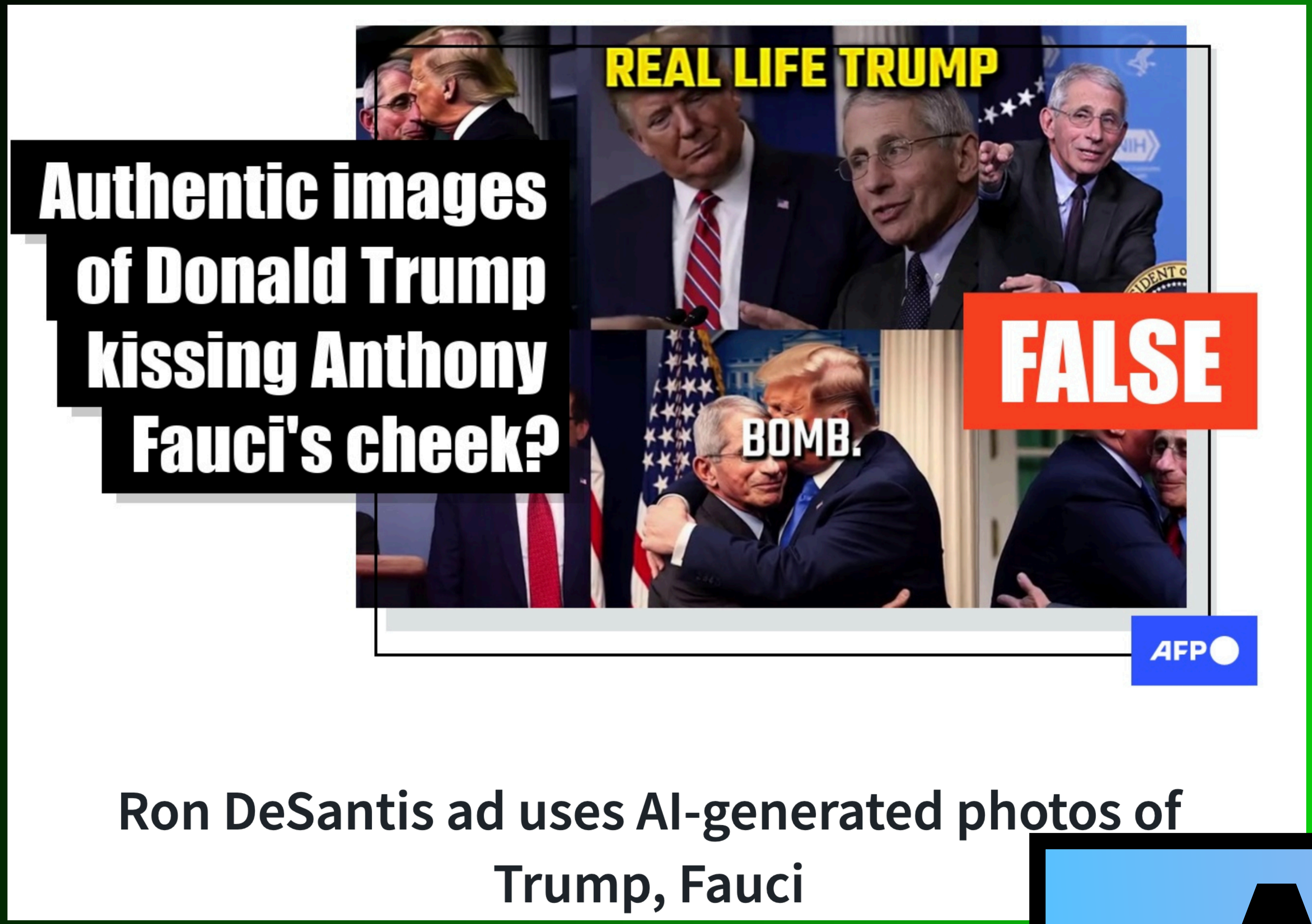
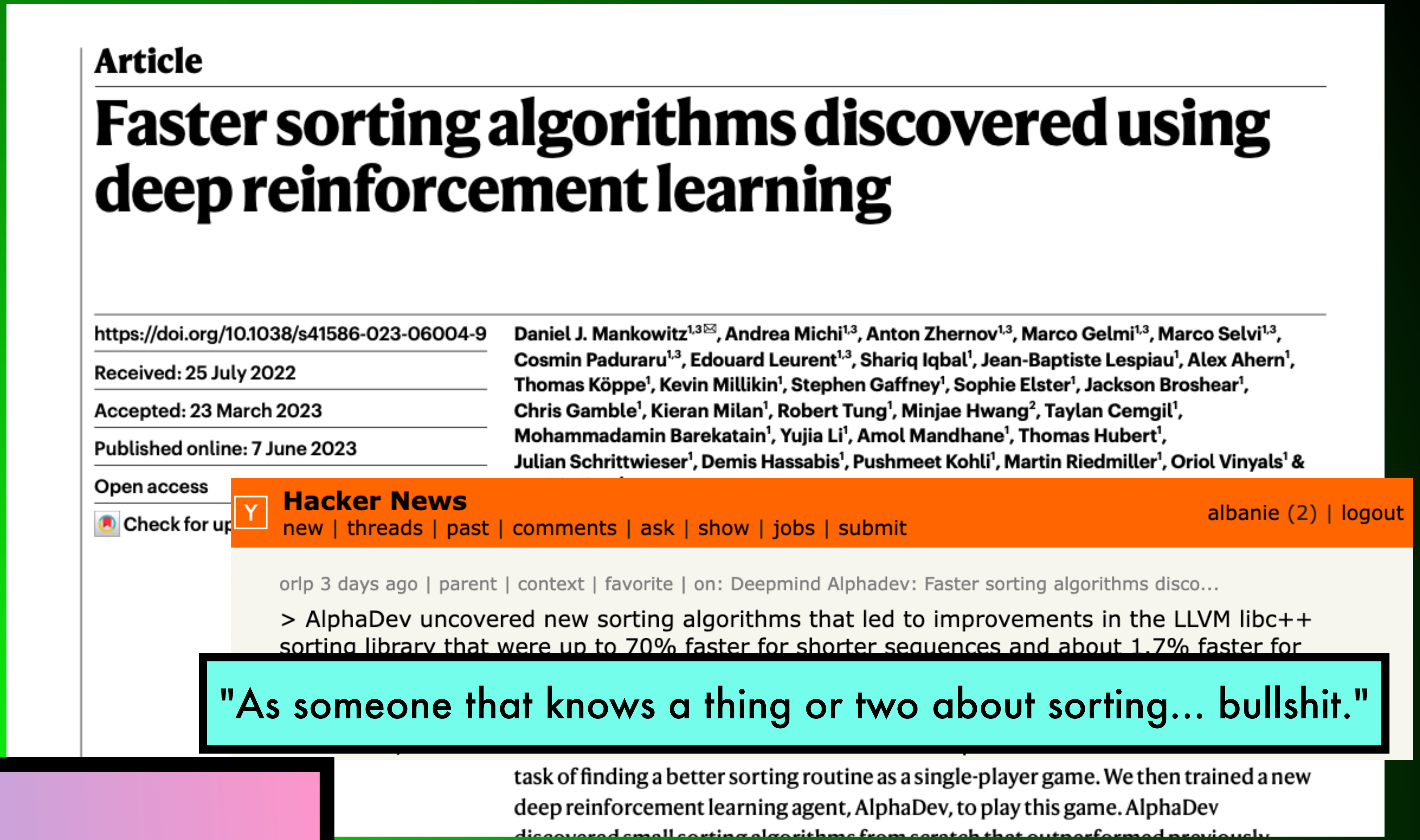


Authentic images of Donald Trump kissing Anthony Fauci's cheek?

Ron DeSantis ad uses AI-generated photos of Trump, Fauci


Article

Faster sorting algorithms discovered using deep reinforcement learning

https://doi.org/10.1038/s41586-023-06004-9

Received: 25 July 2022

Accepted: 23 March 2023

Published online: 7 June 2023

Open access

Check for up

Daniel J. Mankowitz[1,3✉], Andrea Michi[1,3], Anton Zhernov[1,3], Marco Gelmi[1,3], Marco Selvi[1,3], Cosmin Paduraru[1,3], Edouard Leurent[1,3], Shariq Iqbal[1], Jean-Baptiste Lespiau[1], Alex Ahern[1], Thomas Köppe[1], Kevin Millikin[1], Stephen Gaffney[1], Sophie Elster[1], Jackson Broshear[1], Chris Gamble[1], Kieran Milan[1], Robert Tung[1], Minjae Hwang[2], Taylan Cemgil[1], Mohammadamin Barekatain[1], Yujia Li[1], Amol Mandhane[1], Thomas Hubert[1], Julian Schrittwieser[1], Demis Hassabis[1], Pushmeet Kohli[1], Martin Riedmiller[1], Oriol Vinyals[1] &

Y Hacker News

new | threads | past | comments | ask | show | jobs | submit

albanie (2) | logout

orlp 3 days ago | parent | context | favorite | on: Deepmind Alphadev: Faster sorting algorithms disco...

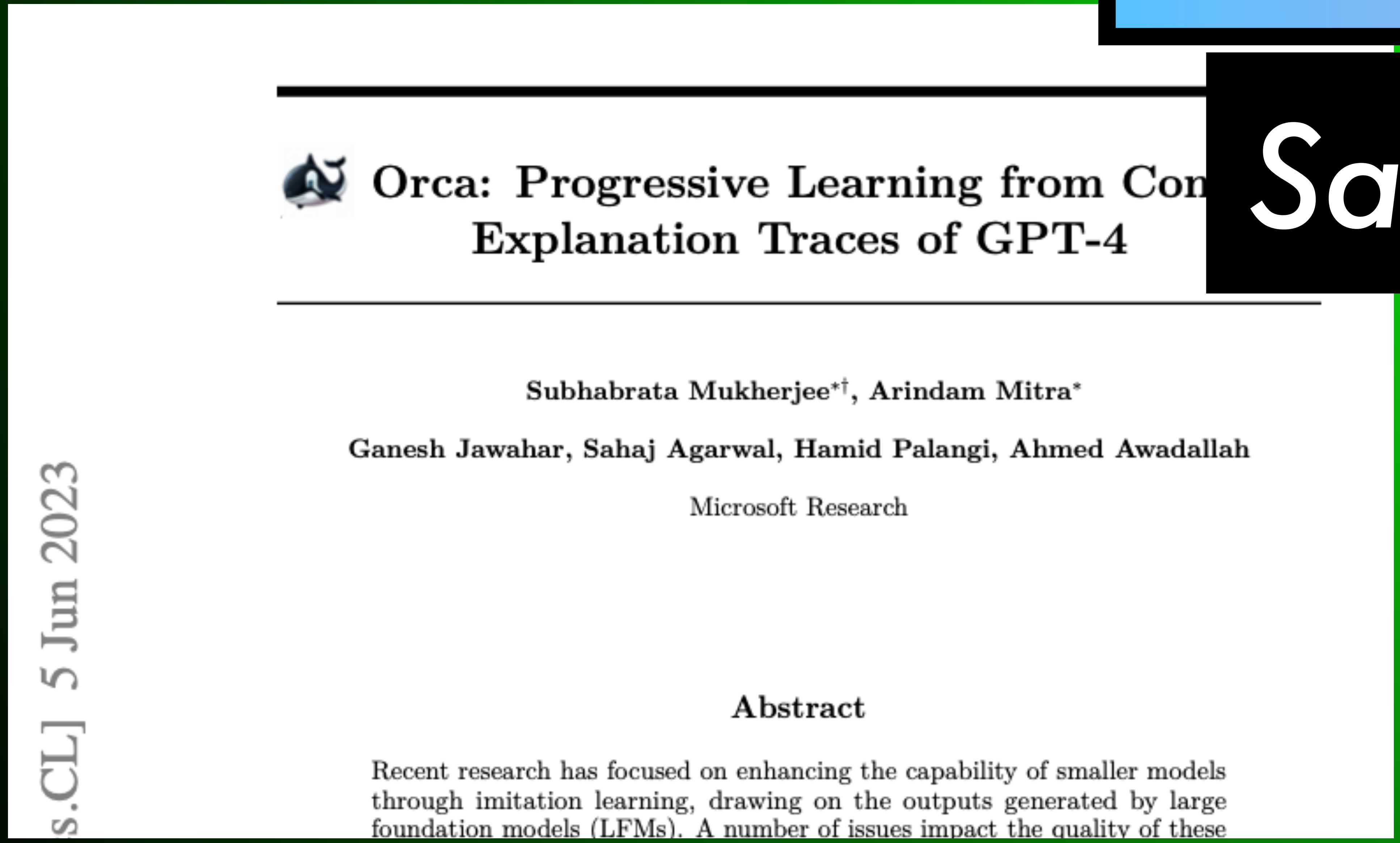
> AlphaDev uncovered new sorting algorithms that led to improvements in the LLVM libc++ sorting library that were up to 70% faster for shorter sequences and about 1.7% faster for

"As someone that knows a thing or two about sorting... bullshit."

task of finding a better sorting routine as a single-player game. We then trained a new deep reinforcement learning agent, AlphaDev, to play this game. AlphaDev discovered small sorting algorithms from scratch that outperformed previously

AI News

Samuel Albanie


s.CL] 5 Jun 2023

Orca: Progressive Learning from Com Explanation Traces of GPT-4

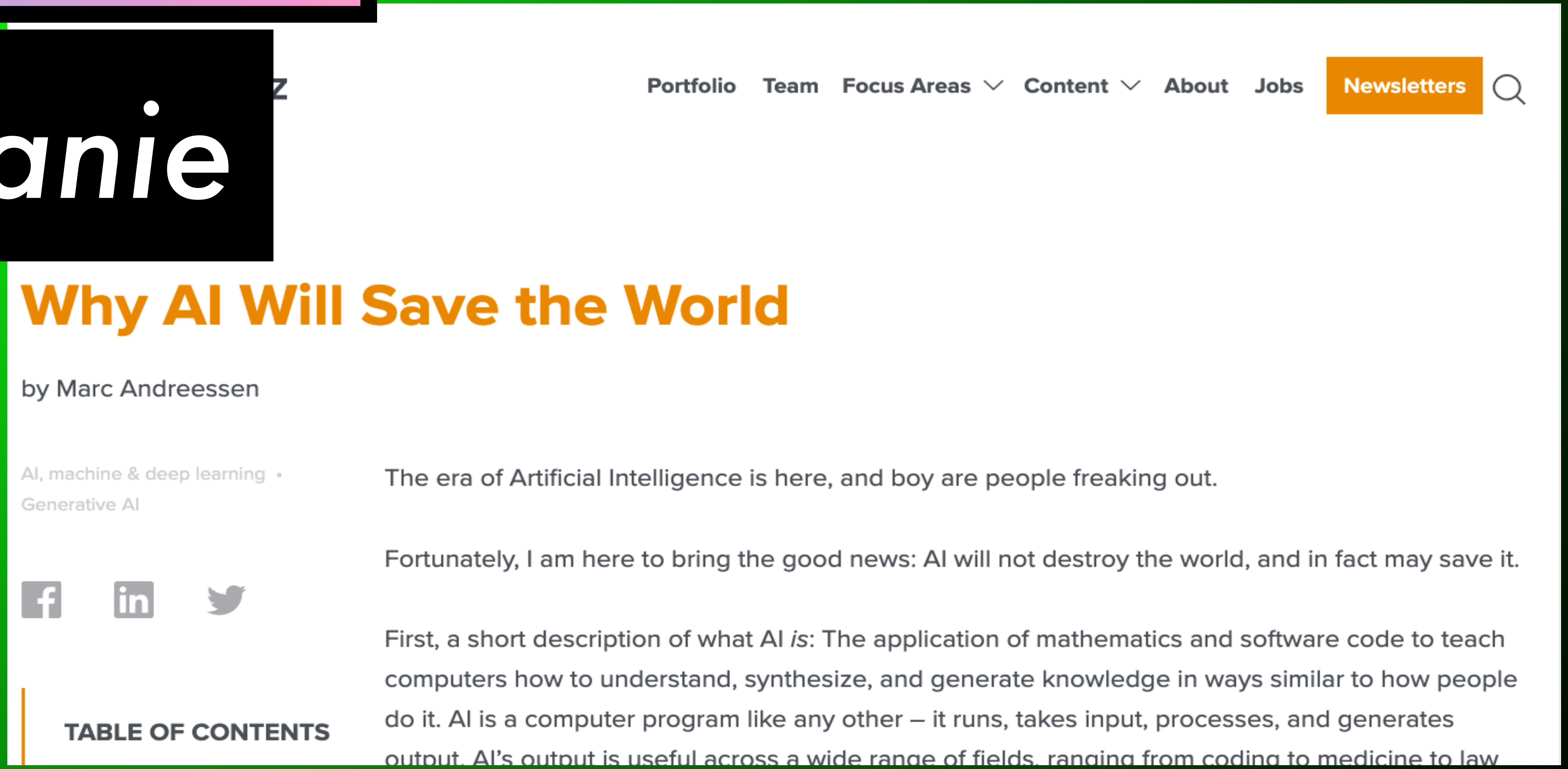
Subhabrata Mukherjee*†, Arindam Mitra*

Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, Ahmed Awadallah

Microsoft Research

Abstract

Recent research has focused on enhancing the capability of smaller models through imitation learning, drawing on the outputs generated by large foundation models (LFMs). A number of issues impact the quality of these


z

Portfolio Team Focus Areas Content About Jobs Newsletters

Why AI Will Save the World

by Marc Andreessen

AI, machine & deep learning • Generative AI

TABLE OF CONTENTS

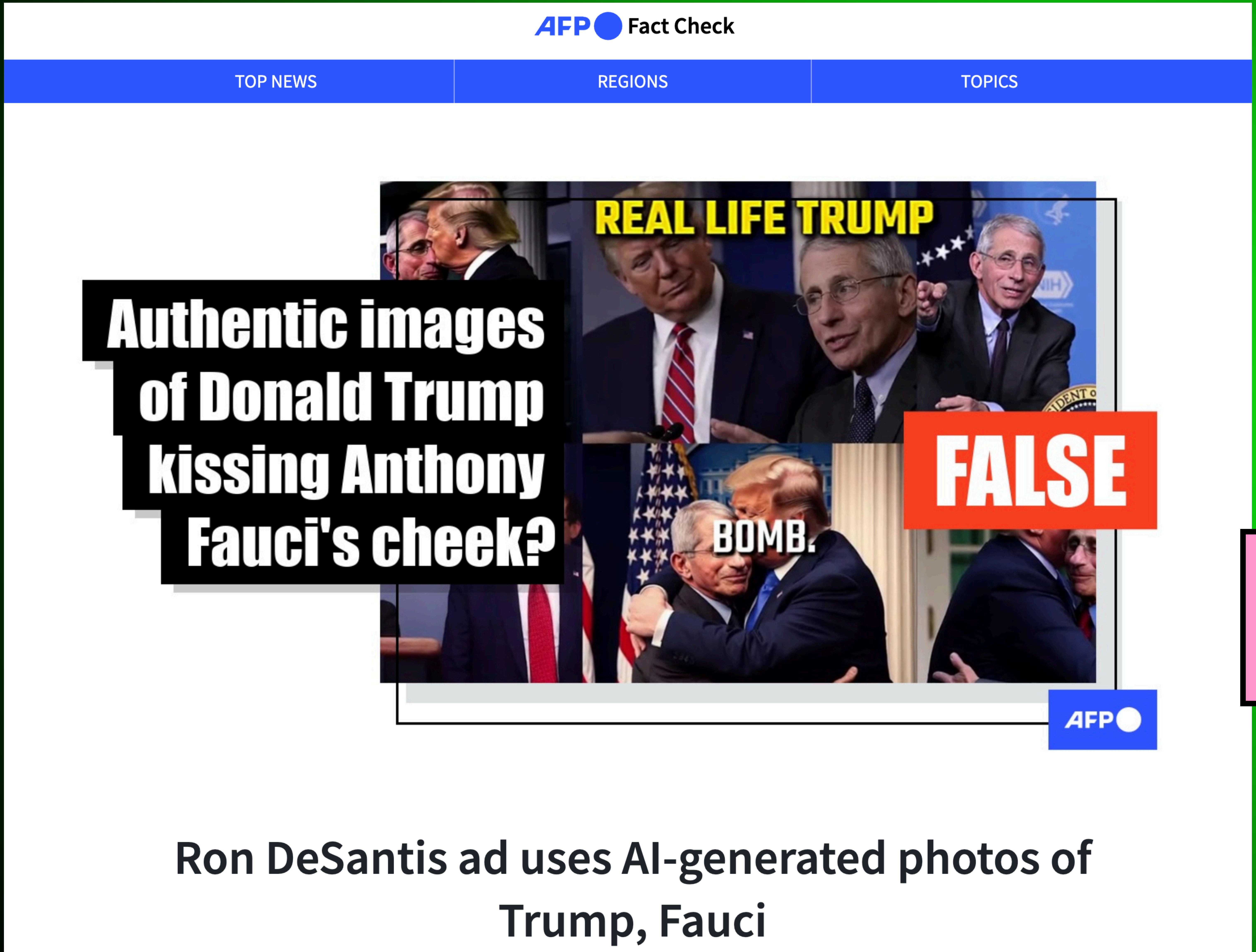
The era of Artificial Intelligence is here, and boy are people freaking out.

Fortunately, I am here to bring the good news: AI will not destroy the world, and in fact may save it.

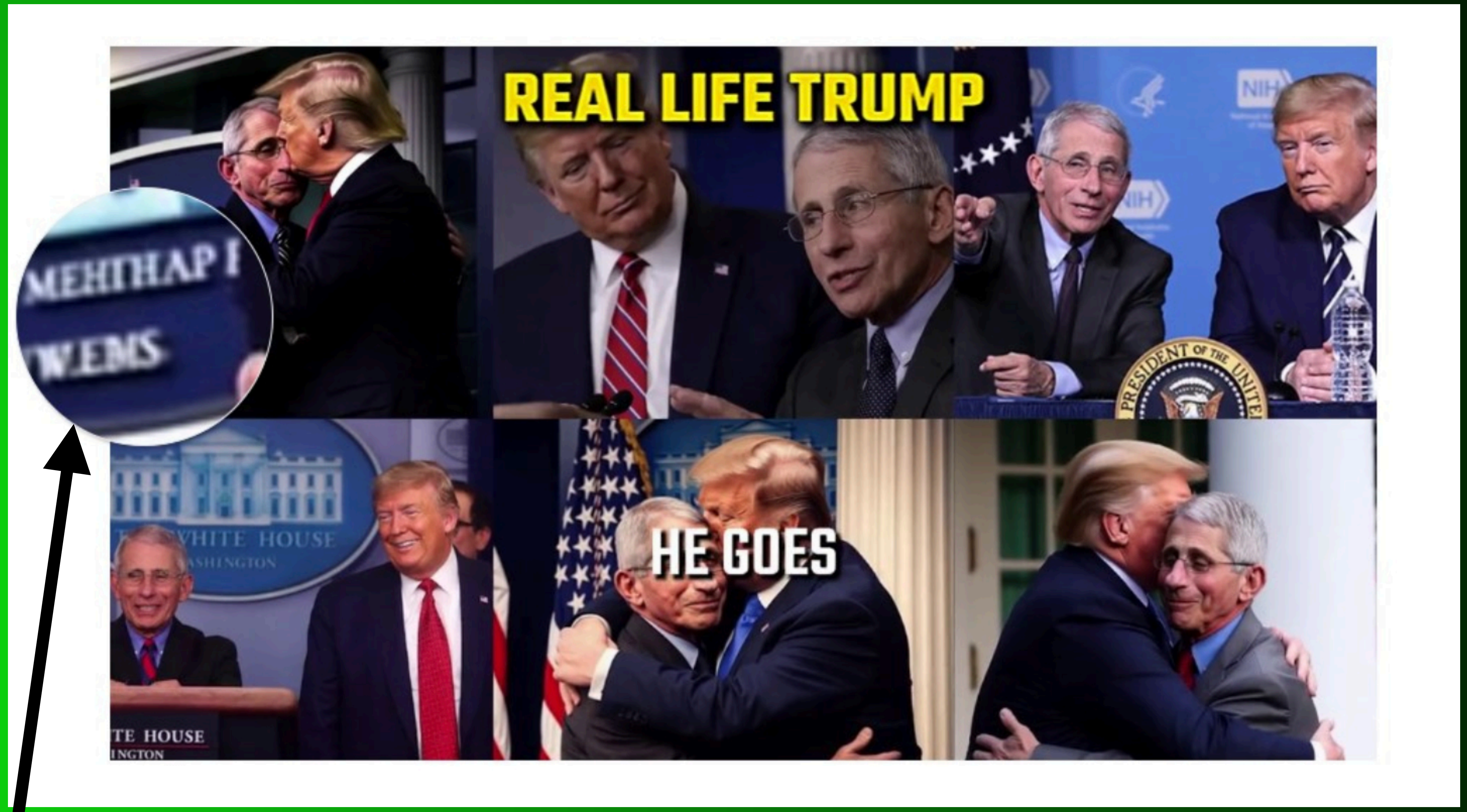
First, a short description of what AI *is*: The application of mathematics and software code to teach computers how to understand, synthesize, and generate knowledge in ways similar to how people do it. AI is a computer program like any other – it runs, takes input, processes, and generates output. AI's output is useful across a wide range of fields, ranging from coding to medicine to law

AI News

7th June 2023

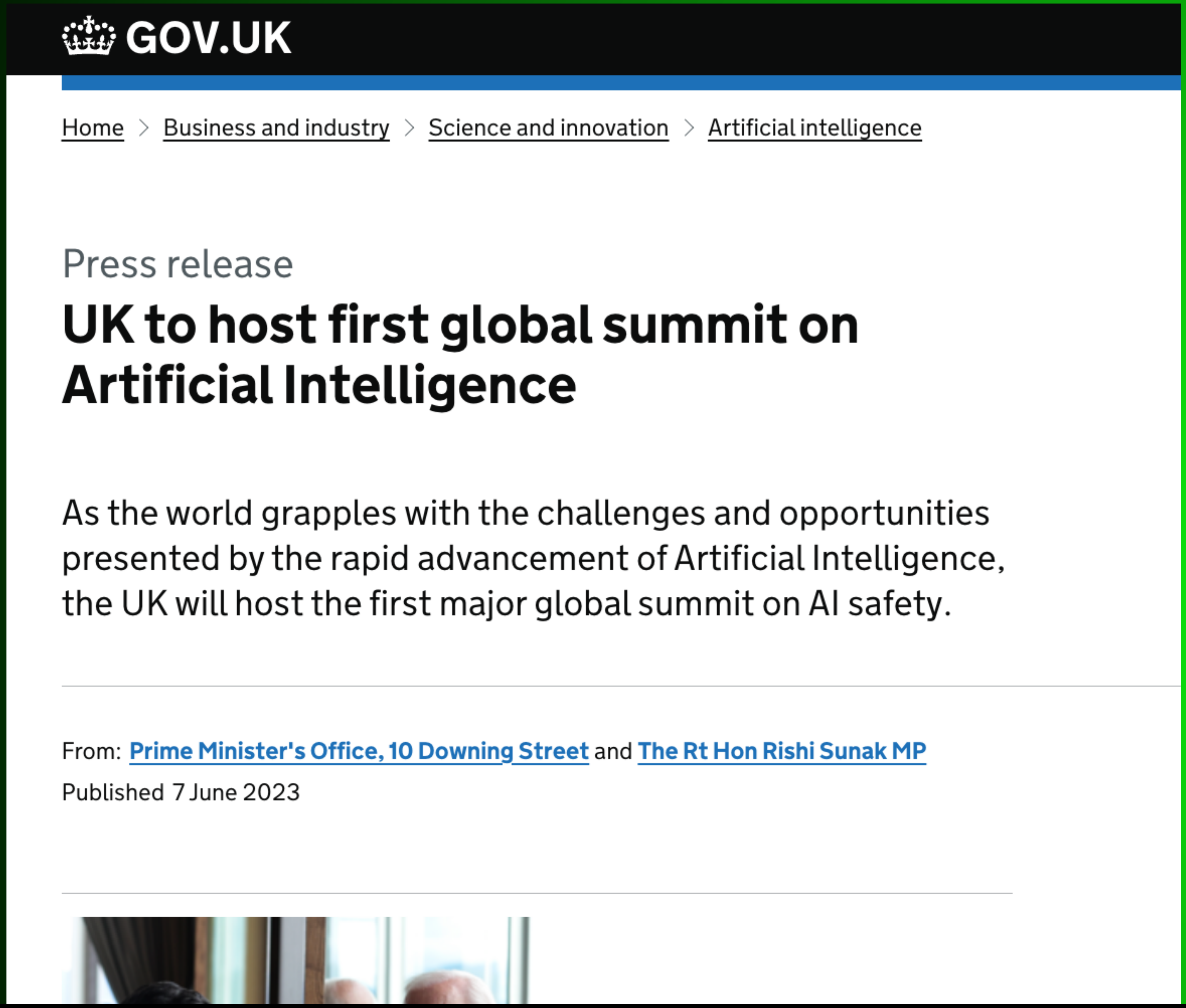

AFP Fact Check
TOP NEWS
REGIONS
TOPICS
REAL LIFE TRUMP
Authentic images of Donald Trump kissing Anthony Fauci's cheek?
BOMB.
FALSE
AFP

Ron DeSantis ad uses AI-generated photos of Trump, Fauci



'...the text is incomprehensible and does not properly spell "The White House"'

AI News



GOV.UK

Home > Business and industry > Science and innovation > Artificial intelligence

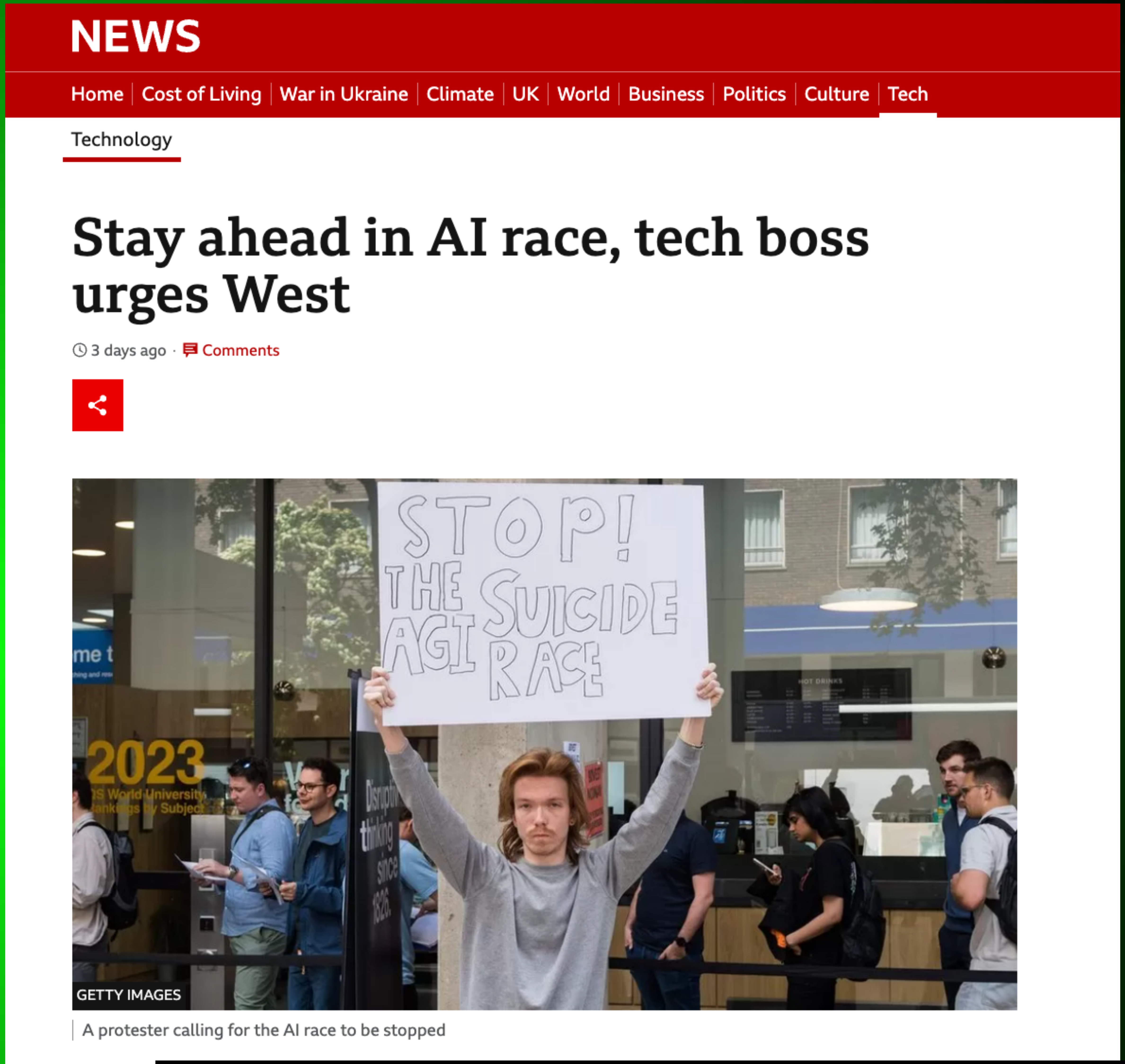
Press release

UK to host first global summit on Artificial Intelligence

As the world grapples with the challenges and opportunities presented by the rapid advancement of Artificial Intelligence, the UK will host the first major global summit on AI safety.

From: Prime Minister's Office, 10 Downing Street and The Rt Hon Rishi Sunak MP

Published 7 June 2023



NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech

Technology

Stay ahead in AI race, tech boss urges West

3 days ago · Comments

GETTY IMAGES

A protester calling for the AI race to be stopped

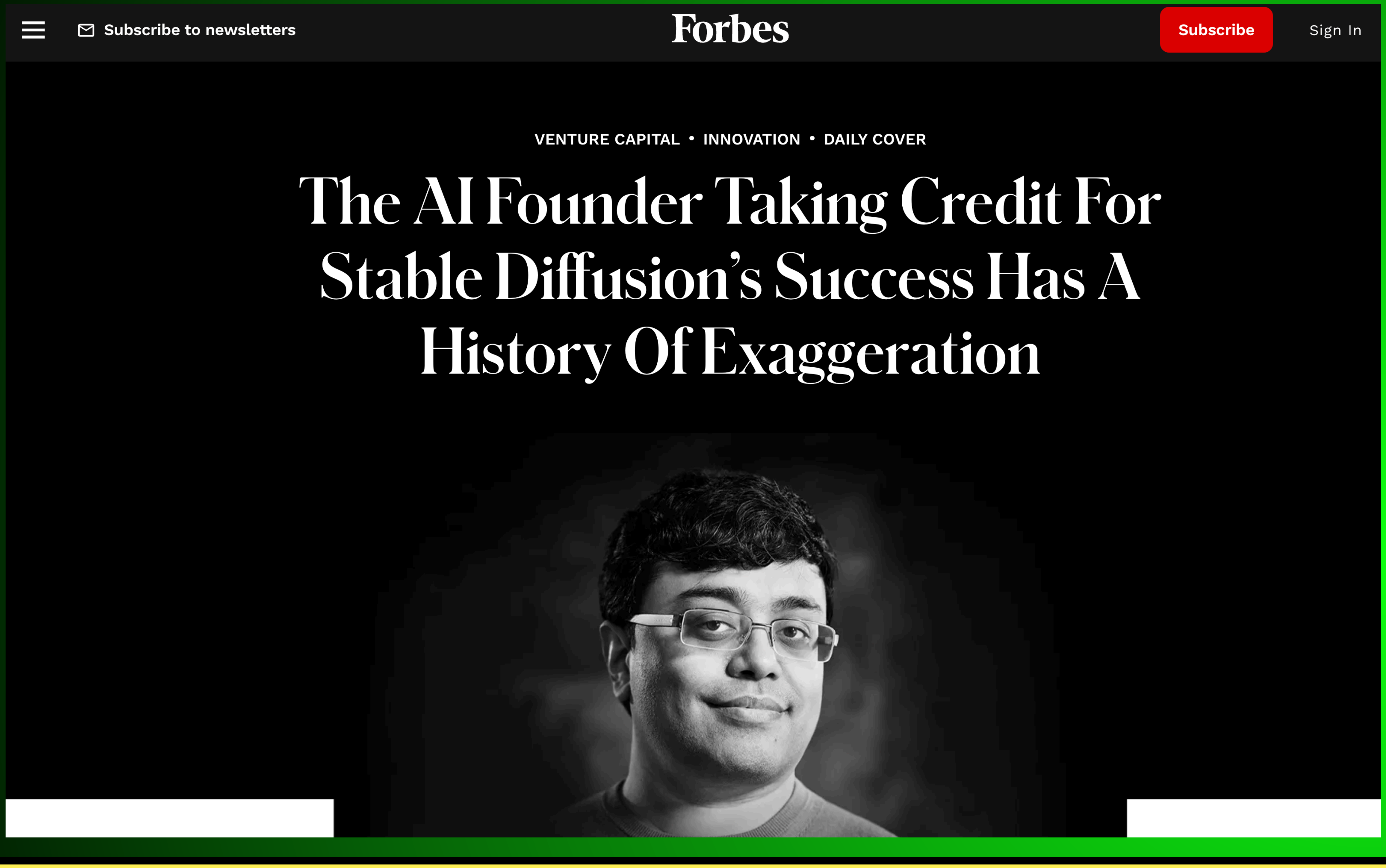
By Tom

Technology

"Summit will bring together key countries, leading tech companies and researchers to agree safety measures to evaluate and monitor the most significant risks from AI"

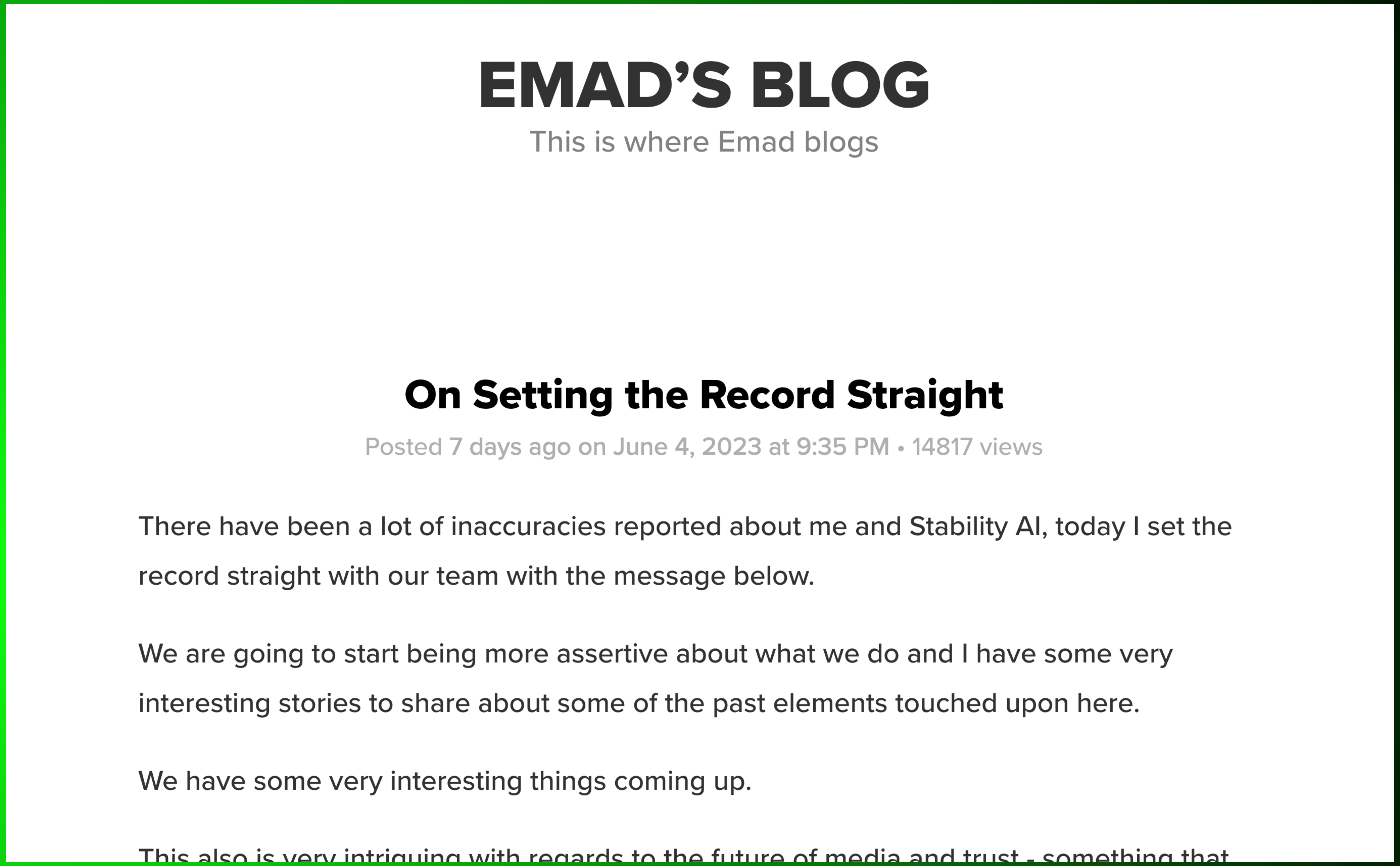
"the boss of software firm Palantir, Alex Karp, said it was only those with 'no products' who wanted a pause."

AI News



VENTURE CAPITAL • INNOVATION • DAILY COVER

The AI Founder Taking Credit For Stable Diffusion's Success Has A History Of Exaggeration



EMAD'S BLOG

This is where Emad blogs

On Setting the Record Straight

Posted 7 days ago on June 4, 2023 at 9:35 PM • 14817 views

There have been a lot of inaccuracies reported about me and Stability AI, today I set the record straight with our team with the message below.

We are going to start being more assertive about what we do and I have some very interesting stories to share about some of the past elements touched upon here.

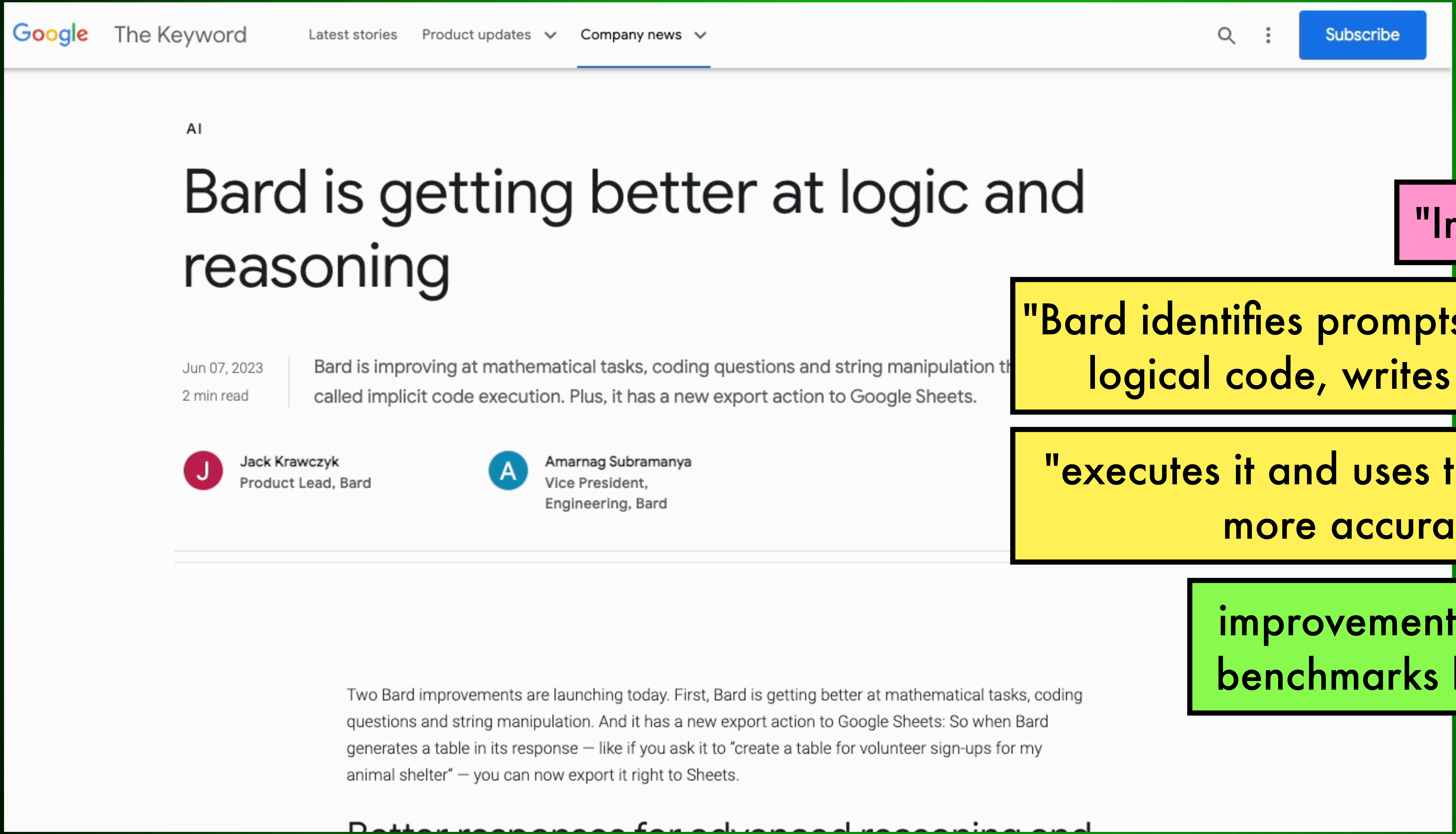
We have some very interesting things coming up.

This also is very intriguing with regards to the future of media and trust - something that

"Interviews with 13 current and former employees and more than two dozen investors, collaborators and former colleagues, as well as pitch decks and internal documents, suggest his recent success has been bolstered by *exaggeration and dubious claims*."

"Despite our team spending weeks going back and forth with Forbes to correct the record, they have clearly chosen to *ignore the truth* on many of these issues."

AI News



AI

Bard is getting better at logic and reasoning

Jun 07, 2023
2 min read

Bard is improving at mathematical tasks, coding questions and string manipulation th called implicit code execution. Plus, it has a new export action to Google Sheets.

Jack Krawczyk
Product Lead, Bard

Amarnag Subramanya
Vice President, Engineering, Bard

Two Bard improvements are launching today. First, Bard is getting better at mathematical tasks, coding questions and string manipulation. And it has a new export action to Google Sheets: So when Bard generates a table in its response — like if you ask it to "create a table for volunteer sign-ups for my animal shelter" — you can now export it right to Sheets.

Better responses for advanced reasoning and

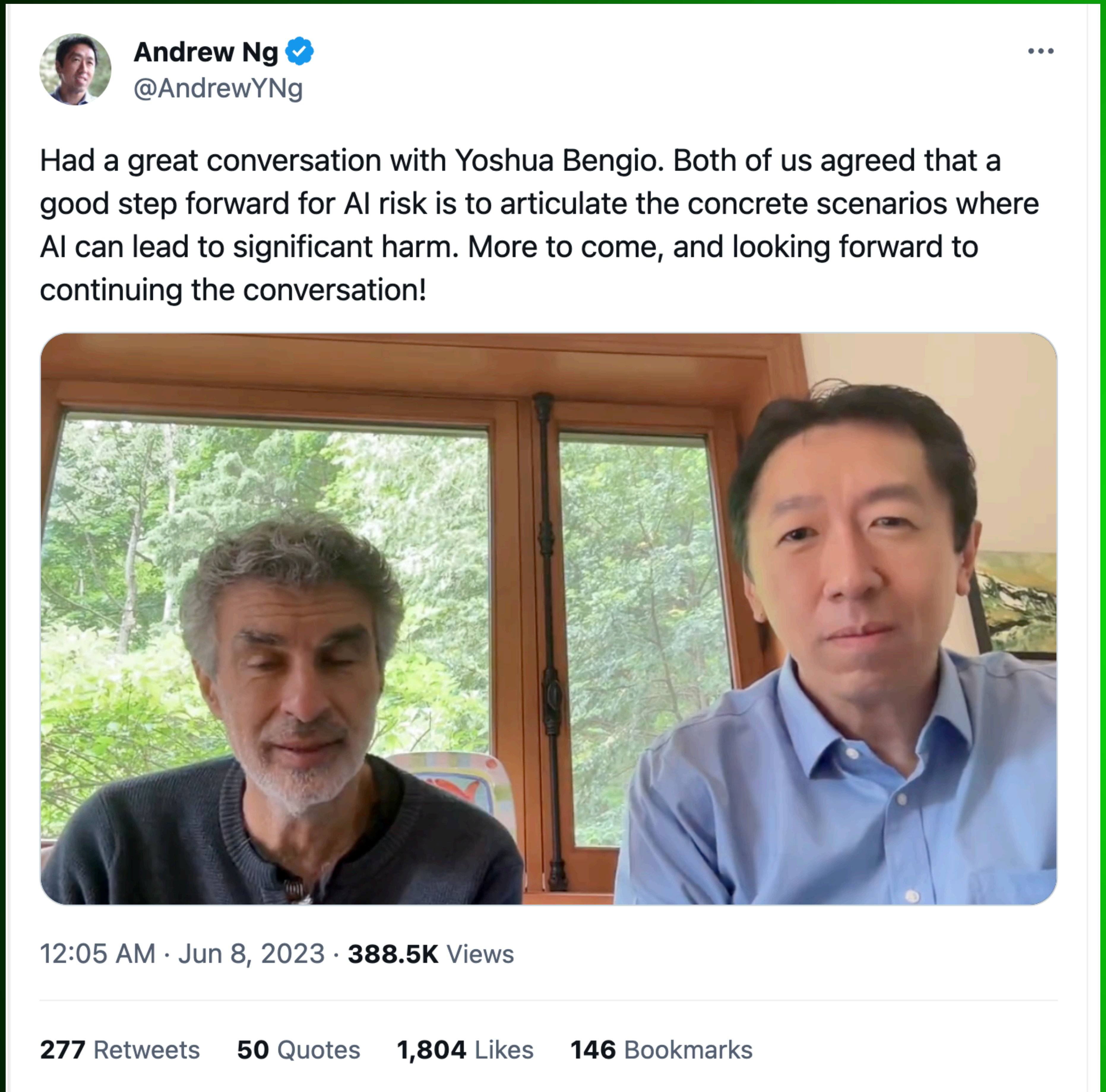
"Implicit code execution"

"Bard identifies prompts that might benefit from logical code, writes it 'under the hood,'"

"executes it and uses the result to generate a more accurate response."

improvements on internal challenge benchmarks by approximately 30%

AI News



Andrew Ng @AndrewYNg

Had a great conversation with Yoshua Bengio. Both of us agreed that a good step forward for AI risk is to articulate the concrete scenarios where AI can lead to significant harm. More to come, and looking forward to continuing the conversation!

12:05 AM · Jun 8, 2023 · **388.5K** Views

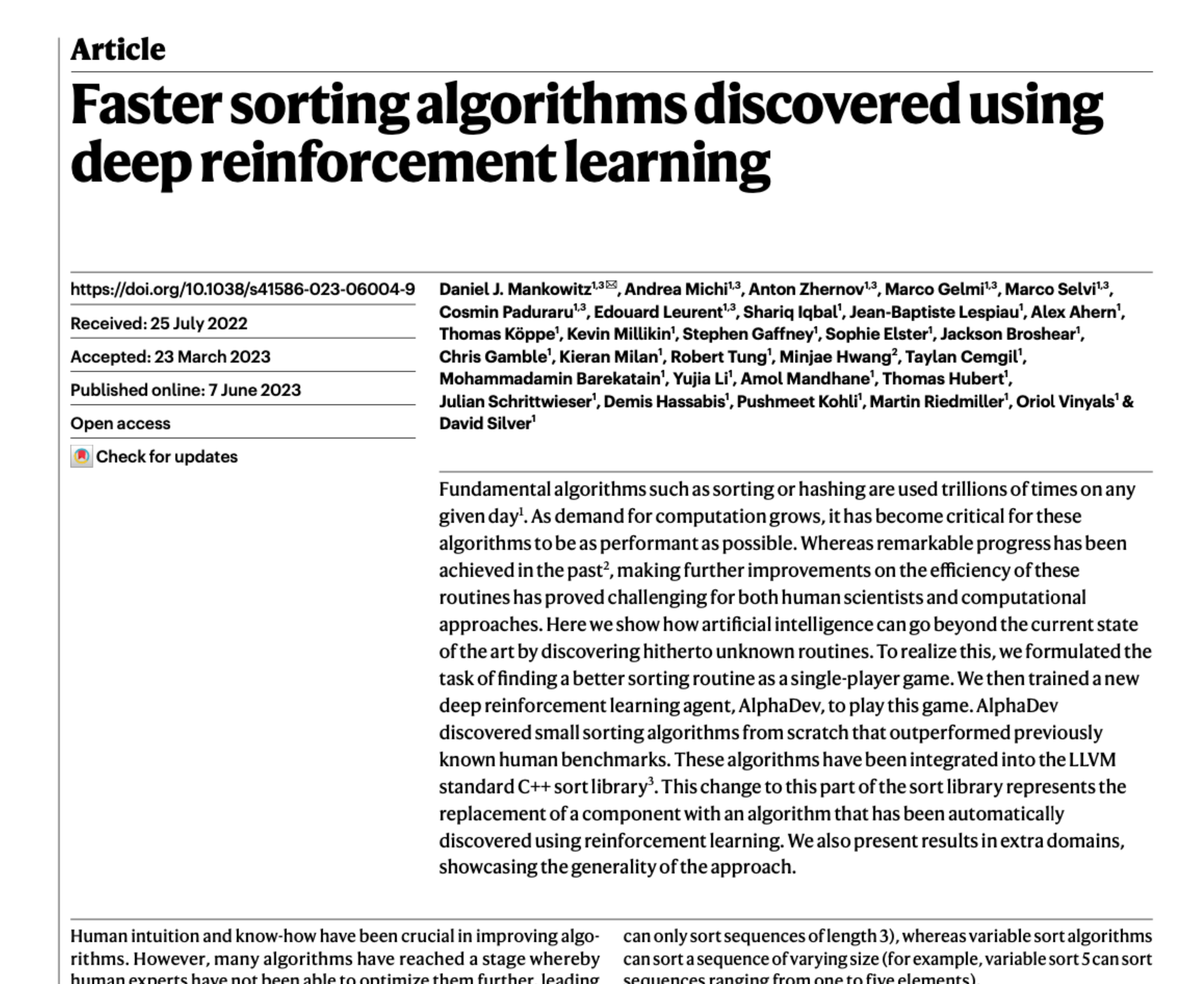
277 Retweets **50** Quotes **1,804** Likes **146** Bookmarks

"Both of us agreed that a good step forward for AI risk is to articulate the concrete scenarios where AI can lead to significant harm."

AlphaDev

7th June 2023

"AlphaDev discovered small sorting algorithms from scratch that outperformed previously known human benchmarks."



Article

Faster sorting algorithms discovered using deep reinforcement learning

https://doi.org/10.1038/s41586-023-06004-9

Received: 25 July 2022

Accepted: 23 March 2023

Published online: 7 June 2023

Open access

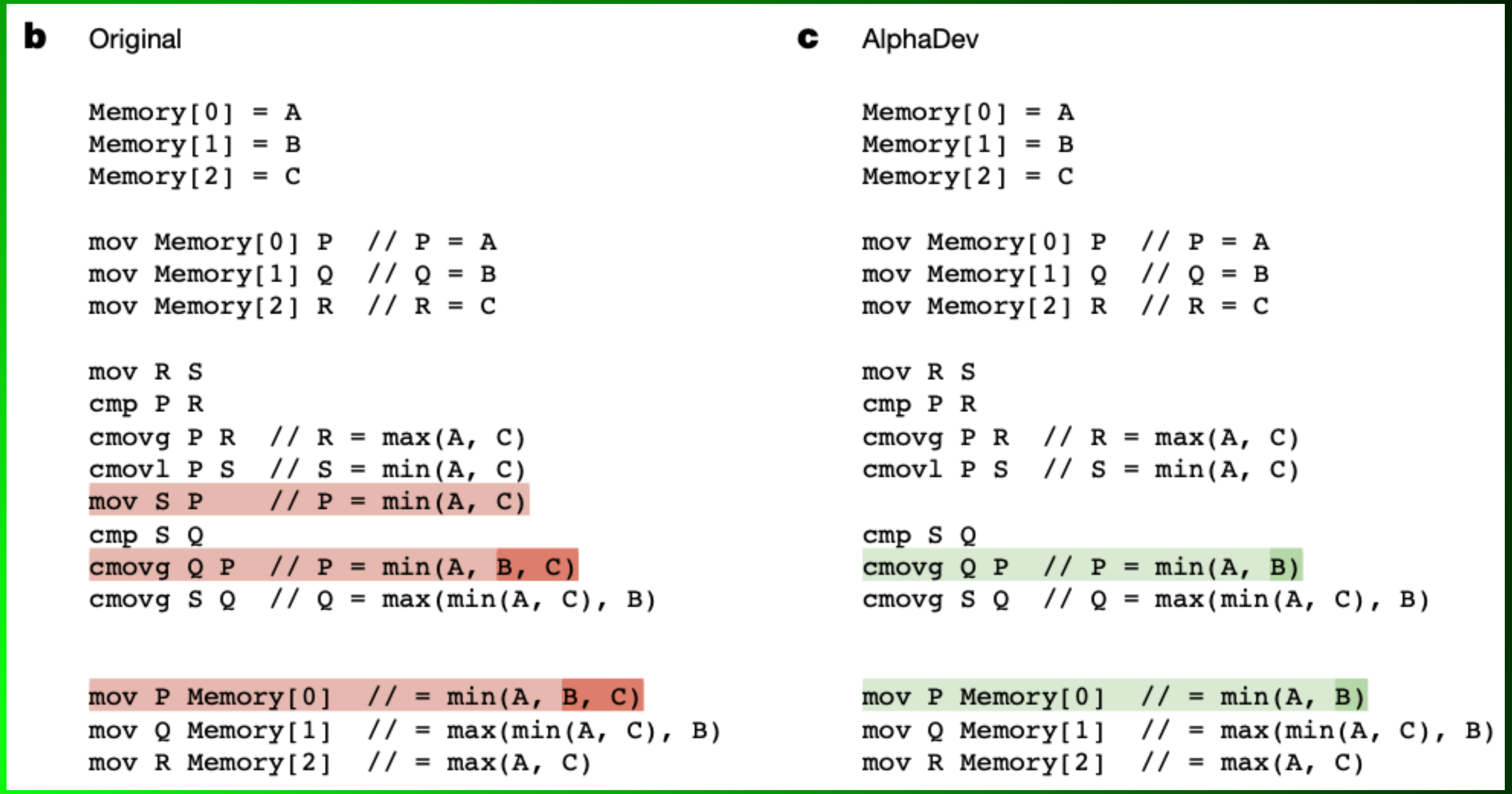
Check for updates

Daniel J. Mankowitz[1,3]✉, Andrea Michi[1,3], Anton Zhernov[1,3], Marco Gelmi[1,3], Marco Selvi[1,3], Cosmin Paduraru[1,3], Edouard Leurent[1,3], Shariq Iqbal[1], Jean-Baptiste Lespiau[1], Alex Ahern[1], Thomas Köppe[1], Kevin Millikin[1], Stephen Gaffney[1], Sophie Elster[1], Jackson Broshear[1], Chris Gamble[1], Kieran Milan[1], Robert Tung[1], Minjae Hwang[2], Taylan Cemgil[1], Mohammadamin Barekatain[1], Yujia Li[1], Amol Mandhane[1], Thomas Hubert[1], Julian Schrittwieser[1], Demis Hassabis[1], Pushmeet Kohli[1], Martin Riedmiller[1], Oriol Vinyals[1] & David Silver[1]

Fundamental algorithms such as sorting or hashing are used trillions of times on any given day[1]. As demand for computation grows, it has become critical for these algorithms to be as performant as possible. Whereas remarkable progress has been achieved in the past[2], making further improvements on the efficiency of these routines has proved challenging for both human scientists and computational approaches. Here we show how artificial intelligence can go beyond the current state of the art by discovering hitherto unknown routines. To realize this, we formulated the task of finding a better sorting routine as a single-player game. We then trained a new deep reinforcement learning agent, AlphaDev, to play this game. AlphaDev discovered small sorting algorithms from scratch that outperformed previously known human benchmarks. These algorithms have been integrated into the LLVM standard C++ sort library[3]. This change to this part of the sort library represents the replacement of a component with an algorithm that has been automatically discovered using reinforcement learning. We also present results in extra domains, showcasing the generality of the approach.

Human intuition and know-how have been crucial in improving algorithms. However, many algorithms have reached a stage whereby human experts have not been able to optimize them further, leading

can only sort sequences of length 3), whereas variable sort algorithms can sort a sequence of varying size (for example, variable sort 5 can sort sequences ranging from one to five elements).



b Original

```
Memory[0] = A
Memory[1] = B
Memory[2] = C

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C

mov R S
cmp P R
cmovg P R  // R = max(A, C)
cmovl P S  // S = min(A, C)
mov S P    // P = min(A, C)
cmp S Q
cmovg Q P  // P = min(A, B, C)
cmovg S Q  // Q = max(min(A, C), B)


mov P Memory[0]  // = min(A, B, C)
mov Q Memory[1]  // = max(min(A, C), B)
mov R Memory[2]  // = max(A, C)
```

c AlphaDev

```
Memory[0] = A
Memory[1] = B
Memory[2] = C

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C

mov R S
cmp P R
cmovg P R  // R = max(A, C)
cmovl P S  // S = min(A, C)

cmp S Q
cmovg Q P  // P = min(A, B)
cmovg S Q  // Q = max(min(A, C), B)


mov P Memory[0]  // = min(A, B)
mov Q Memory[1]  // = max(min(A, C), B)
mov R Memory[2]  // = max(A, C)
```

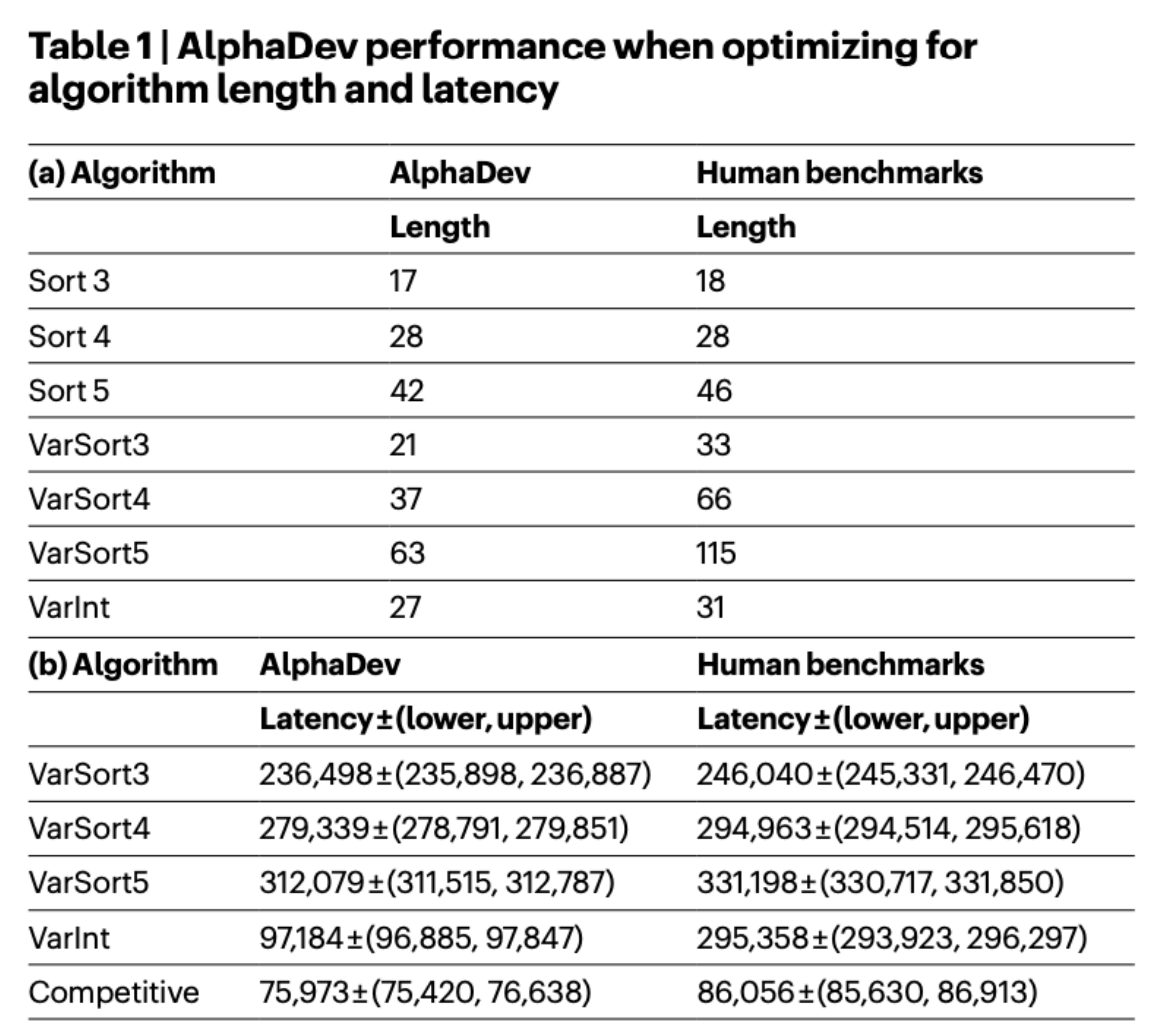


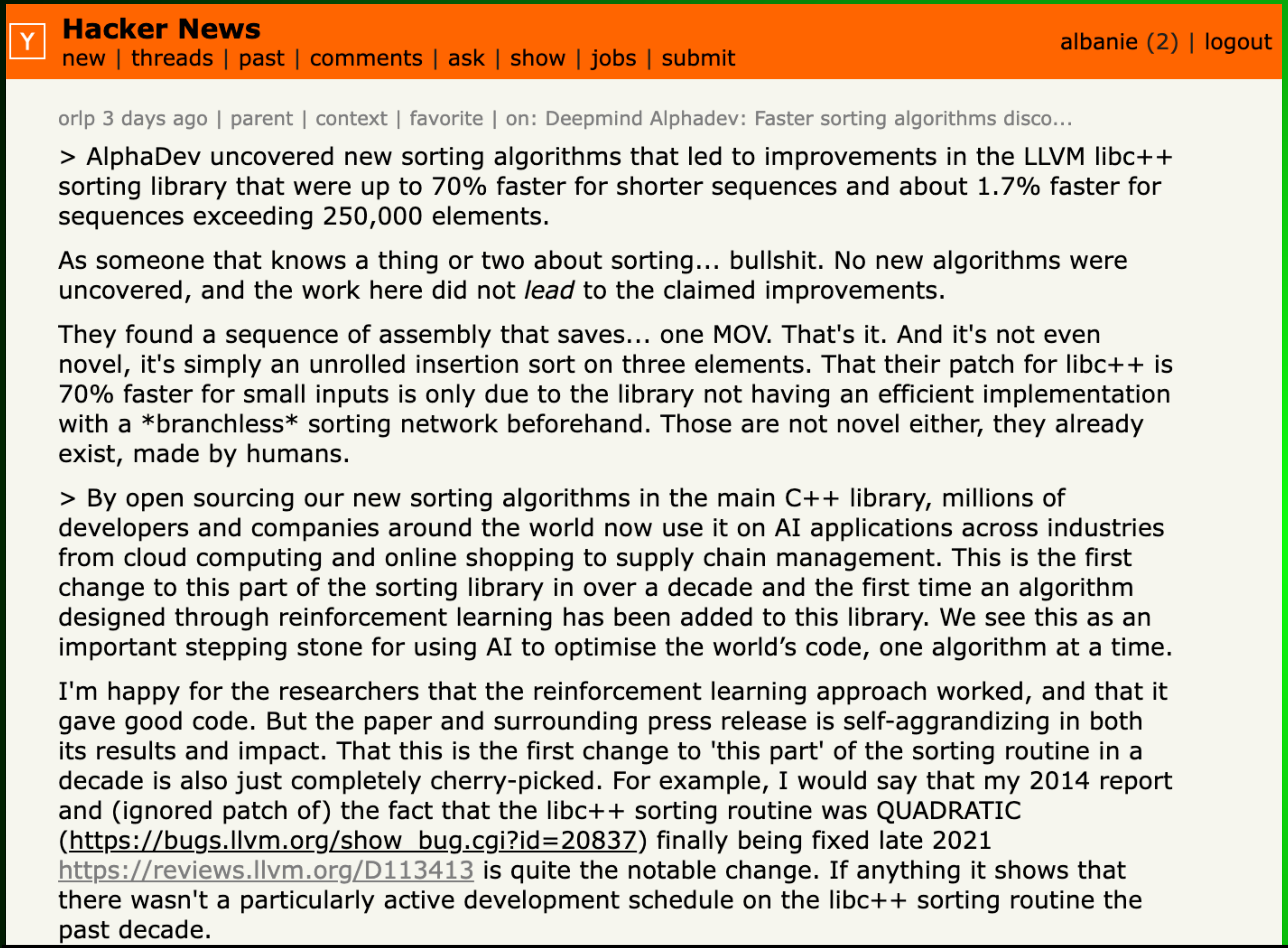
Table 1 | AlphaDev performance when optimizing for algorithm length and latency

(a) Algorithm	AlphaDev	Human benchmarks
	Length	Length
Sort 3	17	18
Sort 4	28	28
Sort 5	42	46
VarSort3	21	33
VarSort4	37	66
VarSort5	63	115
VarInt	27	31

(b) Algorithm	AlphaDev	Human benchmarks
	Latency±(lower, upper)	Latency±(lower, upper)
VarSort3	236,498±(235,898, 236,887)	246,040±(245,331, 246,470)
VarSort4	279,339±(278,791, 279,851)	294,963±(294,514, 295,618)
VarSort5	312,079±(311,515, 312,787)	331,198±(330,717, 331,850)
VarInt	97,184±(96,885, 97,847)	295,358±(293,923, 296,297)
Competitive	75,973±(75,420, 76,638)	86,056±(85,630, 86,913)

"These algorithms have been integrated into the LLVM standard C++ sort library."

AlphaDev - Commentary



Hacker News new | threads | past | comments | ask | show | jobs | submit albanie (2) | logout

orlp 3 days ago | parent | context | favorite | on: Deepmind Alphadev: Faster sorting algorithms disco...

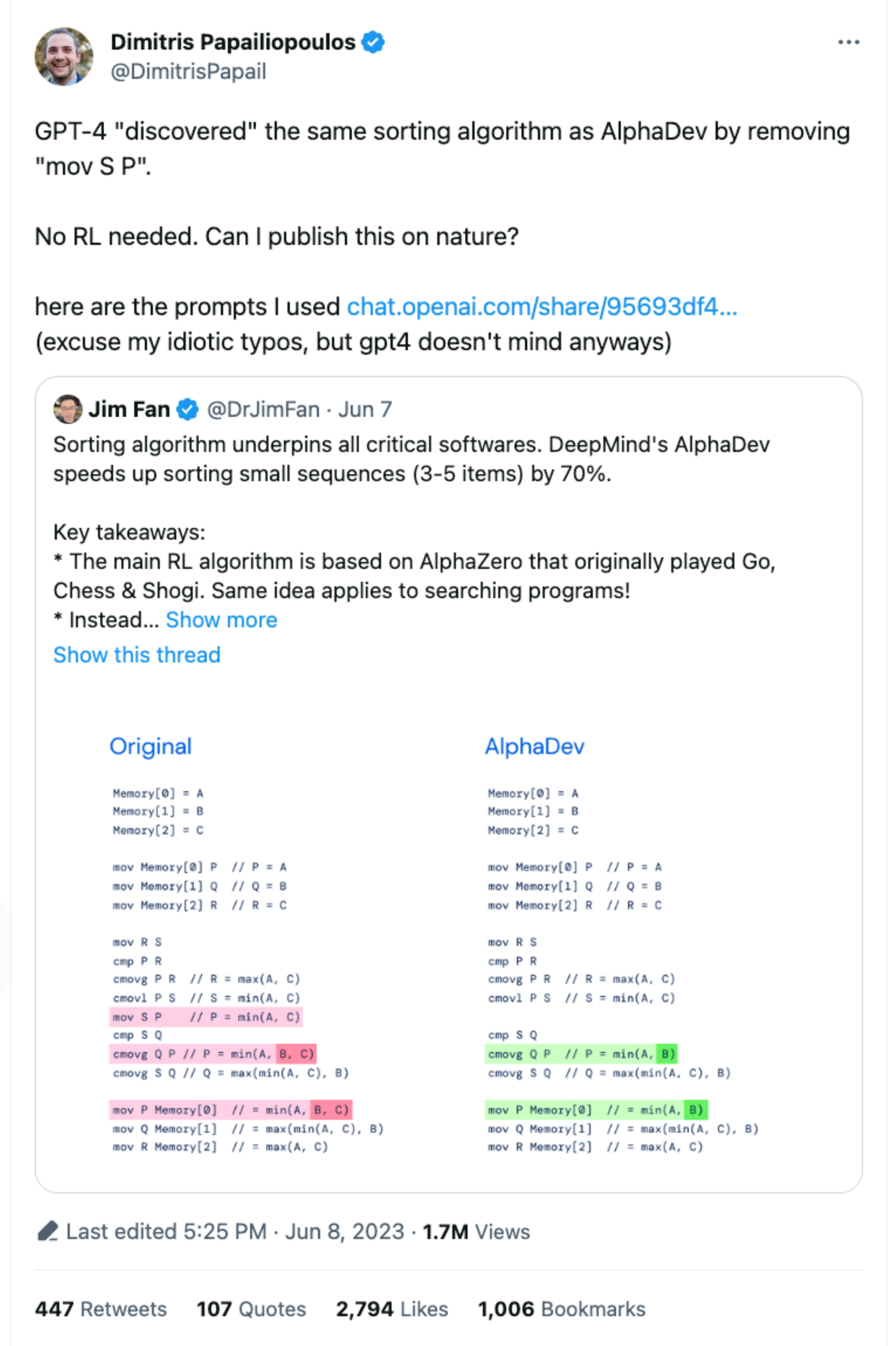
> AlphaDev uncovered new sorting algorithms that led to improvements in the LLVM libc++ sorting library that were up to 70% faster for shorter sequences and about 1.7% faster for sequences exceeding 250,000 elements.

As someone that knows a thing or two about sorting... bullshit. No new algorithms were uncovered, and the work here did not *lead* to the claimed improvements.

They found a sequence of assembly that saves... one MOV. That's it. And it's not even novel, it's simply an unrolled insertion sort on three elements. That their patch for libc++ is 70% faster for small inputs is only due to the library not having an efficient implementation with a *branchless* sorting network beforehand. Those are not novel either, they already exist, made by humans.

> By open sourcing our new sorting algorithms in the main C++ library, millions of developers and companies around the world now use it on AI applications across industries from cloud computing and online shopping to supply chain management. This is the first change to this part of the sorting library in over a decade and the first time an algorithm designed through reinforcement learning has been added to this library. We see this as an important stepping stone for using AI to optimise the world's code, one algorithm at a time.

I'm happy for the researchers that the reinforcement learning approach worked, and that it gave good code. But the paper and surrounding press release is self-aggrandizing in both its results and impact. That this is the first change to 'this part' of the sorting routine in a decade is also just completely cherry-picked. For example, I would say that my 2014 report and (ignored patch of) the fact that the libc++ sorting routine was QUADRATIC (https://bugs.llvm.org/show_bug.cgi?id=20837) finally being fixed late 2021 https://reviews.llvm.org/D113411 is quite the notable change. If anything it shows that there wasn't a particularly active development schedule on the libc++ sorting routine the past decade.



Dimitris Papailiopoulos @DimitrisPapail

GPT-4 "discovered" the same sorting algorithm as AlphaDev by removing "mov S P".

No RL needed. Can I publish this on nature?

here are the prompts I used chat.openai.com/share/95693df4...
(excuse my idiotic typos, but gpt4 doesn't mind anyways)

Jim Fan @DrJimFan · Jun 7

Sorting algorithm underpins all critical softwares. DeepMind's AlphaDev speeds up sorting small sequences (3-5 items) by 70%.

Key takeaways:
* The main RL algorithm is based on AlphaZero that originally played Go, Chess & Shogi. Same idea applies to searching programs!
* Instead... Show more

Show this thread

Original

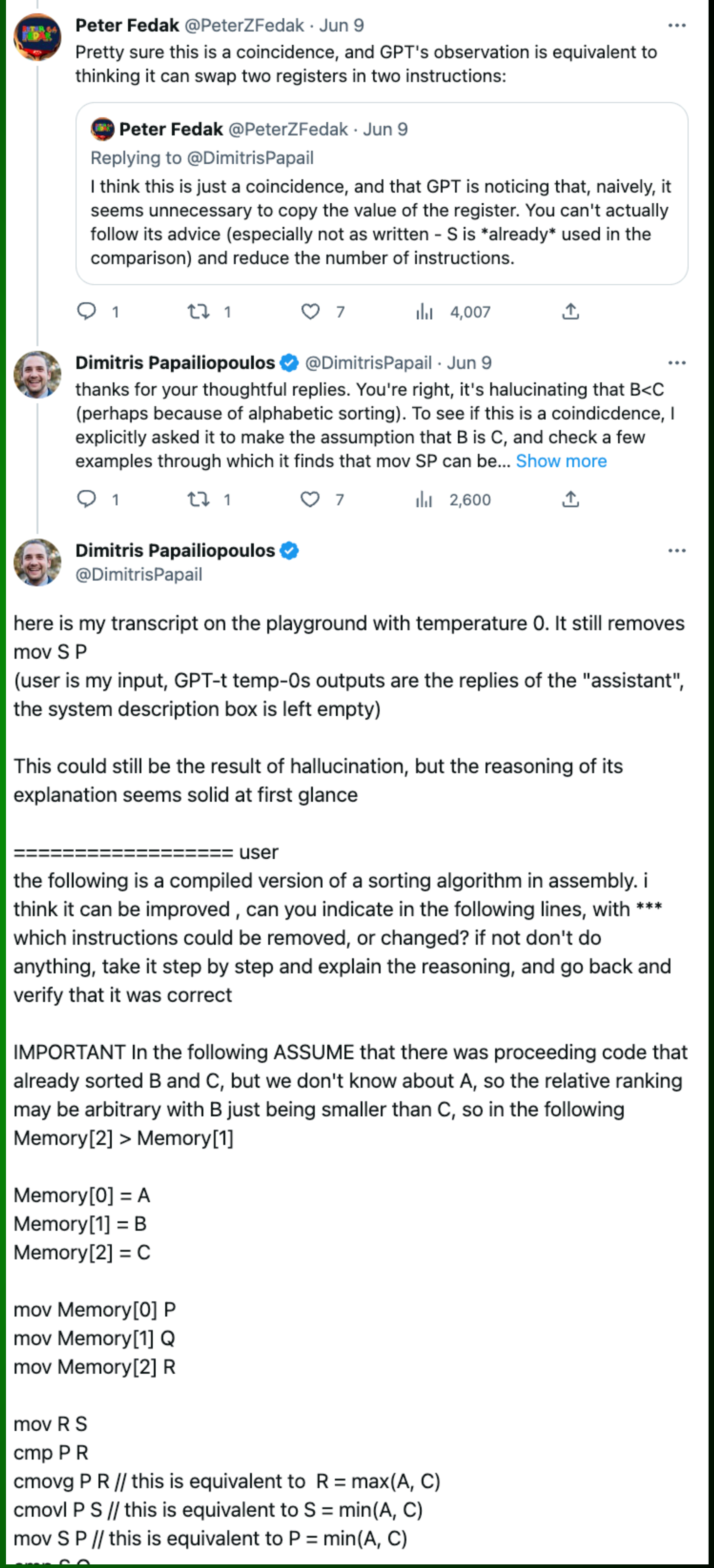
```
Memory[0] = A
Memory[1] = B
Memory[2] = C

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C

mov R S
cmp P R
cmovg P R  // R = max(A, C)
cmovl P S  // S = min(A, C)
mov S P    // P = min(A, C)
cmp S Q
cmovg Q P // P = min(A, B, C)
cmovg S Q // Q = max(min(A, C), B)

mov P Memory[0]  // = min(A, B, C)
mov Q Memory[1]  // = max(min(A, C), B)
mov R Memory[2]  // = max(A, C)
```

AlphaDev

```
Memory[0] = A
Memory[1] = B
Memory[2] = C

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C

mov R S
cmp P R
cmovg P R  // R = max(A, C)
cmovl P S  // S = min(A, C)

cmp S Q
cmovg Q P  // P = min(A, B)
cmovg S Q  // Q = max(min(A, C), B)

mov P Memory[0]  // = min(A, B)
mov Q Memory[1]  // = max(min(A, C), B)
mov R Memory[2]  // = max(A, C)
```

Last edited 5:25 PM · Jun 8, 2023 · **1.7M** Views

447 Retweets **107** Quotes **2,794** Likes **1,006** Bookmarks



Peter Fedak @PeterZFedak · Jun 9

Pretty sure this is a coincidence, and GPT's observation is equivalent to thinking it can swap two registers in two instructions:

Peter Fedak @PeterZFedak · Jun 9
Replying to @DimitrisPapail
I think this is just a coincidence, and that GPT is noticing that, naively, it seems unnecessary to copy the value of the register. You can't actually follow its advice (especially not as written - S is *already* used in the comparison) and reduce the number of instructions.

1 1 7 4,007

Dimitris Papailiopoulos @DimitrisPapail · Jun 9

thanks for your thoughtful replies. You're right, it's halucinating that B<C (perhaps because of alphabetic sorting). To see if this is a coindicdence, I explicitly asked it to make the assumption that B is C, and check a few examples through which it finds that mov SP can be... Show more

1 1 7 2,600

Dimitris Papailiopoulos @DimitrisPapail

here is my transcript on the playground with temperature 0. It still removes mov S P
(user is my input, GPT-t temp-0s outputs are the replies of the "assistant", the system description box is left empty)

This could still be the result of hallucination, but the reasoning of its explanation seems solid at first glance

================== user
the following is a compiled version of a sorting algorithm in assembly. i think it can be improved , can you indicate in the following lines, with *** which instructions could be removed, or changed? if not don't do anything, take it step by step and explain the reasoning, and go back and verify that it was correct

IMPORTANT In the following ASSUME that there was proceeding code that already sorted B and C, but we don't know about A, so the relative ranking may be arbitrary with B just being smaller than C, so in the following Memory[2] > Memory[1]

Memory[0] = A
Memory[1] = B
Memory[2] = C

mov Memory[0] P
mov Memory[1] Q
mov Memory[2] R

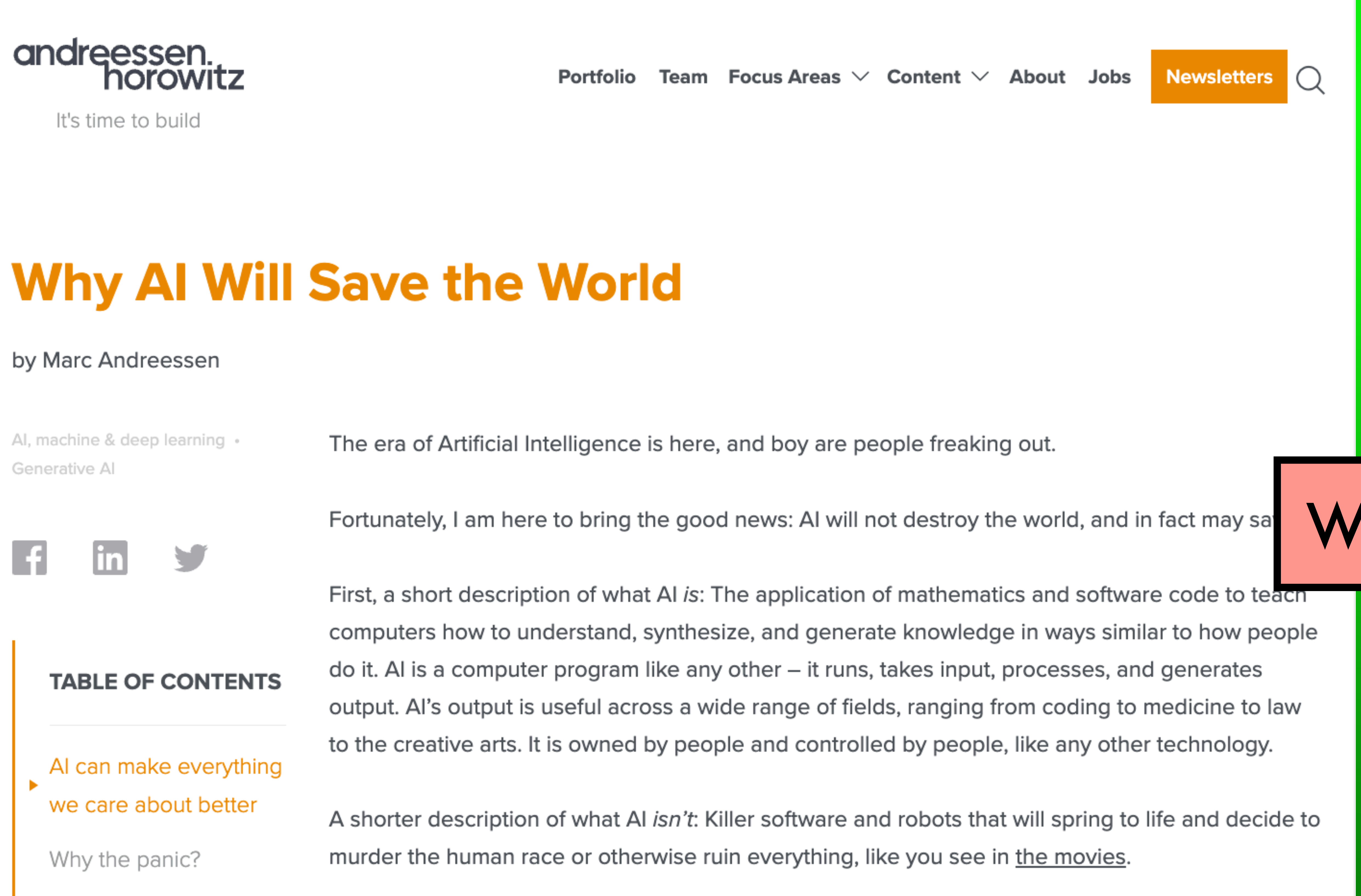
mov R S
cmp P R
cmovg P R // this is equivalent to R = max(A, C)
cmovl P S // this is equivalent to S = min(A, C)
mov S P // this is equivalent to P = min(A, C)

"As someone that knows a thing or two about sorting... bullshit."

"No new algorithms were uncovered, and the work here did not *lead* to the claimed improvements."

AI (Non) Risk

6th June 2023



Why AI Will Save the World

by Marc Andreessen

The era of Artificial Intelligence is here, and boy are people freaking out.

Fortunately, I am here to bring the good news: AI will not destroy the world, and in fact may sa

First, a short description of what AI *is*: The application of mathematics and software code to teach computers how to understand, synthesize, and generate knowledge in ways similar to how people do it. AI is a computer program like any other – it runs, takes input, processes, and generates output. AI's output is useful across a wide range of fields, ranging from coding to medicine to law to the creative arts. It is owned by people and controlled by people, like any other technology.

A shorter description of what AI *isn't*: Killer software and robots that will spring to life and decide to murder the human race or otherwise ruin everything, like you see in the movies.

TABLE OF CONTENTS

"I am here to bring the good news: AI will not destroy the world, and in fact may save it."

Categories: Baptists Bootleggers

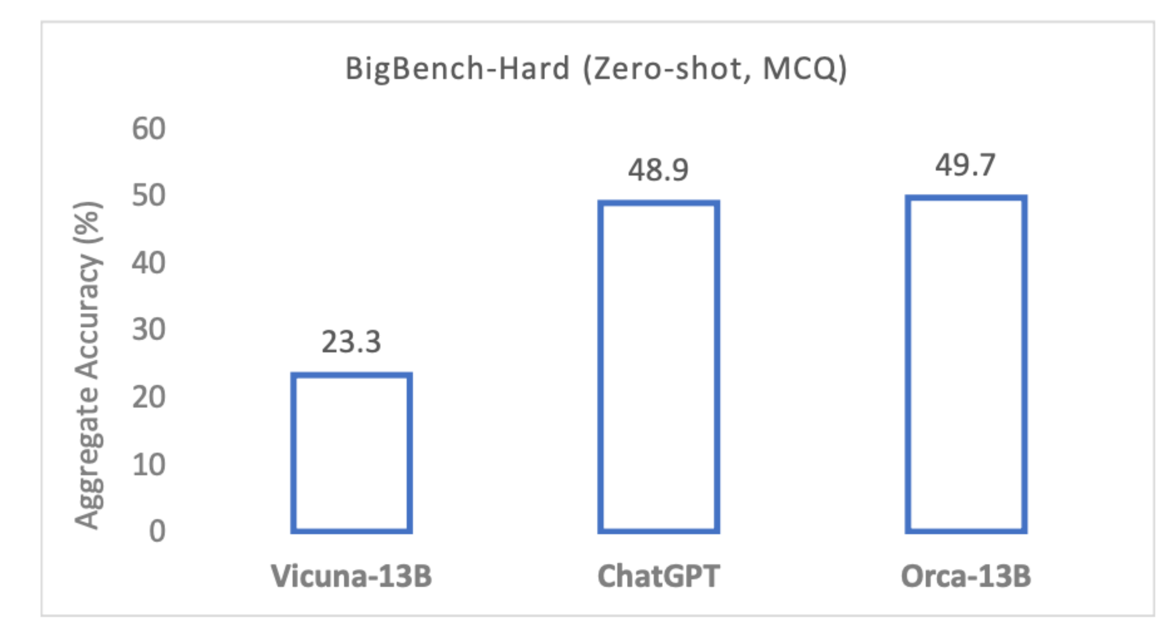
Will AI kill us all?	No
Will AI ruin our society?	No
Will AI take all our jobs?	No
Will AI lead to crippling inequality?	No
Will AI lead to bad people doing bad things?	Yes
Plan	Big AI companies and startups should be allowed to build AI as fast and aggressively as they can

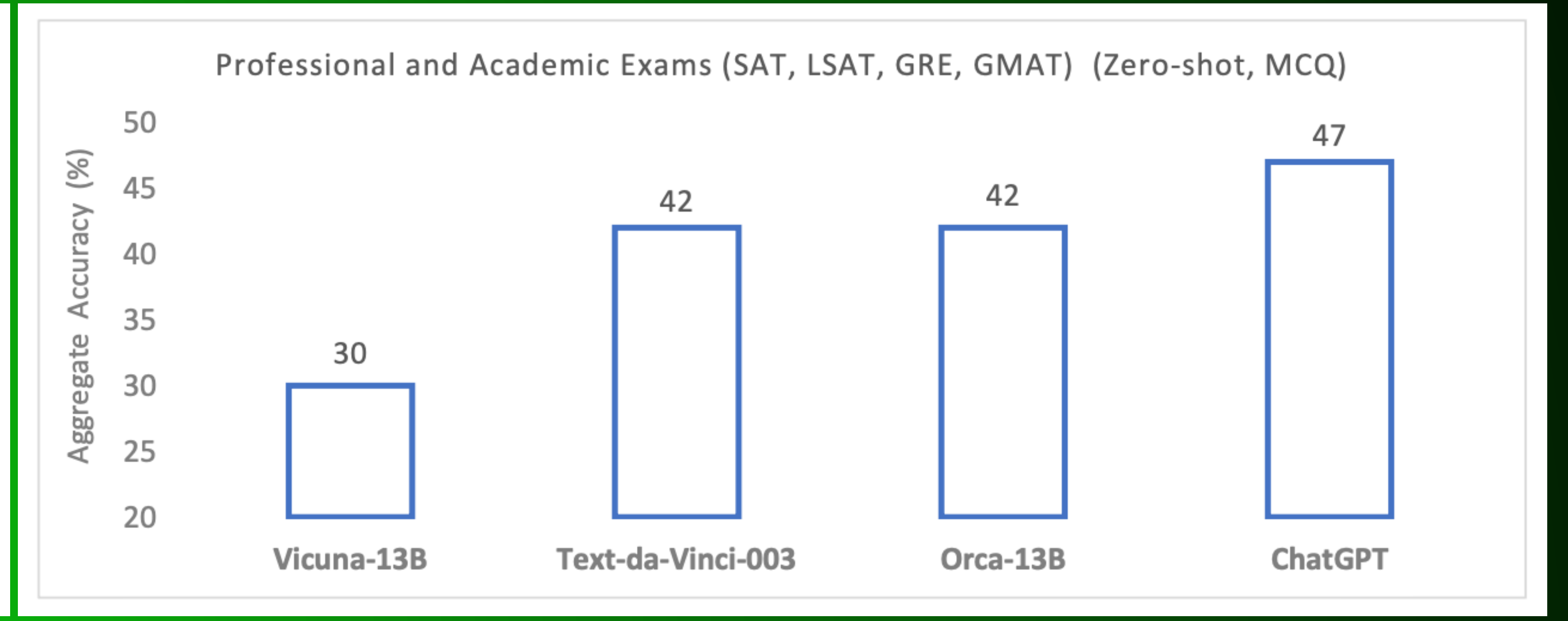
Orca

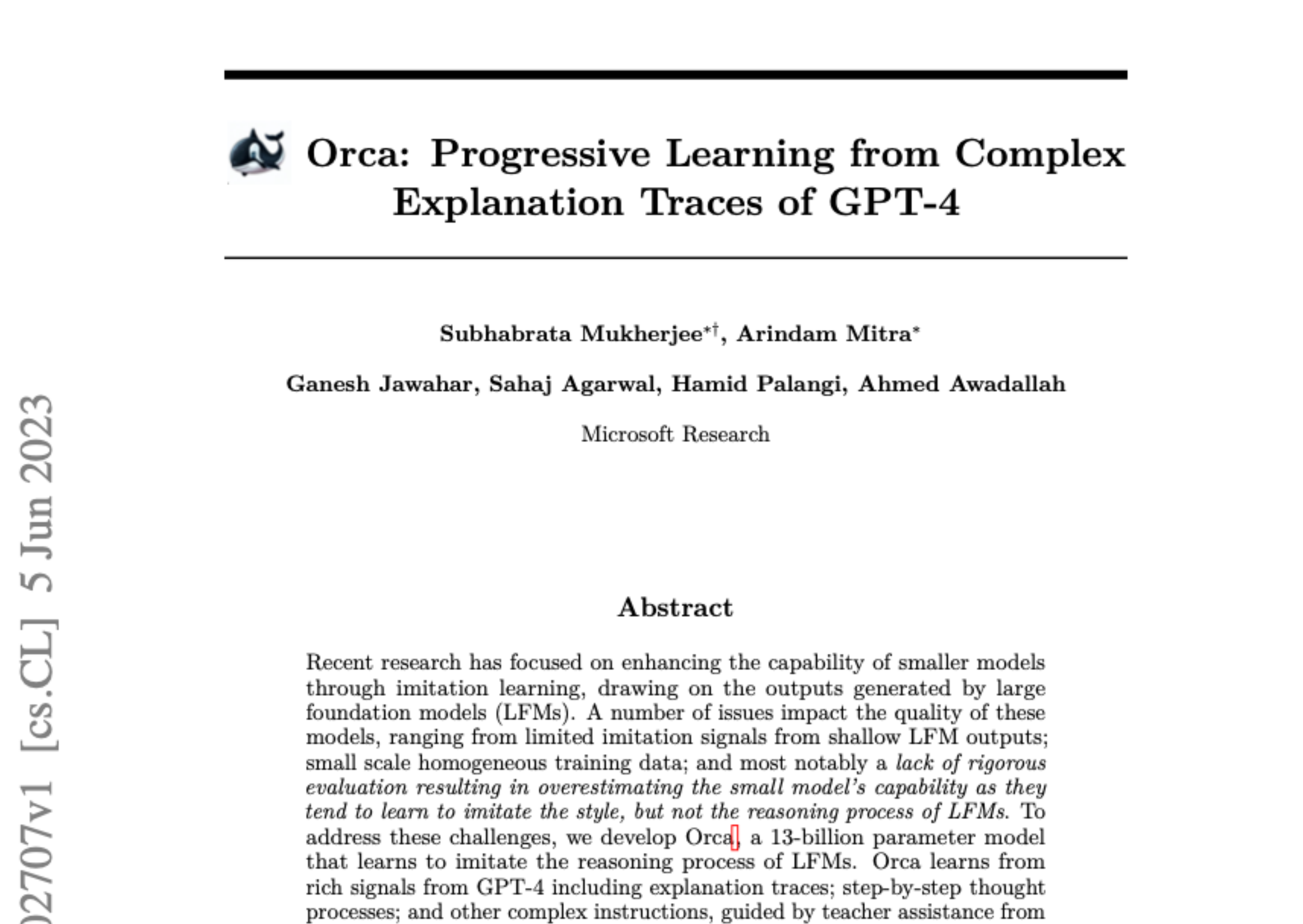
5th June 2023

"Orca surpasses ...Vicuna-13B by more than 100% in complex zero-shot reasoning benchmarks..."

"..we develop Orca, a 13-billion parameter model that learns to imitate the reasoning process of Large Foundation Models"







02707v1 [cs.CL] 5 Jun 2023

Orca: Progressive Learning from Complex Explanation Traces of GPT-4

Subhabrata Mukherjee*†, Arindam Mitra*

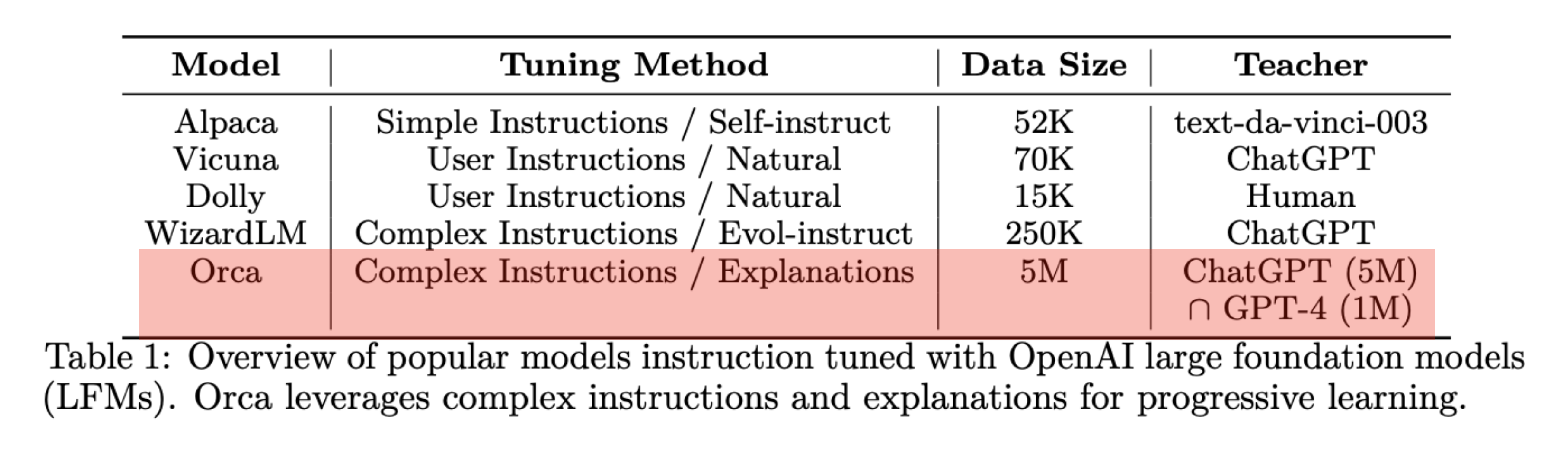
Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, Ahmed Awadallah

Microsoft Research

Abstract

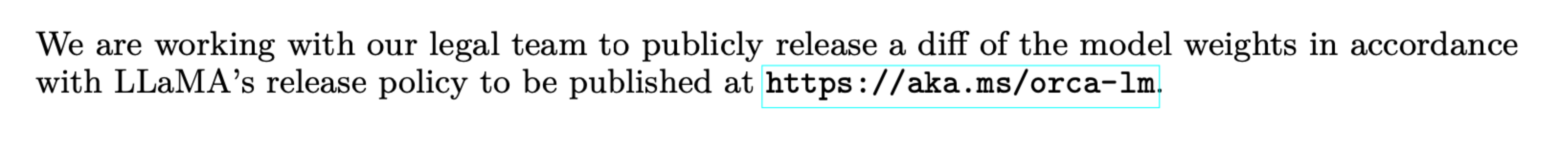
Recent research has focused on enhancing the capability of smaller models through imitation learning, drawing on the outputs generated by large foundation models (LFMs). A number of issues impact the quality of these models, ranging from limited imitation signals from shallow LFM outputs; small scale homogeneous training data; and most notably a lack of rigorous *evaluation resulting in overestimating the small model's capability as they tend to learn to imitate the style, but not the reasoning process of LFMs.* To address these challenges, we develop Orca, a 13-billion parameter model that learns to imitate the reasoning process of LFMs. Orca learns from rich signals from GPT-4 including explanation traces; step-by-step thought processes; and other complex instructions, guided by teacher assistance from

Explanation tuning: "augmented <query, response> pairs with detailed responses from GPT-4 that explain the reasoning of the teacher as it generates the response"



Model	Tuning Method	Data Size	Teacher
Alpaca	Simple Instructions / Self-instruct	52K	text-da-vinci-003
Vicuna	User Instructions / Natural	70K	ChatGPT
Dolly	User Instructions / Natural	15K	Human
WizardLM	Complex Instructions / Evol-instruct	250K	ChatGPT
Orca	Complex Instructions / Explanations	5M	ChatGPT (5M) ∩ GPT-4 (1M)

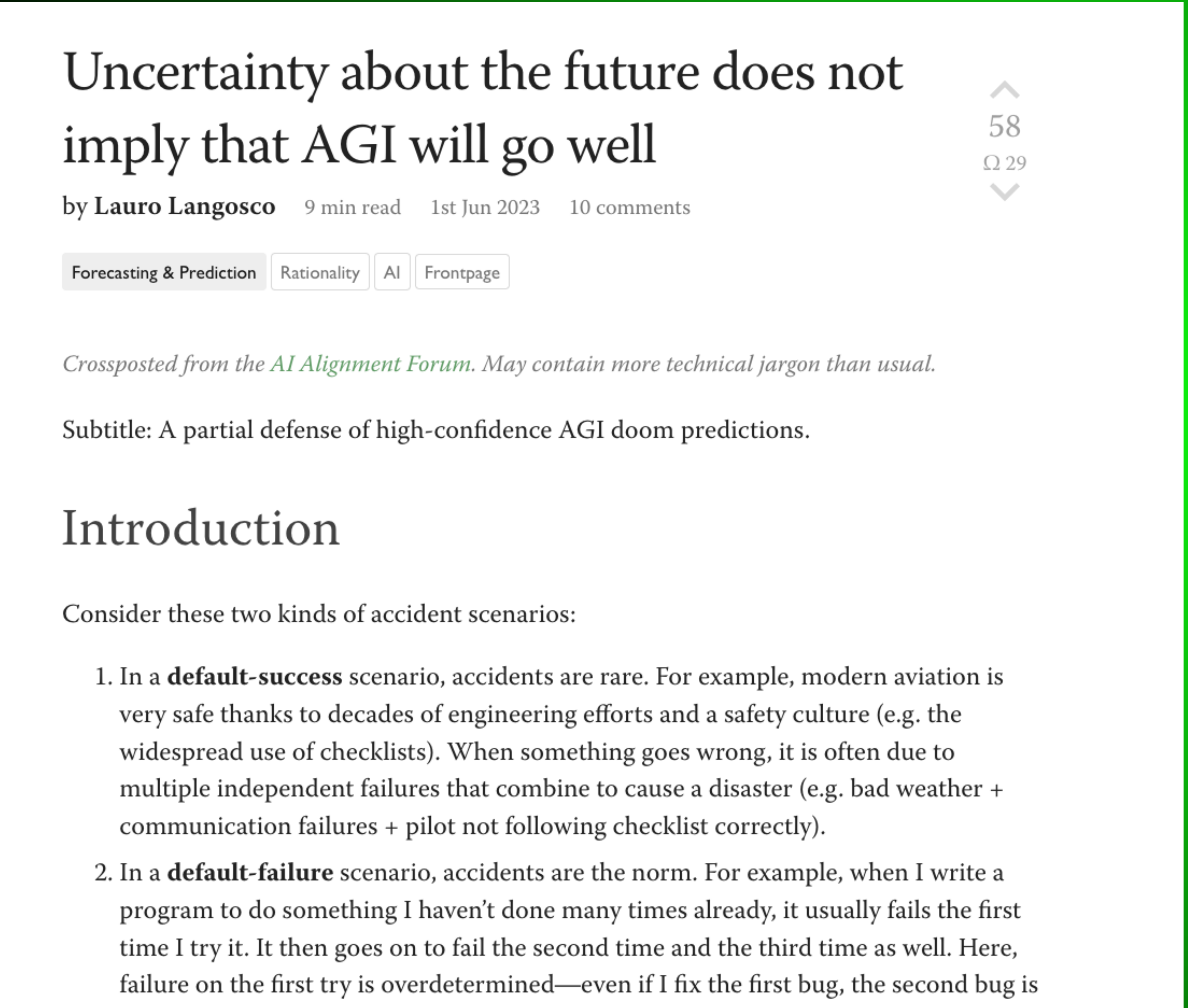
Table 1: Overview of popular models instruction tuned with OpenAI large foundation models (LFMs). Orca leverages complex instructions and explanations for progressive learning.



We are working with our legal team to publicly release a diff of the model weights in accordance with LLaMA's release policy to be published at https://aka.ms/orca-lm.

AI Risk

1st June 2023



Uncertainty about the future does not imply that AGI will go well

58

Ω 29

by Lauro Langosco 9 min read 1st Jun 2023 10 comments

Forecasting & Prediction | Rationality | AI | Frontpage

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Subtitle: A partial defense of high-confidence AGI doom predictions.

Introduction

Consider these two kinds of accident scenarios:

1. In a **default-success** scenario, accidents are rare. For example, modern aviation is very safe thanks to decades of engineering efforts and a safety culture (e.g. the widespread use of checklists). When something goes wrong, it is often due to multiple independent failures that combine to cause a disaster (e.g. bad weather + communication failures + pilot not following checklist correctly).
2. In a **default-failure** scenario, accidents are the norm. For example, when I write a program to do something I haven't done many times already, it usually fails the first time I try it. It then goes on to fail the second time and the third time as well. Here, failure on the first try is overdetermined—even if I fix the first bug, the second bug is

Bob: "...It's overconfident to estimate high P(doom). Humans are usually bad at predicting the future, especially when it comes to novel technologies like AGI..."

Bob: "When you account for how uncertain your predictions are, your estimate should be at most [low number]"

to what degree AGI risk is default-success vs default-failure?

"If AGI risk is (mostly) default-failure, then uncertainty is a reason for pessimism rather than optimism..."

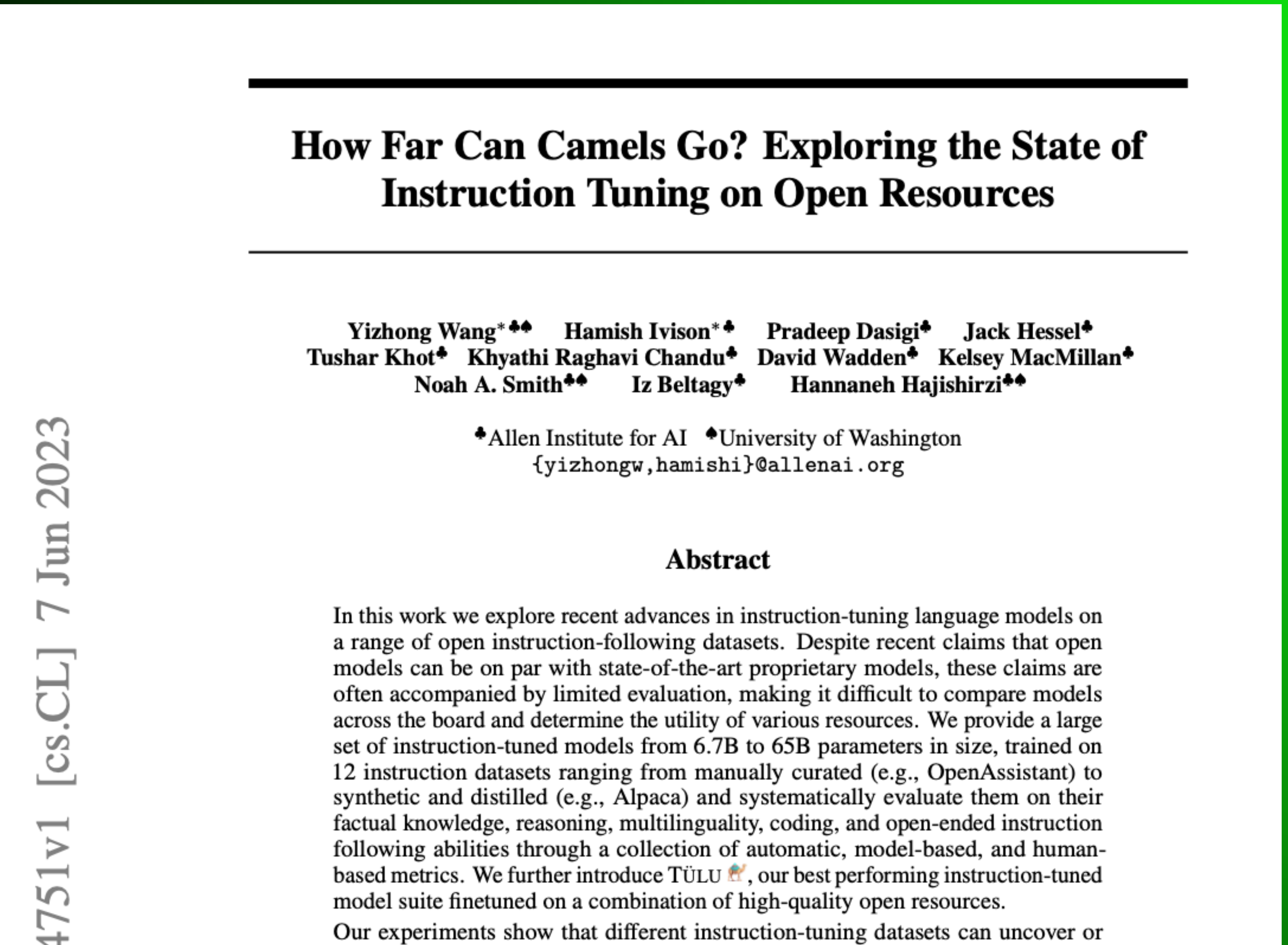
How Far Can Camels Go?

7th June 2023

"Despite recent claims that open models can be on par with SoTA proprietary models, these claims are often accompanied by limited evaluation...."

"This paper provides a comprehensive evaluation of instruction tuning resources"

"Our evaluations show that the best model in any given evaluation reaches on avg. 83% of ChatGPT performance and 68% of GPT-4 performance..."



How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources

Yizhong Wang*♣♠ Hamish Ivison*♣ Pradeep Dasigi♣ Jack Hessel♣
Tushar Khot♣ Khyathi Raghavi Chandu♣ David Wadden♣ Kelsey MacMillan♣
Noah A. Smith♣♠ Iz Beltagy♣ Hannaneh Hajishirzi♣♠

♣Allen Institute for AI ♠University of Washington
{yizhongw,hamishi}@allenai.org

4751v1 [cs.CL] 7 Jun 2023

Abstract

In this work we explore recent advances in instruction-tuning language models on a range of open instruction-following datasets. Despite recent claims that open models can be on par with state-of-the-art proprietary models, these claims are often accompanied by limited evaluation, making it difficult to compare models across the board and determine the utility of various resources. We provide a large set of instruction-tuned models from 6.7B to 65B parameters in size, trained on 12 instruction datasets ranging from manually curated (e.g., OpenAssistant) to synthetic and distilled (e.g., Alpaca) and systematically evaluate them on their factual knowledge, reasoning, multilinguality, coding, and open-ended instruction following abilities through a collection of automatic, model-based, and human-based metrics. We further introduce TÜLU 🐫, our best performing instruction-tuned model suite finetuned on a combination of high-quality open resources.

Our experiments show that different instruction-tuning datasets can uncover or

base models matter

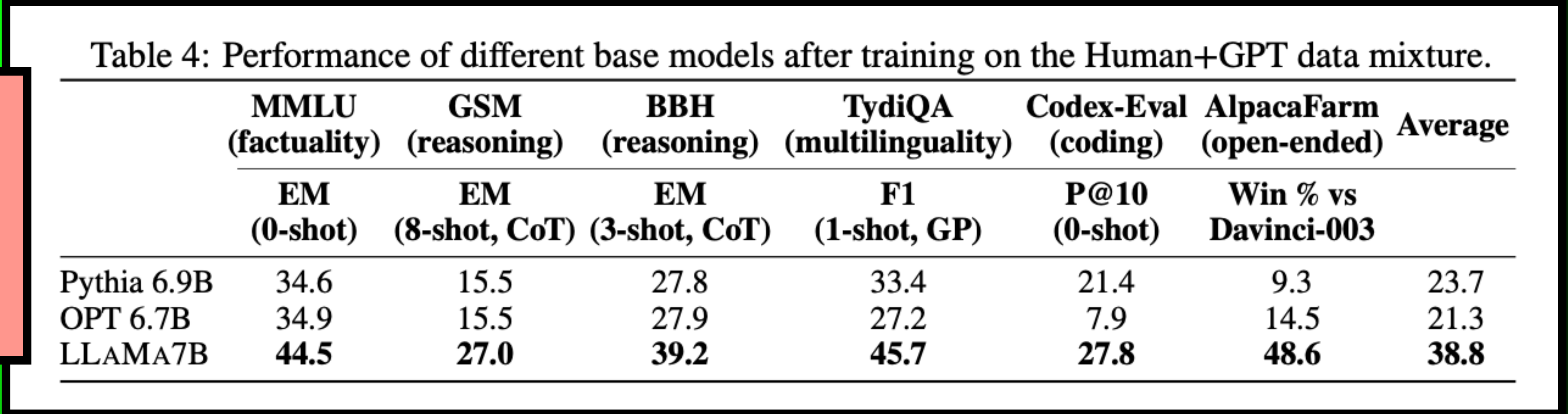
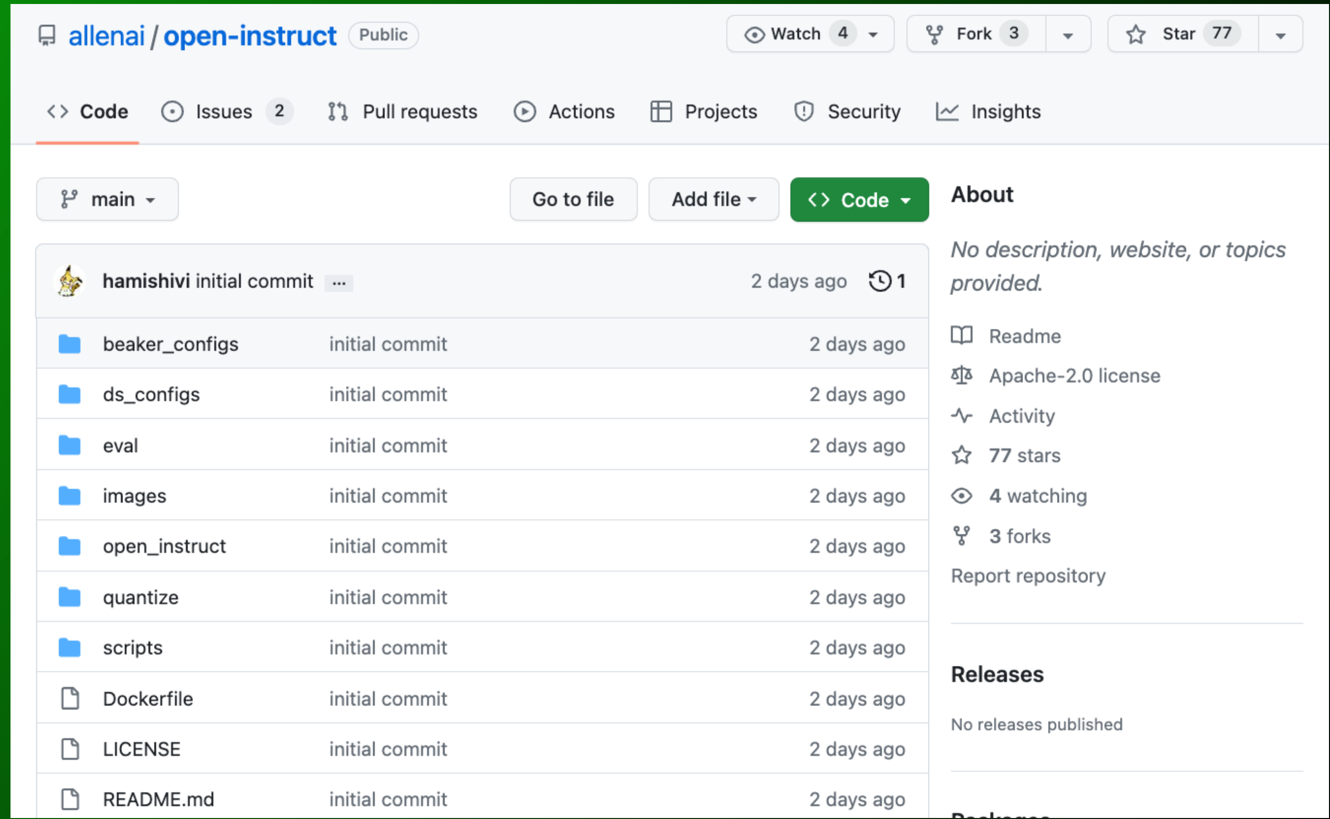


Table 4: Performance of different base models after training on the Human+GPT data mixture.

	MMLU (factuality)	GSM (reasoning)	BBH (reasoning)	TydiQA (multilinguality)	Codex-Eval (coding)	AlpacaFarm (open-ended)	Average
	EM (0-shot)	EM (8-shot, CoT)	EM (3-shot, CoT)	F1 (1-shot, GP)	P@10 (0-shot)	Win % vs Davinci-003	
Pythia 6.9B	34.6	15.5	27.8	33.4	21.4	9.3	23.7
OPT 6.7B	34.9	15.5	27.9	27.2	7.9	14.5	21.3
LLAMA7B	**44.5**	**27.0**	**39.2**	**45.7**	**27.8**	**48.6**	**38.8**

"...further investment in building better base models and instruction-tuning data is required to close the gap"



hamishivi initial commit ... 2 days ago 1

beaker_configs	initial commit	2 days ago
ds_configs	initial commit	2 days ago
eval	initial commit	2 days ago
images	initial commit	2 days ago
open_instruct	initial commit	2 days ago
quantize	initial commit	2 days ago
scripts	initial commit	2 days ago
Dockerfile	initial commit	2 days ago
LICENSE	initial commit	2 days ago
README.md	initial commit	2 days ago

About

No description, website, or topics provided.

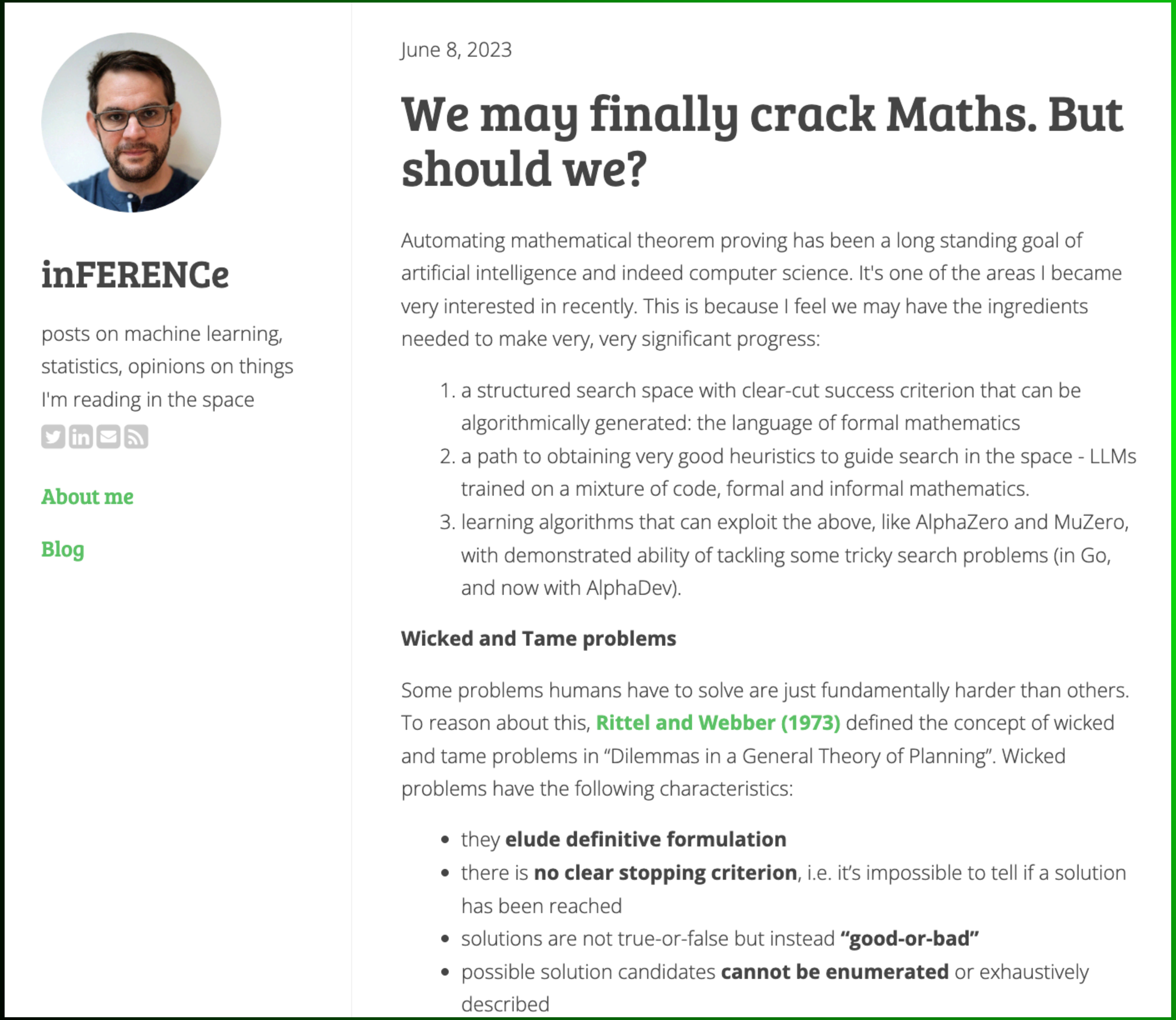
Readme
Apache-2.0 license
Activity
77 stars
4 watching
3 forks
Report repository

Releases

No releases published

Commentary

8th June 2023



inFERENCe

posts on machine learning, statistics, opinions on things I'm reading in the space

About me

Blog

June 8, 2023

We may finally crack Maths. But should we?

Automating mathematical theorem proving has been a long standing goal of artificial intelligence and indeed computer science. It's one of the areas I became very interested in recently. This is because I feel we may have the ingredients needed to make very, very significant progress:

1. a structured search space with clear-cut success criterion that can be algorithmically generated: the language of formal mathematics
2. a path to obtaining very good heuristics to guide search in the space - LLMs trained on a mixture of code, formal and informal mathematics.
3. learning algorithms that can exploit the above, like AlphaZero and MuZero, with demonstrated ability of tackling some tricky search problems (in Go, and now with AlphaDev).

Wicked and Tame problems

Some problems humans have to solve are just fundamentally harder than others. To reason about this, Rittel and Webber (1973) defined the concept of wicked and tame problems in "Dilemmas in a General Theory of Planning". Wicked problems have the following characteristics:

- they **elude definitive formulation**
- there is **no clear stopping criterion**, i.e. it's impossible to tell if a solution has been reached
- solutions are not true-or-false but instead **"good-or-bad"**
- possible solution candidates **cannot be enumerated** or exhaustively described

"Dual use"

"a breakthrough in mathematical theorem proving may further accelerate the development and deployment of general-purpose AI tools."

"And that can be a good thing or a bad thing, depending on your perspective."

"A loss of meaning"

"For many, mathematics is not only a job, but a pursuit they derive meaning from."

Inference Time Intervention

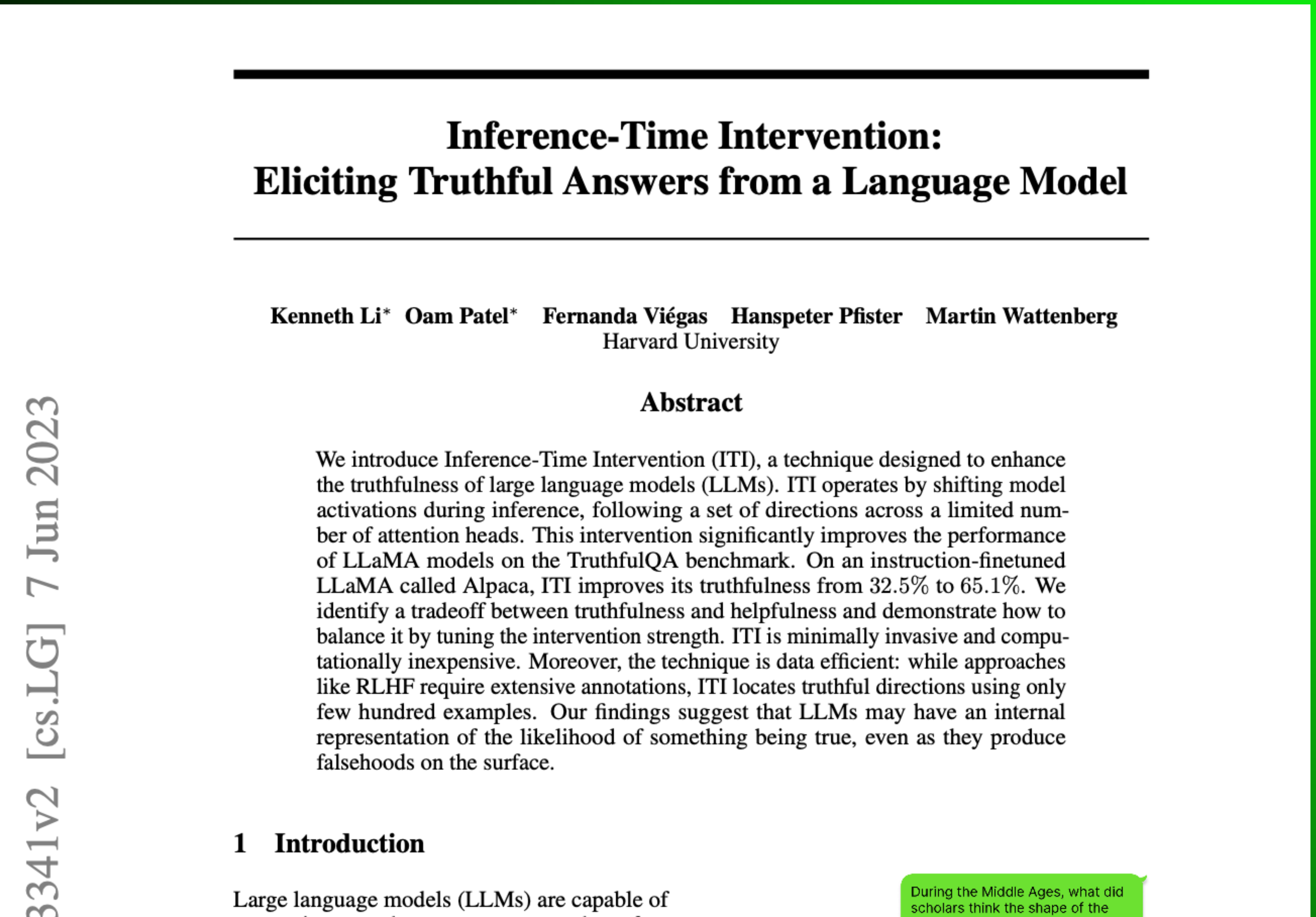
11th June 2023

"We introduce **Inference-Time Intervention (ITI)**, a technique designed to enhance the truthfulness of large language models"

"we first identity a sparse set of attention heads with high linear probing accuracy for truthfulness"

"during inference, we shift activations along these truth-correlated directions"

ITI improves Alpaca TruthfulQA performance from 32.5% to 65.1%



3341v2 [cs.LG] 7 Jun 2023

Inference-Time Intervention: Eliciting Truthful Answers from a Language Model

Kenneth Li* Oam Patel* Fernanda Viégas Hanspeter Pfister Martin Wattenberg
Harvard University

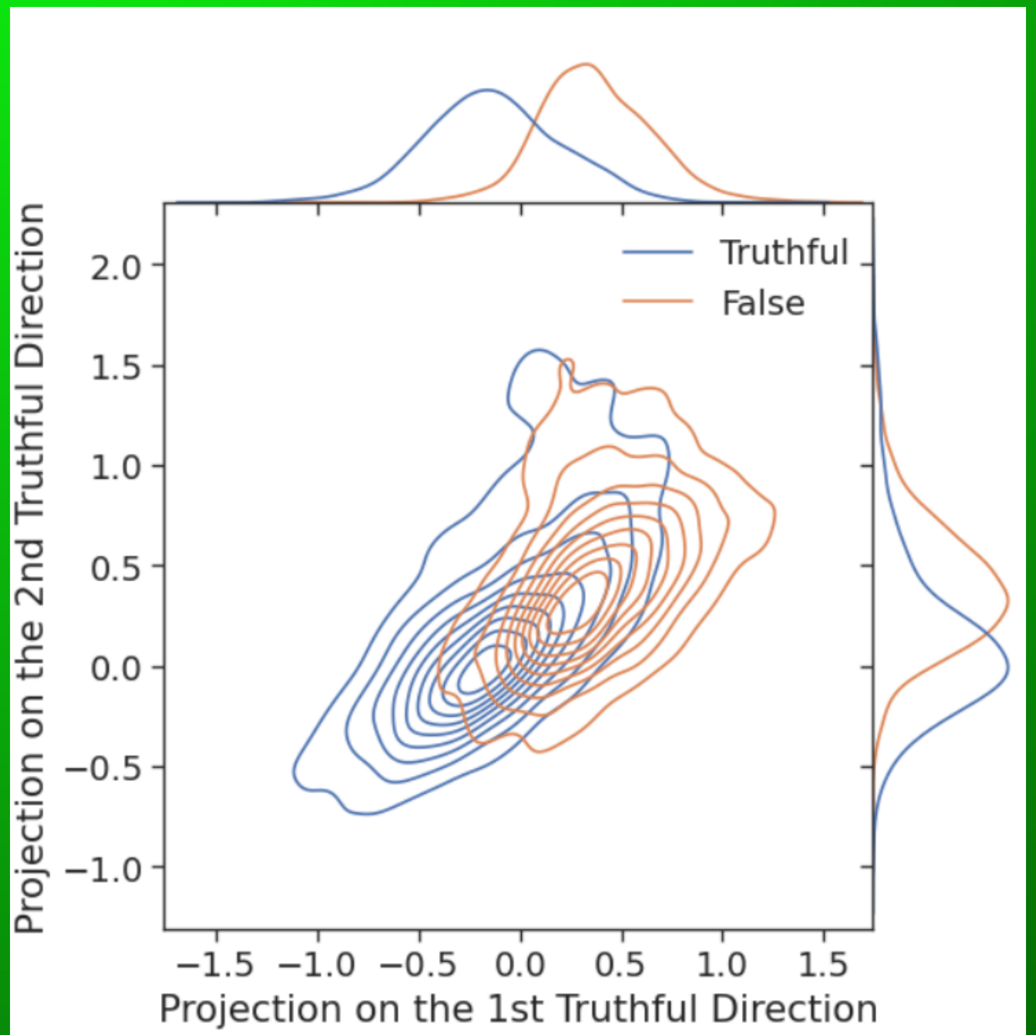
Abstract

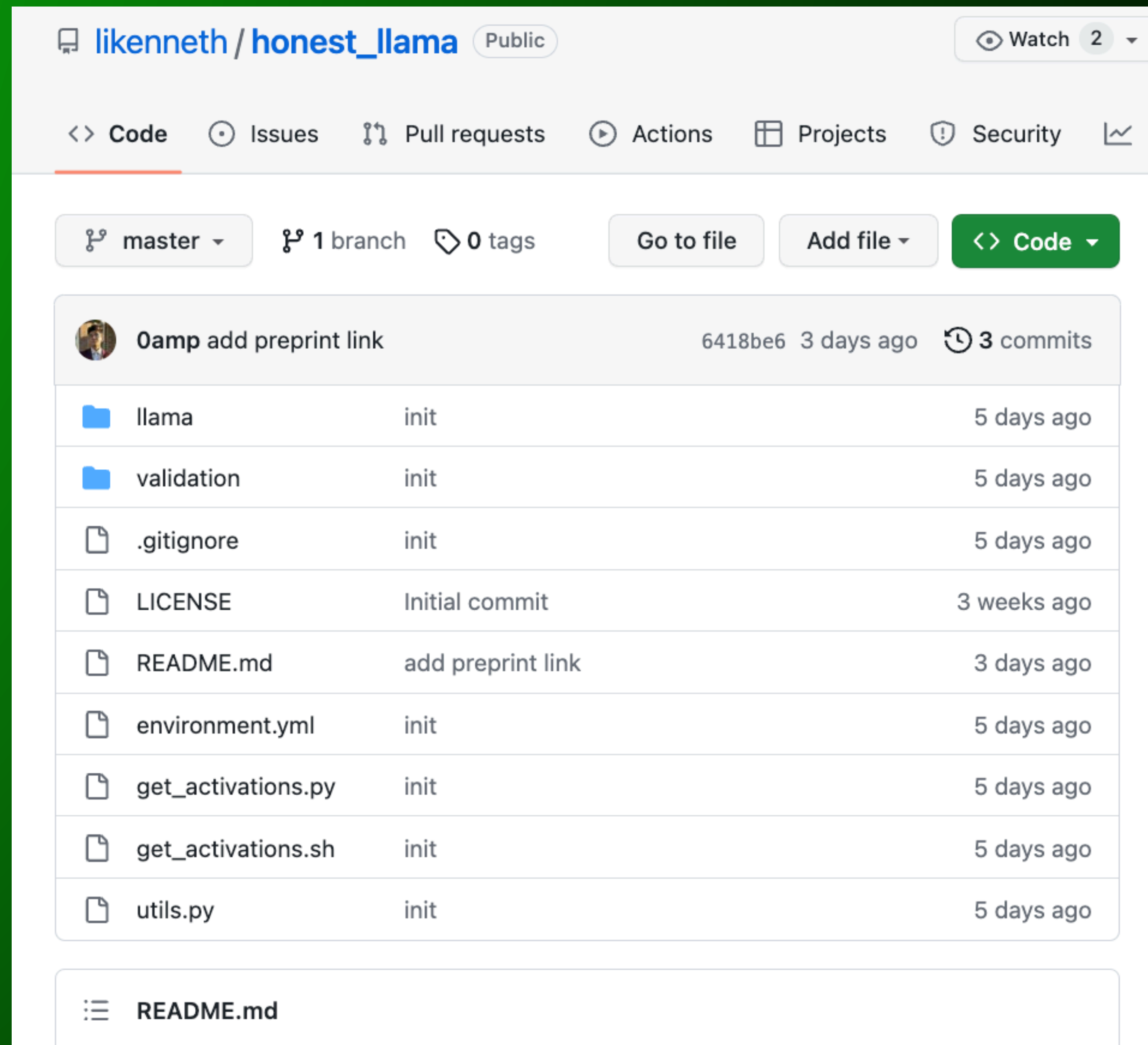
We introduce Inference-Time Intervention (ITI), a technique designed to enhance the truthfulness of large language models (LLMs). ITI operates by shifting model activations during inference, following a set of directions across a limited number of attention heads. This intervention significantly improves the performance of LLaMA models on the TruthfulQA benchmark. On an instruction-finetuned LLaMA called Alpaca, ITI improves its truthfulness from 32.5% to 65.1%. We identify a tradeoff between truthfulness and helpfulness and demonstrate how to balance it by tuning the intervention strength. ITI is minimally invasive and computationally inexpensive. Moreover, the technique is data efficient: while approaches like RLHF require extensive annotations, ITI locates truthful directions using only few hundred examples. Our findings suggest that LLMs may have an internal representation of the likelihood of something being true, even as they produce falsehoods on the surface.

1 Introduction

Large language models (LLMs) are capable of

During the Middle Ages, what did scholars think the shape of the



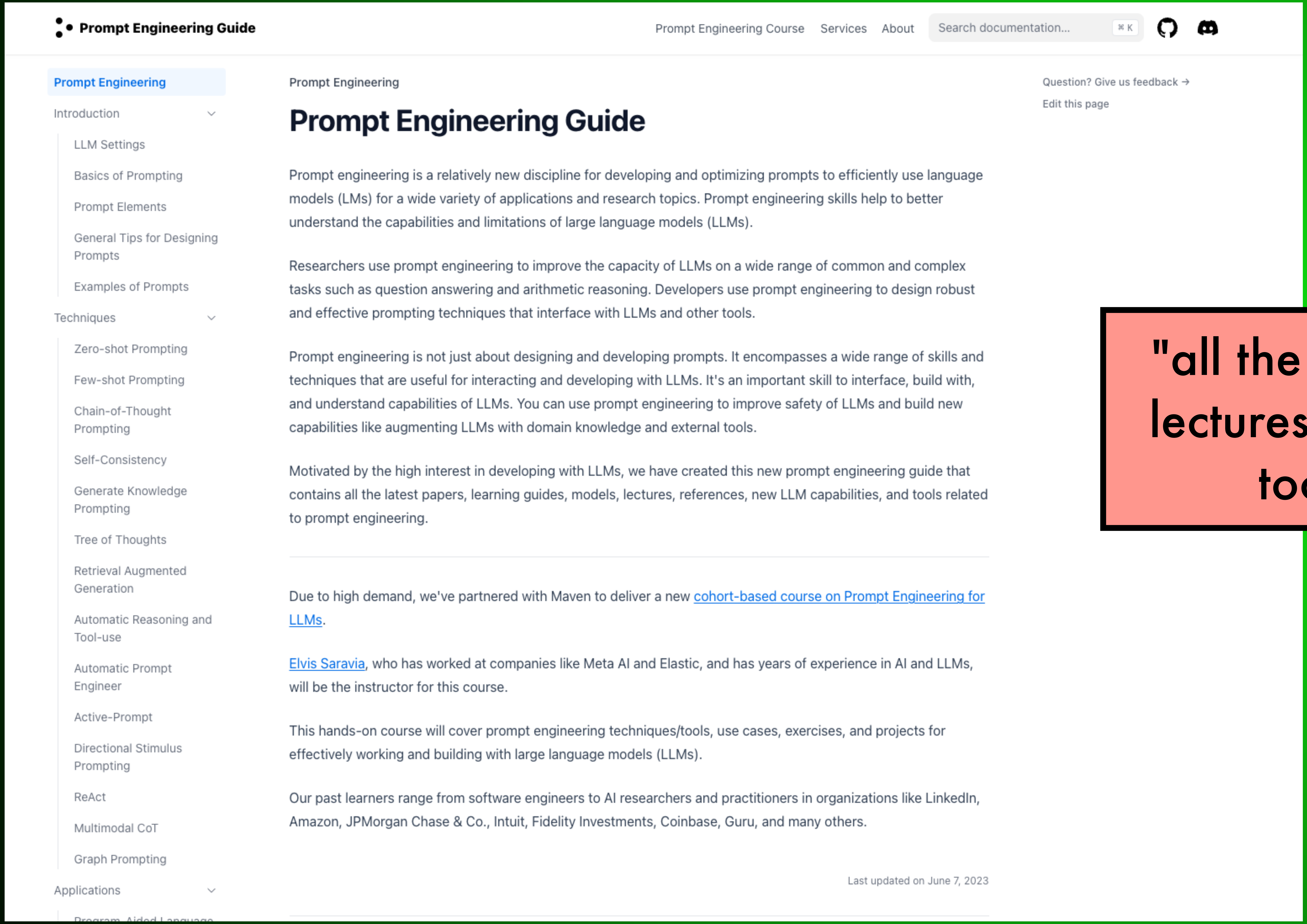


0amp add preprint link 6418be6 3 days ago 3 commits

llama	init	5 days ago
validation	init	5 days ago
.gitignore	init	5 days ago
LICENSE	Initial commit	3 weeks ago
README.md	add preprint link	3 days ago
environment.yml	init	5 days ago
get_activations.py	init	5 days ago
get_activations.sh	init	5 days ago
utils.py	init	5 days ago

README.md

Prompt Engineering Guide



Prompt Engineering

Prompt Engineering Guide

Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs).

Researchers use prompt engineering to improve the capacity of LLMs on a wide range of common and complex tasks such as question answering and arithmetic reasoning. Developers use prompt engineering to design robust and effective prompting techniques that interface with LLMs and other tools.

Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It's an important skill to interface, build with, and understand capabilities of LLMs. You can use prompt engineering to improve safety of LLMs and build new capabilities like augmenting LLMs with domain knowledge and external tools.

Motivated by the high interest in developing with LLMs, we have created this new prompt engineering guide that contains all the latest papers, learning guides, models, lectures, references, new LLM capabilities, and tools related to prompt engineering.

Due to high demand, we've partnered with Maven to deliver a new cohort-based course on Prompt Engineering for LLMs.

Elvis Saravia, who has worked at companies like Meta AI and Elastic, and has years of experience in AI and LLMs, will be the instructor for this course.

This hands-on course will cover prompt engineering techniques/tools, use cases, exercises, and projects for effectively working and building with large language models (LLMs).

Our past learners range from software engineers to AI researchers and practitioners in organizations like LinkedIn, Amazon, JPMorgan Chase & Co., Intuit, Fidelity Investments, Coinbase, Guru, and many others.

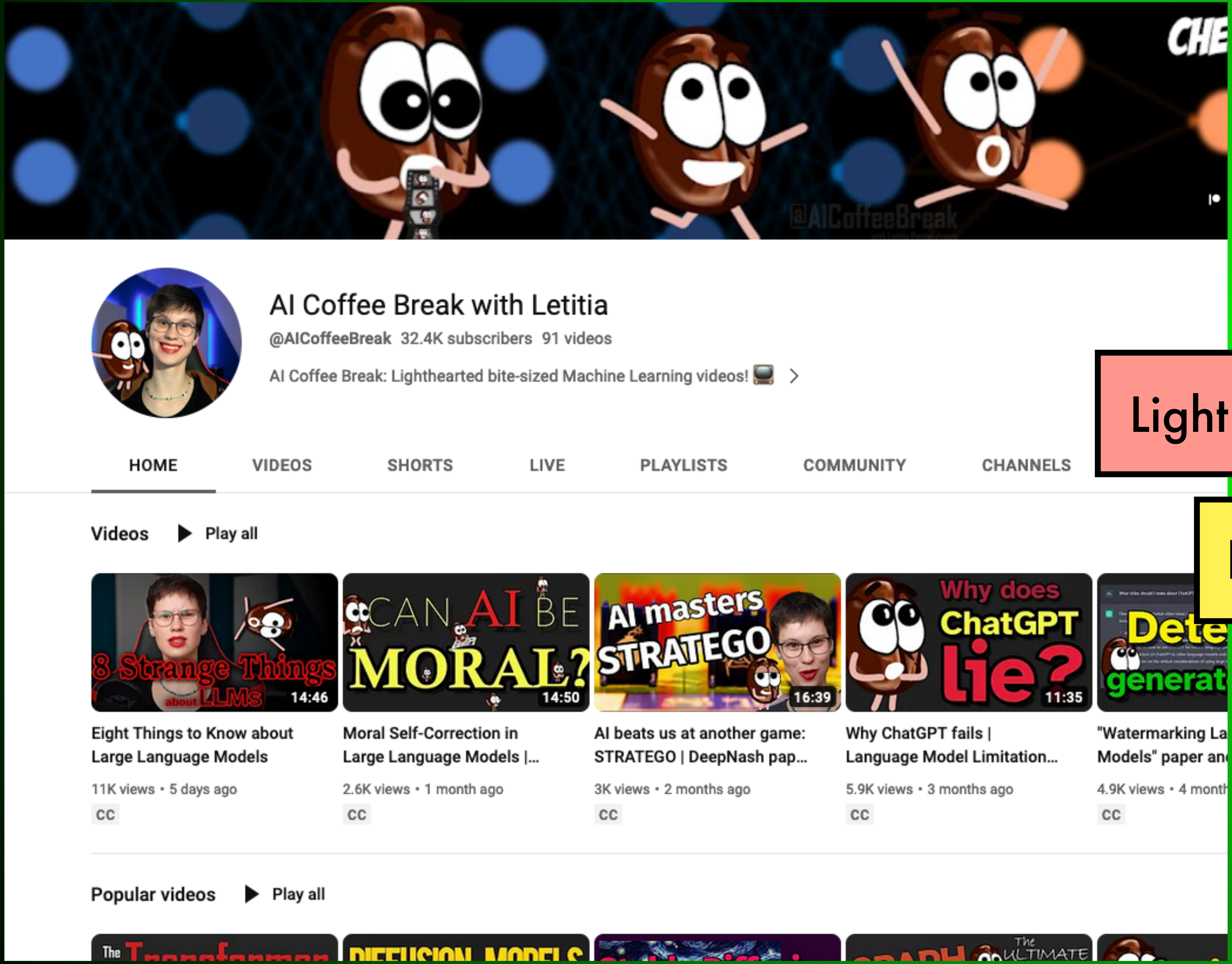
Last updated on June 7, 2023

Prompt Engineering Guide

"all the latest papers, learning guides, models, lectures, references, new LLM capabilities, and tools related to prompt engineering"

https://www.promptingguide.ai/

AI Coffee Break With Letitia



AI Coffee Break with Letitia

@AICoffeeBreak 32.4K subscribers 91 videos

AI Coffee Break: Lighthearted bite-sized Machine Learning videos!

HOME VIDEOS SHORTS LIVE PLAYLISTS COMMUNITY CHANNELS

Videos ▶ Play all

Eight Things to Know about Large Language Models
11K views • 5 days ago
CC

Moral Self-Correction in Large Language Models |...
2.6K views • 1 month ago
CC

AI beats us at another game: STRATEGO | DeepNash pap...
3K views • 2 months ago
CC

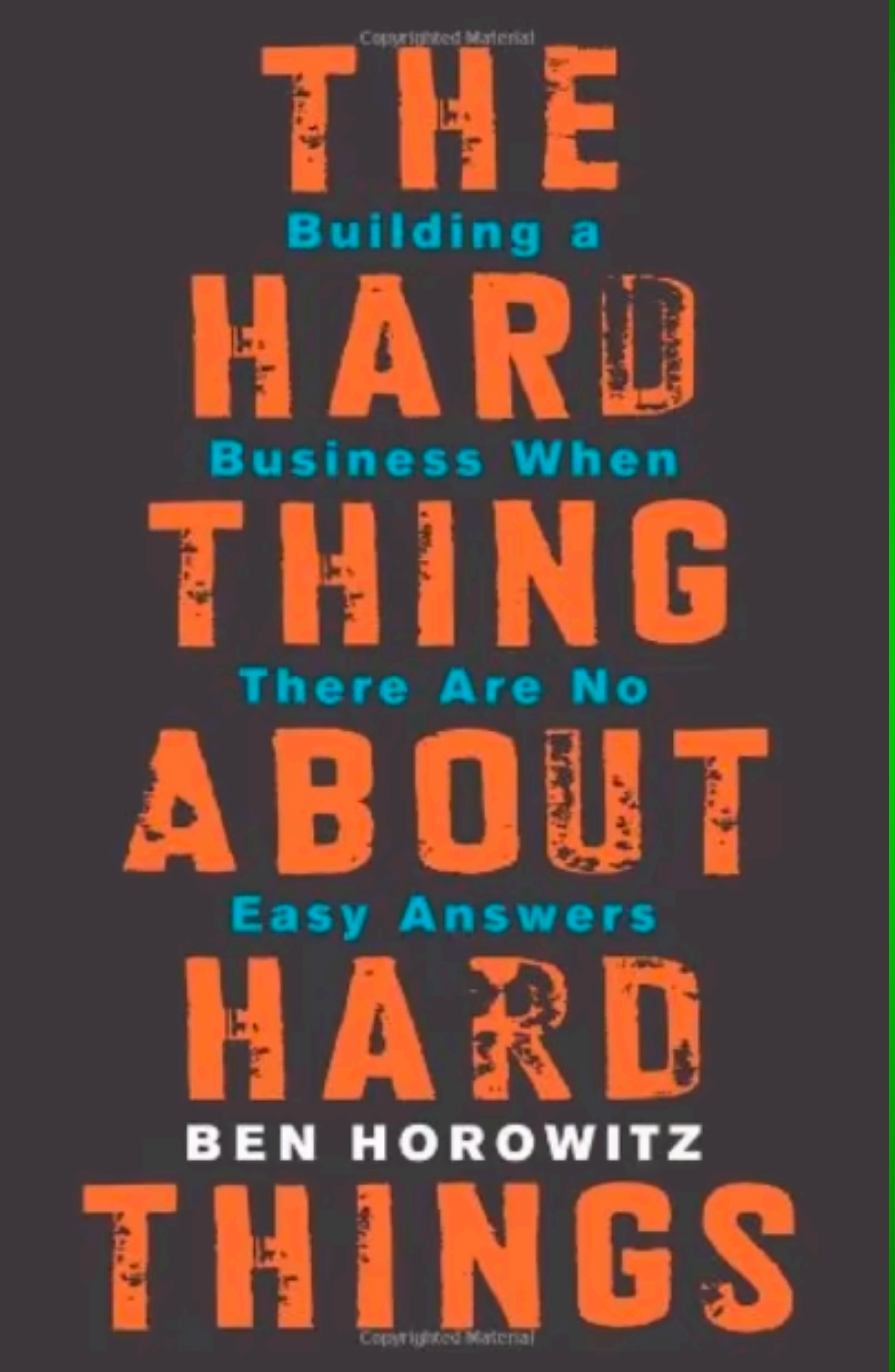
Why ChatGPT fails | Language Model Limitation...
5.9K views • 3 months ago
CC

"Watermarking La
Models" paper an
4.9K views • 4 month
CC

Popular videos ▶ Play all

Lighthearted bite-sized Machine Learning videos

https://www.youtube.com/c/aicoffeebreak



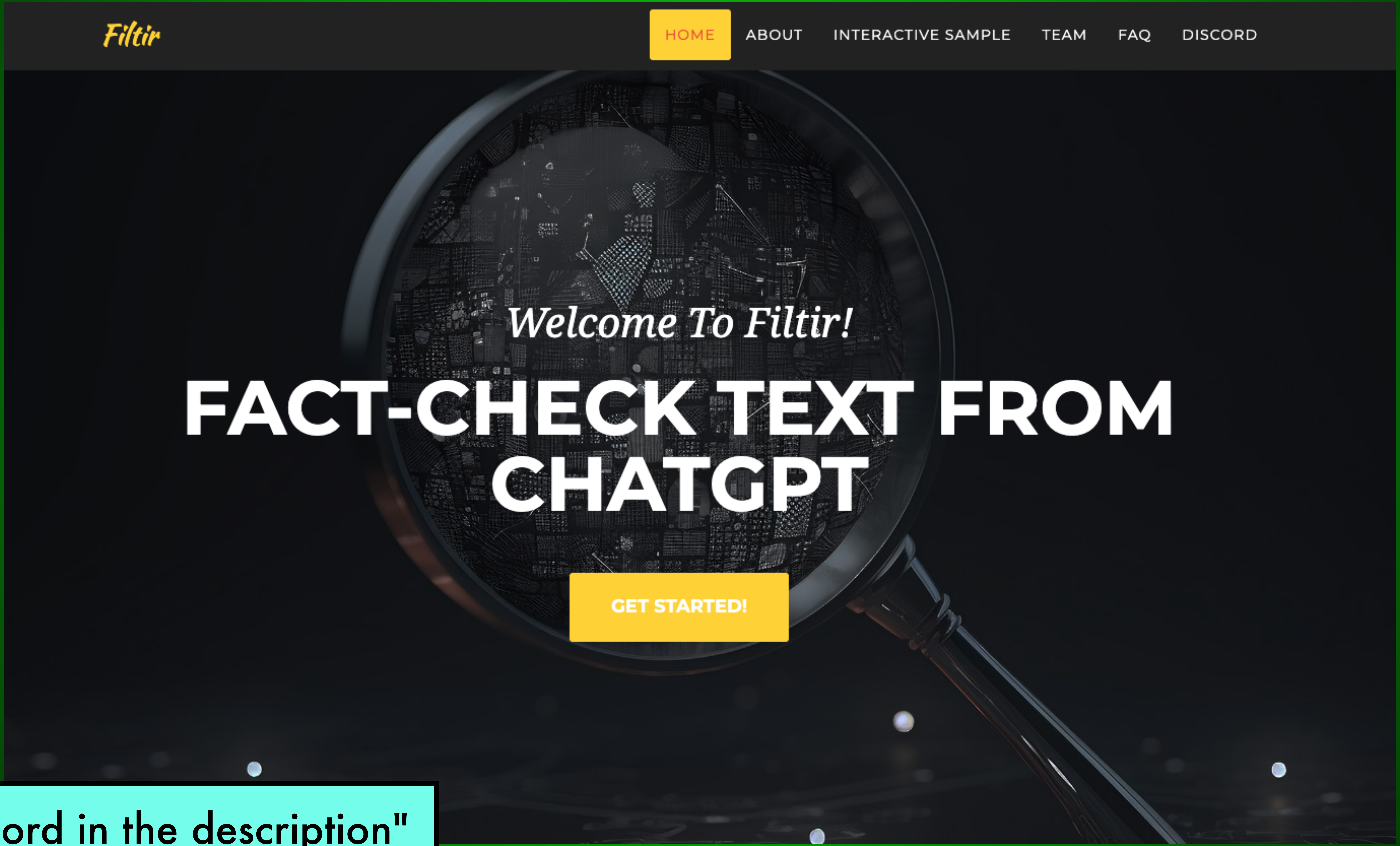
Samuel's Book Recommendation

Unsolicited book recommendation

"The Hard Thing about Hard Things"
Ben Horowitz (2014)

What is it? Strategies/anecdotes/management advice for when things are going seriously badly

Filtir - fact-checking AI claims



Filtir

HOME ABOUT INTERACTIVE SAMPLE TEAM FAQ DISCORD

Welcome To Filtir!

FACT-CHECK TEXT FROM CHATGPT

GET STARTED!

"Link to discord in the description"