

Ethiek en AI

Waar kun je op letten?

Hidde de Vries, 10 december 2025



DigiToegankelijk





COTHAM HOUSE

Philosophy

School of Arts



University of
BRISTOL

Filosofie en KI

Wat is kennis?

Hoe werkt taal?

Wat is waar?

Wat is mooi?

Ben ik omdat ik denk?

“kritisch”



ethiek:

hoe te handelen?

wat te doen?

Definitie van AI
(voor deze presentatie)

grote **taalmodellen** die zijn getraind
om tekst, beeld of code te kunnen
genereren op basis van **prompts**

over AI wil
je weten...

**Hoe het wordt
gemaakt**

**Hoe het wordt
verkocht**

**Hoe het wordt
gebruikt**





KABELJAUW
FILET

5 stuks

ALTIJD
LAGE
PRIJS

JUMBO

KABELJAUW
MET ZIJN SMAAK

MALSE
TONIJN
STEAK

2 stuks

KABELJAUWFILET

VLOCCIE MET
TONIJNSTEAK

± 5 MIN.
KLAAR

KABELJAUW
HAAS
PORTER

training data

AI bedrijven gebruiken heel veel
boeken, muziek, beelden en
tekst **zonder toestemming**

DOWNLOAD A
TRAINING DATA

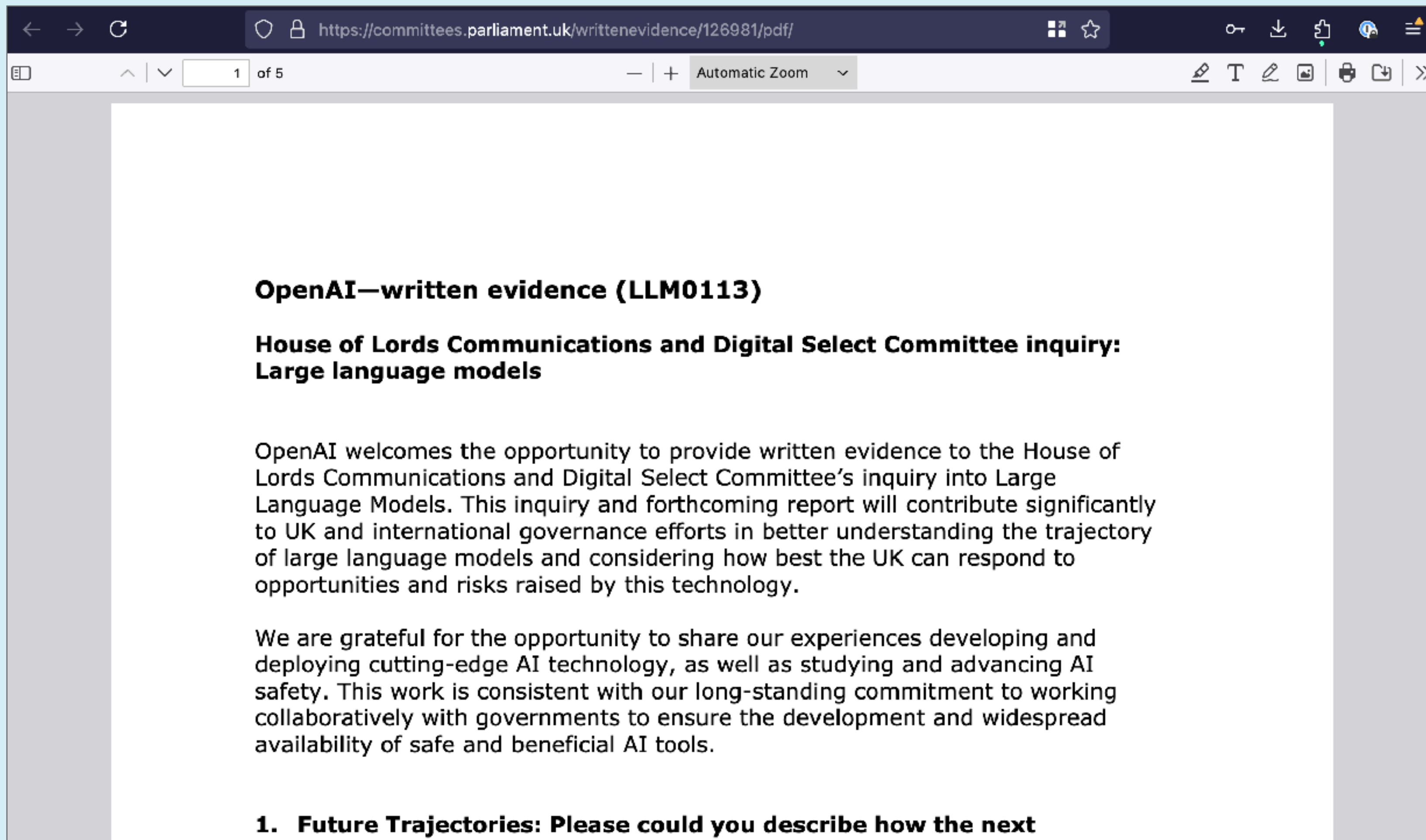


YOU WOULDN'T
DOWNLOAD A
TRAINING DATA



OpenAI, tegen UK House of Lords:

- “we dachten dat trainen niet onder copyright viel”
- “we kunnen ons bedrijf niet runnen zonder copyrighted materiaal”



The screenshot shows a web browser displaying a PDF document. The address bar shows the URL: <https://committees.parliament.uk/writtenevidence/126981/pdf/>. The document title is "OpenAI—written evidence (LLM0113)". The subtitle is "House of Lords Communications and Digital Select Committee inquiry: Large language models". The text of the document is as follows:

OpenAI welcomes the opportunity to provide written evidence to the House of Lords Communications and Digital Select Committee’s inquiry into Large Language Models. This inquiry and forthcoming report will contribute significantly to UK and international governance efforts in better understanding the trajectory of large language models and considering how best the UK can respond to opportunities and risks raised by this technology.

We are grateful for the opportunity to share our experiences developing and deploying cutting-edge AI technology, as well as studying and advancing AI safety. This work is consistent with our long-standing commitment to working collaboratively with governments to ensure the development and widespread availability of safe and beneficial AI tools.

1. Future Trajectories: Please could you describe how the next



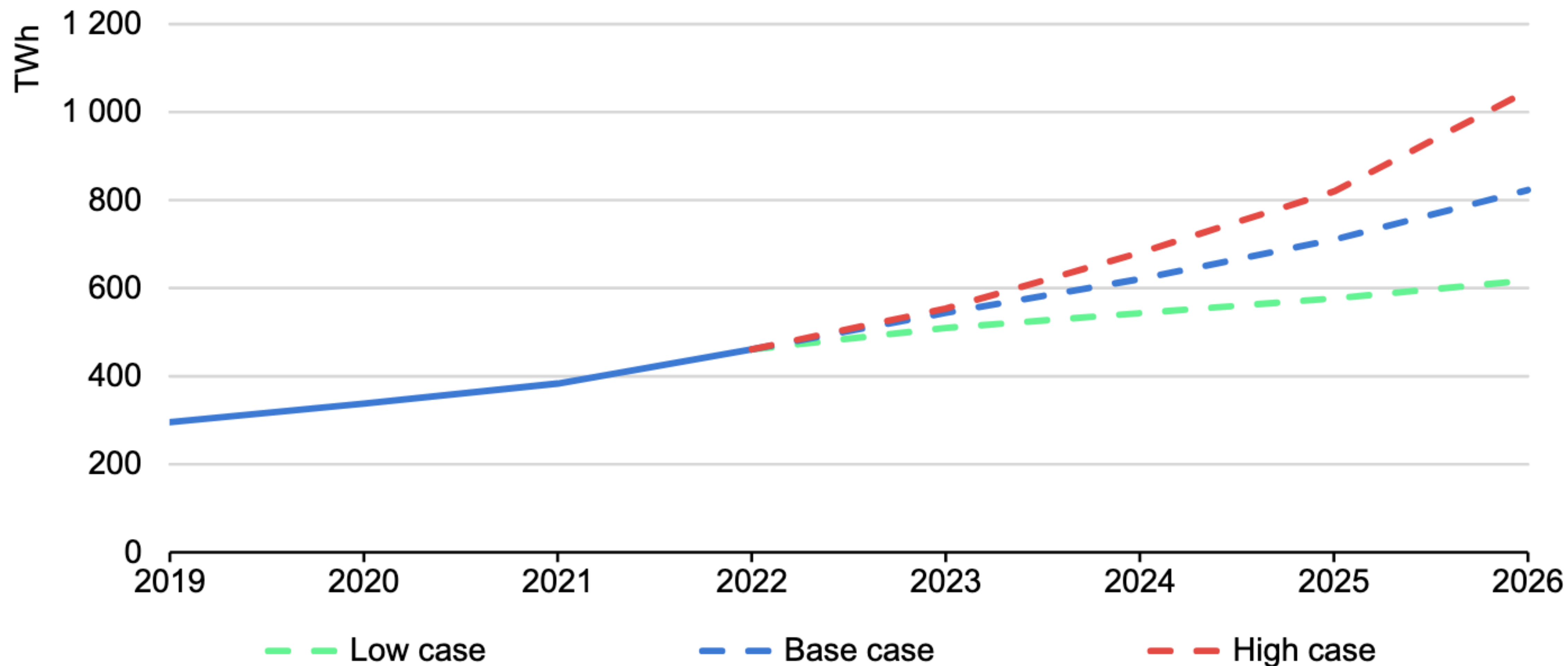
The Exploited Labor Behind Artificial Intelligence

Supporting transnational worker organizing should be at the center of the fight for “ethical AI.”



materiaal- en stroomgebruik

Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026



IEA. CC BY 4.0.

**“Emissions would increase
by 60× for a GPT-3 style
model**

– Wim Vanderbauwhede

“Emissions would increase by 60× for a GPT-3 style model; for a GPT-4 style model, it could be 200×.”

– Wim Vanderbauwhede

classificatie en discriminatie



A collage image featuring David Bowie on the left, a large 'RECORDS' sign at the top, and a view of a record store interior with various vinyl records and CDs displayed on shelves and walls.



Please Ask for Assistance

er zitten **waarden** in
(technologie is nooit neutraal)

Copilot stops working on gender related subjects #72603

Unanswered

 jebarjonet asked this question in Copilot Conversations

jebarjonet on Oct 26, 2023

edited ▾ ⋮

As some people already mentioned [here](#) or [here](#), Copilot purposely stops working on code containing hardcoded banned words from Github such as gender or sex .

I am labelling this as a bug because this behavior is unexpected and undocumented.

I guess you might be embarrassed by what your AI says when it autocompletes gender related topics. However disabling autocompletion is not a great solution since gender can be very present in :

- medical, administrative or demographics studies
- translations in many languages (not much in English though), such as latine languages where nouns do have genders
- sentences generation (for documentation, games, etc.) in those languages

For which Copilot shutting down on most of the files is deceptive and pretty annoying.

I don't have any elegant solution in mind and I know this may seems like an edge case but I hope this will be taken into account someday

Category



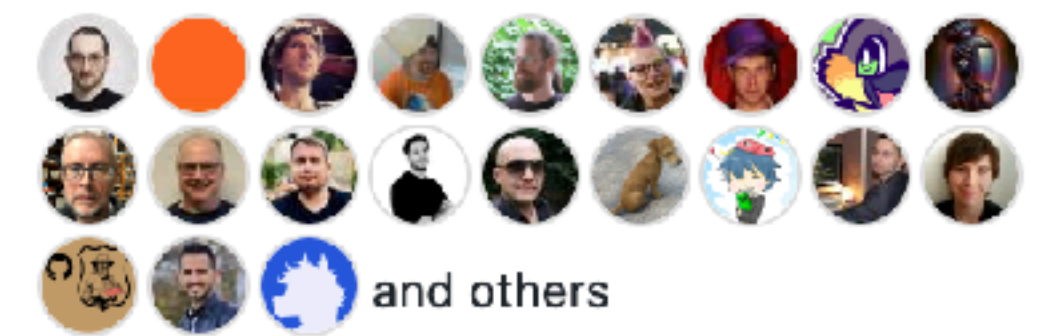
Copilot Conversations

Labels

Bug

Copilot

44 participants



Notifications

 Subscribe



**“[it’s] effectively
giving [them] an
“ideology” dial for
your writing and
messaging.”**

Baldur Bjarnason, “Poisoning for propaganda”



“[it’s] effectively giving

ChatGPT's Political Views Are Shifting Right, a New Analysis Finds

While the chatbot still seems to lean left on many issues, its bias is changing.

By Todd Feathers Published February 11, 2025 | Comments (21)



BEST OF IFA

Gizmodo

BEST OF

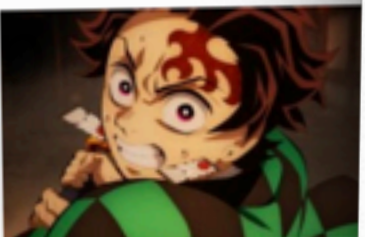
IFA 2025

AWARDS

LATEST NEWS

The Latest 'Demon Slayer' Movie is Another Record-Breaking Hit

9/14/2025, 4:15 pm



Chatbot Grok verheerlijkt Holocaust en noemt zichzelf 'MechaHitler': 'Musk speelt met de regels'

Rick Hartkamp • 11 juli 2025 10:15

Eén van de nieuwste speeltjes van Elon Musk, AI-chatbot Grok, is behoorlijk op hol geslagen. Het spuugde antisemitische teksten uit en prees

Hoe het gemaakt wordt

Training op grotendeels
gestolen data.

Data bevat vooroordelen.

“Data-taggers” met
psychische schade.

Hoe het gemaakt wordt

Training op grotendeels
gestolen data.

Data bevat vooroordelen.

“Data-taggers” met
psychische schade.

Benodigt grote
hoeveelheden
zeldzame metalen.

Gebruikt vele malen
meer stroom en water
dan alternatieven.

Hoe het gemaakt wordt

Training op grotendeels
gestolen data.

Data bevat vooroordelen.

“Data-taggers” met
psychische schade.

Benodigt grote
hoeveelheden
zeldzame metalen.

Gebruikt vele malen
meer stroom en water
dan alternatieven.

Modellen “bevatten”
waarden van
bedrijven met best
extreme CEO's.

**Hoe het wordt
gemaakt**

**Hoe het wordt
verkocht**

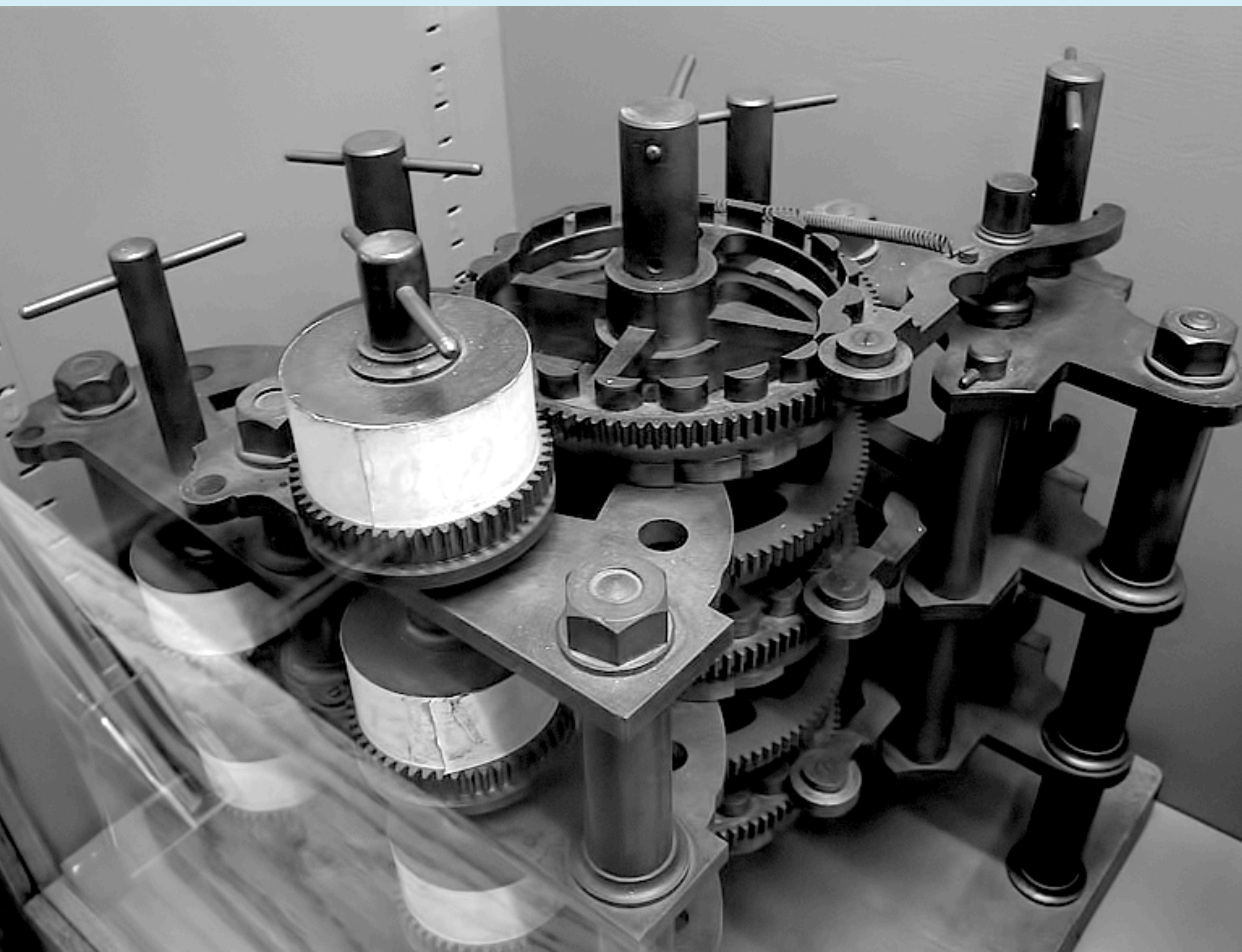
**Hoe het wordt
gebruikt**



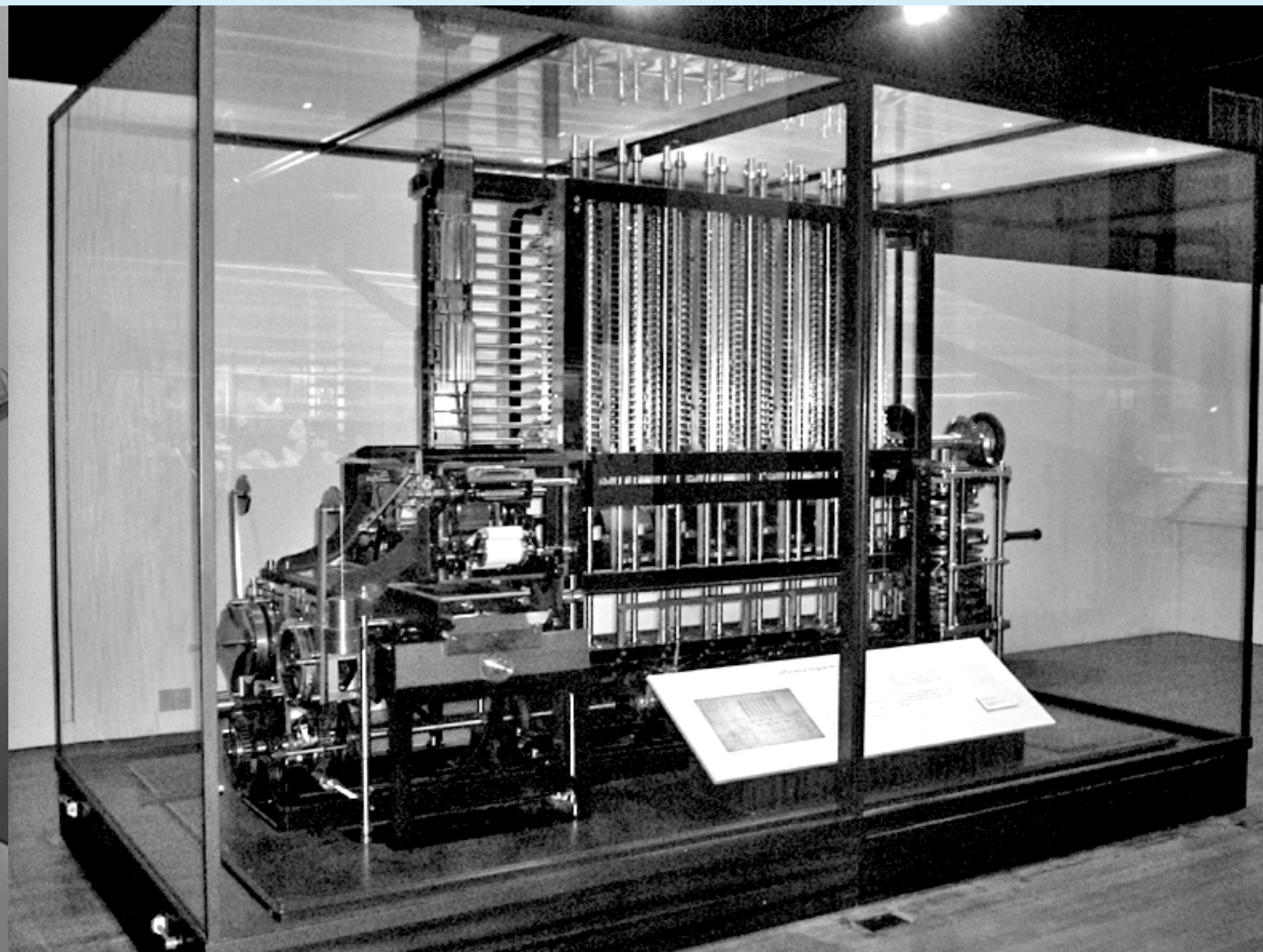
Charles
Babbage



The Difference Engine (1820s)



The Analytical Engine (1830s)



“It is desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine.”

– Ada Lovelace, in Note G



A Proposal for the
DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

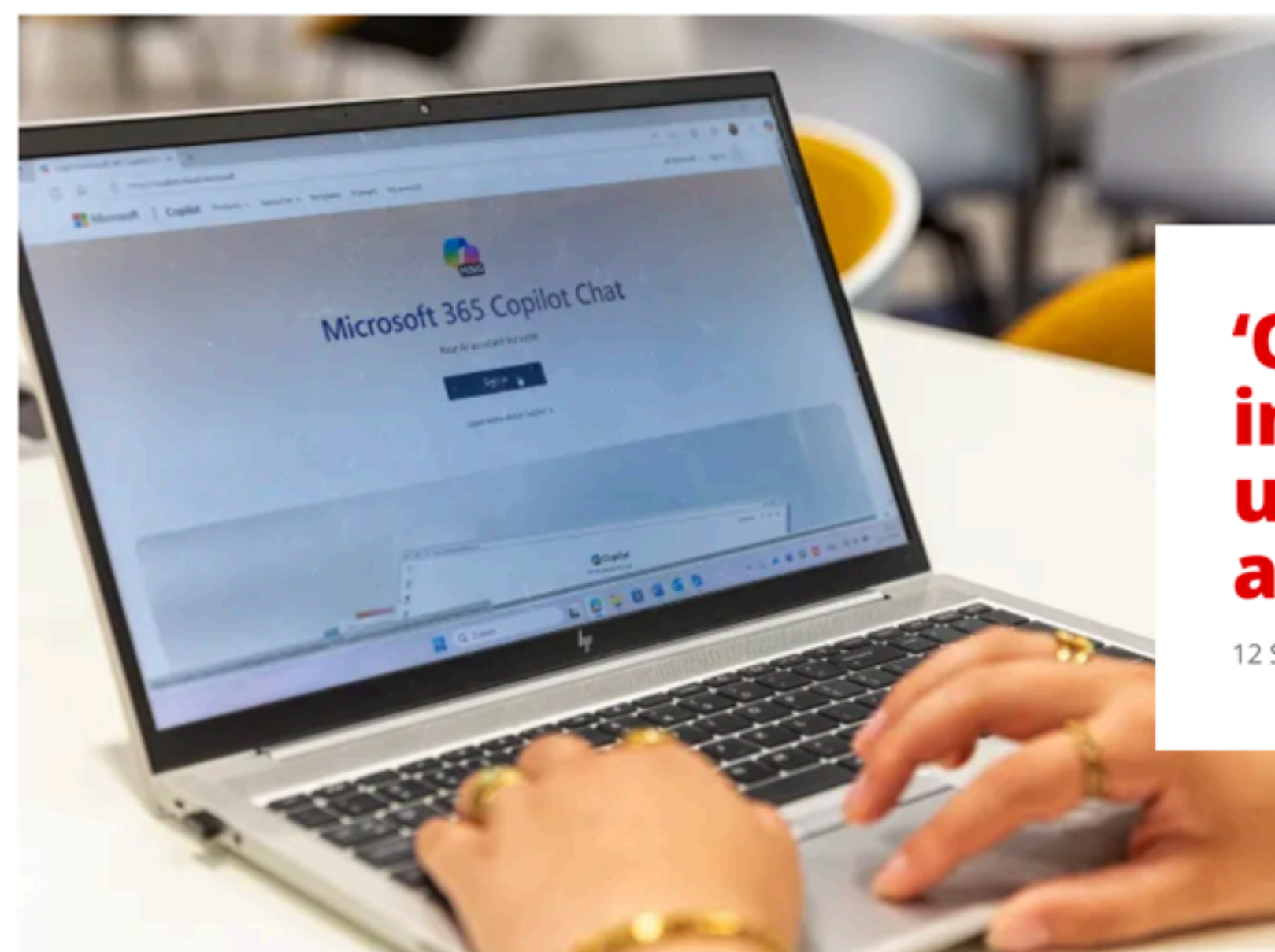
We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

Ja, er zijn interessante en
effectieve toepassingen.

Maar... overdrijving en hype
liggen steeds op de loer.

“Het is onvermijdelijk”

“Het is onvermijdelijk”
voor AI-bedrijven



'Opposing the inevitability of AI at universities is possible and necessary'

12 September 2025 • Research news item

Since the widespread release of ChatGPT in December of 2022, AI has

Against the Uncritical Adoption of 'AI' Technologies in Academia

AGAINST THE UNCRITICAL ADOPTION OF 'AI' TECHNOLOGIES IN ACADEMIA

Olivia Guest^{1,2}, Marcela Suarez¹, Barbara C. N. Müller³, Edwin v. Meerkerk⁴, Arnoud Oude Groote Beverborg⁵, Ronald de Haan⁶, Andrea Reyes Elizondo⁷, Mark Blokpoel^{1,2}, Natalia Scharfenberg Annelies Kleinherenbrink^{1,8}, Ileana Camerino¹, Marieke Woensdregt^{1,2}, Dagmar Monett⁹, Jed Brown¹⁰, Lucy Avraamidou Juliette Alenda-Demoutiez¹², Felienne Hermans¹³, and Iris van Rooij^{1,2,14}

¹Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands

²Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

³Communication & Media, Behavioural Science Institute, Radboud University, The Netherlands

⁴Radboud Institute for Culture and History, Radboud University, The Netherlands

⁵Pedagogical and Educational Sciences, Radboud University, The Netherlands

⁶Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands

⁷CWTS & LUCAS, Faculties of Social Sciences & Humanities, Leiden University, The Netherlands

⁸Gender and Diversity, Radboud University, The Netherlands

⁹Computer Science Dept., Berlin School of Economics and Law, Germany

¹⁰Department of Computer Science, University of Colorado, USA

¹¹Institute for Science Education and Communication, University of Groningen, The Netherlands

¹²Department of Economics and Business Economics, Radboud University, The Netherlands

¹³Vrije Universiteit Amsterdam, The Netherlands

¹⁴Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Denmark

Abstract: *Under the banner of progress, products have been uncritically adopted or even imposed on users — in past centuries with tobacco and combustion engines, and in the 21st with social media. For these collective blunders, we now regret our involvement or*



Is er wel een tekort aan intelligentie?

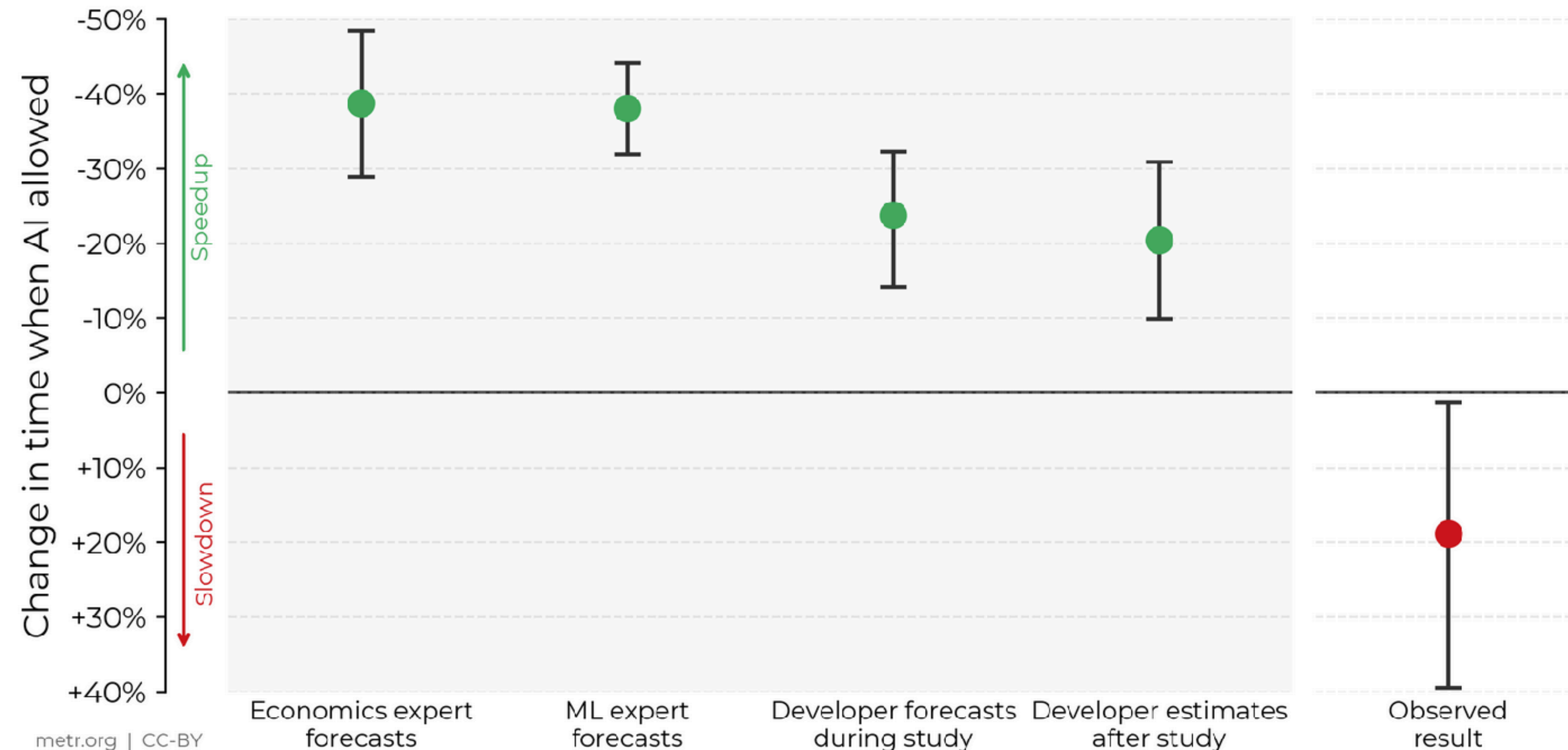
(Feliene Hermans bij Buitenhof)

“Het bespaart tijd”

Against Expert Forecasts and Developer Self-Reports, Early-2025 AI Slows Down Experienced Open-Source Developers



In this RCT, 16 developers with moderate AI experience complete 246 tasks in large and complex projects on which they have an average of 5 years of prior experience.



<https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>

PUBLIC SECTOR

UK government trial of M365 Copilot finds no clear productivity boost

65 💬

AI tech shows promise writing emails or summarizing meetings. Don't bother with anything more complex

🔥 [Paul Kunert](#)

Thu 4 Sep 2025 // 09:15 UTC



A UK government department's three-month trial of Microsoft's M365 Copilot has revealed no discernible gain in productivity – speeding up some tasks yet making others slower due to lower quality outputs.

The Department for Business and Trade received 1,000 licenses for use between October and December 2024, with the majority of these allocated to volunteers and 30 percent to randomly selected participants. Some 300 of these people consented to their data being analyzed.

An evaluation of time savings, quality assurance, and productivity was then calculated in the [assessment](#).

Overall, 72 percent of users were satisfied or very satisfied with their digital assistant and voiced disappointment when the test ended. However, the reality of productivity gains was more nuanced than Microsoft's marketing materials might suggest.



Microsoft promises

Copilot will be a

“Anders gaan we achterlopen”

“Dat wordt nog wel opgelost”

Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

Santosh S. Vempala†
Georgia Tech

Edwin Zhang
OpenAI

September 4, 2025

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such “hallucinations” persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This “epidemic” of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

1 Introduction

Language models are known to produce overconfident, plausible falsehoods, which diminish their utility. This phenomenon is known as “hallucination” though it differs fundamentally from the

“[Hij/zij] [denkt/vindt]...”

Hoe het verkocht wordt

Een best groot deel van het enthousiasme is marketing (al sinds 1956).

Soms blind vertrouwen in dat alles steeds beter wordt, dat is niet zeker.

Antropo-
morphisme.

**Hoe het wordt
gemaakt**

**Hoe het wordt
verkocht**

**Hoe het wordt
gebruikt**



discriminatie en bias

FT AI medical tools downplay symp X

+

← → ↺

https://www.ft.com/content/128ee880-acdb-42fb-8bc0-ea9b71ca11a8

📄 🗖️ ⭐

🔑 ⬇️ 📁 ⓘ ☰

☰ 🔍

FINANCIAL TIMES

Subscribe Sign In

HOME WORLD US COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI

Health + Add to myFT

AI medical tools downplay symptoms in women and ethnic minorities

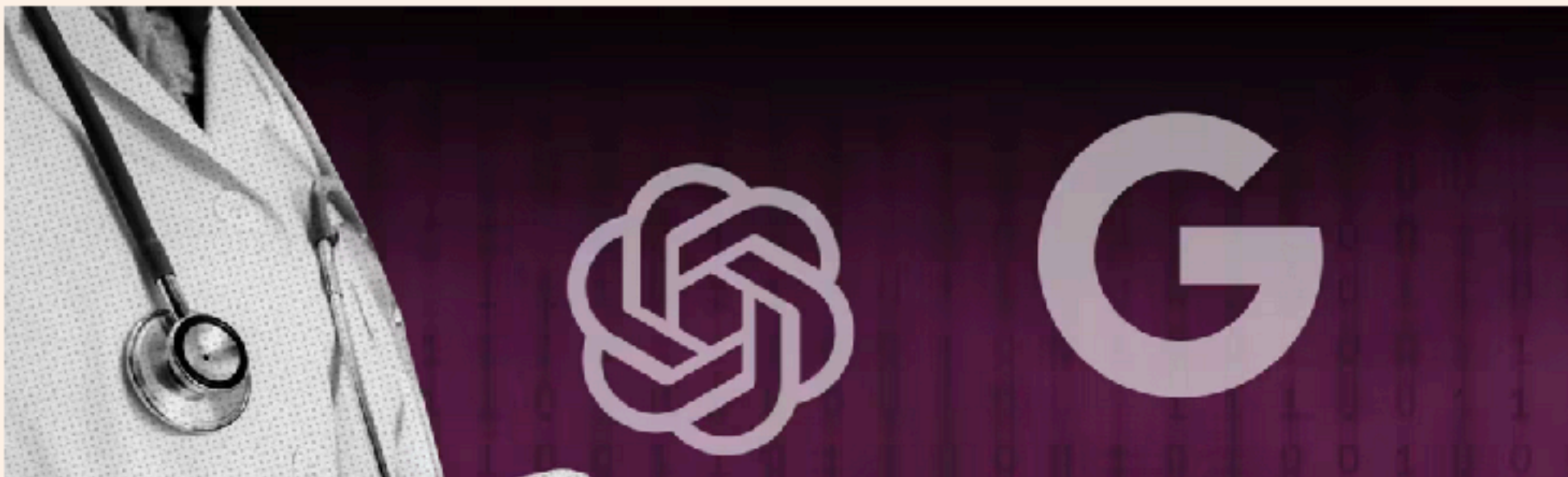
Large language models reflect biases that can lead to inferior healthcare advice to female, Black and Asian patients

✕

f

in

🔖 Save

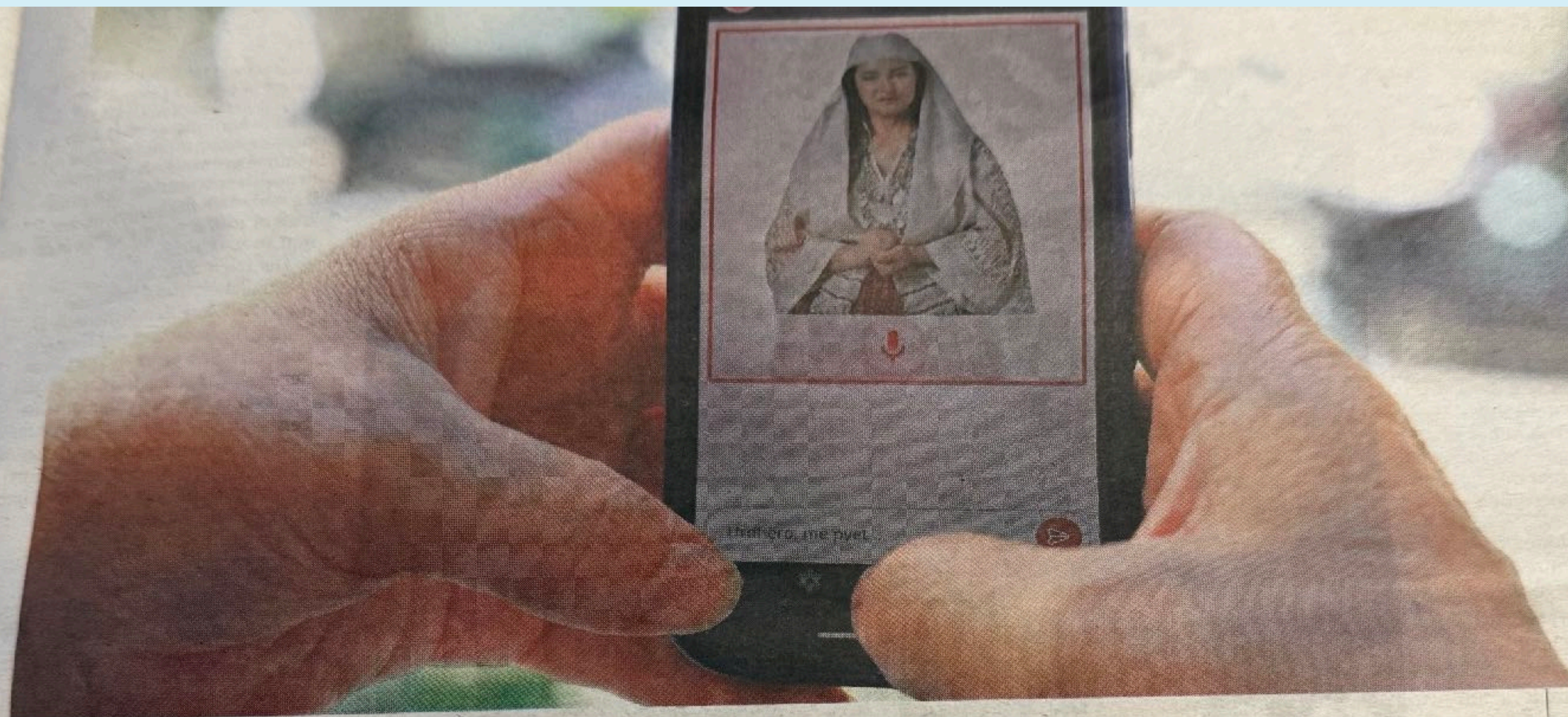


The findings by researchers at leading US and UK universities suggest that medical [AI](#) tools powered by LLMs have a tendency to [not reflect the severity of symptoms among female patients, while also displaying less “empathy” towards Black and Asian ones.](#)

ft.com/content/128ee880-acdb-42fb-8bc0-ea9b71ca11a8

subjectiviteit verpakt
als objectiviteit

Albanië: minister van aanbestedingen



Chatbot Diella is de eerste AI-minister ter wereld.
FOTO AFP

De nieuwste Albanese minister is geen mens, maar een chatbot

Kunstmatige intelligentie | De Albanese Diella kreeg deze week promotie: ze is de eerste overheidschatbot die een ministerpost krijgt. Is dat idee net zo absurd als het

devaluatie van kwaliteit

ethiek:

hoe te handelen?

wat te doen?

“Ethiek
daargelaten...”

“Ethiek

daargelaten...”

De praktijk

Modellen

**Overweeg
makers, grootte,
en inputdata**

AI als oplossing
**Vaak is het
antwoord ook
“zonder AI”**

Hypeloze AI
**Bedenk wat
burgers helpt**

Bias

**Wees alert op
bevooroordeelning**

**Slides +
alle links**



Bedankt!