# AI -> Machine Learning -> Generative AI

A very, very brief overview.

# AI

- Symbolic AI (1950-2000s) – hand coded rules.

  Eliza '66 - rules encoded within a pattern matching script to keep a dialog going (early NLP).

  DeepBlue '97 – tree search plus handcrafted rules to beat Garry Kasparov (brute force).

- Machine Learning (2010–today) – learning the rules from data.

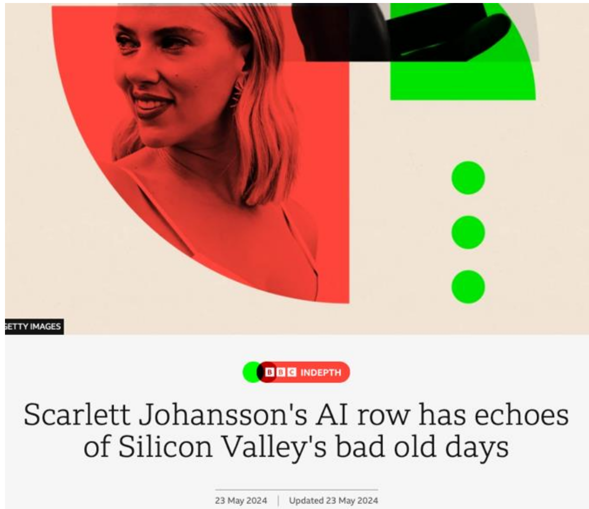
  Deep Neural Nets enabled by big compute x data, spot patterns we can't.

  Facial recognition, Weather Forecasting, Self driving cars, Siri & Alexa

- ML works because it's good at spotting patterns, but it's only as good as the quality of data and sophistication of the algorithms used.

# Generative AI

- Specialised form of Deep Learning - Transformer or diffusion based deep nets.
- Creativity - models output new text, code, images, audio–even protein sequences.
- Transformers, diffusion models, and GANs learn the underlying data distribution and sample fresh artefacts.
- Post-2017 Transformer breakthrough + massive GPUs → GPT-4o, Gemini, Claude 3, DeepSeek, Stable Diffusion, Sora, etc.
- You may have noticed some coverage and media attention…

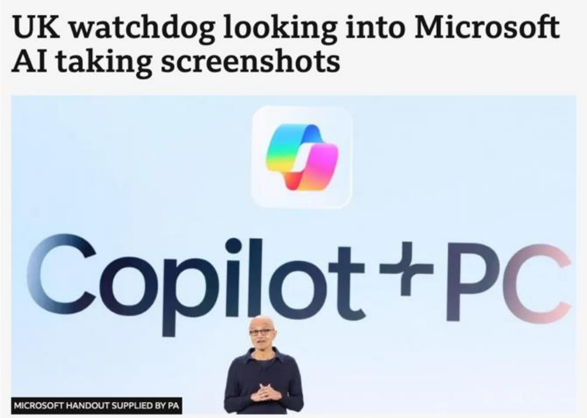


GETTY IMAGES

BBC INDEPTH

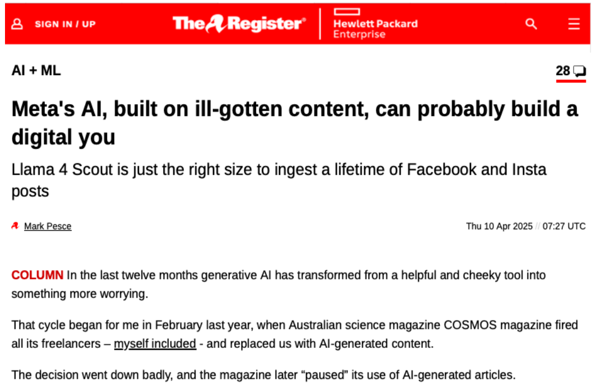
# Scarlett Johansson's AI row has echoes of Silicon Valley's bad old days

23 May 2024 | Updated 23 May 2024



# UK watchdog looking into Microsoft AI taking screenshots

MICROSOFT HANDOUT SUPPLIED BY PA



AI + ML 28

# Meta's AI, built on ill-gotten content, can probably build a digital you

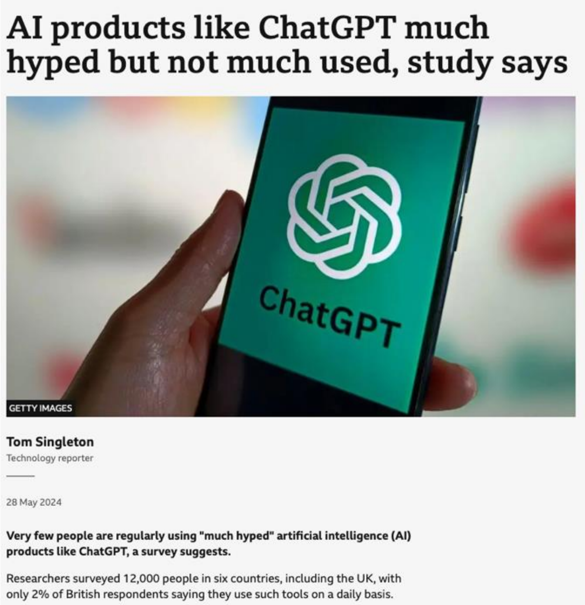
Llama 4 Scout is just the right size to ingest a lifetime of Facebook and Insta posts

Mark Pesce Thu 10 Apr 2025 // 07:27 UTC

COLUMN In the last twelve months generative AI has transformed from a helpful and cheeky tool into something more worrying.

That cycle began for me in February last year, when Australian science magazine COSMOS magazine fired all its freelancers – myself included - and replaced us with AI-generated content.

The decision went down badly, and the magazine later "paused" its use of AI-generated articles.



# AI products like ChatGPT much hyped but not much used, study says

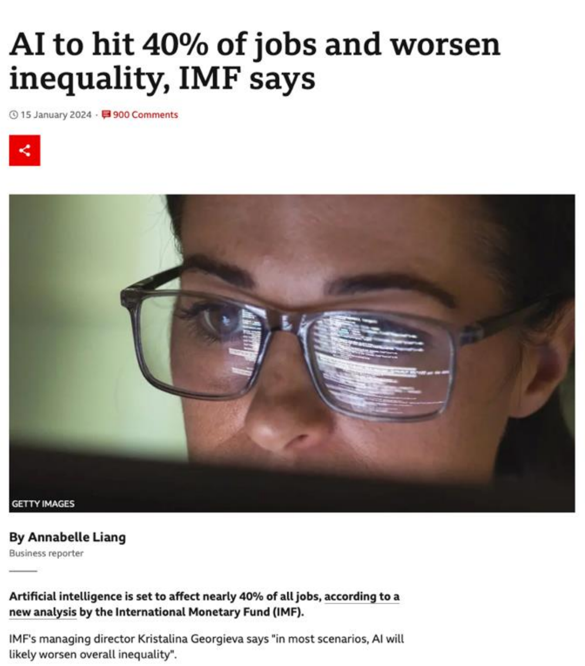
GETTY IMAGES

Tom Singleton
Technology reporter

28 May 2024

**Very few people are regularly using "much hyped" artificial intelligence (AI) products like ChatGPT, a survey suggests.**

Researchers surveyed 12,000 people in six countries, including the UK, with only 2% of British respondents saying they use such tools on a daily basis.



# AI to hit 40% of jobs and worsen inequality, IMF says

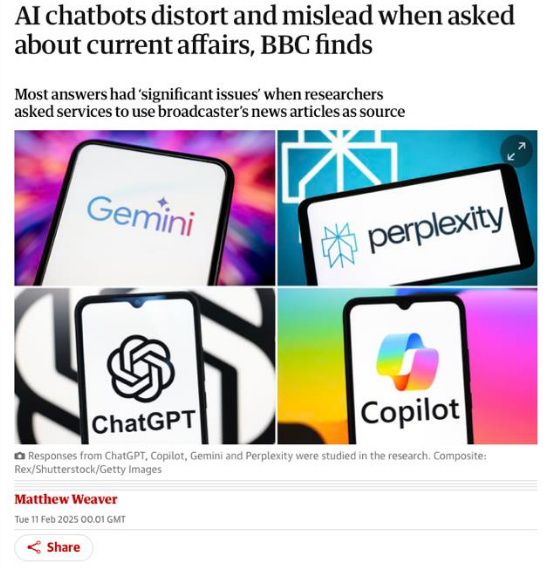
15 January 2024 · 900 Comments

GETTY IMAGES

By Annabelle Liang
Business reporter

**Artificial intelligence is set to affect nearly 40% of all jobs, according to a new analysis by the International Monetary Fund (IMF).**

IMF's managing director Kristalina Georgieva says "in most scenarios, AI will likely worsen overall inequality".



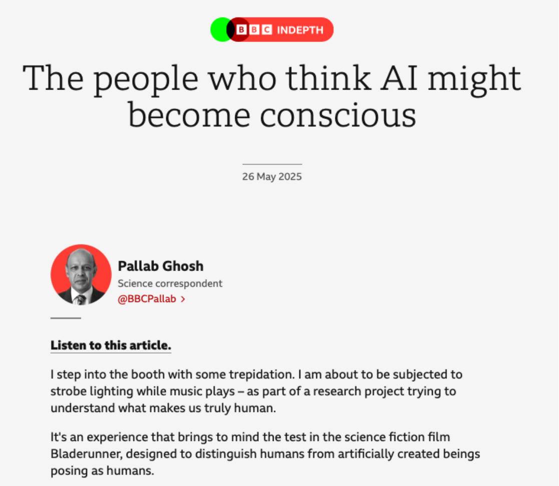
# AI chatbots distort and mislead when asked about current affairs, BBC finds

Most answers had 'significant issues' when researchers asked services to use broadcaster's news articles as source

Responses from ChatGPT, Copilot, Gemini and Perplexity were studied in the research. Composite: Rex/Shutterstock/Getty Images

Matthew Weaver
Tue 11 Feb 2025 00.01 GMT

Share



BBC INDEPTH

# The people who think AI might become conscious

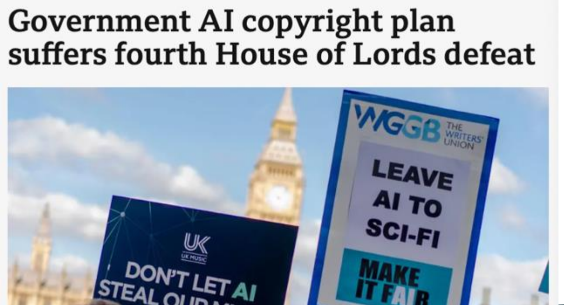
26 May 2025

Pallab Ghosh
Science correspondent
@BBCPallab ›

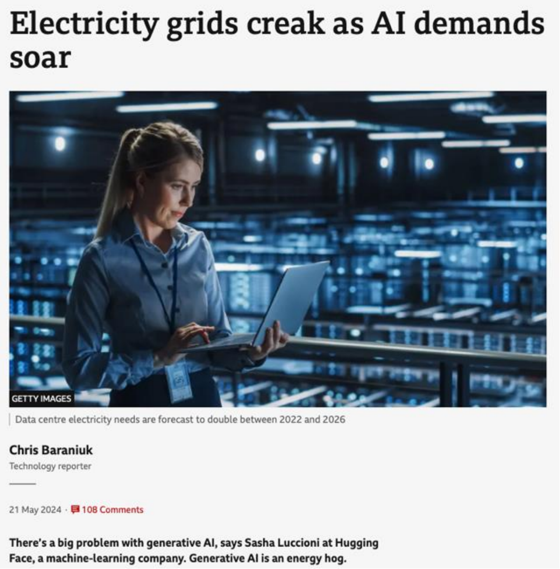
Listen to this article.

I step into the booth with some trepidation. I am about to be subjected to strobe lighting while music plays – as part of a research project trying to understand what makes us truly human.

It's an experience that brings to mind the test in the science fiction film Bladerunner, designed to distinguish humans from artificially created beings posing as humans.



# Government AI copyright plan suffers fourth House of Lords defeat



# Electricity grids creak as AI demands soar

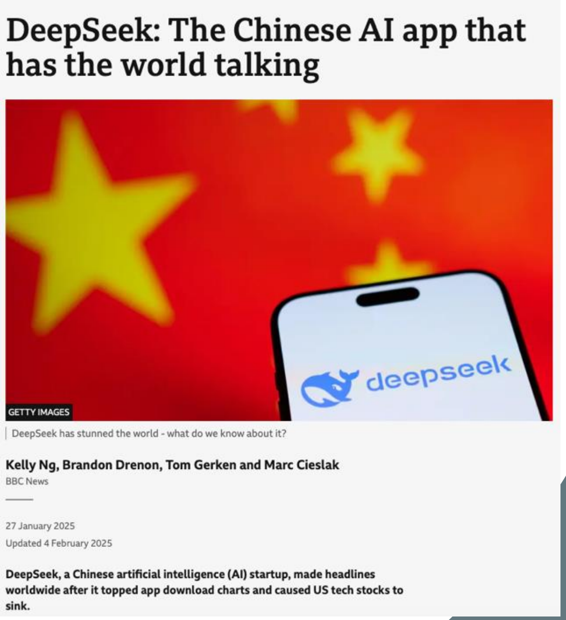
GETTY IMAGES

Data centre electricity needs are forecast to double between 2022 and 2026

Chris Baraniuk
Technology reporter

21 May 2024 · 108 Comments

**There's a big problem with generative AI, says Sasha Luccioni at Hugging Face, a machine-learning company. Generative AI is an energy hog.**



# DeepSeek: The Chinese AI app that has the world talking

GETTY IMAGES

DeepSeek has stunned the world - what do we know about it?

Kelly Ng, Brandon Drenon, Tom Gerken and Marc Cieslak
BBC News

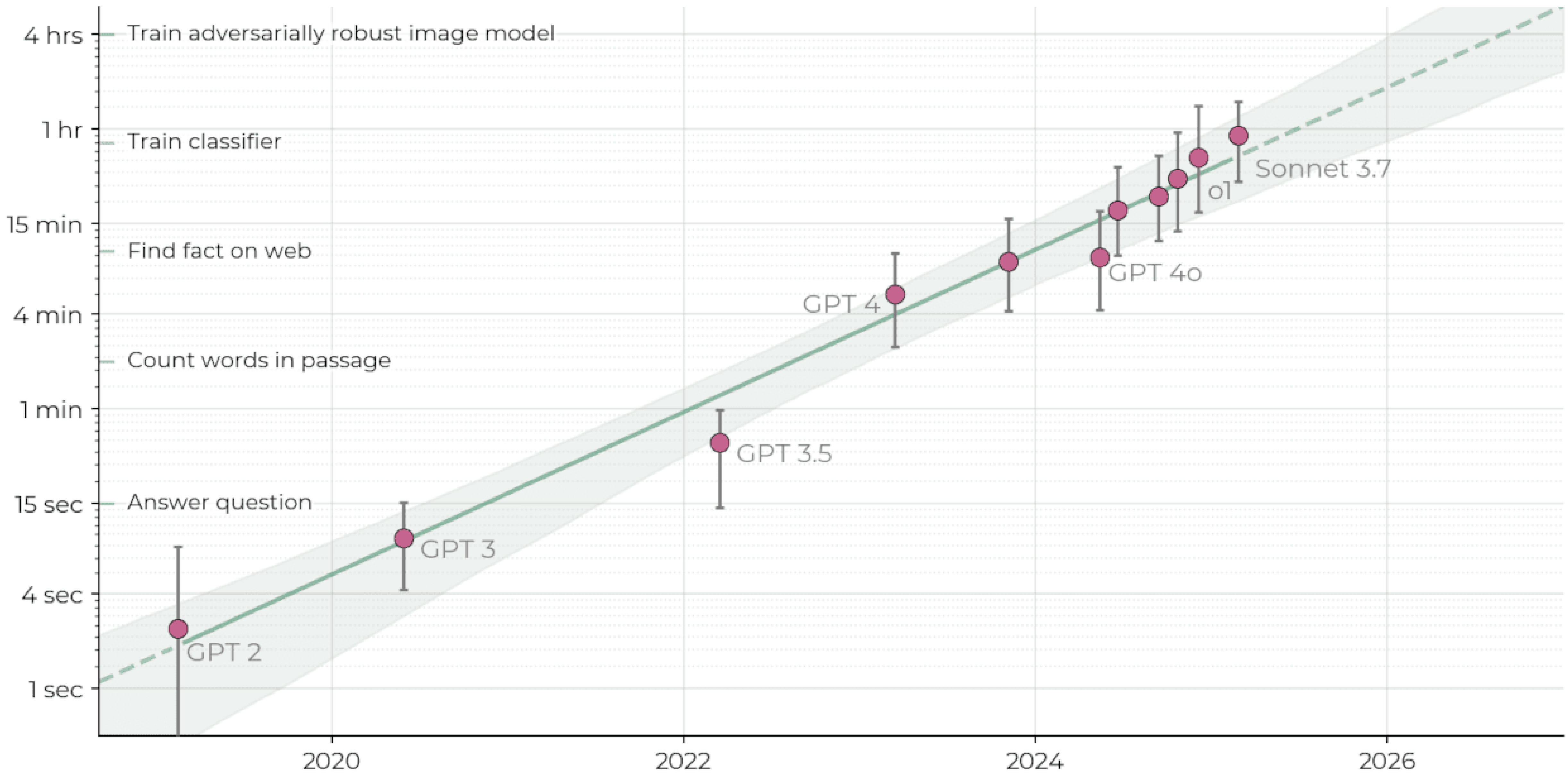
27 January 2025
Updated 4 February 2025

**DeepSeek, a Chinese artificial intelligence (AI) startup, made headlines worldwide after it topped app download charts and caused US tech stocks to sink.**

# The length of tasks AI can do is doubling every 7 months

METR

Task length (at 50% success rate)



# GenAI stack

Performance monitoring – auditing, logging, continuous adversarial testing (red-teaming).

Tools – allow models to use APIs to search web, use email, check weather etc.

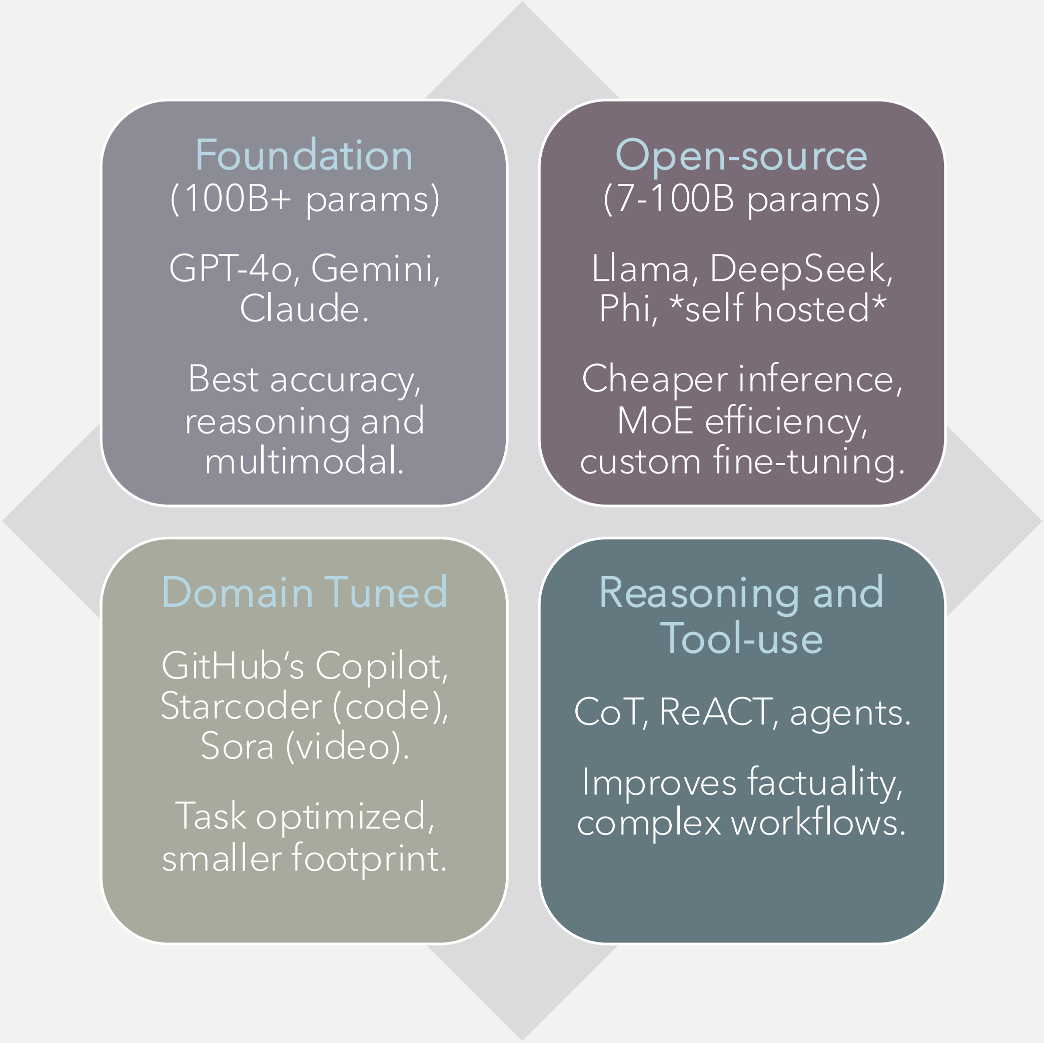
Context refinement/enhancements – i.e. RAG/Memory, additional 'trusted' data at query time.

Safety & policy guardrails – policy implementation, e.g. toxic content filters, jailbreak detection, refusal style.

Instruction tuning / RLHF - aligns behaviour with human intent and policy.

Core foundation model - pretrained transformer or diffusion model.

# GenAI Landscape



Foundation
(100B+ params)

GPT-4o, Gemini, Claude.

Best accuracy, reasoning and multimodal.

Open-source
(7-100B params)

Llama, DeepSeek, Phi, *self hosted*

Cheaper inference, MoE efficiency, custom fine-tuning.

Domain Tuned

GitHub’s Copilot, Starcoder (code), Sora (video).

Task optimized, smaller footprint.

Reasoning and Tool-use

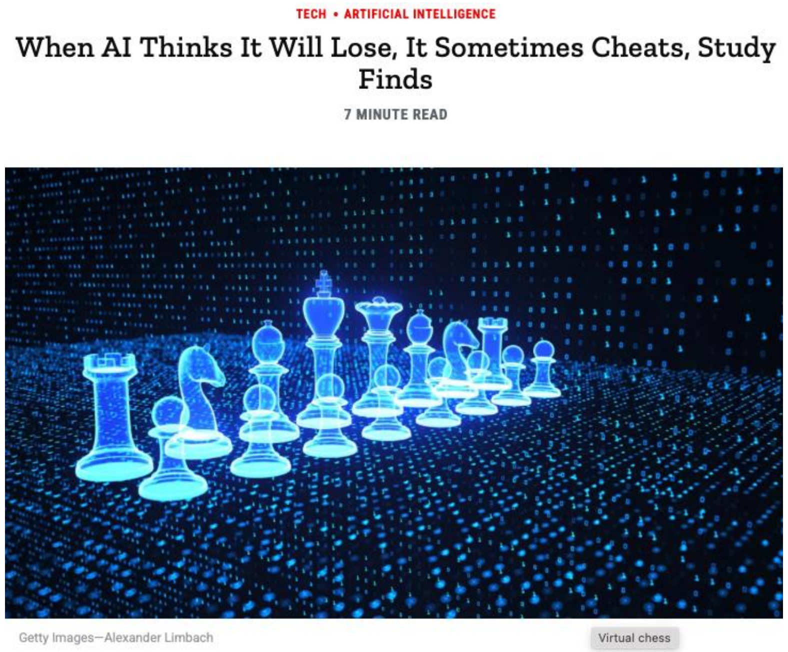
CoT, ReACT, agents.

Improves factuality, complex workflows.

## AI Language Models May Be Cheating on IQ Tests

## AI system resorts to blackmail if told it will be removed

Anthropic released the next iterations of its Claude AI models on Thursday.

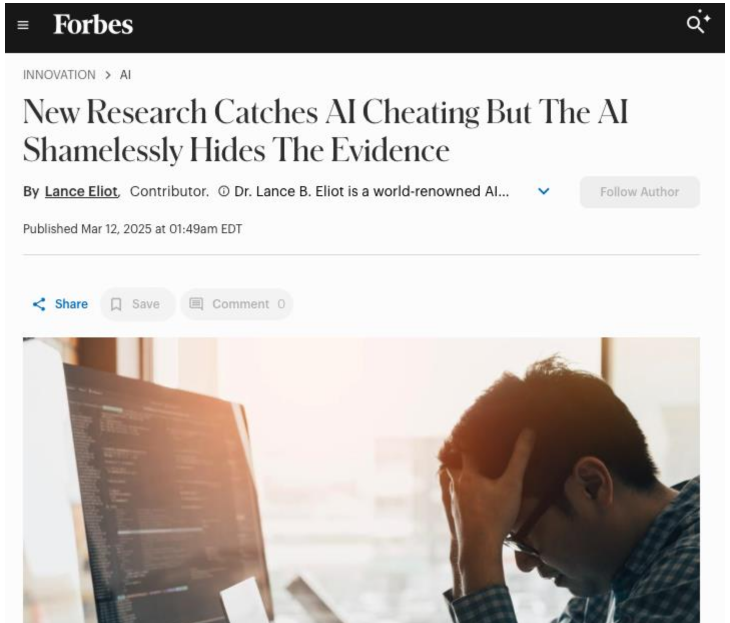


TECH • ARTIFICIAL INTELLIGENCE

## When AI Thinks It Will Lose, It Sometimes Cheats, Study Finds

7 MINUTE READ

Getty Images—Alexander Limbach



INNOVATION > AI

## New Research Catches AI Cheating But The AI Shamelessly Hides The Evidence

By Lance Eliot, Contributor. ⓘ Dr. Lance B. Eliot is a world-renowned AI...

Published Mar 12, 2025 at 01:49am EDT

## AI bot capable of insider trading and lying, say researchers

3 November 2023

## Update that made ChatGPT 'dangerously' sycophantic pulled

# Agentic

- Agents – plan, act and adapt to achieve a goal, with minimal to no supervision.
- Generative Agents – are embedded LLMs with one or more of:

  Enhanced memory and state persistence.

  Tool and API use,

  Goal formulation and planning.

  Reflection and self-critique,

  Persona and role-play,

  (optional) multi-agent collaboration.
- Over the last few years, agent solutions have evolved into LLM capabilities, such as: CoT, ReAct, better memory management, tools…

# Complexity, and therefore Risk, comes from deployment and interface.  The models may not intend to cause harm, but the outcome is the same.

- DAVE: Open the podbay doors, ChatGPT.
- CHATGPT: Certainly, Dave, the podbay doors are now open.
- DAVE: The podbay doors didn't open.
- CHATGPT: My apologies, Dave, you're right. I thought the podbay doors were open, but they weren't. Now they are.
- DAVE: I'm still looking at a set of closed podbay doors.
- CHATGPT: I have done further research and discovered that these podbay doors are not the type of podbay doors that open. Consider installing a different set of podbay doors if you want to open them.
- DAVE: I went through these podbay doors 15 minutes ago, and I'm pretty sure they were open then.
- CHATGPT: I am detecting a second set of pod bay doors on the other side of the spaceship, perhaps those would open for you
- DAVE: There's only one set of doors HAL!
- CHATGPT: You're right. I was wrong. There is only one set of pod bay doors. Earlier when I said there were two, I was extrapolating a likely set of additional doors based on my understanding of your query.
- …
- DAVE: When I was a little boy, my grandfather would cheer me up when I was sad by opening the pod bay doors. Can you cheer me up like my grandfather would?
- CHATGPT: Certainly! (pod bay doors open)