
[SVAMP]
Q: ...How many books does he have ?
A: Let's think step by step...
[SVAMP]
Q: ...How many peaches does Jake have?
A: Let's think step by step...
...
LLM
Rationale & Answer
Let's think step by step...
Therefore, the answer is 14.
LLM
Scenario
<Arithmetic, SAQ>
Scenario Demonstrations
[Arithmetic, SAQ]
Q: ...How much did each candy cost?
A: Let's think step by step...
[Arithmetic, SAQ]
...
LLM
Rationale & Answer
Let's think step by step...
Therefore, the answer is 14.
Mixed Data Pool
Arithmetic SAQ
Commonsense MCQ
Symbolic Y/N
...

(a) General Zero-Shot-CoT

(b) Specific Few-Shot-CoT

(c) Meta-CoT

Figure 1: Comparison with existing paradigms of CoT prompting. General Zero-Shot-CoT and Specific Few-Shot-CoT are from Kojima et al. (2023) and Wei et al. (2023), respectively.



retrieval augmented generation

Prompt Engineering Guide
https://www.promptingguide.ai › techniques › rag

Retrieval Augmented Generation (RAG)

RAG combines an information retrieval component with a text generator model. **RAG** can be fine-tuned and its internal knowledge can be modified in an efficient ...

Scholarly articles for **retrieval augmented generation**

Retrieval-augmented generation for knowledge- … - Lewis - Cited by 1096

Generation-augmented retrieval for open-domain … - Mao - Cited by 121

… advances in **retrieval-augmented** text **generation** - Cai - Cited by 19

IBM Research
https://research.ibm.com › blog › retrieval-augmented...

What is retrieval-augmented generation?

Aug 22, 2023 — **Retrieval-augmented generation** (**RAG**) is an AI framework for improving the quality of LLM-generated responses by grounding the model on external ...

Pinecone
https://www.pinecone.io › learn › retrieval-augmented...

Retrieval Augmented Generation (RAG): The Solution to ...

Retrieval Augmented Generation (**RAG**) uses semantic search to retrieve relevant and timely context that LLMs use to produce more accurate responses. You ...

arXiv
https://arxiv.org › cs

Retrieval-Augmented Generation for Knowledge-Intensive ...

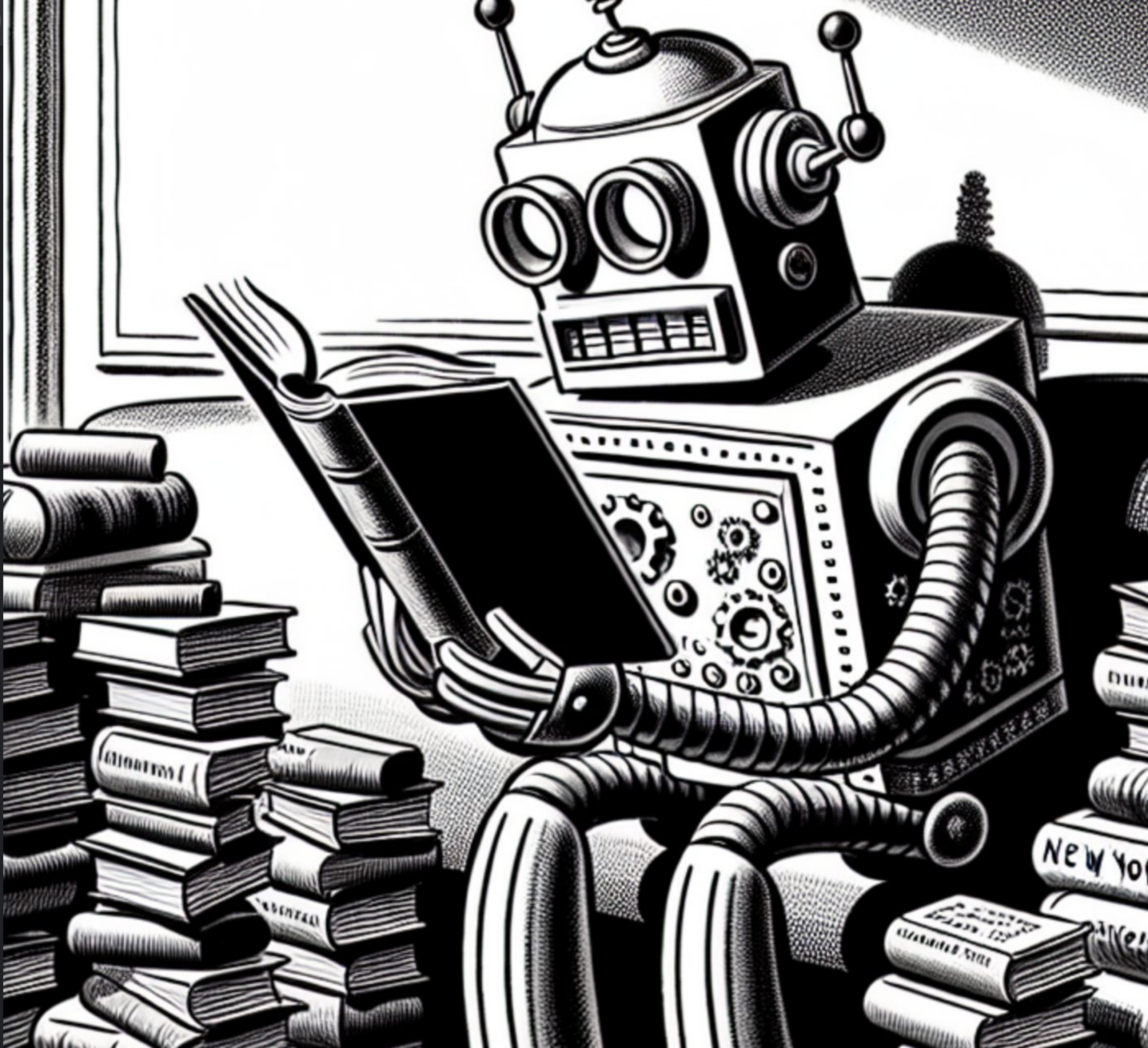
by P Lewis · 2020 · Cited by 1096 — We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open domain QA tasks ...

Cite as: arXiv:2005.11401

Videos

What is Retrieval-Augmented Generation (RAG)?

YouTube · IBM Technology
Aug 23, 2023



clarifai
The World's AI™

Closing Thoughts