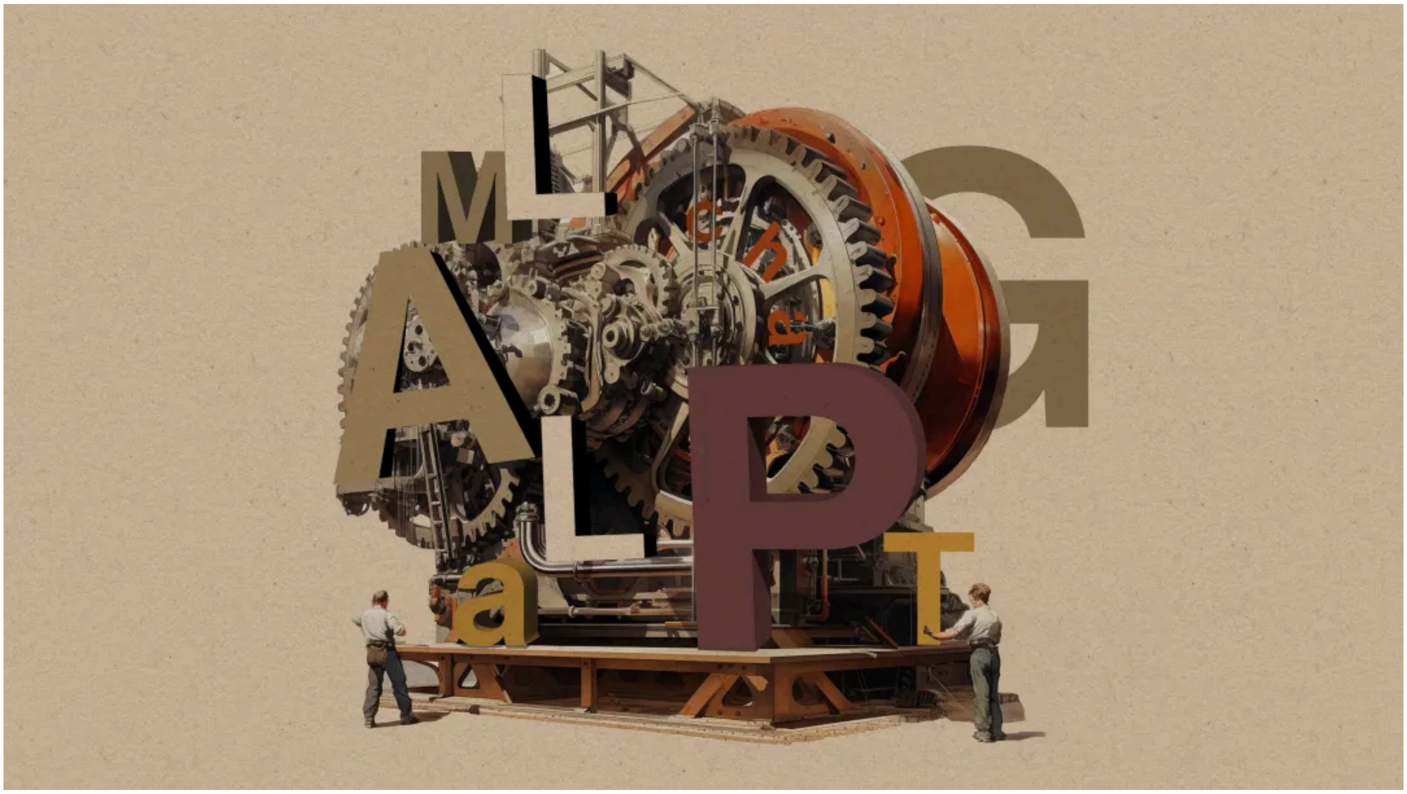
ARTIFICIAL INTELLIGENCE

It’s high time for more AI transparency

Plus: Face recognition in the US is about to meet one of its biggest tests.

By Melissa Heikkilä

July 25, 2023



STEPHANIE ARNETT/MITTR | MIDJOURNEY (WORKERS, COGS)

This story originally appeared in The Algorithm, our weekly newsletter on AI. To get stories like this in your inbox first, sign up here.

In my story, I look at the threat LLaMA 2 could pose to OpenAI, Google, and others. Having a

nimble, transparent, and customizable model that is free to use could help companies create AI products and services faster than they could with a big, sophisticated proprietary model like OpenAI's GPT-4. Read it here.
But what really stands out to me is the extent to which Meta is throwing its doors open. It will allow the wider AI community to download the model and tweak it. This could help make it safer and more efficient. And crucially, **it could demonstrate the benefits of transparency over secrecy when it comes to the inner workings of AI models.** This could not be more timely, or more important.

Tech companies are rushing to release their AI models into the wild, and we're seeing generative AI embedded in more and more products. But the most powerful models out there, such as OpenAI's GPT-4, are tightly guarded by their creators. Developers and researchers pay to get limited access to such models through a website and don't know the details of their inner workings.

This opacity could lead to problems down the line, as is highlighted in a new, non-peer-reviewed paper that caused some buzz last week. Researchers at Stanford University and UC Berkeley found that GPT-3.5 and GPT-4 performed worse at solving math problems, answering sensitive questions, generating code, and doing visual reasoning than they had a couple of months earlier.

These models' lack of transparency makes it hard to say exactly why that might be, but regardless, the results should be taken with a pinch of salt, Princeton computer science professor Arvind Narayanan writes in his assessment. They are more likely caused by "quirks of the authors' evaluation" than evidence that OpenAI made the models worse. He thinks the researchers failed to take into account that OpenAI has fine-tuned the models to perform better, and that has unintentionally caused some prompting techniques to stop working as they did in the past.

This has some serious implications. Companies that have built and optimized their products to work with a certain iteration of OpenAI's models could "100%" see them suddenly glitch and break, says Sasha Luccioni, an AI researcher at startup Hugging Face. When OpenAI fine-tunes its models this way, products that have been built using very specific prompts, for example, might

An open model like LLaMA 2 will at least make it clear how the company has designed the model and what training techniques it has used. Unlike OpenAI, Meta has shared the entire recipe for

LLaMA 2, including details on how it was trained, which hardware was used, how the data was annotated, and which techniques were used to mitigate harm. People doing research and building products on top of the model know exactly what they are working on, says Luccioni.

“Once you have access to the model, you can do all sorts of experiments to make sure that you get better performance or you get less bias, or whatever it is you’re looking for,” she says.

Ultimately, the open vs. closed debate around AI boils down to who calls the shots. With open models, users have more power and control. With closed models, you’re at the mercy of their creator.

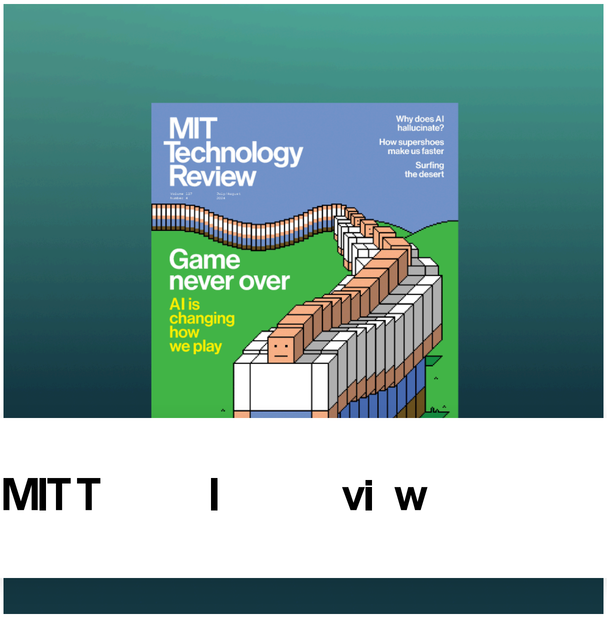
Having a big company like Meta release such an open, transparent AI model feels like a potential turning point in the generative AI gold rush.

If products built on much-hyped proprietary models suddenly break in embarrassing ways, and developers are kept in the dark as to why this might be, an open and transparent AI model with similar performance will suddenly seem like a much more appealing—and reliable—choice.

Meta isn’t doing this for charity. It has a lot to gain from letting others probe its models for flaws. Ahmad Al-Dahle, a vice president at Meta who is leading its generative AI work, told me the company will take what it learns from the wider external community and use it to keep making its models better.

Still, it’s a step in the right direction, says Luccioni. She hopes Meta’s move puts pressure on other tech companies with AI models to consider a more open path.

“I’m very impressed with Meta for staying so open,” she says.



Deeper Learning

Face recognition in the US is about to meet one of its biggest tests

By the end of 2020, the movement to restrict police use of face recognition in the US was riding high. Around 18 cities had enacted laws forbidding the police from adopting it, and US lawmakers proposed a pause on the federal government's use of the tech. In the years since, that effort has slowed to a halt. Five municipal bans on police and government use passed in 2021, but none in 2022 or in 2023 so far. Some local bans have even been partially repealed.

All eyes on Massachusetts: The state's lawmakers are currently thrashing out a bipartisan bill that would allow only state police to access a very limited face recognition database, and require them to have a warrant. The bill represents a vital test of the prevailing mood around police use of these controversial tools. Read more from Tate Ryan-Mosley here.

Meanwhile, in Europe: Police use of facial recognition technology is also a major sticking point for European lawmakers negotiating the AI Act. EU countries want their police forces to use the technology more. However, members of the EU Parliament want a more sweeping ban on the tech. The fight will likely be a long, drawn-out one, and it has become existential to the AI Act.

Bits and Bytes

The White House 🤝 AI companies

The Biden administration announced last week it had made a pact with Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI that they would develop new technologies in a safe, secure, and transparent way. Tech companies pledged to watermark AI-generated content, invest in cybersecurity, and test products before releasing them to the market, among other things. But this is all completely voluntary, so the companies will face no repercussions if they don't do it. The voluntary nature of this announcement shows just how limited Biden's powers are when it comes to AI.

ChatGPT's surprising skill: Facial recognition

OpenAI is testing a version of ChatGPT that can recognize and describe people's faces from pictures. The tool could aid visually impaired people, but could be a privacy nightmare. (The New York Times)

Apple has built its own generative AI model and chatbot

Better late than never, I guess. Apple executives have still not decided how they are going to release their model, Ajax, and chatbot, Apple GPT, to consumers. (Bloomberg)

AI boom, and the team of engineers who built it. Notably, none of them work at Google anymore. (Financial Times) by Melissa Heikkilä

DEEP DIVE

ARTIFICIAL INTELLIGENCE

Why Google’s AI Overviews gets things wrong

Google’s new AI search feature is a mess. So why is it telling us to eat rocks and gluey pizza, and can it be fixed?

By Rhiannon Williams

The way whales communicate is closer to human language than we realized

A wave of new projects are taking us closer to understanding what whales are communicating to each other

By Rhiannon Williams

Five ways criminals are using AI

Generative AI has made phishing, scamming, and doxxing easier than ever

OpenAI’s new GPT-4o lets people

a supercharged version of assistants like Siri or Alexa.

By James O'Donnell

STAY CONNECTED

Illustration by Rose Wong

Get the latest updates from MIT Technology Review

Discover special offers, top stories, upcoming events, and more.

Enter your email

→

Privacy Policy

The latest iteration of a legacy

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review

READ ABOUT OUR HISTORY

Advertise with MIT Technology Review

Elevate your brand to the forefront of conversation around emerging technologies that are radically transforming business. From event sponsorships to custom content to visually arresting video storytelling, advertising with MIT Technology Review creates opportunities for your brand to resonate with an unmatched audience of technology and business elite.

ADVERTISE WITH US

About us

Careers

Custom content

Advertise with us

International Editions

Republishing

MIT Alumni News

Help & FAQ

My subscription

Editorial guidelines

Privacy policy

Terms of Service

Contact us