# We did the math on AI's energy footprint. Here's the story you haven't heard.

## MIT Technology Review

**The emissions from individual AI text, image, and video queries seem small—until you add up what the industry isn't tracking and consider where it's heading next.**

AI's integration into our lives is the most significant shift in online life in more than a decade. Hundreds of millions of people now regularly turn to chatbots for help with homework, research, coding, or to create images and videos. But what's powering all of that?

**Today, new analysis by *MIT Technology Review* provides an unprecedented and comprehensive look at how much energy the AI industry uses—down to a single query—to trace where its carbon footprint stands now, and where it's headed, as AI barrels towards billions of daily users.**

We spoke to two dozen experts measuring AI's energy demands, evaluated different AI models and prompts, pored over hundreds of pages of projections and reports, and questioned top AI model makers about their plans. **Ultimately, we found that the common understanding of AI's energy consumption is full of holes.**

We started small, as the question of how much a single query costs is vitally important to understanding the bigger picture. That's because those queries are being built into ever more applications beyond standalone chatbots: from search, to agents, to the mundane daily apps we use to track our fitness, shop online, or book a flight. **The energy resources required to power this artificial-intelligence revolution are staggering, and the world's biggest tech companies have made it a top priority to harness ever more of that energy, aiming to reshape our energy grids in the process.**

Meta and Microsoft are working to fire up new nuclear power plants. OpenAI and President Donald Trump announced the Stargate initiative, which aims to spend $500 billion—more than the Apollo space program—to build as many as 10 data centers (each of which could require five gigawatts, more than the total power demand from the state of New Hampshire). Apple announced plans to spend $500 billion on manufacturing and data centers in the US over the next four years. Google expects to spend $75 billion on AI infrastructure alone in 2025.

**This isn't simply the norm of a digital world. It's unique to AI, and a marked departure from Big Tech's electricity appetite in the recent past. From 2005 to 2017, the amount of electricity going to data centers remained quite flat thanks to increases in efficiency, despite the construction of armies of new data centers to serve the rise of cloud-**

**based online services, from Facebook to Netflix. In 2017, AI began to change everything. Data centers started getting built with energy-intensive hardware designed for AI, which led them to double their electricity consumption by 2023. The latest reports show that 4.4% of all the energy in the US now goes toward data centers.**

**The carbon intensity of electricity used by data centers was 48% higher than the US average.**

Given the direction AI is headed—more personalized, able to reason and solve complex problems on our behalf, and everywhere we look—it's likely that our AI footprint today is the smallest it will ever be. **According to new projections published by Lawrence Berkeley National Laboratory in December, by 2028 more than half of the electricity going to data centers will be used for AI. At that point, AI alone could consume as much electricity annually as 22% of all US households.**

Meanwhile, data centers are expected to continue trending toward using dirtier, more carbon-intensive forms of energy (like gas) to fill immediate needs, leaving clouds of emissions in their wake. And all of this growth is for a new technology that's still finding its footing, and in many applications—education, medical advice, legal analysis—might be the wrong tool for the job or at least have a less energy-intensive alternative.

Tallies of AI's energy use often short-circuit the conversation—either by scolding individual behaviour, or by triggering comparisons to bigger climate offenders. Both reactions dodge the point: **AI is unavoidable, and even if a single query is low-impact, governments and companies are now shaping a much larger energy future around AI's needs.**

We're taking a different approach with an accounting meant to inform the many decisions still ahead: where data centers go, what powers them, and how to make the growing toll of AI visible and accountable.

ChatGPT is now estimated to be the fifth-most visited website in the world, just after Instagram and ahead of X.

That's because despite the ambitious AI vision set forth by tech companies, utility providers, and the federal government, details of how this future might come about are murky. Scientists, federally funded research facilities, activists, and energy companies argue that leading AI companies and data center operators disclose too little about their activities. Companies building and deploying AI models are largely quiet when it comes to answering a central question: **Just how much energy does interacting with one of these models use? And what sorts of energy sources will power AI's future?**

This leaves even those whose job it is to predict energy demands forced to assemble a puzzle with countless missing pieces, making it nearly impossible to plan for AI's future impact on energy grids and emissions. Worse, the deals that

utility companies make with the data centers will likely transfer the costs of the AI revolution to the rest of us, in the form of higher electricity bills.

It's a lot to take in. To describe the big picture of what that future looks like, we have to start at the beginning.

**Part One: Making the model|**

Before you can ask an AI model to help you with travel plans or generate a video, the model is born in a data center.

**Racks of servers hum along for months, ingesting training data, crunching numbers, and performing computations. This is a time-consuming and expensive process—it's estimated that training OpenAI's GPT-4 took over $100 million and consumed 50 gigawatt-hours of energy, enough to power San Francisco for three days. It's only after this training, when consumers or customers "inference" the AI models to get answers or generate outputs, that model makers hope to recoup their massive costs and eventually turn a profit.**

"For any company to make money out of a model—that only happens on inference," says Esha Choukse, a researcher at Microsoft Azure who has studied how to make AI inference more efficient.

As conversations with experts and AI companies made clear, inference, not training, represents an increasing majority of AI's energy demands and will continue to do so in the near future. It's now estimated that 80–90% of computing power for AI is used for inference.

All this happens in data centers. There are roughly 3,000 such buildings across the United States that house servers and cooling systems and are run by cloud providers and tech giants like Amazon or Microsoft, but used by AI startups too. A growing number—though it's not clear exactly how many, since information on such facilities is guarded so tightly—are set up for AI inferencing.

**Part Two: A Query**

If you've seen a few charts estimating the energy impact of putting a question to an AI model, you might think it's like measuring a car's fuel economy or a dishwasher's energy rating: a knowable value with a shared methodology for calculating it. You'd be wrong.

In reality, the type and size of the model, the type of output you're generating, and countless variables beyond your control—like which energy grid is connected to the data center your request is sent to and what time of day it's processed—can make one query thousands of times more energy-intensive and emissions-producing than another.

And when you query most AI models, whether on your phone within an app like Instagram or on the web interface for ChatGPT, much of what happens after

your question is routed to a data center remains a secret. Factors like which data center in the world processes your request, how much energy it takes to do so, and how carbon-intensive the energy sources used are tend to be knowable only to the companies that run the models.

This is true for most of the name-brand models you're accustomed to, like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, which are referred to as "closed." The key details are held closely by the companies that make them, guarded because they're viewed as trade secrets (and also possibly because they might result in bad PR). These companies face few incentives to release this information, and so far they have not.

"The closed AI model providers are serving up a total black box," says Boris Gamazaychikov, head of AI sustainability at Salesforce, who has led efforts with researchers at Hugging Face, an AI platform provider of tools, models, and libraries for individuals and companies, to make AI's energy demands more transparent. Without more disclosure from companies, it's not just that we don't have good estimates—we have little to go on at all.

Without more disclosure from companies, it's not just that we don't have good estimates—we have little to go on at all.

So where can we turn for estimates? So-called open-source models can be downloaded and tweaked by researchers, who can access special tools to measure how much energy the H100 GPU requires for a given task. Such models are also incredibly popular; Meta announced in April that its Llama models have been downloaded more than 1.2 billion times, and many companies use open-source models when they want more control over outputs than they can get using something like ChatGPT.

But even if researchers can measure the power drawn by the GPU, that leaves out the power used up by CPUs, fans, and other equipment. A 2024 paper by Microsoft analyzed energy efficiencies for inferencing large language models and found that doubling the amount of energy used by the GPU gives an approximate estimate of the entire operation's energy demands.

So for now, measuring leading open-source models (and adding estimates for all these other pieces) gives us the best picture we have of just how much energy is being used for a single AI query. However, keep in mind that the ways people use AI today—to write a grocery list or create a surrealist video—are far simpler than the ones we'll use in the autonomous, agentic future that AI companies are hurling us toward. More on that later.

Here's what we found.

**Text models**

Let's start with models where you type a question and receive back a response in words. One of the leading groups evaluating the energy demands of AI is at the University of Michigan, led by PhD candidate Jae-Won Chung and

associate professor Mosharaf Chowdhury, who publish energy measurements on their ML.Energy leaderboard. We worked with the team to focus on the energy demands of one of the most widely adopted open-source models, Meta's Llama.

The smallest model in our Llama cohort, Llama 3.1 8B, has 8 billion parameters—essentially the adjustable "knobs" in an AI model that allow it to make predictions. When tested on a variety of different text-generating prompts, like making a travel itinerary for Istanbul or explaining quantum computing, the model required about 57 joules per response, or an estimated 114 joules when accounting for cooling, other computations, and other demands. This is tiny—about what it takes to ride six feet on an e-bike, or run a microwave for one-tenth of a second.

The largest of our text-generation cohort, Llama 3.1 405B, has 50 times more parameters. More parameters generally means better answers but more energy required for each response. On average, this model needed 3,353 joules, or an estimated 6,706 joules total, for each response. That's enough to carry a person about 400 feet on an e-bike or run the microwave for eight seconds.

So model size is a huge predictor of energy demand. One reason is that once a model gets to a certain size, it has to be run on more chips, each of which adds to the energy required. The largest model we tested has 405 billion parameters, but others, such as DeepSeek, have gone much further, with over 600 billion parameters. The parameter counts for closed-source models are not publicly disclosed and can only be estimated. GPT-4 is estimated to have over 1 trillion parameters.

But in all these cases, the prompt itself was a huge factor too. Simple prompts, like a request to tell a few jokes, frequently used nine times less energy than more complicated prompts to write creative stories or recipe ideas.

**Generating an image**

AI models that generate images and videos work with a different architecture, called diffusion. Rather than predicting and generating words, they learn how to transform an image of noise into, let's say, a photo of an elephant. They do this by learning the contours and patterns of pictures in their training data and storing this information across millions or billions of parameters. Video-generator models learn how to do this across the dimension of time as well.

The energy required by a given diffusion model doesn't depend on your prompt—generating an image of a skier on sand dunes requires the same amount of energy as generating one of an astronaut farming on Mars. The energy requirement instead depends on the size of the model, the image resolution, and the number of "steps" the diffusion process takes (more steps lead to higher quality but need more energy).

Generating a standard-quality image (1024 x 1024 pixels) with Stable Diffusion 3 Medium, the leading open-source image generator, with 2 billion parameters,

requires about 1,141 joules of GPU energy. With diffusion models, unlike large language models, there are no estimates of how much GPUs are responsible for the total energy required, but experts suggested we stick with the “doubling” approach we’ve used thus far because the differences are likely subtle. That means an estimated 2,282 joules total. Improving the image quality by doubling the number diffusion steps to 50 just about doubles the energy required, to about 4,402 joules. That’s equivalent to about 250 feet on an e-bike, or around five and a half seconds running a microwave. That’s still less than the largest text model.

This might be surprising if you imagined generating images to require more energy than generating text. “Large [text] models have a lot of parameters,” says Chung, who performed the measurements on open-source text and image generators featured in this story. “Even though they are generating text, they are doing a lot of work. ” Image generators, on the other hand, often work with fewer parameters.

**Making a video**

Videos generated by CogVideoX, an open-source model.

Last year, OpenAI debuted Sora, its dazzling tool for making high-fidelity videos with AI. Other closed-source video models have come out as well, like Google Veo2 and Adobe’s Firefly.

Given the eye-watering amounts of capital and content it takes to train these models, it’s no surprise that free-to-use, open-source models generally lag behind in quality. Still, according to researchers at Hugging Face, one of the best is CogVideoX, made by a Chinese AI startup called Zhipu AI and researchers from Tsinghua University in Beijing.

Sasha Luccioni, an AI and climate researcher at Hugging Face, tested the energy required to generate videos with the model using a tool called Code Carbon.

An older version of the model, released in August, made videos at just eight frames per second at a grainy resolution—more like a GIF than a video. Each one required about 109,000 joules to produce. But three months later the company launched a larger, higher-quality model that produces five-second videos at 16 frames per second (this frame rate still isn’t high definition; it’s the one used in Hollywood’s silent era until the late 1920s). The new model uses more than 30 times more energy on each 5-second video: about 3.4 million joules, more than 700 times the energy required to generate a high-quality image. This is equivalent to riding 38 miles on an e-bike, or running a microwave for over an hour.

It’s fair to say that the leading AI video generators, creating dazzling and hyperrealistic videos up to 30 seconds long, will use significantly more energy. As these generators get larger, they’re also adding features that allow you to tweak particular elements of videos and stitch multiple shots together into

scenes—all of which add to their energy demands. A note: AI companies have defended these numbers saying that generative video has a smaller footprint than the film shoots and travel that go into typical video production. That claim is hard to test and doesn't account for the surge in video generation that might follow if AI videos become cheap to produce.

**All in a day's prompt**

So what might a day's energy consumption look like for one person with an AI habit?

There is a significant caveat to this math. These numbers cannot serve as a proxy for how much energy is required to power something like ChatGPT 4o. We don't know how many parameters are in OpenAI's newest models, how many of those parameters are used for different model architectures, or which data centers are used and how OpenAI may distribute requests across all these systems. You can guess, as many have done, but those guesses are so approximate that they may be more distracting than helpful.

"We should stop trying to reverse-engineer numbers based on hearsay," Luccioni says, "and put more pressure on these companies to actually share the real ones." Luccioni has created the AI Energy Score, a way to rate models on their energy efficiency. But closed-source companies have to opt in. Few have, Luccioni says.

**Part Three: Fuel and emissions|**

Now that we have an estimate of the total energy required to run an AI model to produce text, images, and videos, we can work out what that means in terms of emissions that cause climate change.

First, a data center humming away isn't necessarily a bad thing. If all data centers were hooked up to solar panels and ran only when the sun was shining, the world would be talking a lot less about AI's energy consumption. That's not the case. Most electrical grids around the world are still heavily reliant on fossil fuels. So electricity use comes with a climate toll attached.

"AI data centers need constant power, 24-7, 365 days a year," says Rahul Mewawalla, the CEO of Mawson Infrastructure Group, which builds and maintains high-energy data centers that support AI.

That means data centers can't rely on intermittent technologies like wind and solar power, and on average, they tend to use dirtier electricity. One preprint study from Harvard's T.H. Chan School of Public Health found that the carbon intensity of electricity used by data centers was 48% higher than the US average. Part of the reason is that data centers currently happen to be clustered in places that have dirtier grids on average, like the coal-heavy grid in the mid-Atlantic region that includes Virginia, West Virginia, and Pennsylvania. They also run constantly, including when cleaner sources may not be available.

**Tech companies like Meta, Amazon, and Google have responded to this fossil fuel issue by announcing goals to use more nuclear power. Those three have joined a pledge to triple the world's nuclear capacity by 2050. But today, nuclear energy only accounts for 20% of electricity supply in the US, and powers a fraction of AI data centers' operations—natural gas accounts for more than half of electricity generated in Virginia, which has more data centers than any other US state, for example. What's more, new nuclear operations will take years, perhaps decades, to materialize.**

In 2024, fossil fuels including natural gas and coal made up just under 60% of electricity supply in the US. Nuclear accounted for about 20%, and a mix of renewables accounted for most of the remaining 20%.

Gaps in power supply, combined with the rush to build data centers to power AI, often mean shortsighted energy plans. In April, Elon Musk's X supercomputing center near Memphis was found, via satellite imagery, to be using dozens of methane gas generators that the Southern Environmental Law Center alleges are not approved by energy regulators to supplement grid power and are violating the Clean Air Act.

The key metric used to quantify the emissions from these data centers is called the carbon intensity: how many grams of carbon dioxide emissions are produced for each kilowatt-hour of electricity consumed. Nailing down the carbon intensity of a given grid requires understanding the emissions produced by each individual power plant in operation, along with the amount of energy each is contributing to the grid at any given time. Utilities, government agencies, and researchers use estimates of average emissions, as well as real-time measurements, to track pollution from power plants.

This intensity varies widely across regions. The US grid is fragmented, and the mixes of coal, gas, renewables, or nuclear vary widely. California's grid is far cleaner than West Virginia's, for example.

Time of day matters too. For instance, data from April 2024 shows that California's grid can swing from under 70 grams per kilowatt-hour in the afternoon when there's a lot of solar power available to over 300 grams per kilowatt-hour in the middle of the night.

This variability means that the same activity may have very different climate impacts, depending on your location and the time you make a request. Take that charity marathon runner, for example. The text, image, and video responses they requested add up to 2.9 kilowatt-hours of electricity. In California, generating that amount of electricity would produce about 650 grams of carbon dioxide pollution on average. But generating that electricity in West Virginia might inflate the total to more than 1,150 grams.

**AI around the corner**

What we've seen so far is that the energy required to respond to a query can be relatively small, but it can vary a lot, depending on the type of query and the

model being used. The emissions associated with that given amount of electricity will also depend on where and when a query is handled. But what does this all add up to?

ChatGPT is now estimated to be the fifth-most visited website in the world, just after Instagram and ahead of X. In December, OpenAI said that ChatGPT receives 1 billion messages every day, and after the company launched a new image generator in March, it said that people were using it to generate 78 million images per day, from Studio Ghibli–style portraits to pictures of themselves as Barbie dolls.

Given the direction AI is headed—more personalized, able to reason and solve complex problems on our behalf, and everywhere we look—it's likely that our AI footprint today is the smallest it will ever be.

One can do some very rough math to estimate the energy impact. In February the AI research firm Epoch AI published an estimate of how much energy is used for a single ChatGPT query—an estimate that, as discussed, makes lots of assumptions that can't be verified. Still, they calculated about 0.3 watt-hours, or 1,080 joules, per message. This falls in between our estimates for the smallest and largest Meta Llama models (and experts we consulted say that if anything, the real number is likely higher, not lower).

One billion of these every day for a year would mean over 109 gigawatt-hours of electricity, enough to power 10,400 US homes for a year. If we add images and imagine that generating each one requires as much energy as it does with our high-quality image models, it'd mean an additional 35 gigawatt-hours, enough to power another 3,300 homes for a year. This is on top of the energy demands of OpenAI's other products, like video generators, and that for all the other AI companies and startups.

But here's the problem: These estimates don't capture the near future of how we'll use AI. In that future, we won't simply ping AI models with a question or two throughout the day, or have them generate a photo. Instead, leading labs are racing us toward a world where AI "agents" perform tasks for us without our supervising their every move. We will speak to models in voice mode, chat with companions for 2 hours a day, and point our phone cameras at our surroundings in video mode. We will give complex tasks to so-called "reasoning models" that work through tasks logically but have been found to require 43 times more energy for simple problems, or "deep research" models that spend hours creating reports for us. We will have AI models that are "personalized" by training on our data and preferences.

This future is around the corner: OpenAI will reportedly offer agents for $20,000 per month and will use reasoning capabilities in all of its models moving forward, and DeepSeek catapulted "chain of thought" reasoning into the mainstream with a model that often generates nine pages of text for each response. AI models are being added to everything from customer service phone lines to doctor's offices, rapidly increasing AI's share of national energy consumption.

“The precious few numbers that we have may shed a tiny sliver of light on where we stand right now, but all bets are off in the coming years,” says Luccioni.

**Every researcher we spoke to said that we cannot understand the energy demands of this future by simply extrapolating from the energy used in AI queries today. And indeed, the moves by leading AI companies to fire up nuclear power plants and create data centers of unprecedented scale suggest that their vision for the future would consume far more energy than even a large number of these individual queries.**

“The precious few numbers that we have may shed a tiny sliver of light on where we stand right now, but all bets are off in the coming years,” says Luccioni. “Generative AI tools are getting practically shoved down our throats and it’s getting harder and harder to opt out, or to make informed choices when it comes to energy and climate.”

To understand how much power this AI revolution will need, and where it will come from, we have to read between the lines.

**Part four: The future ahead|**

A report published in December by the Lawrence Berkeley National Laboratory, which is funded by the Department of Energy and has produced 16 Nobel Prizes, attempted to measure what AI’s proliferation might mean for energy demand.

**In analyzing both public and proprietary data about data centers as a whole, as well as the specific needs of AI, the researchers came to a clear conclusion. Data centers in the US used somewhere around 200 terawatt-hours of electricity in 2024, roughly what it takes to power Thailand for a year. AI-specific servers in these data centers are estimated to have used between 53 and 76 terawatt-hours of electricity. On the high end, this is enough to power more than 7.2 million US homes for a year.**

If we imagine the bulk of that was used for inference, it means enough electricity was used on AI in the US last year for every person on Earth to have exchanged more than 4,000 messages with chatbots. In reality, of course, average individual users aren’t responsible for all this power demand. Much of it is likely going toward startups and tech giants testing their models, power users exploring every new feature, and energy-heavy tasks like generating videos or avatars.

**Data centers in the US used somewhere around 200 terawatt-hours of electricity in 2024, roughly what it takes to power Thailand for a year.**

**By 2028, the researchers estimate, the power going to AI-specific purposes will rise to between 165 and 326 terawatt-hours per year. That’s more than all electricity currently used by US data centers for all purposes; it’s enough to power 22% of US households each year. That**

**could generate the same emissions as driving over 300 billion miles—over 1,600 round trips to the sun from Earth.**

The researchers were clear that adoption of AI and the accelerated server technologies that power it has been the primary force causing electricity demand from data centers to skyrocket after remaining stagnant for over a decade. Between 2024 and 2028, the share of US electricity going to data centers may triple, from its current 4.4% to 12%.

This unprecedented surge in power demand for AI is in line with what leading companies are announcing. SoftBank, OpenAI, Oracle, and the Emirati investment firm MGX intend to spend $500 billion in the next four years on new data centers in the US. The first has started construction in Abilene, Texas, and includes eight buildings that are each the size of a baseball stadium. In response to a White House request for information, Anthropic suggested that the US build an additional 50 gigawatts of dedicated power by 2027.

**AI companies are also planning multi-gigawatt constructions abroad, including in Malaysia, which is becoming Southeast Asia's data center hub. In May OpenAI announced a plan to support data-center buildouts abroad as part of a bid to "spread democratic AI." Companies are taking a scattershot approach to getting there—inking deals for new nuclear plants, firing up old ones, and striking massive deals with utility companies.**

*MIT Technology Review* sought interviews with Google, OpenAI, and Microsoft about their plans for this future, and for specific figures on the energy required to inference leading AI models. OpenAI declined to provide figures or make anyone available for an interview but provided a statement saying that it prioritizes efficient use of computing resources and collaborates with partners to support sustainability goals, and that AI might help discover climate solutions. The company said early sites for its Stargate initiative will be natural gas and solar powered and that the company will look to include nuclear and geothermal wherever possible.

Microsoft discussed its own research on improving AI efficiencies but declined to share specifics of how these approaches are incorporated into its data centers.

Google declined to share numbers detailing how much energy is required at inference time for its AI models like Gemini and features like AI Overviews. The company pointed to information about its TPUs—Google's proprietary equivalent of GPUs—and the efficiencies they've gained.

**The Lawrence Berkeley researchers offered a blunt critique of where things stand, saying that the information disclosed by tech companies, data center operators, utility companies, and hardware manufacturers is simply not enough to make reasonable projections about the unprecedented energy demands of this future or estimate the emissions it will create. They offered ways that companies could disclose more**

**information without violating trade secrets, such as anonymized data-sharing arrangements, but their report acknowledged that the architects of this massive surge in AI data centers have thus far not been transparent, leaving them without the tools to make a plan.**

“Along with limiting the scope of this report, this lack of transparency highlights that data center growth is occurring with little consideration for how best to integrate these emergent loads with the expansion of electricity generation/transmission or for broader community development,” they wrote. The authors also noted that only two other reports of this kind have been released in the last 20 years.

We heard from several other researchers who say that their ability to understand the emissions and energy demands of AI are hampered by the fact that AI is not yet treated as its own sector. The US Energy Information Administration, for example, makes projections and measurements for manufacturing, mining, construction, and agriculture, but detailed data about AI is simply nonexistent.

“Why should we be paying for this infrastructure? Why should we be paying for their power bills?”

Individuals may end up footing some of the bill for this AI revolution, according to new research published in March. The researchers, from Harvard’s Electricity Law Initiative, analyzed agreements between utility companies and tech giants like Meta that govern how much those companies will pay for power in massive new data centers. They found that discounts utility companies give to Big Tech can raise the electricity rates paid by consumers. In some cases, if certain data centers fail to attract the promised AI business or need less power than expected, ratepayers could still be on the hook for subsidizing them. A 2024 report from the Virginia legislature estimated that average residential ratepayers in the state could pay an additional $37.50 every month in data center energy costs.

“It’s not clear to us that the benefits of these data centers outweigh these costs,” says Eliza Martin, a legal fellow at the Environmental and Energy Law Program at Harvard and a coauthor of the research. “Why should we be paying for this infrastructure? Why should we be paying for their power bills?”

When you ask an AI model to write you a joke or generate a video of a puppy, that query comes with a small but measurable energy toll and an associated amount of emissions spewed into the atmosphere. Given that each individual request often uses less energy than running a kitchen appliance for a few moments, it may seem insignificant.

But as more of us turn to AI tools, these impacts start to add up. And increasingly, you don’t need to go looking to use AI: It’s being integrated into every corner of our digital lives.

**Crucially, there's a lot we don't know; tech giants are largely keeping quiet about the details. But to judge from our estimates, it's clear that AI is a force reshaping not just technology but the power grid and the world around us.**

At each of these centers, AI models are loaded onto clusters of servers containing special chips called graphics processing units, or GPUs, most notably a particular model made by Nvidia called the H100.

This chip started shipping in October 2022, just a month before ChatGPT launched to the public. Sales of H100s have soared since, and are part of why Nvidia regularly ranks as the most valuable publicly traded company in the world.

Other chips include the A100 and the latest Blackwells. What all have in common is a significant energy requirement to run their advanced operations without overheating.

A single AI model might be housed on a dozen or so GPUs, and a large data center might have well over 10,000 of these chips connected together.

Wired close together with these chips are CPUs (chips that serve up information to the GPUs) and fans to keep everything cool.

Some energy is wasted at nearly every exchange through imperfect insulation materials and long cables in between racks of servers, and many buildings use millions of gallons of water (often fresh, potable water) per day in their cooling operations.

Depending on anticipated usage, these AI models are loaded onto hundreds or thousands of clusters in various data centers around the globe, each of which have different mixes of energy powering them.

They’re then connected online, just waiting for you to ping them with a question.