ChatGPT has enormous hidden costs that could throttle AI development

washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute

Will Oremus June 5, 2023

AI chatbots lose money every time you use them. That is a problem.

The cost of operating the systems is so high that companies aren’t deploying their best versions to the public

By Will Oremus
June 5, 2023 at 6:00 a.m. EDT

ChatGPT, running on a smartphone in March. (Gabby Jones/Bloomberg News)

AI chatbots have a problem: They lose money on every chat.

The enormous cost of running today’s large language models, which underpin tools like ChatGPT and Bard, is limiting their quality and threatening to throttle the global AI boom they’ve sparked.

Tech is not your friend. We are. Sign up for The Tech Friend newsletter.→
Their expense, and the limited availability of the computer chips they require, are also constraining which companies can afford to run them and pressuring even the world’s richestcompanies to turn chatbots into moneymakers sooner than they may be ready to.

“The models being deployed right now, as impressive as they seem, are really not the best models available,” said Tom Goldstein, a computer science professor at the University of Maryland. “So as a result, the models you see have a lot of weaknesses” that might be avoidable if cost were no object — such as a propensity to spit out biased results or blatant falsehoods.

What happens when ChatGPT lies about real people?

The tech giants staking their future on AI rarely discuss the technology’s cost. OpenAI (the maker of ChatGPT), Microsoft and Google all declined to comment. But experts say it’s the most glaring obstacle to Big Tech’s vision of generative AI zipping its way across every industry, slicing head counts and boosting efficiency.

The intensive computing AI requires is why OpenAI has held back its powerful new language model, GPT-4, from the free version of ChatGPT, which is still running a weaker GPT-3.5 model. ChatGPT’s underlying data set was last updated in September 2021, making it useless for researching or discussing recent events. And even those who pay $20 per month for GPT-4 can send only 25 messages every three hours because it’s so expensive to run. (It’s also much slower to respond.)

Those costs may also be one reason Google has yet to build an AI chatbot into its flagship search engine, which fields billions of queries every day. When Google released its Bard chatbot in March, it opted not to use its largest language model. Dylan Patel, chief analyst at the semiconductor research firm SemiAnalysis, estimated that a single chat with ChatGPT could cost up to 1,000 times as much as a simple Google search.

In a recent report on artificial intelligence, the Biden administration pinpointed the computational costs of generative AI as a national concern. The White House wrote that the technology is expected to “dramatically increase computational demands and the associated environmental impacts,” and that there’s an “urgent need” to design more sustainable systems.

Even more than other forms of machine learning, generative AI requires dizzying amounts of computational power and specialized computer chips, known as GPUs, that only the wealthiest of companies can afford. The intensifying battle for access to those chips has helped to make their leading providers into tech giants in their own right, giving them the keys to what has become the technology industry’s most prized asset.

Why Nvidia is suddenly one of the most valuable companies in the world

Silicon Valley came to dominate the internet economy in part by offering services like online search, email and social media to the world free, losing money initially but eventually turning hefty profits on personalized advertising. And ads are probably coming to AI chatbots. But analysts say ads alone probably won't be enough to make cutting-edge AI tools profitable anytime soon.

In the meantime, the companies offering AI models for consumer use must balance their desire to win market share with the financial losses they're racking up.

The search for more reliable AI also is likely to drive profits primarily to the chipmakers and cloud computing giants that already control much of the digital space — along with the chipmakers whose hardware they need to run the models.

It's no accident that the companies building the leading AI language models are either among the largest cloud computing providers, as with Google and Microsoft, or have close partnerships with them, as OpenAI does with Microsoft. Companies that buythose firms' AI tools don't realize they're being locked into a heavily subsidized service that costs much more than what they're currently paying, said Clem Delangue, CEO of Hugging Face, an open-source AI company.

OpenAI CEO Sam Altman indirectly acknowledged the problem at a Senate hearing last month, when Sen. Jon Ossoff (D-Ga.) warned that if OpenAI were to try to make ChatGPT addictive in a way that harms kids, Congress "will look very harshly" on it. Altman said Ossoff needn't worry: "We try to design systems that do not maximize for engagement. In fact, we're so short on GPUs, the less people use our products, the better."

The expense of AI language models starts with developing and training them, which requires gargantuan amounts of data and software to identify patterns in language. AI companies also typically hire star researchers whose salaries can rival those of pro athletes.That presents an initial barrier to any company hoping to build its own model, though a few well-funded start-ups have succeeded — including Anthropic AI, which OpenAI alumni founded with financial backing from Google.

Then, each query to a chatbot like ChatGPT, Microsoft's Bing or Anthropic's Claude is routed to data centers, where supercomputers crunch the models and perform numerous high-speed calculations at the same time — first, interpreting the user's prompt, then working to predict the most plausible response, one "token," or four-letter sequence, at a time.

That sort of computational power requires GPUs, or graphics processing units, that were first made for video games but were found to be the only chips that could handle such heavy computer tasks as large language models. Currently, just one company, Nvidia, sells the best of those, for which it charges tens of thousands of dollars. Nvidia's valuation recentlyrocketed to $1 trillion on the anticipated sales. The Taiwan-based company that manufactures many of those chips, TSMC, has likewise soared in value.

"GPUs at this point are considerably harder to get than drugs," Elon Musk, who recently purchased some 10,000 GPUs for his own AI start-up, told a May 23 Wall Street Journal summit.

Those computing requirements also help to explain why OpenAI is no longer the nonprofit it was founded to be.

Started in 2015 with the stated mission of developing AI "in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return," by 2019, it had switched to a for-profit model to attract investors, including Microsoft, which pumped in $1 billion and became OpenAI's exclusive computing provider. (Microsoft has since poured in $10 billion more and integrated OpenAI's technology with Bing, Windows and other products.)

Exactly how much chatbots like ChatGPT cost to run is a moving target, as companies work to make them more efficient.

In December, not long after its launch, Altman estimated the cost of ChatGPT at “probably single-digits cents per chat.” That might not sound like much, until you multiply it by upward of 10 million users per day, as analysts have estimated. In February, SemiAnalysis calculated that ChatGPT was costing OpenAI some $700,000 per day in computing costs alone, based on the processing needed to run GPT-3.5, the default model at the time.

Multiply those computing costs by the 100 million people per day who useMicrosoft’s Bing search engineor the more than 1 billion who reportedly use Google, and one can begin to see why the tech giants are reluctant to make the best AI models available to the public.

The new Bing told our reporter it ‘can feel or think things’

“This is not a sustainable equation for the democratization or wide availability of generative AI, the economy or the environment,” said Sid Sheth, founder and CEO of d-Matrix, a start-up working to build more efficient chips for AI.

Google said in its February announcement of Bard that it would initially run on a “lightweight” version of the company’s LaMDA language model because it required “significantly less computing power, enabling us to scale to more users.” In other words, even a company as wealthy as Google wasn’t prepared to foot the bill of putting its most powerful AI technology into a free chatbot.

Perspective: What Google’s new AI gets right, wrong and weird.

The cost-cutting took a toll: Bard stumbled over basic facts in its launch demonstration, shearing $100 billion from the value of Google’s shares. Bing, for its part, went off the rails early on, prompting Microsoft to scale back both its personality and the number of questions users could ask it in a given conversation.

Such errors, sometimes called “hallucinations,” have become a major concern with AI language models as both individuals and companies increasingly rely on them. Experts say they’re a function of the models’ basic design: They’re built to generate likely sequences of words, not true statements.

Another Google chatbot, called Sparrow, was designed by the company’s DeepMind subsidiary to search the internet and cite its sources, with the goal of reducing falsehoods. But Google has not released that one so far.

ChatGPT ‘hallucinates.’ Some researchers worry it isn’t fixable.

Meanwhile, each of the major players is racing for ways to make AI language models cheaper.

Running a query on OpenAI’s new, lightweight GPT-3.5 Turbo model costs less than one-tenth as much as its top-of-the-line GPT-4. Google is making its own AI chips, which it claims are more efficient than Nvidia’s, as are start-ups like d-Matrix. And numerous start-ups are building on open-source language models, such as Meta’s LLaMA, so that they don’t have to pay OpenAI or Google to use theirs — even though those models don’t yet perform as well and may lack guardrails to prevent abuse.

The push for smaller, cheaper models marks a sudden reversal for the industry, said Maryland’s Goldstein.

“We spent the last four years just trying to make the biggest models we could,” he said. But that was when the goal was to publish research papers, not release AI chatbots to the public. “Now, just within the last few months, there’s been a complete turnaround in the community, and suddenly everyone’s trying to build the smallest model they can to control the costs.”

For consumers,that could mean the days of unfettered access to powerful, general-purpose AI models are numbered.

Microsoft is already experimenting with building advertisements into its AI-powered Bing results. At the Senate hearing, OpenAI’s Altman wouldn’t rule out doing the same, although he said he prefers a paid subscription model.

Both companies say they're confident the economics will eventually pencil out. Altman told the tech blog Stratechery in February, "There's so much value here, it's inconceivable to me that we can't figure out how to ring the cash register on it."

Yet critics note that generative AI also comes with costs to society.

"All this processing has implications for greenhouse gas emissions," said Bhaskar Chakravorti, dean of global business at Tufts University's Fletcher School. The computing requires energy that could be used for other purposes — including other computing tasks that are less trendy than AI language models. That "could even slow down the development and application of AI for other, more meaningful uses, such as in health care, drug discovery, cancer detection, etc.," Chakravorti said.

Based on estimates of ChatGPT's usage and computing needs, data scientist Kasper Groes Albin Ludvigsen estimated that it may have used as much electricity in January as 175,000 people — the equivalent of a midsize city.

For now, the tech giants are willing to lose money in a bid to win market share with their AI chatbots, Goldstein said. But if they can't make them profitable? "Eventually you come to the end of the hype curve, and the only thing your investors are going to look at, at that point, is your bottom line."

Still, Goldstein predicted many people and companies will find generative AItools hard to resist,even with all their flaws. "Even though it's expensive," he said, "it's still far less expensive than human labor."

Nitasha Tiku contributed to this report.