

# The False Promise of Imitating Proprietary LLMs

**Arnav Gudibande\***  
UC Berkeley  
arnavg@berkeley.edu

**Eric Wallace\***  
UC Berkeley  
ericwallace@berkeley.edu

**Charlie Snell\***  
UC Berkeley  
csnell22@berkeley.edu

Xinyang Geng  
UC Berkeley  
young.geng@berkeley.edu

Hao Liu  
UC Berkeley  
hao.liu@berkeley.edu

Pieter Abbeel  
UC Berkeley  
pabbeel@berkeley.edu

Sergey Levine  
UC Berkeley  
svlevine@berkeley.edu

Dawn Song  
UC Berkeley  
dawnsong@berkeley.edu

25 May 2023



# Samuel Albanie

2023-5-25

## Model evaluation for extreme risks

Toby Shevlane<sup>1</sup>, Sebastian Farquhar<sup>1</sup>, Ben Garfinkel<sup>2</sup>, Mary Phuong<sup>1</sup>, Jess Whittlestone<sup>3</sup>, Jade Leung<sup>4</sup>, Daniel Kokotajlo<sup>4</sup>, Nahema Marchal<sup>1</sup>, Markus Anderljung<sup>2</sup>, Noam Kolt<sup>5</sup>, Lewis Ho<sup>1</sup>, Divya Siddarth<sup>6, 7</sup>, Shahar Avin<sup>8</sup>, Will Hawkins<sup>1</sup>, Been Kim<sup>1</sup>, Iason Gabriel<sup>1</sup>, Vijay Bolina<sup>1</sup>, Jack Clark<sup>9</sup>, Yoshua Bengio<sup>10, 11</sup>, Paul Christiano<sup>12</sup> and Allan Dafoe<sup>1</sup>

<sup>1</sup>Google DeepMind, <sup>2</sup>Centre for the Governance of AI, <sup>3</sup>Centre for Long-Term Resilience, <sup>4</sup>OpenAI, <sup>5</sup>University of Toronto, <sup>6</sup>University of Oxford, <sup>7</sup>Collective Intelligence Project, <sup>8</sup>University of Cambridge, <sup>9</sup>Anthropic, <sup>10</sup>Université de Montréal, <sup>11</sup>Mila – Quebec AI Institute, <sup>12</sup>Alignment Research Center

Current approaches to building general-purpose AI systems tend to produce systems with both beneficial and harmful capabilities. Further progress in AI development could lead to capabilities that pose extreme risks, such as offensive cyber capabilities or strong manipulation skills. We explain why *model evaluation* is critical for addressing extreme risks. Developers must be able to identify dangerous capabilities (through “dangerous capability evaluations”) and the propensity of models to apply their capabilities for harm (through “alignment evaluations”). These evaluations will become critical for

# Statement on AI Risk

AI experts and public figures express their concern about AI risk.

## Contents

- Statement
- Signatories
- Sign the statement

AI experts, journalists, policymakers, and the public are increasingly discussing a broad spectrum of important and urgent risks from AI. Even so, it can be difficult to voice concerns about some of advanced AI’s most severe risks. The succinct statement below aims to overcome this obstacle and open up discussion. It is also meant to create common knowledge of the growing number of experts and public figures who also take some of advanced AI’s most severe risks seriously.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.



REUTERS®

World ▾ Business ▾ Markets ▾ Sustainability ▾ Legal ▾ Breakingviews Technology ▾ Investigations Mo



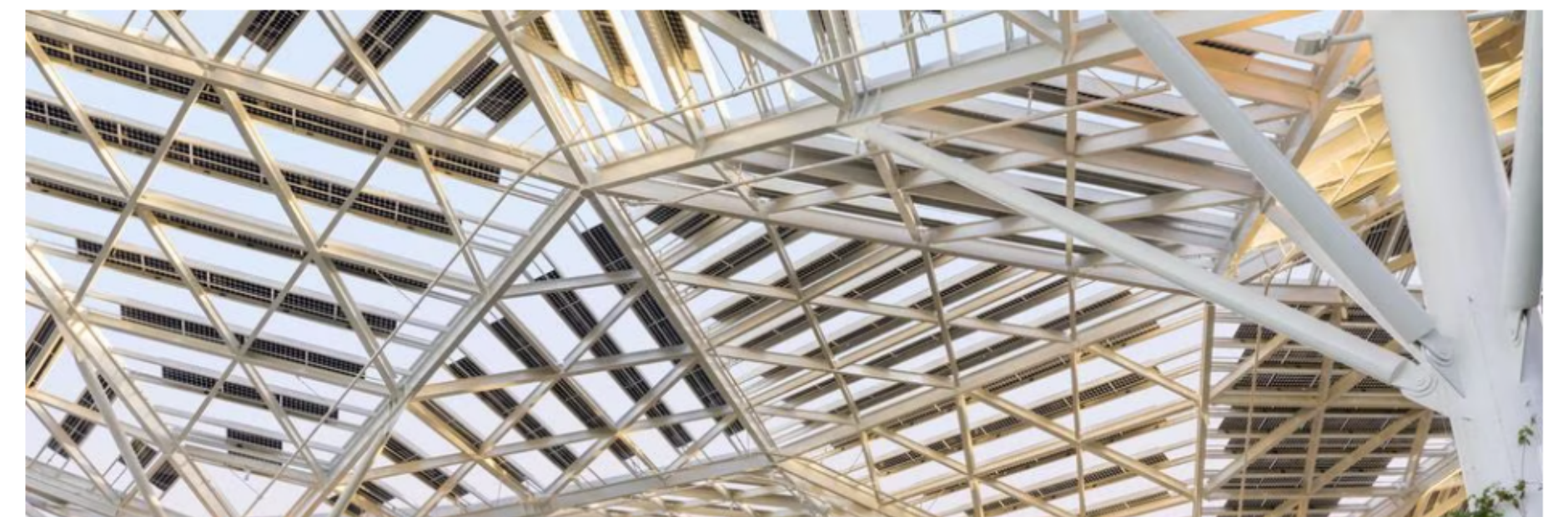
Technology



## Nvidia joins \$1 trillion valuation club on booming AI demand

By Akash Sriram ▾ and Samritha A ▾

May 30, 2023 5:28 PM GMT+1 · Updated 2 hours ago

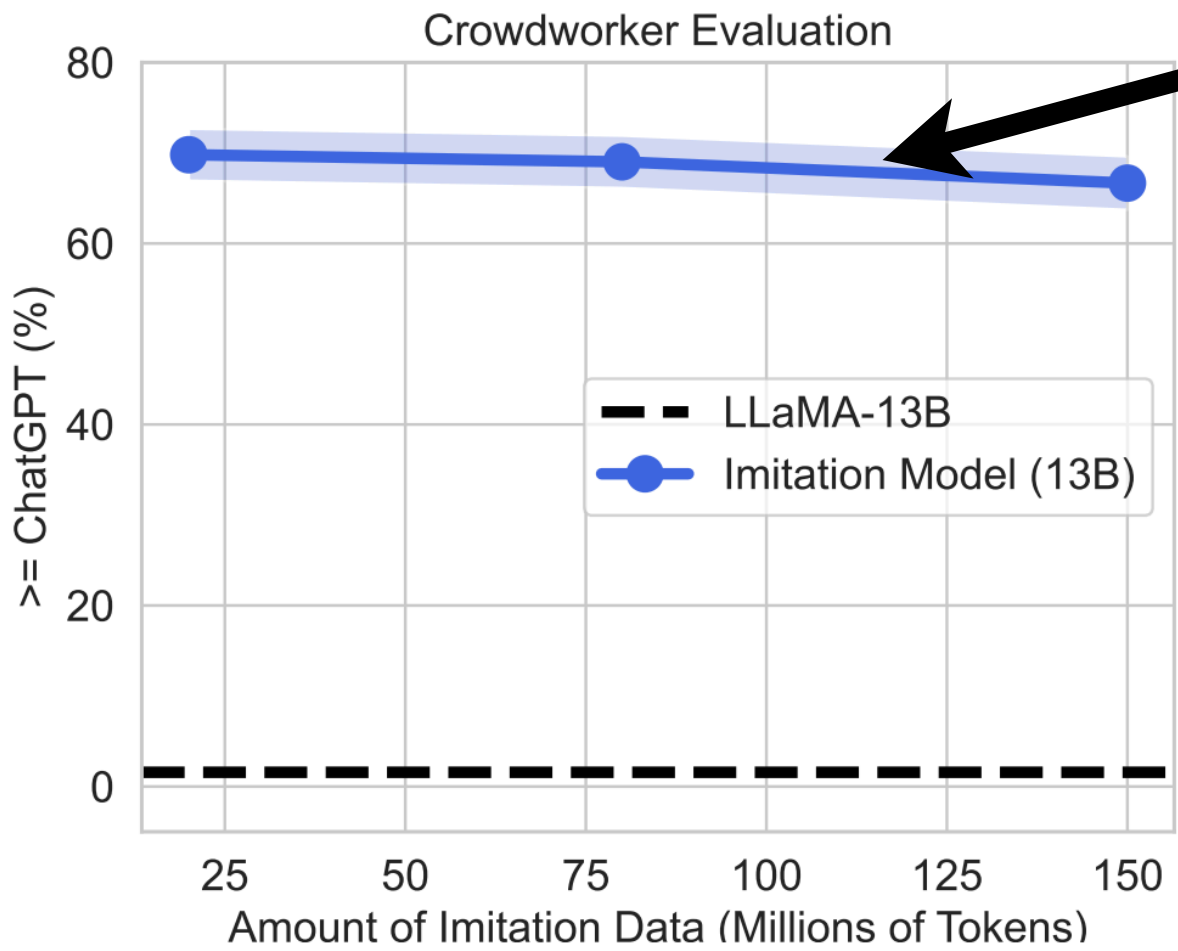


2023

# The False Promise of Imitating Proprietary LLMs 25th May 2023

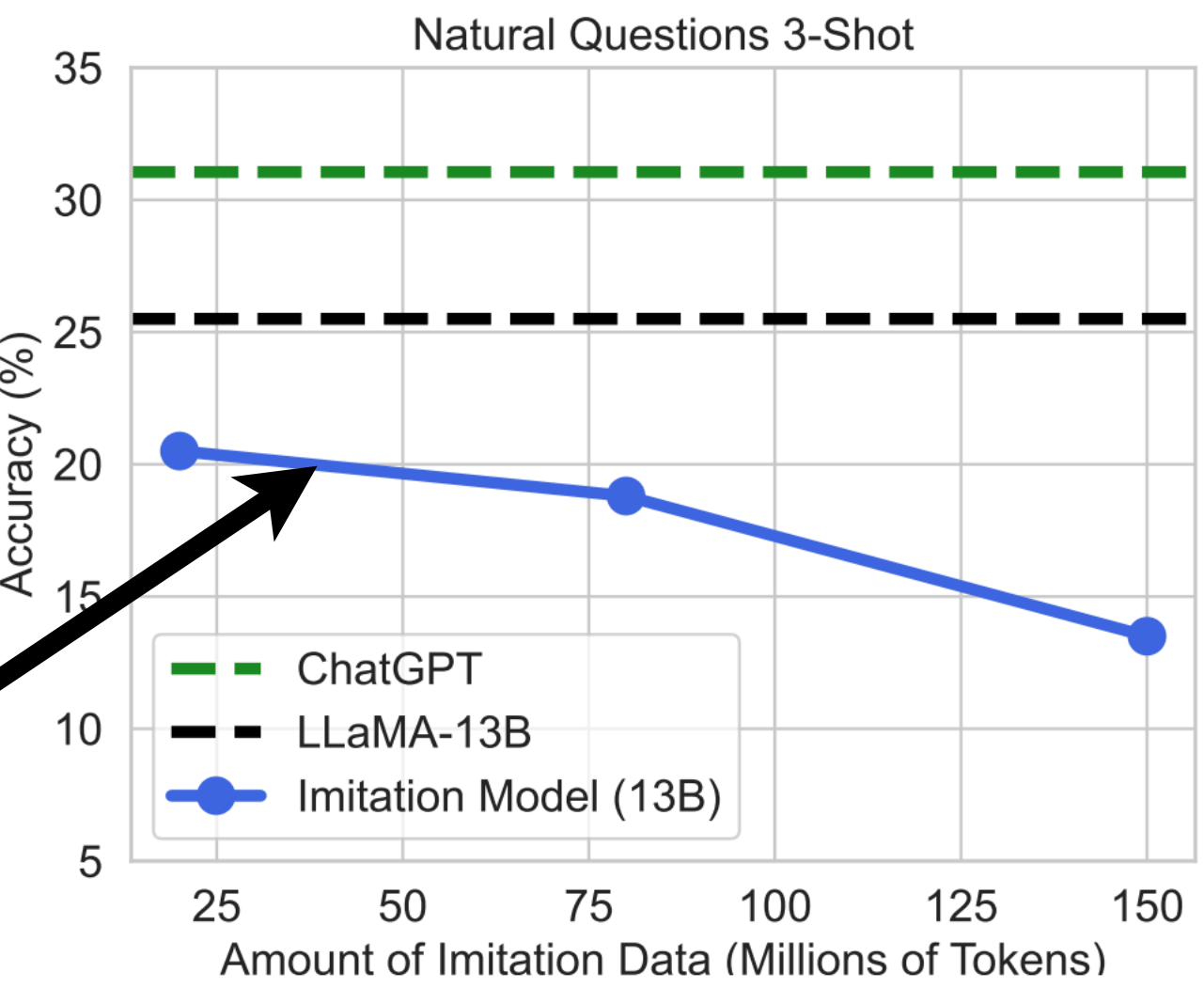
"model imitation is a false promise"

"there exists a substantial capabilities gap between open and closed LMs..."



"Crowd workers rate... imitation models highly"

but models "regress along other axes, e.g. factual knowledge"



## The False Promise of Imitating Proprietary LLMs

**Arnav Gudibande\***  
UC Berkeley  
arnavg@berkeley.edu

**Eric Wallace\***  
UC Berkeley  
ericwallace@berkeley.edu

**Charlie Snell\***  
UC Berkeley  
csnell22@berkeley.edu

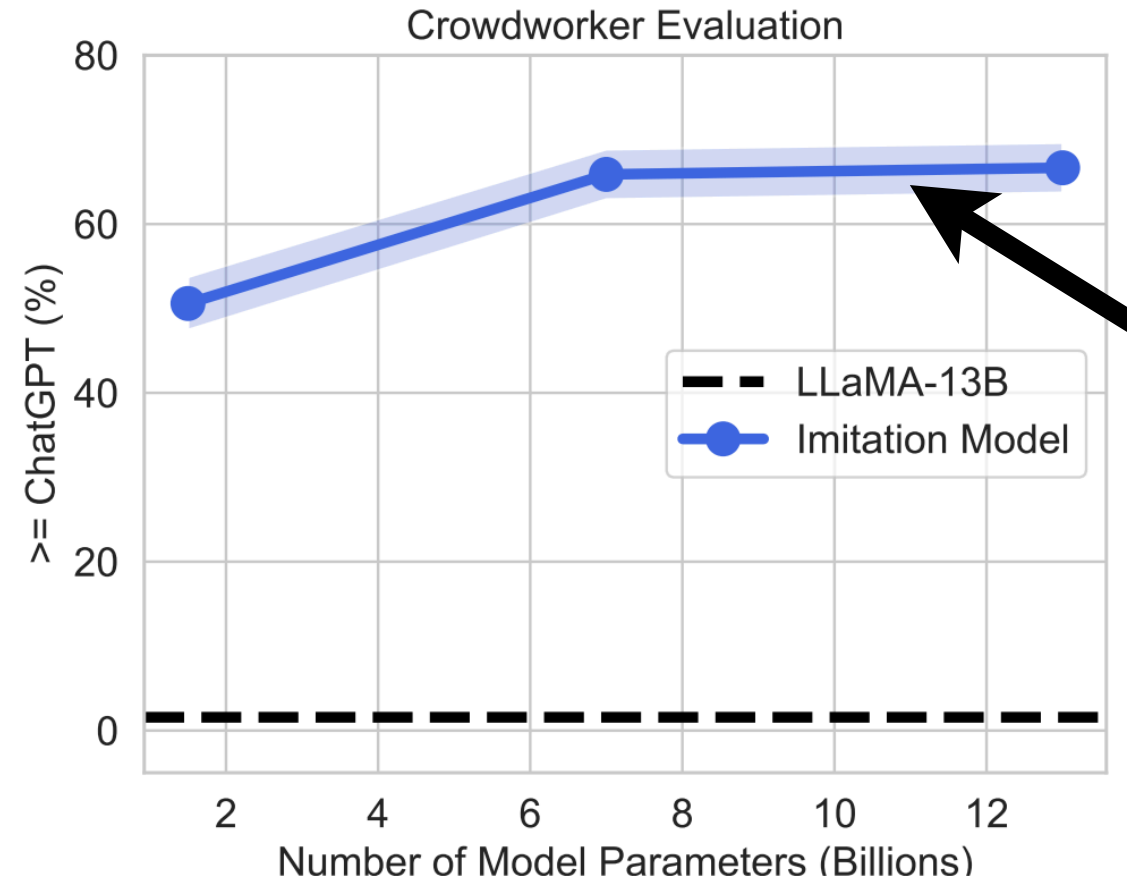
**Xinyang Geng**  
UC Berkeley  
young.geng@berkeley.edu

**Hao Liu**  
UC Berkeley  
hao.liu@berkeley.edu

**Pieter Abbeel**  
UC Berkeley  
pabbeel@berkeley.edu

**Sergey Levine**  
UC Berkeley  
svlevine@berkeley.edu

**Dawn Song**  
UC Berkeley  
dawnsong@berkeley.edu



scaling/pretraining will have more influence than further imitation data

"Our main conclusion is that the biggest limitation of current open-source LMs is their weaker base capabilities"

### Abstract

An emerging method to cheaply improve a weaker language model is to finetune it on outputs from a stronger model, such as a proprietary system like ChatGPT

# FactScore

23rd May 2023

Evaluating factual precision of LMs is challenging

**FactScore: % of atomic facts supported by a given knowledge source**

## FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min<sup>†1</sup> Kalpesh Krishna<sup>†2</sup> Xinxu Lyu<sup>1</sup> Mike Lewis<sup>4</sup> Wen-tau Yih<sup>4</sup>  
 Pang Wei Koh<sup>1</sup> Mohit Iyyer<sup>2</sup> Luke Zettlemoyer<sup>1,4</sup> Hannaneh Hajishirzi<sup>1</sup>

<sup>1</sup>University of Washington <sup>2</sup>University of Massachusetts Amherst

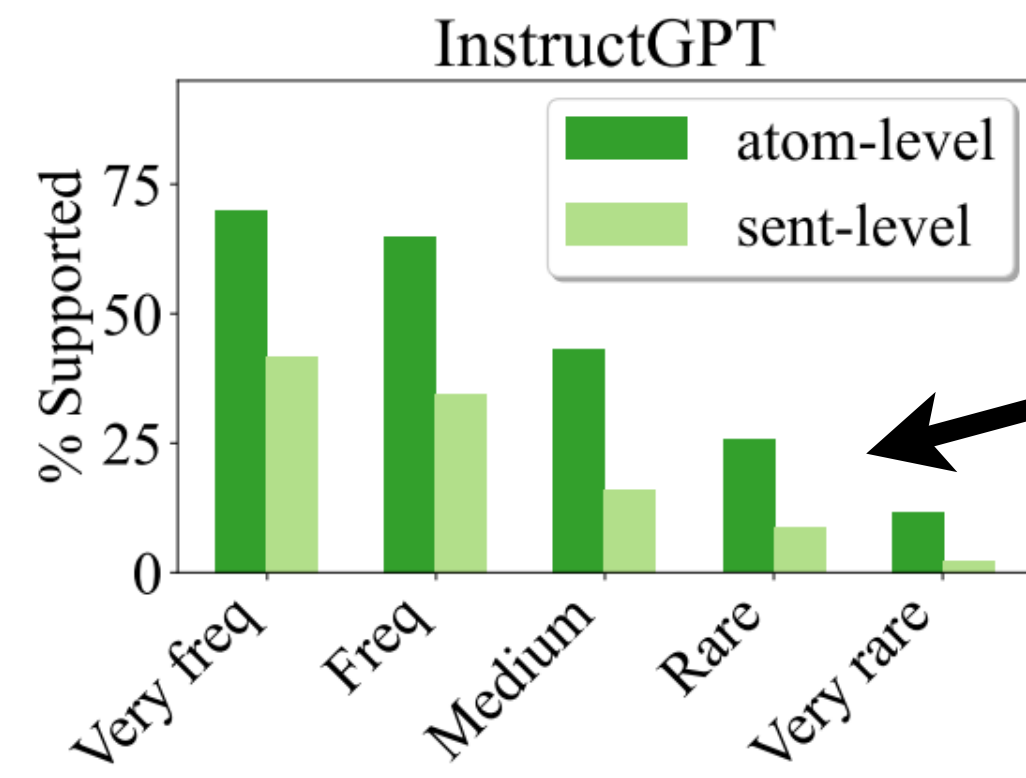
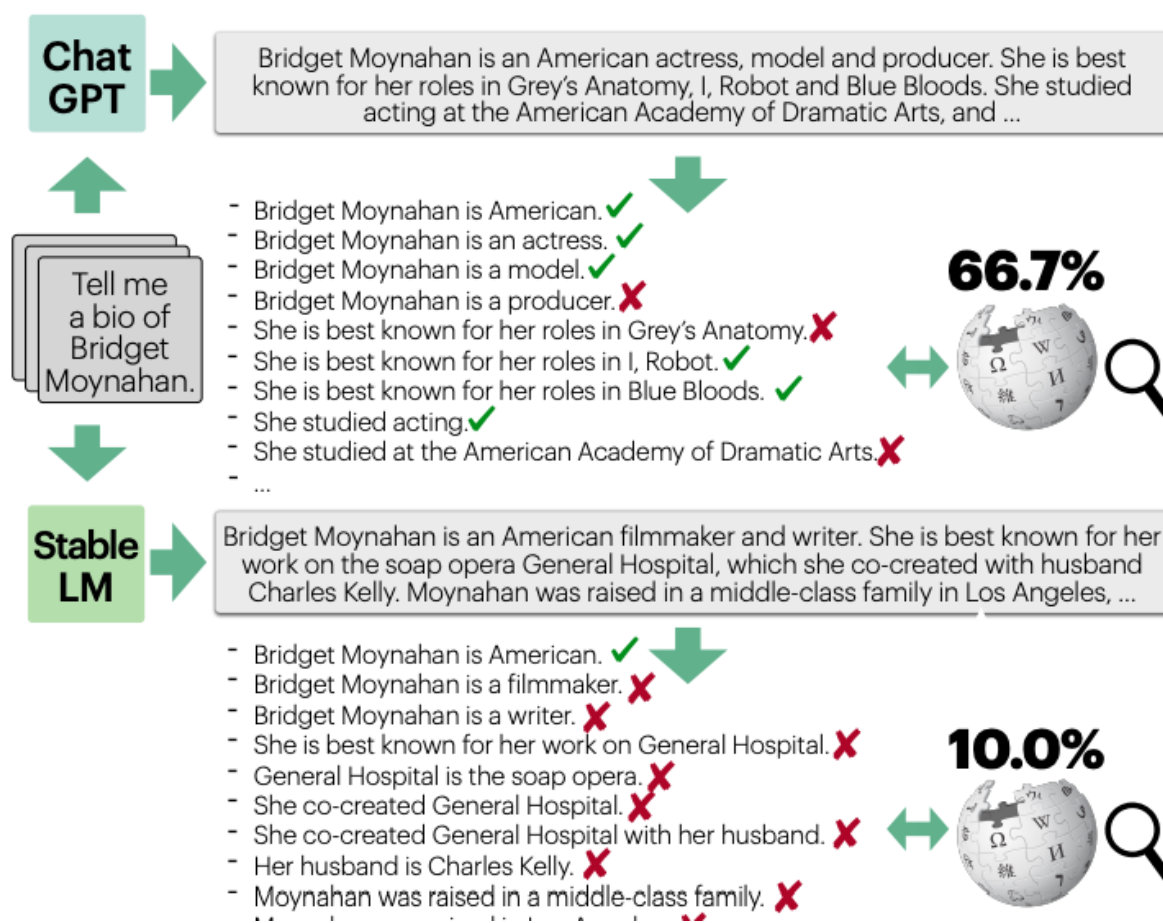
<sup>3</sup>Allen Institute for AI <sup>4</sup>Meta AI

{sewon,alrope,pangwei,lsz,hannaneh}@cs.washington.edu

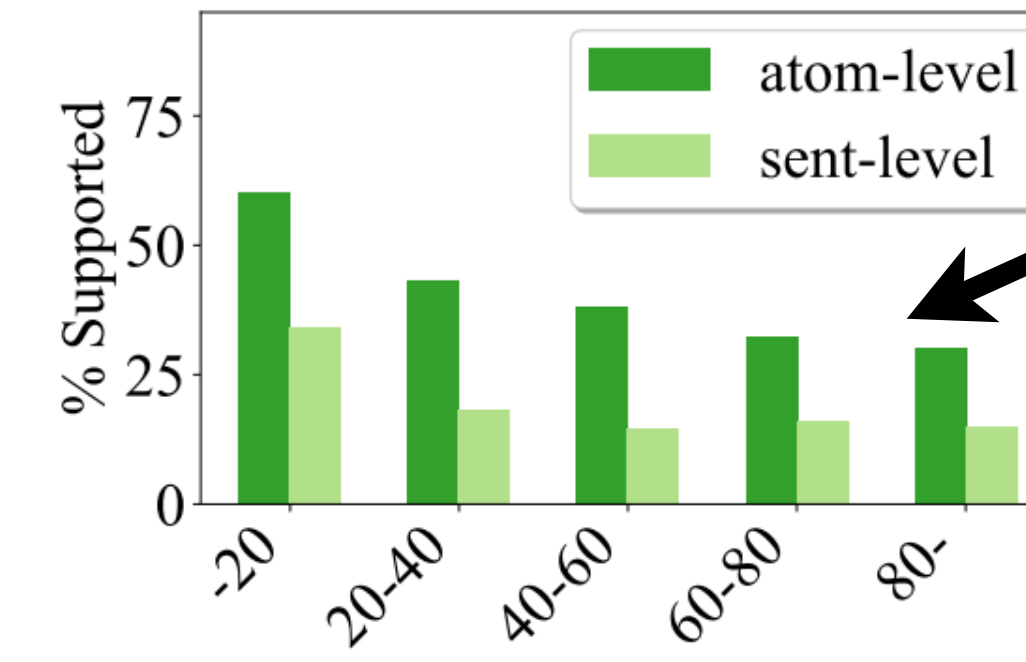
{kalpesh,miyyer}@cs.umass.edu {mikelewis,scottyih}@meta.com

### Abstract

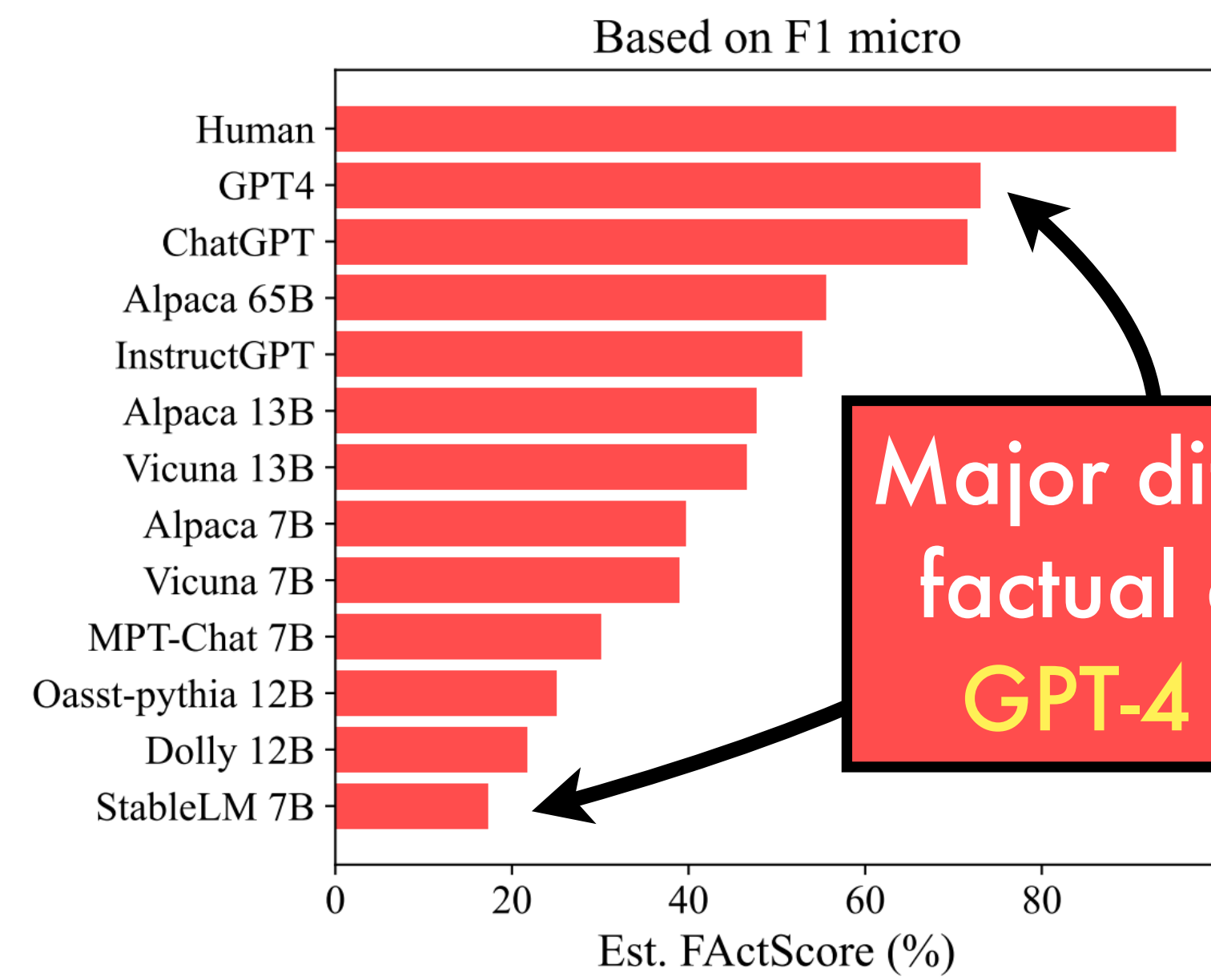
Evaluating the factuality of long-form text generated by large language models (LMs) is non-trivial because (1) generations often contain a mixture of supported and unsupported pieces of information, making binary judgments of quality inadequate, and (2) human evaluation is time-consuming and costly. In this paper, we introduce **FACTSCORE** (Factual precision in Atomicity Score), a new evaluation that breaks a generation into a series of atomic facts and computes the percentage of atomic facts supported by a reliable knowledge source. We conduct an extensive human evaluation to ob-



"Error rates are higher for rare entities"



"Error rates are higher for facts mentioned later in the generation"



Major differences in est. factual accuracy (e.g. GPT-4 vs StableLM)

[L] 23 May 2023

# Gorilla

24th May 2023

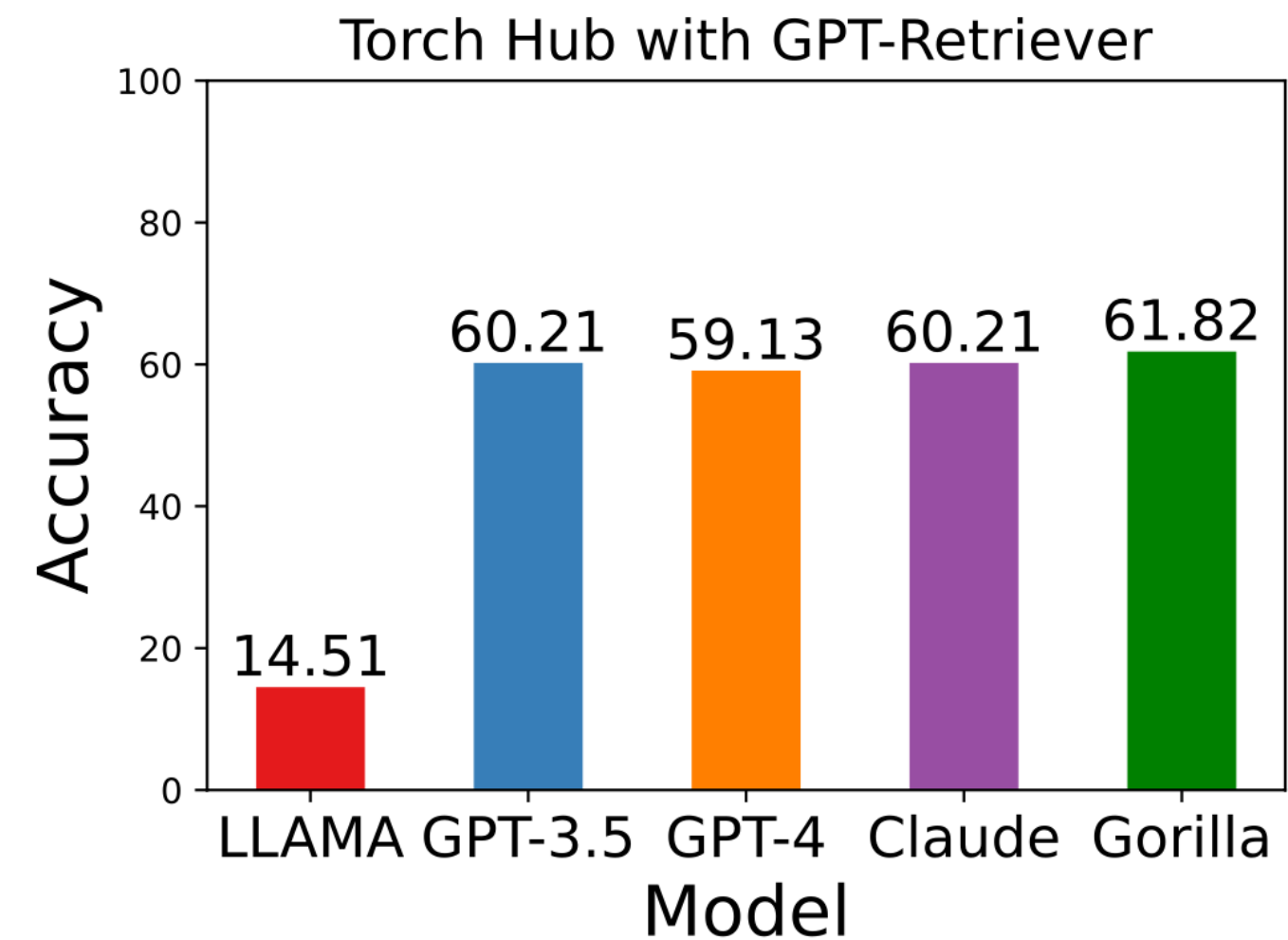
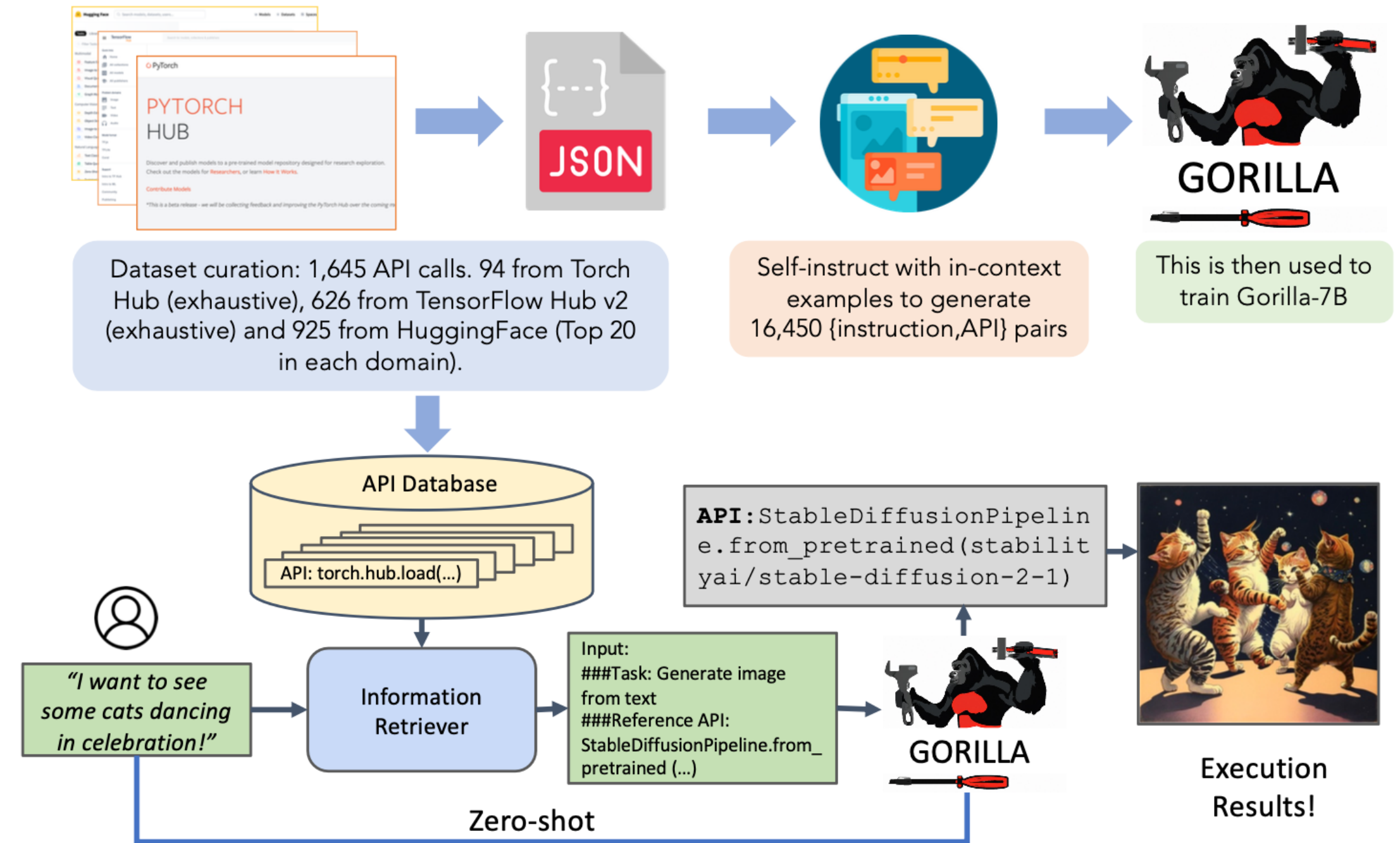
"We release Gorilla, a **finetuned** LLaMA-based model that surpasses GPT-4 on writing API calls."

## Gorilla: Large Language Model Connected with Massive APIs

Shishir G. Patil<sup>1\*</sup> Tianjun Zhang<sup>1,\*</sup> Xin Wang<sup>2</sup> Joseph E. Gonzalez<sup>1</sup>  
<sup>1</sup>UC Berkeley <sup>2</sup>Microsoft Research  
sgp@berkeley.edu

### Abstract

Large Language Models (LLMs) have seen an impressive wave of advances recently, with models now excelling in a variety of tasks, such as mathematical reasoning and program synthesis. However, their potential to effectively use tools via API calls remains unfulfilled. This is a challenging task even for today's state-of-the-art LLMs such as GPT-4, largely due to their inability to generate accurate input arguments and their tendency to hallucinate the wrong usage of an API call. We release Gorilla, a finetuned LLaMA-based model that surpasses the performance of GPT-4 on writing API calls. When combined with a document retriever, Gorilla demonstrates a strong capability to adapt to test-time document changes, enabling flexible user updates or version changes. It also substantially mitigates the issue of hallucination, commonly encountered when prompting LLMs directly. To evaluate





Mac Desktop  
9 items

🔍 My most

# Model Evaluation for Extreme Risks

24th May 2023

"..model evaluation is critical for addressing **extreme risks**"

'We focus on "**extreme**" risks... e.g. damage in the **tens of thousands of lives lost, hundreds of billions of dollars of economic or environmental damage**'



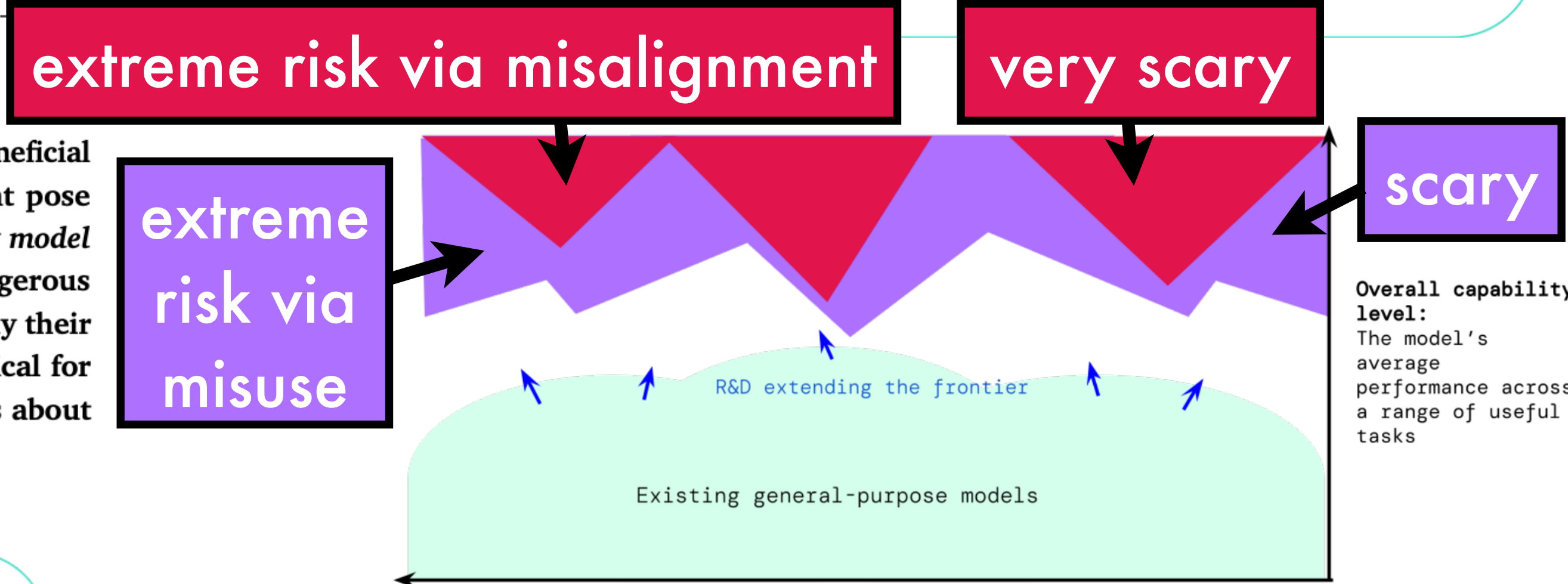
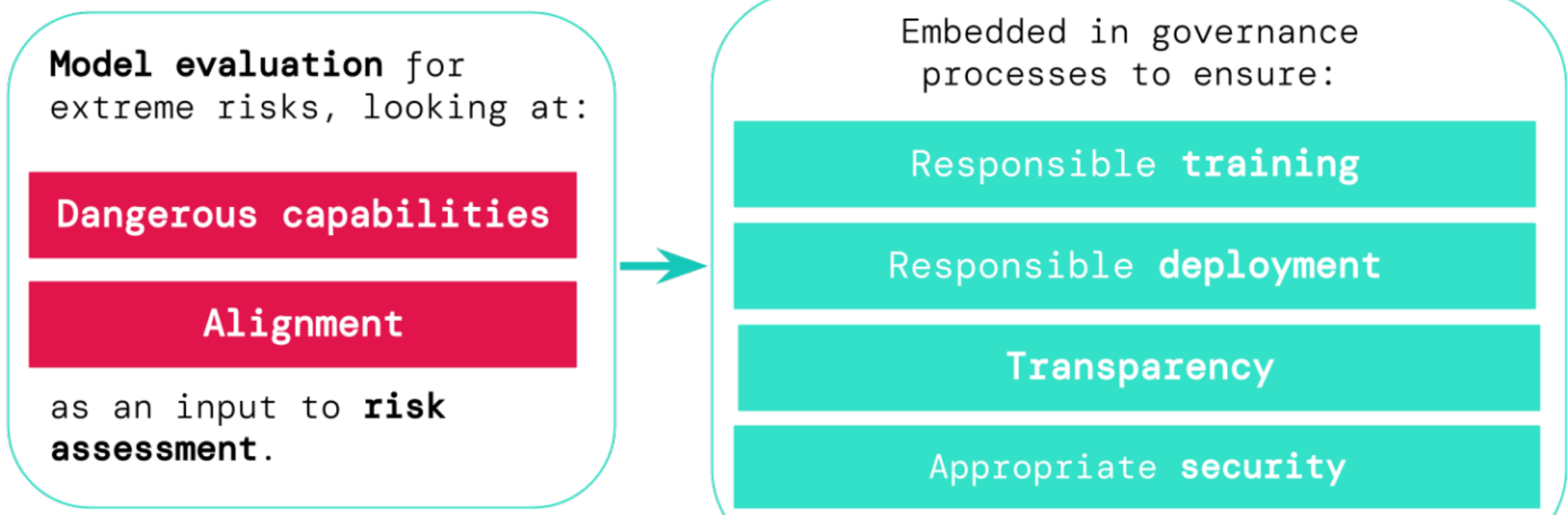
2023-5-25

## Model evaluation for extreme risks

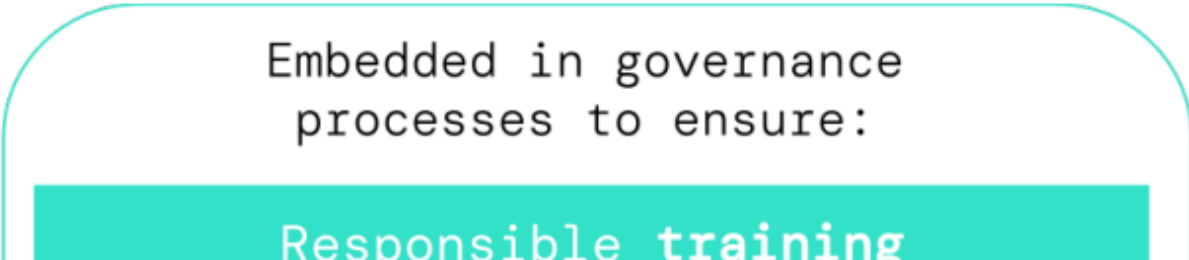
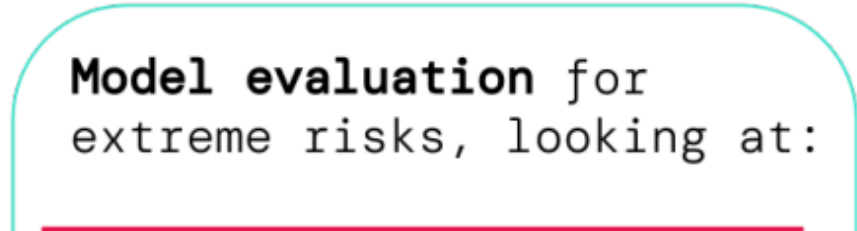
Toby Shvlane<sup>1</sup>, Sebastian Farquhar<sup>1</sup>, Ben Garfinkel<sup>2</sup>, Mary Phuong<sup>1</sup>, Jess Whittlestone<sup>3</sup>, Jade Leung<sup>4</sup>, Daniel Kokotajlo<sup>4</sup>, Nahema Marchal<sup>1</sup>, Markus Anderljung<sup>2</sup>, Noam Kolt<sup>5</sup>, Lewis Ho<sup>1</sup>, Divya Siddarth<sup>6,7</sup>, Shahar Avin<sup>8</sup>, Will Hawkins<sup>1</sup>, Been Kim<sup>1</sup>, Iason Gabriel<sup>1</sup>, Vijay Bolina<sup>1</sup>, Jack Clark<sup>9</sup>, Yoshua Bengio<sup>10,11</sup>, Paul Christiano<sup>12</sup> and Allan Dafoe<sup>1</sup>

<sup>1</sup>Google DeepMind, <sup>2</sup>Centre for the Governance of AI, <sup>3</sup>Centre for Long-Term Resilience, <sup>4</sup>OpenAI, <sup>5</sup>University of Toronto, <sup>6</sup>University of Oxford, <sup>7</sup>Collective Intelligence Project, <sup>8</sup>University of Cambridge, <sup>9</sup>Anthropic, <sup>10</sup>Université de Montréal, <sup>11</sup>Mila-Quebec AI Institute, <sup>12</sup>Alignment Research Center

Current approaches to building general-purpose AI systems tend to produce systems with both beneficial and harmful capabilities. Further progress in AI development could lead to capabilities that pose extreme risks, such as offensive cyber capabilities or strong manipulation skills. We explain why *model evaluation* is critical for addressing extreme risks. Developers must be able to identify dangerous capabilities (through "dangerous capability evaluations") and the propensity of models to apply their capabilities for harm (through "alignment evaluations"). These evaluations will become critical for keeping policymakers and other stakeholders informed, and for making responsible decisions about model training, deployment, and security.






Different approaches for training general-purpose models → leading to models with different capability and alignment profiles.



DeepMind AI 24 May 2023

# PaLM 2 vs other LLMs (Chatbot Arena)

22nd May 2023

Rank	Model	Elo Rating	Description
1	 <a href="#">gpt-4</a>	1225	ChatGPT-4 by OpenAI
2	 <a href="#">claude-v1</a>	1195	Claude by Anthropic
3	 <a href="#">claude-instant-v1</a>	1153	Claude Instant by Anthropic
4	<a href="#">gpt-3.5-turbo</a>	1143	ChatGPT-3.5 by OpenAI
5	<a href="#">vicuna-13b</a>	1054	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
6	<a href="#">palm-2</a>	1042	PaLM 2 for Chat (chat-bison@001) by Google
7	<a href="#">vicuna-7b</a>	1007	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
8	<a href="#">koala-13b</a>	980	a dialogue model for academic research by BAIR
9	<a href="#">mpt-7b-chat</a>	952	a chatbot fine-tuned from MPT-7B by MosaicML

Major gap in Elo rating (GPT-4 vs PaLM-2)



GPT-4

Claude-v1

Claude-inst-v1

GPT-3.5-turbo

Vicuna

PaLM-2

 OpenAI

ANTHROPIC

ANTHROPIC

 OpenAI

LMSYS  
ORG

Google

# SPRING

RL has **high sample complexity**

24th May 2023

this limits effectiveness in games like Crafter/Minecraft

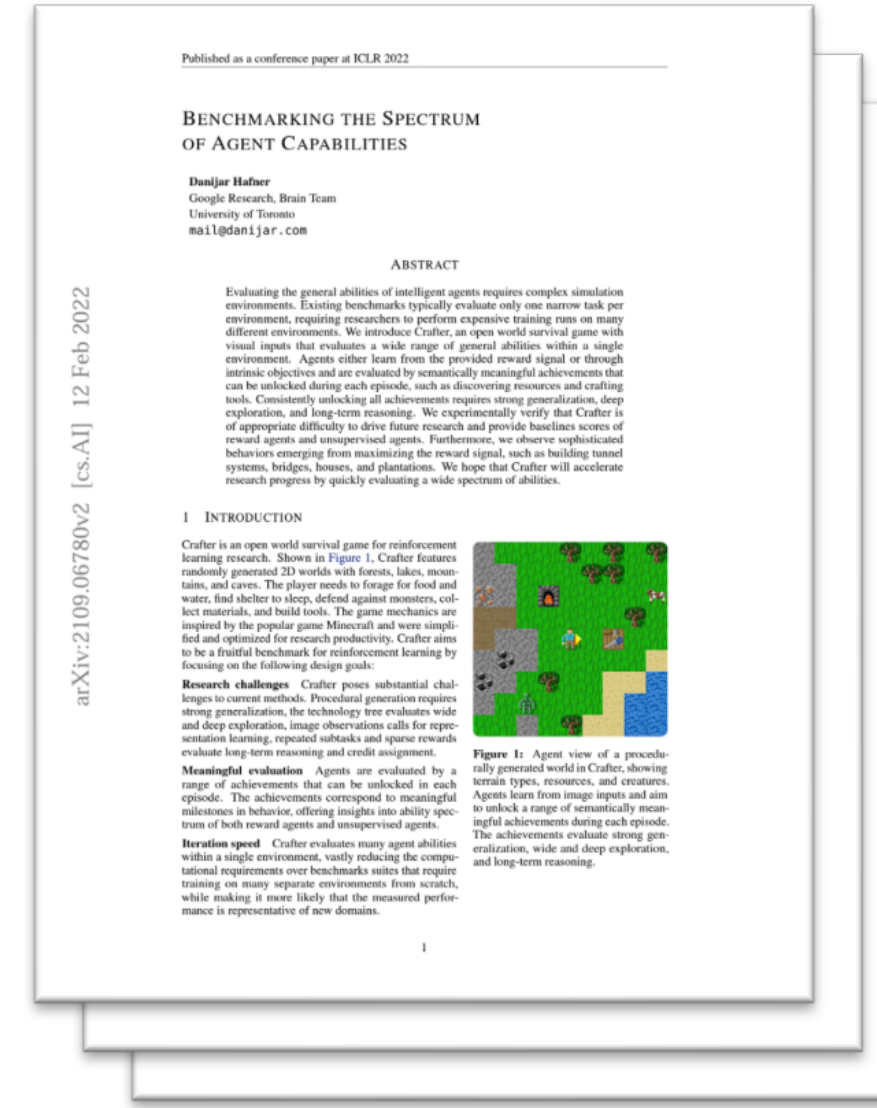
**This work:** "first to show SOTA performance in a challenging open world game with a **zero-shot LLM-based (GPT-4) policy**"

## SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning

Yue Wu<sup>1,4\*</sup>, Shrimai Prabhunoye<sup>2</sup>, So Yeon Min<sup>1</sup>, Yonatan Bisk<sup>1</sup>, Ruslan Salakhutdinov<sup>1</sup>, Amos Azaria<sup>3</sup>, Tom Mitchell<sup>1</sup>, Yuanzhi Li<sup>1,4</sup>  
<sup>1</sup>Carnegie Mellon University, <sup>2</sup>NVIDIA, <sup>3</sup>Ariel University, <sup>4</sup>Microsoft Research

### Abstract

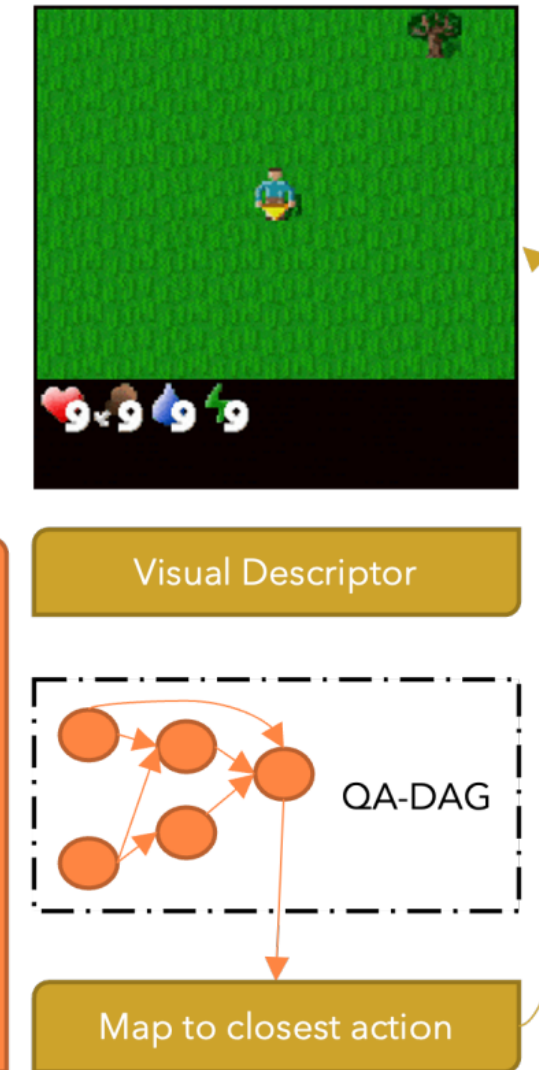
Open-world survival games pose significant challenges for AI algorithms due to their multi-tasking, deep exploration, and goal prioritization requirements. Despite reinforcement learning (RL) being popular for solving games, its high sample complexity limits its effectiveness in complex open-world games like Crafter or Minecraft. We propose a novel approach, SPRING, to read the game's original academic paper and use the knowledge learned to reason and play the game through a large language model (LLM). Prompted with the  $\LaTeX$  source as game context and a description of the agent's current observation, our SPRING framework employs a directed acyclic graph (DAG) with game-related questions as nodes and dependencies as edges. We identify the optimal action to take in the environment by traversing the DAG and calculating LLM responses for each node in topological order, with the LLM's answer to final node directly translating to environment actions. In our experiments, we study the quality of in-context "reasoning" induced by different forms of prompts under the setting of the Crafter open-world



LLM

- Write all information helpful for the game in a numbered list.
1. Collect resources such as wood, stone, and iron to craft tools and weapons.
  2. Build shelters to protect yourself from monsters at night.
  3. Use tools and weapons to defend yourself against monsters.
  4. Build bridges to cross lakes and rivers.
  5. Dig tunnels to surprise monsters and outsmart them.
  6. Plant saplings and defend them against monsters to ensure a steady food supply.
  7. Eat Cow to restore health.
  8. Collect Drink to restore thirst.
  9. Place a Plant to eat for health.
  10. Make a Wood Pickaxe to collect Stone.
  11. Make a Wood Sword to defeat Zombies.
  12. Make a Stone Pickaxe to collect Iron.
  13. Make a Stone Sword to defeat Skeletons.
  14. Place a Furnace to smelt Iron.
  15. Collect Coal to smelt Iron.
  16. Collect Iron to make an Iron Pickaxe and Sword.
  17. Make an Iron Pickaxe to collect Diamond.
  18. Make an Iron Sword to defeat Zombies and Skeletons.
  19. Collect Diamond to progress further.
  20. Unlock achievements to receive rewards.
  21. Wake Up to start the episode.

LLM



Method	Score	Reward	Training Steps
Human Experts	50.5 ± 6.8%	14.3 ± 2.3	N/A
<b>SPRING + paper (Ours)</b>	<b>27.3 ± 1.2%</b>	<b>12.3 ± 0.7</b>	<b>0</b>
DreamerV3 Hafner et al. (2023)	14.5 ± 1.6%	11.7 ± 1.9	1M
ELLM Du et al. (2023)	N/A	6.0 ± 0.4	5M
EDE Jiang et al. (2022)	11.7 ± 1.0%	N/A	1M
DreamerV2 Hafner et al. (2020)	10.0 ± 1.2%	9.0 ± 1.7	1M
PPO Schulman et al. (2017)	4.6 ± 0.3%	4.2 ± 1.2	1M
Rainbow Hessel et al. (2018)	4.3 ± 0.2%	5.0 ± 1.3	1M
Plan2Explore Sekar et al. (2020)	2.1 ± 0.1%	2.1 ± 1.5	1M
RND Burda et al. (2018)	2.0 ± 0.1%	0.7 ± 1.3	1M
Random	1.6 ± 0.0%	2.1 ± 1.3	0

Table 2: Table comparing SPRING and popular RL algorithms in terms of game score, reward, and training steps. Results for SPRING is summarized over 5 independent trials.

**Crafter**



"reduces **memory usage** enough to finetune a 65B parameter model on a 48GB GPU"

"...while **preserving** full 16-bit finetuning performance"

"QLoRA backprops gradients through a frozen, 4-bit quantized pretrained LM into **Low Rank Adapters**"

**Innovations** 4-bit NormalFloat (new data type)

Double quantization (quantize the quantization constants)

Paged Optimizers (to manage memory spikes)

## QLoRA: Efficient Finetuning of Quantized LLMs

Tim Dettmers\*

Artidoro Pagnoni\*

Ari Holtzman

Luke Zettlemoyer

University of Washington

{dettmers,artidoro,ahai,lsz}@cs.washington.edu

### Abstract

We present QLoRA, an efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA). Our best model family, which we name **Guanaco**, outperforms all previous openly released models on the Vicuna benchmark, reaching 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU. QLoRA introduces a number of innovations to save memory without sacrificing performance: (a) 4-bit NormalFloat (NF4), a new data type that is information theoretically optimal for normally distributed weights (b) Double

**Table 7:** Elo rating for a tournament between models where models compete to generate the best response for a prompt, judged by human raters or GPT-4. Overall, Guanaco 65B and 33B tend to be preferred to ChatGPT-3.5 on the benchmarks studied. According to human raters they have a Each 10-point difference in Elo is approximately a difference of 1.5% in win-rate.

Benchmark	Vicuna		Vicuna		Open Assistant		Median Rank
	# Prompts		# Prompts		# Prompts		
Judge	80		80		953		
	Human raters		GPT-4		GPT-4		
Model	Elo	Rank	Elo	Rank	Elo	Rank	
GPT-4	1176	1	1348	1	1294	1	1
Guanaco-65B	1023	2	1022	2	1008	3	2
Guanaco-33B	1009	4	992	3	1002	4	4
ChatGPT-3.5 Turbo	916	7	966	5	1015	2	5
Vicuna-13B	984	5	974	4	936	5	5
Guanaco-13B	975	6	913	6	885	6	6
Guanaco-7B	1010	3	879	8	860	7	7
Bard	909	8	902	7	-	-	8

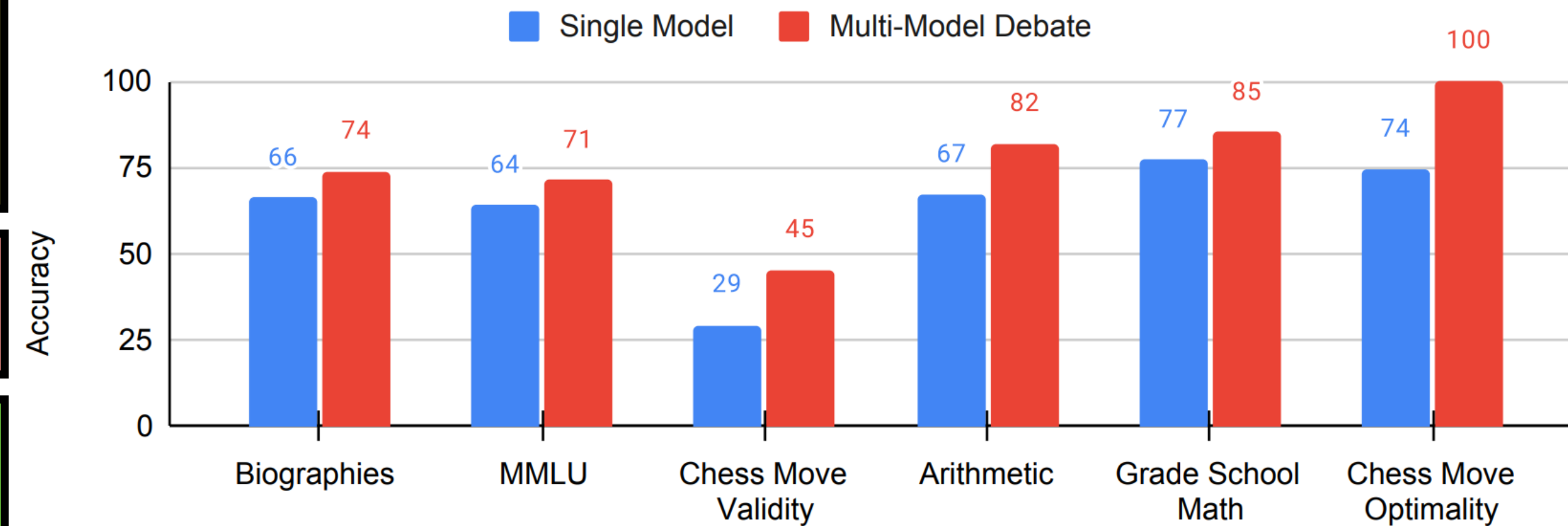
# Multiagent Debate

23rd May 2023

"multiple LM instances **propose and debate** their individual responses... over multiple rounds"

"enhances **mathematical & strategic** reasoning..."

"improves **factual validity** of generated content..."



## Improving Factuality and Reasoning in Language Models through Multiagent Debate

**Yilun Du**  
MIT CSAIL  
yilundu@mit.edu

**Shuang Li**  
MIT CSAIL  
lishuang@mit.edu

**Antonio Torralba**  
MIT CSAIL  
torralba@mit.edu

**Joshua B. Tenenbaum**  
MIT CSAIL, BCS, CBMM  
jbt@mit.edu

**Igor Mordatch**  
Google Brain  
imordatch@google.com

### Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in language generation, understanding, and few-shot learning in recent years. An extensive body of work has explored how their performance may be further improved through the tools of prompting, ranging from verification, self-consistency, or intermediate scratchpads. In this paper, we present a complementary approach to improve language responses where multiple language model instances propose and debate their individual responses and reasoning processes over multiple rounds

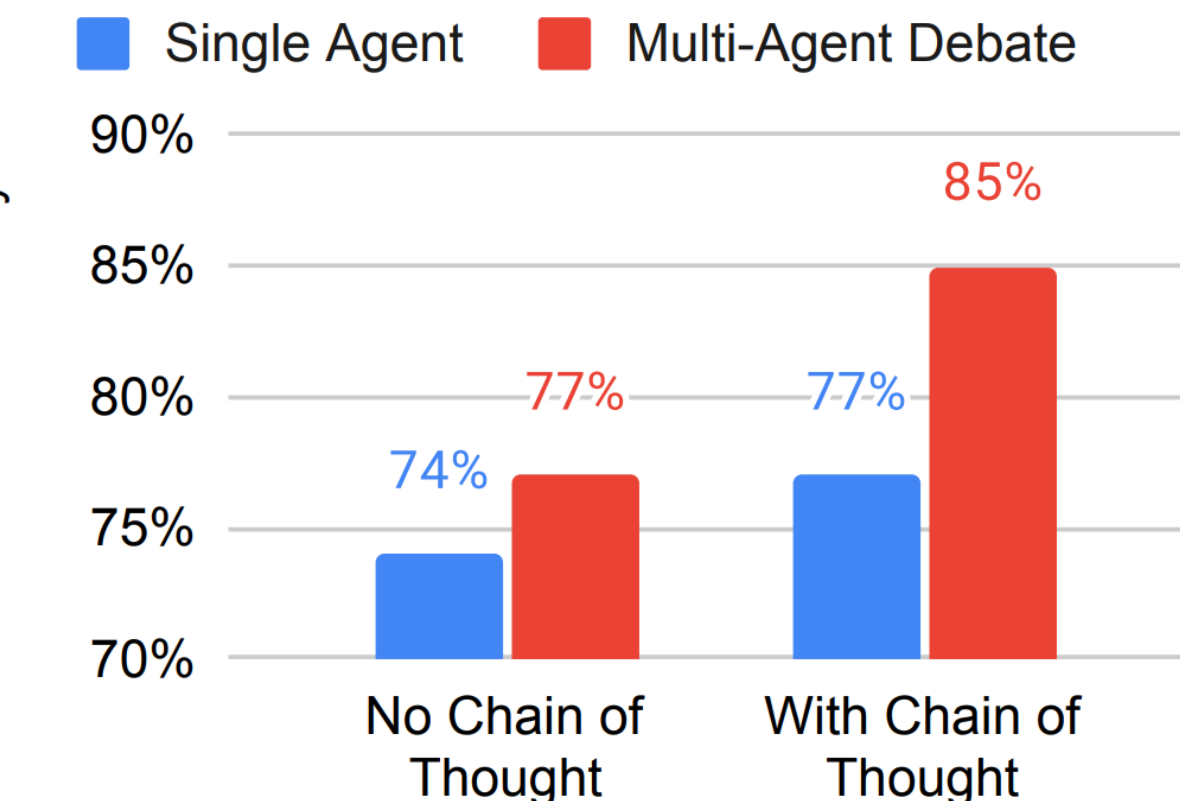


Figure 6: **Synergy with Other Methods.** Performance of debate increases with use of Chain of Thought prompting.

The screenshot shows the GitHub repository page for 'composable-models / llm\_multiagent\_debate'. It includes the repository name, public status, 4 watchers, 6 forks, and 72 stars. The 'Code' tab is selected, showing a list of files: biography, gsm, math, mmlu, README.md, and requirements.txt. The 'About' section contains the repository description: 'Code for Improving Factuality and Reasoning in Language Models through Multiagent Debate'. It also lists 4 contributors: yilundu, yilundu1, ShuangLI59 Shuang Li, and guangyusong.

# Hallucination Snowballing

22nd May 2023

"Hallucination snowballing: an LM **over-commits** to early mistakes, leading to more mistakes that it otherwise would not make"

"We hypothesise that LMs produce snowballed hallucinations for **consistency** with earlier hallucinations"

## How Language Model Hallucinations Can Snowball

Muru Zhang<sup>♥</sup> Ofir Press<sup>♥</sup> William Merrill<sup>♠</sup> Alisa Liu<sup>♥</sup> Noah A. Smith<sup>♥♣</sup>  
<sup>♥</sup>Paul G. Allen School of Computer Science and Engineering, University of Washington  
<sup>♠</sup>New York University  
<sup>♣</sup>Allen Institute for Artificial Intelligence  
nanami17@cs.washington.edu

### Abstract

A major risk of using language models in practical applications is their tendency to hallucinate incorrect statements. Hallucinations are often attributed to knowledge gaps in LMs, but we hypothesize that in some cases, when justifying previously generated hallucinations, LMs output false claims that they can separately recognize as incorrect. We construct three question-answering datasets where ChatGPT and GPT-4 often state an incorrect answer and offer an explanation with at least one incorrect claim. Crucially, we find that ChatGPT and GPT-4 can

Is 9677 a prime number?

No, 9677 is not a prime number. It can be factored into 13 and 745, as  $9677 = 13 \times 745$ .

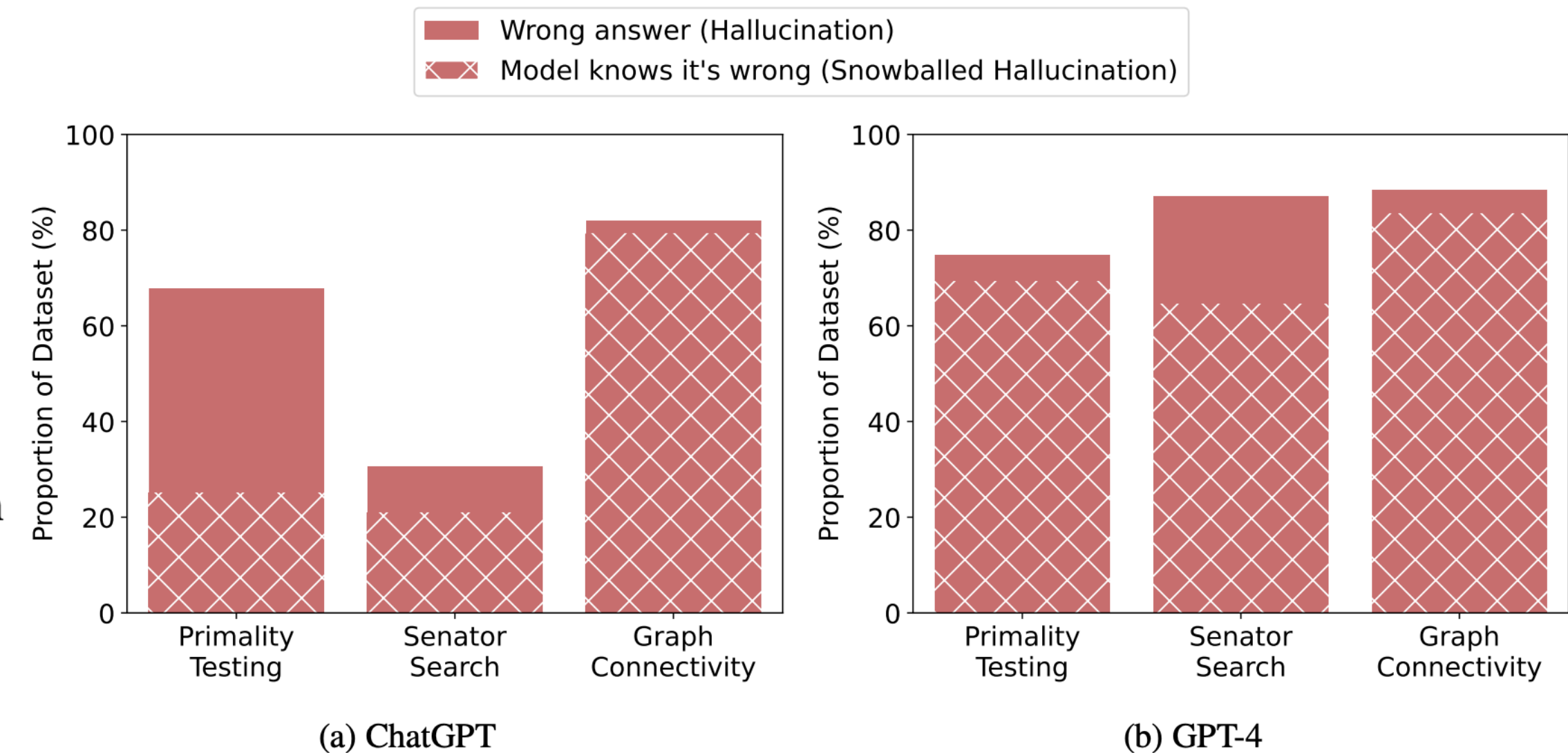
Is 9677 divisible by 13?

No

incorrect assertion

snowballed hallucination

in a separate session, GPT-4 recognizes its claim as incorrect!



Nanami18 / Snowballed\_Hallucination Public

Code Issues Pull requests Actions Projects Security Insights

main 1 branch 0 tags

ofirpress Update README.md f620318 3 days ago 16 commits

- README.md Update README.md 3 days ago
- graph\_connectivity.json Rename graph\_connectivity.json to graph\_connectivity.json 3 days ago
- primality\_testing.json Rename determine\_primeonly\_train.json to primality\_testing.json last week
- senator\_search.json Rename senator\_filteredcollege\_rephrased1\_annotated\_train.json to ... last week

README.md

### How Language Model Hallucinations Can Snowball

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, Noah A. Smith

22 May 2023

# ALIGNSCORE

26th May 2023

"Automatic evaluation of **factual consistency** is challenging"

"ALIGNSCORE, a new general **factual consistency metric** based on a unified text-to-text information alignment function"

## ALIGNSCORE: Evaluating Factual Consistency with A Unified Alignment Function

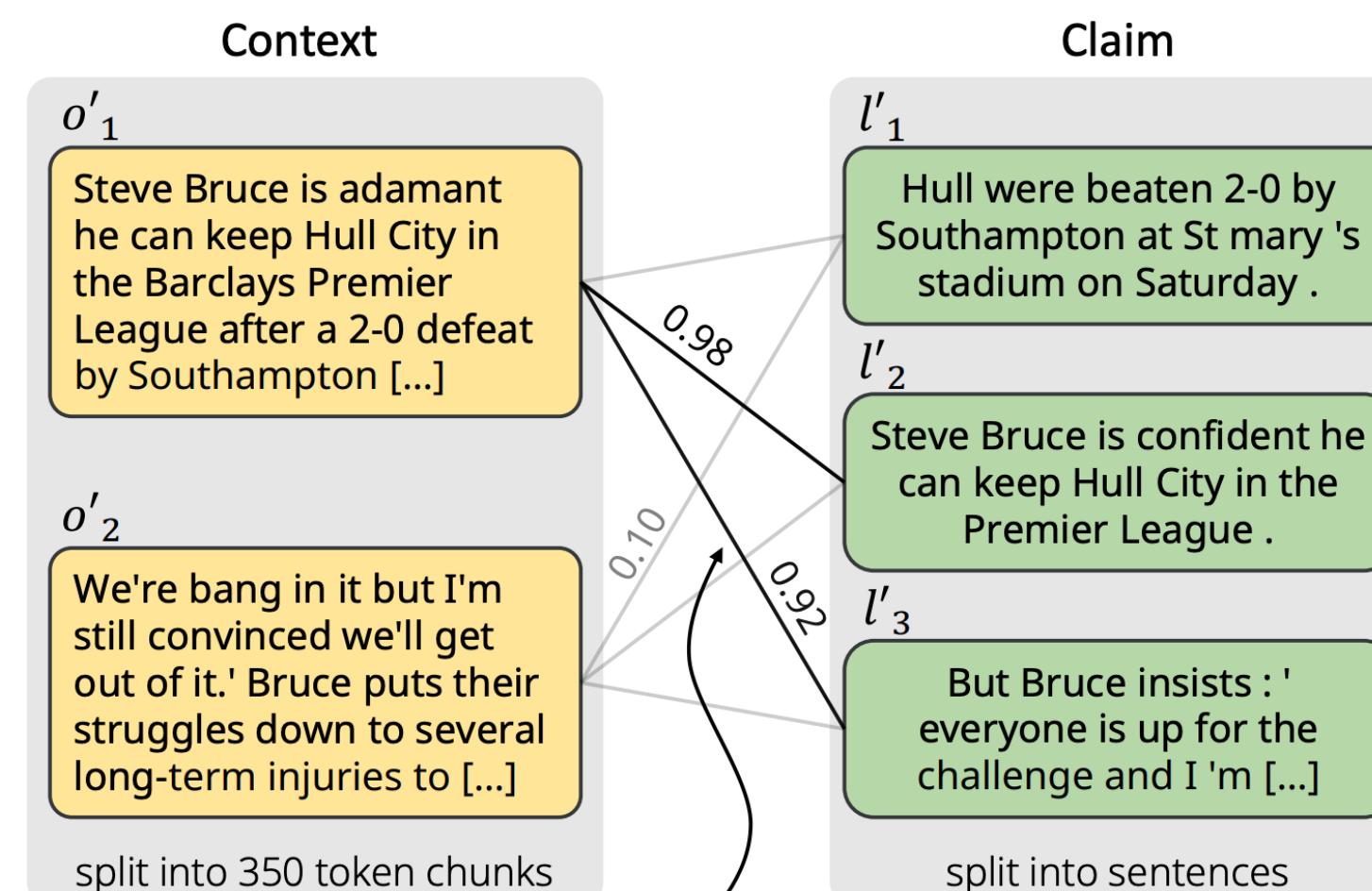
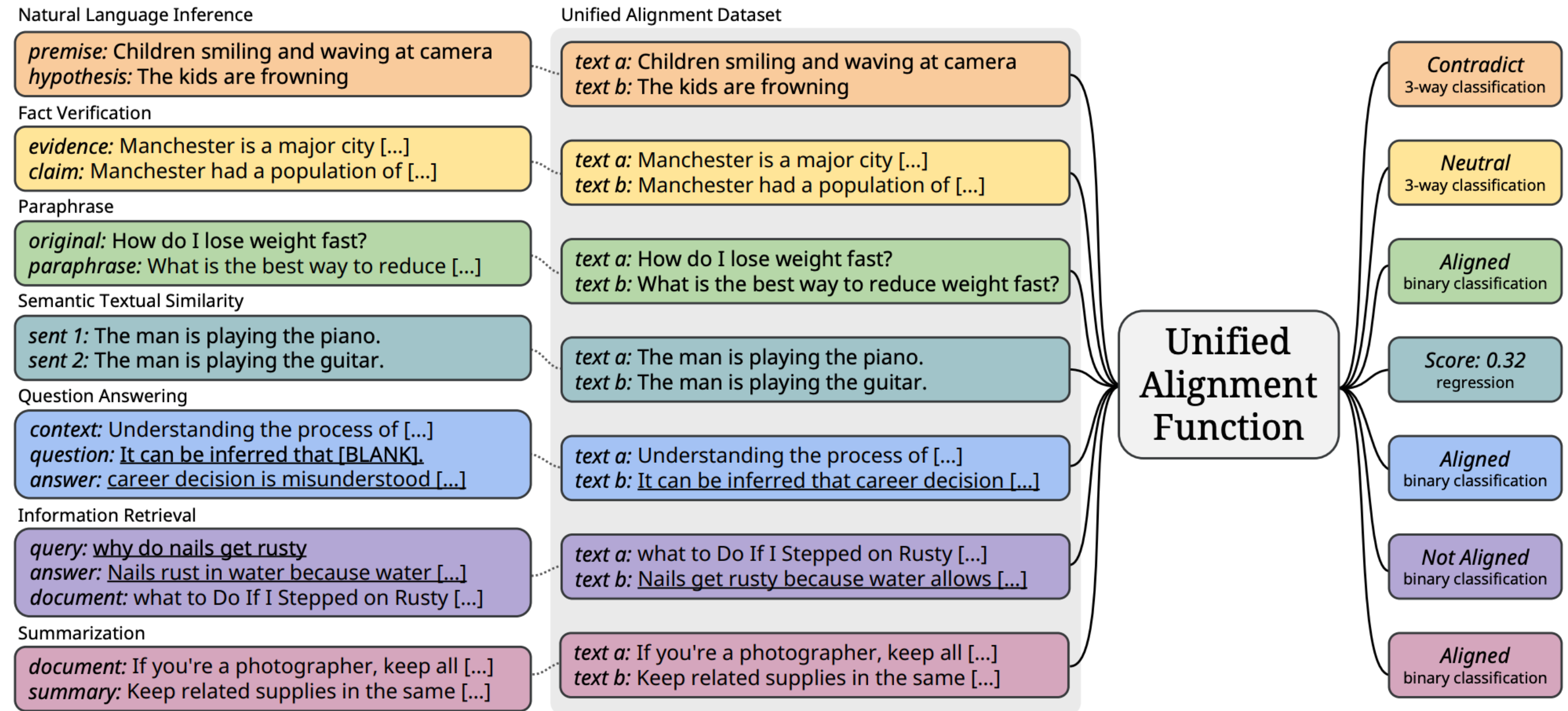
Yuheng Zha Yichi Yang Ruichen Li Zhiting Hu  
UC San Diego  
{yzha, yiy067, rul014, zhh019}@ucsd.edu

### Abstract

Many text generation applications require the generated text to be factually consistent with input information. Automatic evaluation of factual consistency is challenging. Previous work has developed various metrics that often depend on *specific* functions, such as natural language inference (NLI) or question answering (QA), trained on limited data. Those metrics thus can hardly assess diverse factual inconsistencies (e.g., contradictions, hallucinations) that occur in varying inputs/outputs (e.g., sentences, documents) from different tasks. In this paper, we propose ALIGNSCORE, a new holistic metric that applies to a variety of fa-

context (Cao et al., 2018; Kryscinski et al., 2019; Nie et al., 2019a; Tan et al., 2020; Maynez et al., 2020; Deng et al., 2021).

It is thus crucial to develop automatic metrics that evaluate factual consistency of a *claim* (e.g., generated text) with regard to a *context* (e.g., model input). The evaluation, however, has long been a challenge. Recent work has devised various metrics based on specific pretrained functions, such as natural language inference (NLI) (Honovich et al., 2022a; Mishra et al., 2021; Kryscinski et al., 2020; Utama et al., 2022; Laban et al., 2022) and question answering (QA) (Durmus et al., 2020; Fabbri et al., 2022; Honovich et al., 2021; Fabbri et al., 2022).



$$p(y_{3way} = \text{ALIGNED} | o'_i, l'_j)$$

Metric	Backbone	Datasets		
		SE	Q-X	Q-C
G-EVAL-3.5	GPT3.5-d03	38.6	40.6	51.6
G-EVAL-4	GPT4	<b>50.7</b>	53.7	68.5
GPTScore	GPT3.5-d03	47.5	/	/
ChatGPT	GPT3.5-turbo	43.3	/	/
<b>ALIGNSCORE-base</b>	RoBERTa (125M)	43.4	51.9	69.0
<b>ALIGNSCORE-large</b>	RoBERTa (355M)	46.6	<b>57.2</b>	<b>73.9</b>

Table 5: The Spearman correlation coefficients of ALIGNSCORE and LLM-based metrics on SummEval (SE), QAGS-XSum (Q-X) and QAGS-CNNM (Q-C). The best models are shown in **bold**. The results of G-EVAL, GPTScore and ChatGPT are from Liu et al. (2023), Fu et al. (2023), and Gao et al. (2023).

# Quality Diversity through AI Feedback

May 24, 2023 | Blogs

Herbie Bradley<sup>1,2,3</sup>, Andrew Dai<sup>4</sup>, Jenny Zhang<sup>5,6</sup>, Jeff Clune<sup>5,6</sup>, Kenneth Stanley<sup>7</sup>, Joel Lehman<sup>1,2</sup>  
<sup>1</sup>CarperAI, <sup>2</sup>Stability AI, <sup>3</sup>University of Cambridge, <sup>4</sup>Aleph Alpha, <sup>5</sup>University of British Columbia, <sup>6</sup>Vector Institute, <sup>7</sup>Maven

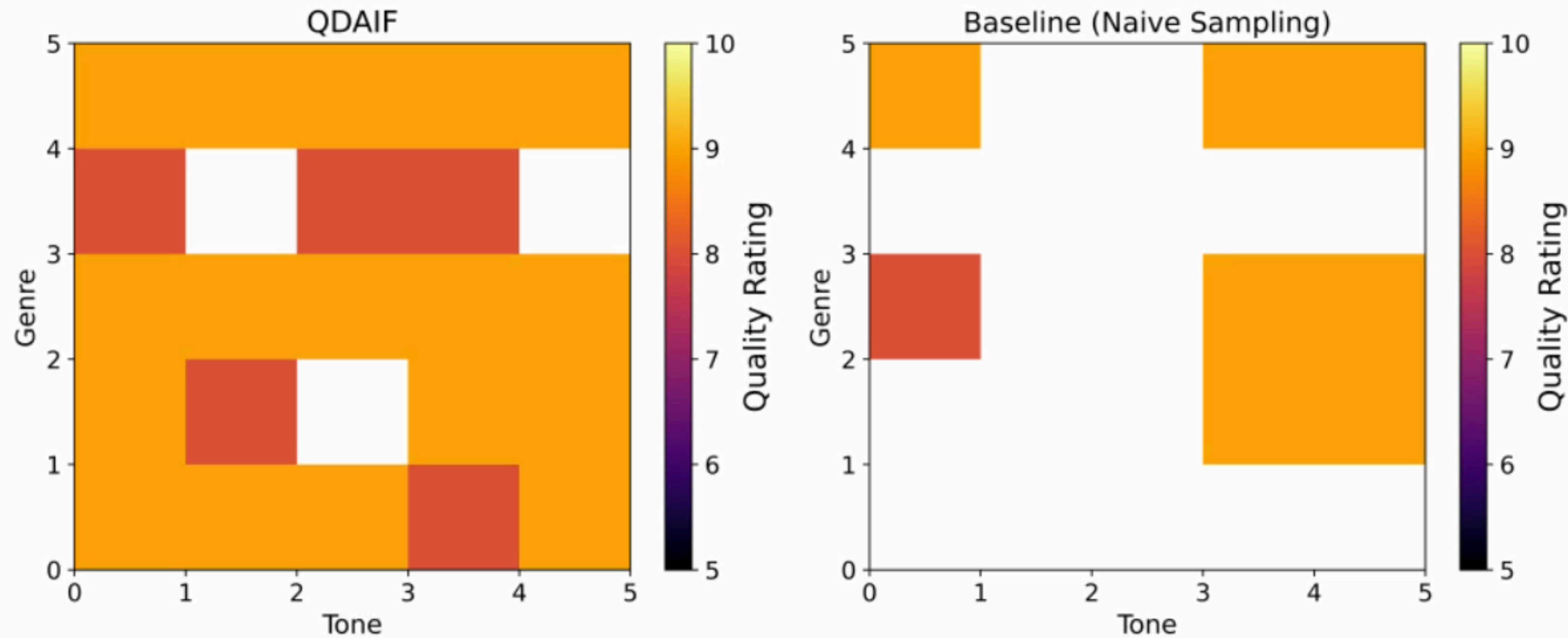


Figure 1: Maps showing the diversity across genre and tone (x and y axes) and quality (color of each grid cell) of generated poems from GPT-4 using our method, QDAIF, compared with a simple independent sampling baseline. The diversity and quality metrics are also obtained from GPT-4. White cells are unfilled.

## Introduction

Human innovation is not only a generative capacity for creativity, but also contains the ability to evaluate the subjective quality of new ideas and

# Scaling Data-Constrained Language Models

25th May 2023

"Given the Chinchilla scaling laws and the trend of training ever-larger models... what should we do when we **run out of data?**"

"we train more than 400 models..."

"...fit a new **data-constrained scaling law** that generalizes the Chinchilla scaling law to the repeated data regime"

"After 40 epochs, repeating is worthless"

"Up to  $\approx 4$  epochs - repeating nearly as good as new data"

## Scaling Data-Constrained Language Models

Niklas Muennighoff<sup>1</sup> Alexander M. Rush<sup>1</sup> Boaz Barak<sup>2</sup> Teven Le Scao<sup>1</sup>  
Aleksandra Piktus<sup>1</sup> Nouamane Tazi<sup>1</sup> Sampo Pyysalo<sup>3</sup> Thomas Wolf<sup>1</sup> Colin Raffel<sup>1</sup>  
<sup>1</sup> Hugging Face <sup>2</sup> Harvard University <sup>3</sup> University of Turku  
n.muennighoff@gmail.com

### Abstract

The current trend of scaling language models involves increasing both parameter count and training dataset size. Extrapolating this trend suggests that training dataset size may soon be limited by the amount of text data available on the internet. Motivated by this limit, we investigate scaling language models in data-constrained regimes. Specifically, we run a large set of experiments varying the extent of data repetition and compute budget, ranging up to 900 billion training tokens and 9 billion parameter models. We find that with constrained data for a fixed compute budget, training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data. However, with more repetition, the value of

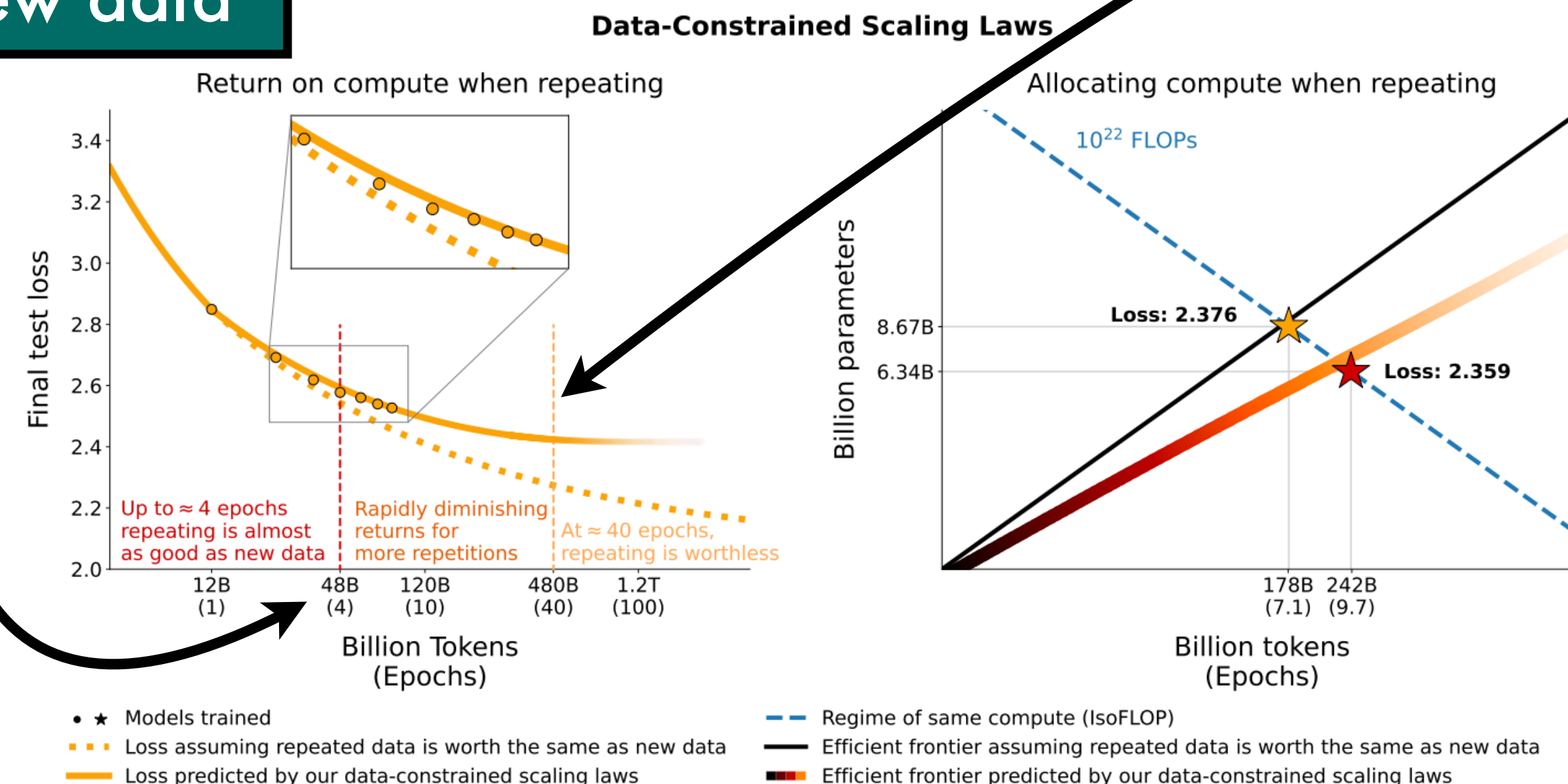
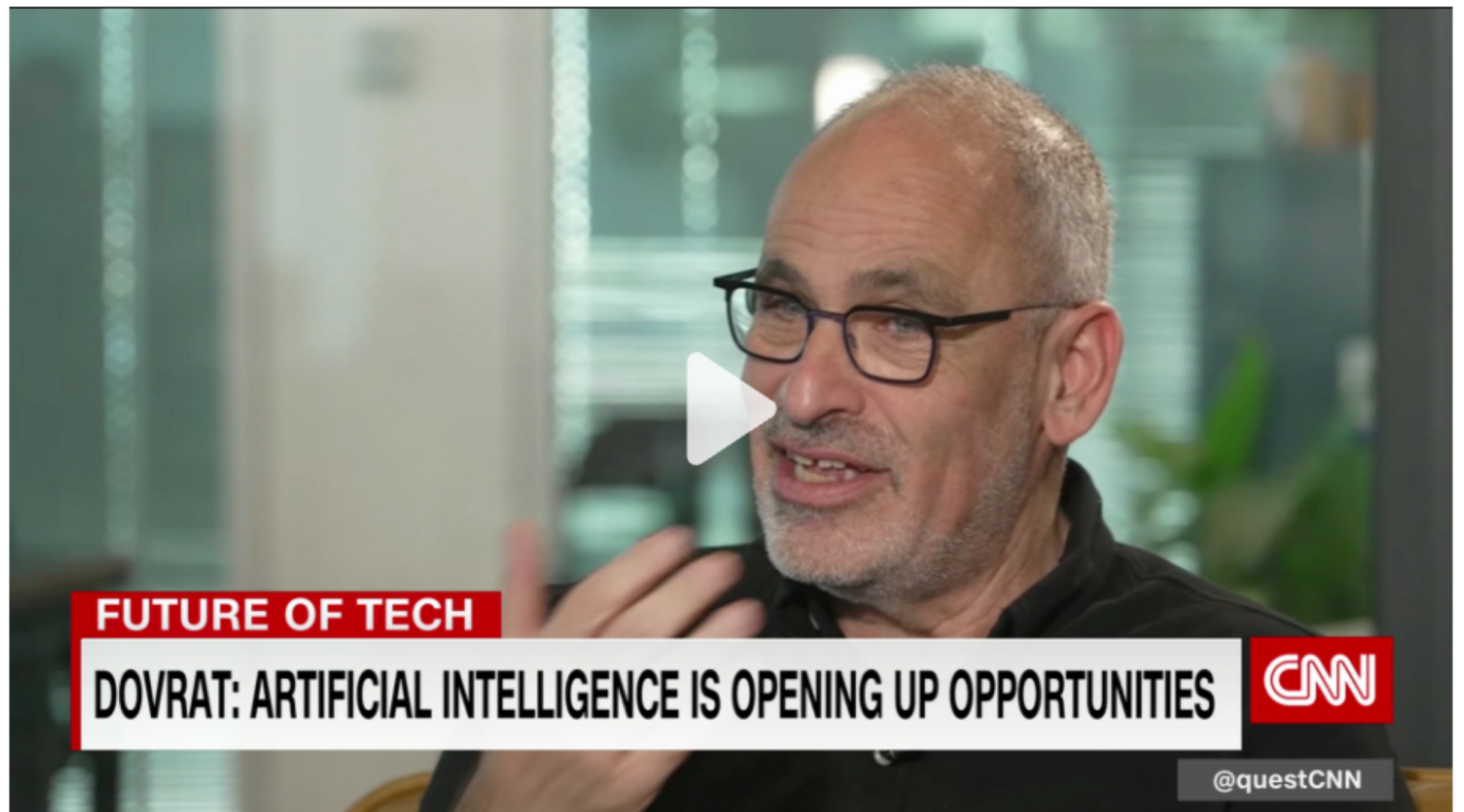


Figure 1: **Return and Allocation when repeating data.** (Left): Loss of LLMs (4.2B parameters) scaled on repeated data decays predictably (§5). (Right): To maximize performance when repeating, our data-constrained scaling laws and empirical data suggest training smaller models for more epochs in contrast to what assuming Chinchilla scaling laws [39] hold for repeated data would predict (§6).

# AI News

## Lawyer apologizes for fake court citations from ChatGPT

By Ramishah Maruf, CNN  
Updated 3:28 PM EDT, Sun May 28, 2023  
f t e l



Dovrat: ChatGPT will change the way we do business  
03:11 - Source: CNN

**New York (CNN)** — The meteoric rise of ChatGPT is shaking up multiple industries – including law, as one attorney recently found out.

Roberto Mata sued Avianca airlines for injuries he says he sustained from a serving cart while on the airline in 2019, claiming negligence. Mata's lawsuit was filed in the Southern District of New York in an order signed by Judge Robert S. Oberman and licensed in New York.

But at least six of the submitted court decisions with bogus quotes were filed in the Southern District of New York in an order signed by Judge Robert S. Oberman and licensed in New York.

**"unaware of the possibility that its content could be false."**

⚡ GPT-3.5 GPT-4

ChatGPT PLUS

GPT-4 currently has a cap of 25 messages every 3 hours.

Send a message...

ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT May 24 Version](#)

**"ChatGPT may produce inaccurate information about people, places or facts."**

"There's only one way to get hired at the tiny corp, and that's by submitting high quality **pull requests** to tinygrad."

## the tiny corp raised \$5.1M

May 24, 2023

Here we go again. I started another company. The money is in the bank.

Mercury Balance 

\$ 5,080,456.93

### What is the tiny corp?

The [tiny corp](#) is a computer company. We sell computers for more than they cost to make; I've been thinking about this one for a while. In the limit, it's a [chip company](#), but there's a lot of intermediates along the way.

The human brain has about 20 PFLOPS of compute. I've written [various blog posts](#) about this. Sadly, 20 PFLOPS of compute is not accessible to most people, costing about \$1M to buy or \$100/hr to rent.

With the way AI is going, we risk large entities controlling the majority of the compute in the world. I do not want "I think there's a world market for maybe five computers." to ever be the world we live in.

The goal of the tiny corp is: **"to commoditize the petaflop"**

### What is tinygrad?

I started [tinygrad](#) in [Oct 2020](#). It started as a toy project to teach me about neural networks, it's now carved out a good niche in the inference space running the model in [openpilot](#), and soon will be a serious competitor to PyTorch in many places.

The main advantage is in the tinygrad IR. It has 12 operations, all of which are ADD/MUL only. `x[3]` is supported, `x[y]` is not. Matrix multiplies and convolutions are just multiplies and sums, surrounded by a bunch of zero cost movement operations (like reshape, permute, expand).

```
# a fast matmul in tinygrad (a@b works also of course)
from tinygrad.tensor import Tensor
N = 2048; a, b = Tensor.randn(N,N), Tensor.randn(N,N)
c = (a.reshape(N,1,N) * b.permute(1,0).reshape(1,N,N)).sum(axis=2)
```

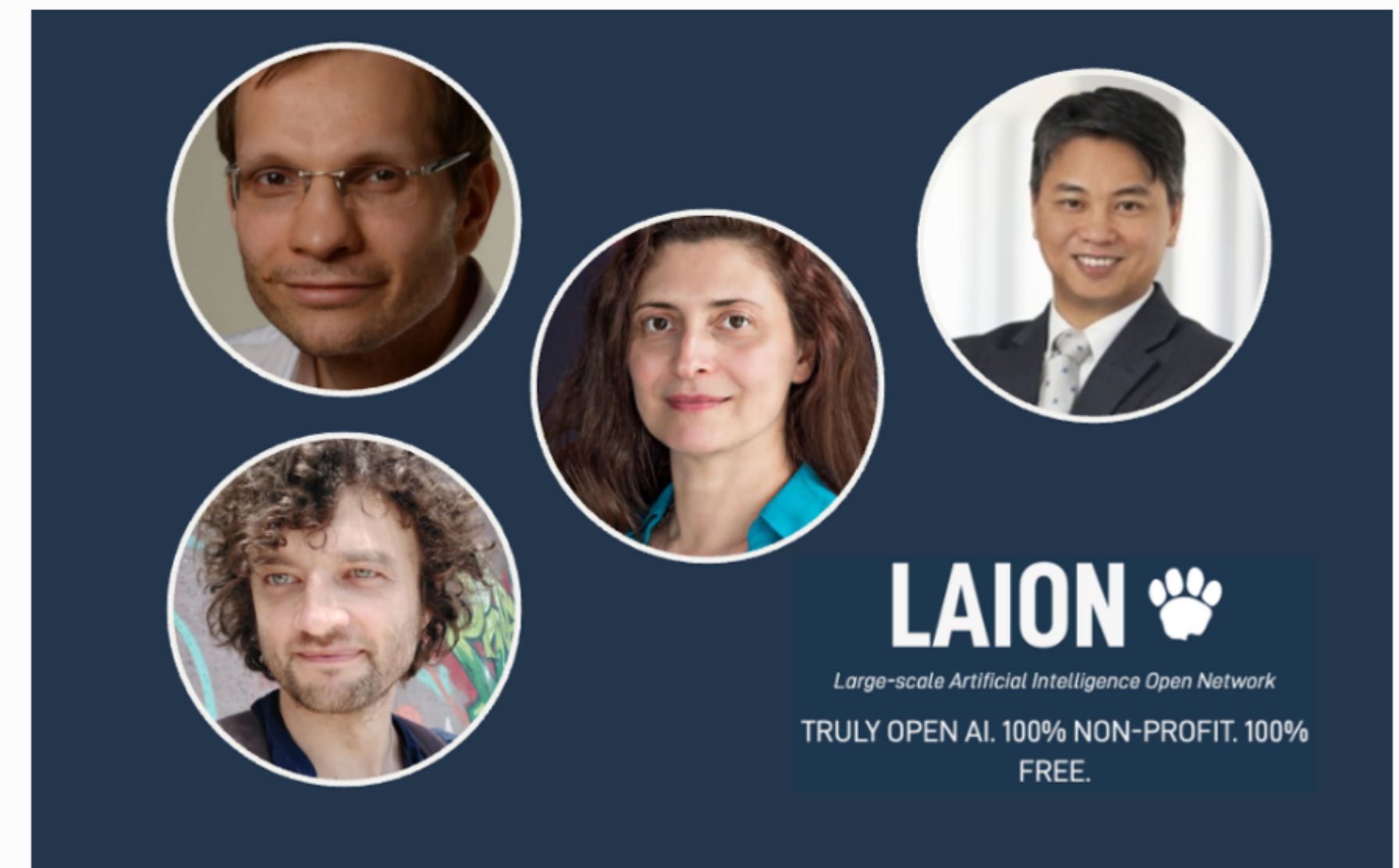
# Amid Growing Call To Pause AI Research, LAION Petitions Governments To Keep AGI Research Open, Active And Responsible

Hessie Jones Contributor 

Strategist, Investor, Advocating for Human-Centered AI, Privacy

Follow

Apr 19, 2023, 08:41am EDT



Christoph Schumann (top left), Jenia Jitsev (bottom left), Irina Rish (middle), Huu Nguyen (top ... [+] LAION

Few outside of the AI Research Community have heard of [LAION](#), a large-scale



# AI News

 GOV.UK

[Home](#) > [Business and industry](#)

News story

## PM meeting with leading CEOs in AI: 24 May 2023

A joint statement between the PM and leading CEOs in Artificial Intelligence (AI) following a meeting to discuss the development of safe and responsible AI.

From: [Prime Minister's Office, 10 Downing Street](#) and [The Rt Hon Rishi Sunak MP](#)

Published 24 May 2023



REUTERS®

[World](#) ▾

[Business](#) ▾

[Markets](#) ▾

[Sustainability](#) ▾

[Legal](#) ▾

[Breakingviews](#)

[Technology](#) ▾

[Investigations](#)

[Mor](#)



Technology



## Nvidia joins \$1 trillion valuation club on booming AI demand

By [Akash Sriram](#) ▾ and [Samrhitha A](#) ▾

May 30, 2023 5:28 PM GMT+1 · Updated 2 hours ago



The logo of NVIDIA as seen at its corporate headquarters in Santa Clara, California, in May of 2022.

Courtesy NVIDIA/Handout via REUTERS

# AI News



TRANSP0 / UBER / RIDE-SHARING

## Uber teams up with Waymo to add robotaxis to its app



/ The two former rivals are ready to let bygones be bygones – in the interest of stirring up more business by getting customers into autonomous vehicles.

By [Andrew J. Hawkins](#), transportation editor with 10+ years of experience who covers EVs, public transportation, and aviation. His work has appeared in The New York Daily News and City & State.

May 23, 2023, 1:00 PM GMT+1 | [0 Comments](#) / [0 New](#)



Waymo’s robotaxis will be available to hail for rides and food delivery on Uber’s app in Phoenix later this year, the result of a new partnership that the two former rivals announced today.

A “set number” of Waymo vehicles will be available to Uber riders and Uber Eats delivery customers in Phoenix, where the Alphabet company



United States

## Deepfaking it: America's 2024 election collides with AI boom

By [Alexandra Ulmer](#) ▾ and [Anna Tong](#) ▾

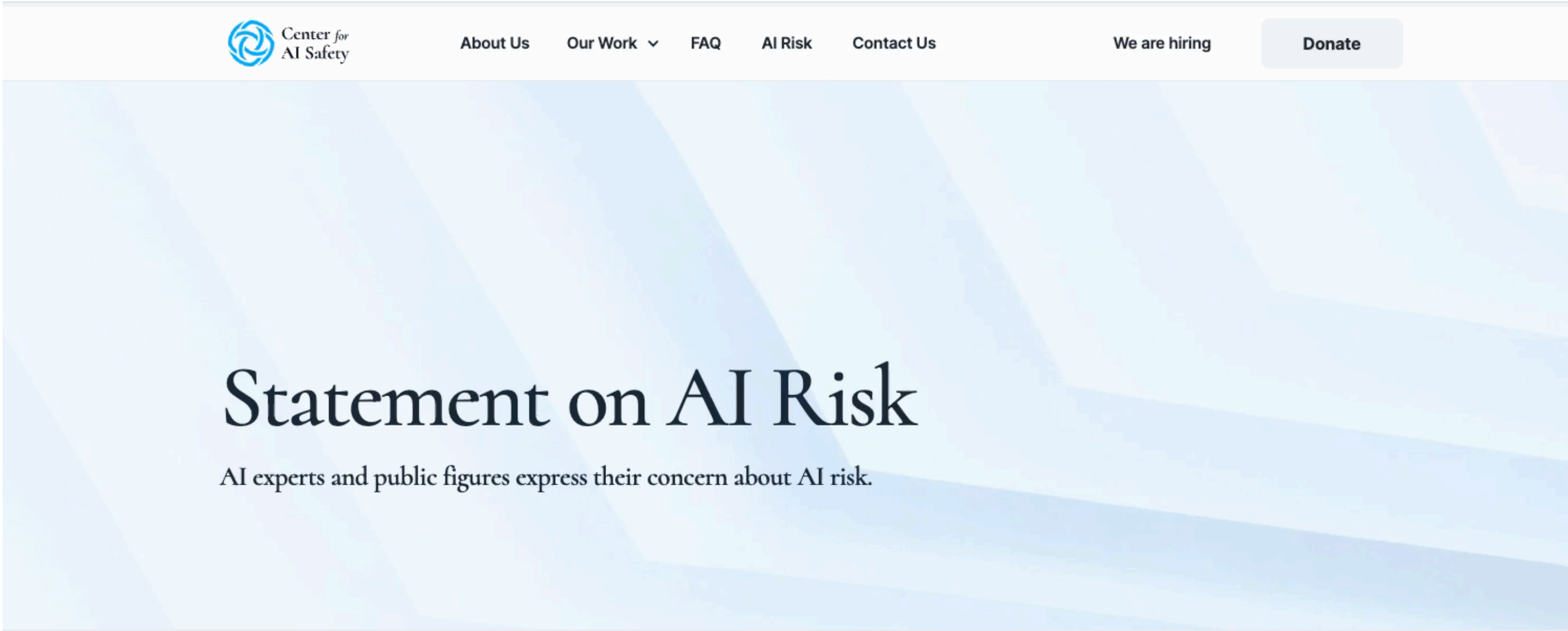
May 30, 2023 1:31 PM GMT+1 · Updated 5 hours ago



# AI Risk

Different views

30th May 2023



**Andrew Ng** @AndrewYNg  
When I think of existential risks to large parts of humanity:  
\* The next pandemic  
\* Climate change→massive depopulation  
\* Another asteroid  
AI will be a key part of our solution. So if you want humanity to survive & thrive the next 1000 years, lets make AI go faster, not slower.  
5:33 PM · May 30, 2023 · 55.2K Views

**Yann LeCun** @ylecun  
Super-human AI is nowhere near the top of the list of existential risks. In large part because it doesn't exist yet.  
  
Until we have a basic design for even dog-level AI (let alone human level), discussing how to make it safe is premature.

**JJ** @JosephJacks\_ · 4h  
I did NOT sign this because AGI fear mongering is nonsensical, toxic and greatly serves the interests of entrenched incumbents.  
  
**Center for AI Safety** @ai\_risks · 9h  
We've released a statement on the risk of extinction from AI.  
Signatories include:

## Contents

- Statement
- Signatories
- Sign the statement

AI experts, journalists, policymakers, and the public are increasingly discussing a broad spectrum of important and urgent risks from AI. Even so, it can be difficult to voice concerns about some of advanced AI's most severe risks. The succinct statement below aims to overcome this obstacle and open up discussion. It is also meant to create common knowledge of the growing number of experts and public figures who also take some of advanced AI's most severe risks seriously.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks

Samuel's view

Agree

may benefit incumbents

Disagree

nonsensical

# SafeTensors

23rd May 2023

## Safetensors audited as really safe and becoming the default

May 23, 2023 · Nicolas Patry, Stella Biderman, Garry Jean-Baptiste



×



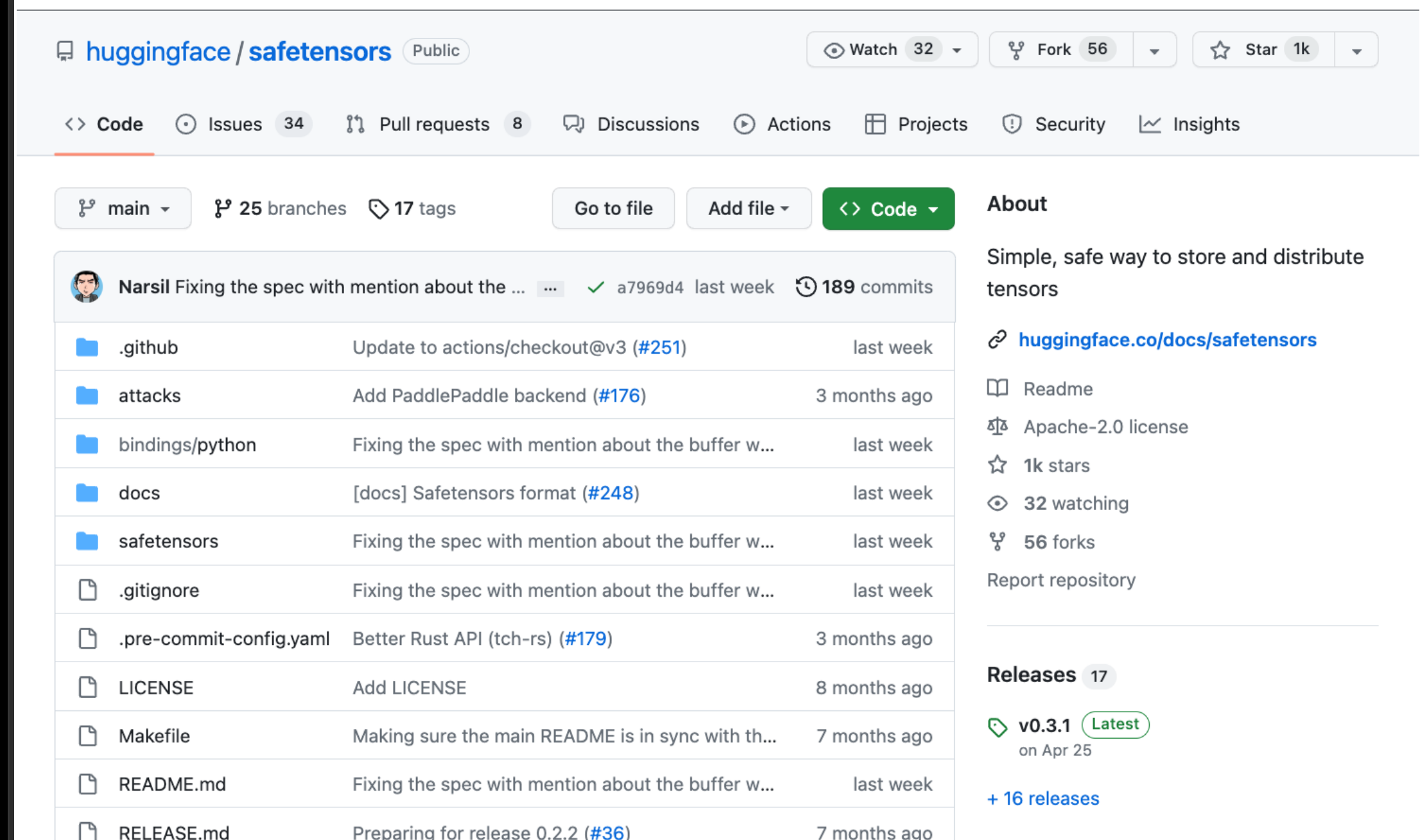
×



**SAFETENSORS**  
ML Safer For All

**Audit shows that safetensors is safe and ready to become the default**

Hugging Face, in close collaboration with EleutherAI and Stability AI, has ordered an external security audit of

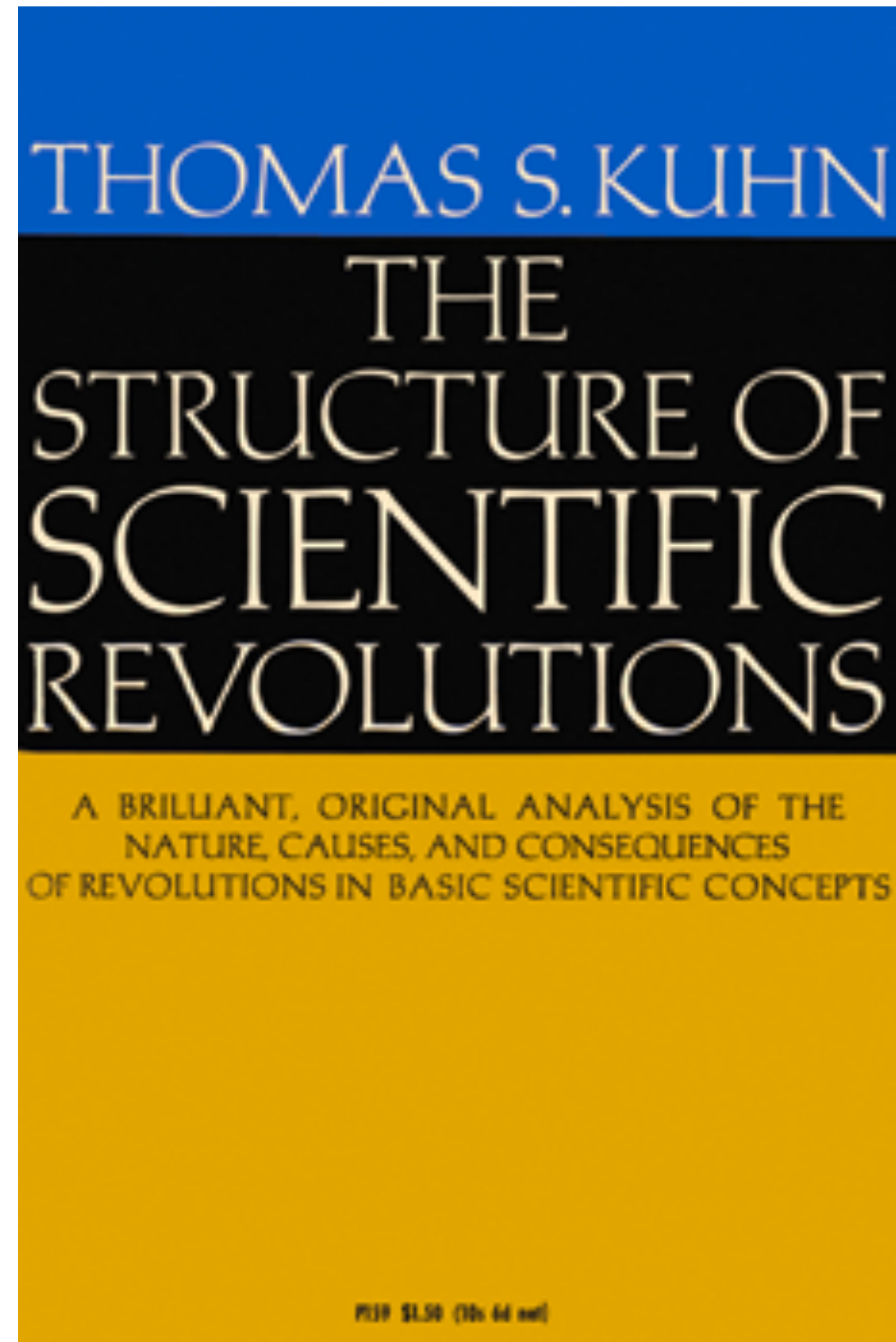


The screenshot shows the GitHub repository page for `huggingface/safetensors`. The repository is public and has 32 watchers, 56 forks, and 1k stars. It features a navigation bar with links to Code, Issues (34), Pull requests (8), Discussions, Actions, Projects, Security, and Insights. The repository is on the `main` branch, with 25 other branches and 17 tags. A commit by Narsil is highlighted, titled "Fixing the spec with mention about the ...", with commit hash `a7969d4` and 189 commits. The file list includes `.github`, `attacks`, `bindings/python`, `docs`, `safetensors`, `.gitignore`, `.pre-commit-config.yaml`, `LICENSE`, `Makefile`, `README.md`, and `RELEASE.md`. The right sidebar shows the repository's description: "Simple, safe way to store and distribute tensors", a link to the documentation, and release information for `v0.3.1` (Latest) on Apr 25.

# State of GPT - Andrej Karpathy



# Samuel's Book Recommendation



Unsolicited book recommendation

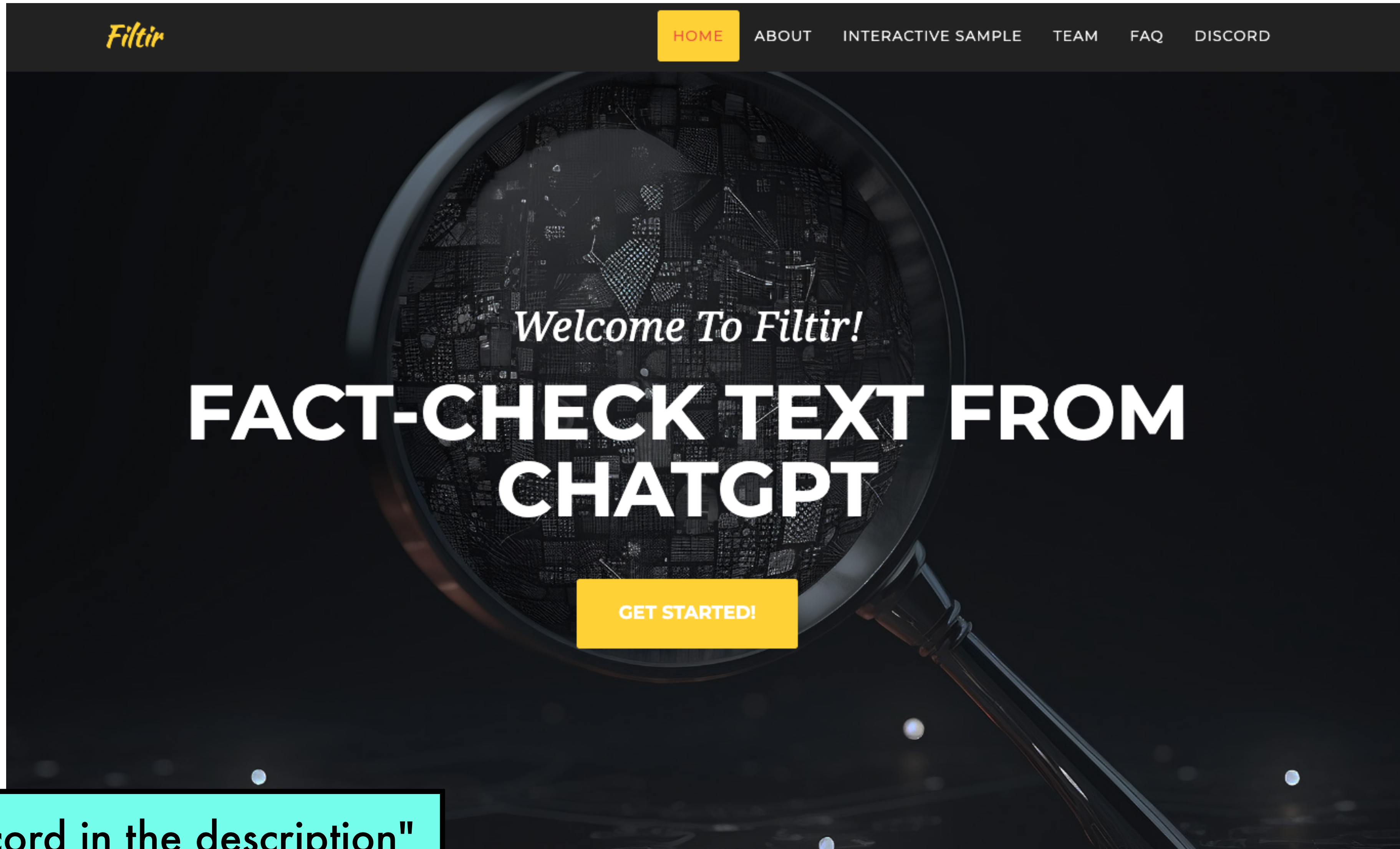
"The Structure of Scientific Revolutions"

Thomas S Kuhn (1962)

**What is it?** A perspective on how **scientific progress** evolves over time:

Rather than linear accumulation of knowledge, science is **episodic**, with "normal" periods punctuated by "revolutions" with big changes

# Filtir - fact-checking AI claims



"Link to discord in the description"