How to Stop Your Data From Being Used to Train AI

If you've ever posted something to the internet—a pithy tweet, a 2009 blog post, a scornful review, or a selfie on Instagram—it has most likely been slurped up and used to help train the current wave of generative AI. Large language models, like ChatGPT, and image creators are powered by vast reams of our data. And even if it's not powering a chatbot, the data can be used for other machine-learning features.

Tech companies have scraped vast swathes of the web to gather the data they claim is needed to create generative AI—with little regard for content creators, copyright laws, or privacy. On top of this, increasingly, firms with reams of people's posts are looking to get in on the AI gold rush by selling or licensing that information. Looking at you, Reddit.

However, as the lawsuits and investigations around generative AI and its opaque data practices pile up, there have been small moves to give people more control over what happens to what they post online. Some companies now let individuals and business customers opt out of having their content used in AI training or being sold for training purposes. Here's what you can—and can't—do.

There's a Limit

Before we get to how you can opt out, it's worth setting some expectations. Many companies building AI have already scraped the web, so anything you've posted is probably already in their systems. Companies are also secretive about what they have actually scraped, purchased, or used to train their systems. "We honestly don't know that much," says Niloofar Mireshghallah, a researcher who focuses on AI privacy at the University of Washington. "In general, everything is very black-box."

Mireshghallah explains that companies can make it complicated to opt out of having data used for AI training, and even where it is possible, many people don't have a "clear idea" about the permissions they've agreed to or how data is being used. That's before various laws, such as copyright protections and Europe's strong privacy laws, are taken into consideration. Facebook, Google, X, and other companies have written into their privacy policies that they may use your data to train AI.

While there are various technical ways AI systems could have data removed from them or "unlearn," Mireshghallah says, there's very little that's known about the processes that are in place. The options can be buried or labor-intensive. Getting posts removed from AI training data is likely to be an uphill battle. Where companies are starting to allow opt-outs for future scraping or data sharing, they are almost always making users opt-in by default.

"Most companies add the friction because they know that people aren't going to go looking for it," says Thorin Klosowski, a security and privacy activist at the Electronic Frontier Foundation. "Opt-in would be a purposeful action, as opposed to opting out, where you have to know it's there."

While less common, some companies building AI tools and machine learning models don't automatically opt-in customers. "We do not train our models on user-submitted data by default. We may use user prompts and outputs to train Claude where the user gives us express permission to do so, such as clicking a thumbs up or down signal on a specific Claude output to provide us feedback," says Jennifer Martinez, a spokesperson for Anthropic. In this situation, the most recent iteration of the company's Claude chatbot is built on public information online and third-party data—content people posted elsewhere online—but not user information.

The majority of this guide deals with opt-outs for text, but artists have also been using "Have I Been Trained?" to signal their images shouldn't be used for training. Run by startup Spawning, the service allows people to see if their creations have been scraped and then opt out of any future training. "Anything with a URL can be opted out. Our search engine only searches images, but our browser extension lets you opt out any media type," says Jordan Meyer, cofounder and CEO of Spawning. Stability AI, the startup behind a text-to-image tool called Stable Diffusion, is among companies that say they are honoring the system.

The list below only includes companies currently with an opt-out process. For example, Microsoft's Copilot does not offer users with personal accounts the option to have their prompts not used to improve the software. "A portion of the total number of user prompts in Copilot and Copilot Pro responses are used to fine-tune the experience," says Donny Turnbaugh, a spokesperson for Copilot. "Microsoft takes steps to deidentify data before it is used, helping to protect consumer identity." Even if the data is deidentified, privacy-minded users may want more potential control over their information.

How to Opt Out of AI Training

Adobe

If you store your files in Adobe's Creative Cloud, the company may use them to train its machine-learning algorithm. "When we analyze your content for product improvement and development purposes, we first aggregate your content with other content and then use the aggregated content to train our algorithms and thus improve our products and services," reads the company's FAQ. This doesn't apply to any files stored only on your device.

If you're using a personal Adobe account, it's easy to opt out. Open up Adobe's **privacy page**, scroll down to the **Content analysis** section, and click the toggle to turn it off. For business or school accounts, the opt-out process is not available on the individual level, and you'll have to reach out to your administrator.

Amazon: AWS

AI services from Amazon Web Services, like Amazon Rekognition or Amazon CodeWhisperer, may save customer data to improve the company's tools. Head's up, this is the most complicated opt-out process included in the roundup, so you likely need help from an IT professional at your company or an AWS representative to perform it successfully. Outlined on this support page from Amazon, the process includes enabling the option for your organization, creating a policy, and attaching that policy where necessary.

Google: Gemini

For users of Google's chatbot, Gemini, conversations may sometimes be selected for human review to improve the AI model. Opting out is simple, though. Open up Gemini in your browser, click on **Activity**, and select the **Turn Off** drop-down menu. Here you can just turn off the Gemini Apps Activity, or you can opt out as well as delete your conversation data. While this does mean in most cases that future chats won't be seen for human review, already selected data is not erased through this process. According to Google's privacy hub for Gemini, these chats may stick around for three years.

Grammarly

Grammarly does not currently offer an opt-out process for personal accounts, but self-serve business accounts can choose to opt out from having their data used to train Grammarly's machine-learning model. Turn it off by opening up your **Account Settings**, clicking on the **Data Settings** tab, and toggling off **Product Improvement & Training**. If you have a managed business account, which includes accounts for classroom education and accounts bought through a Grammarly sales representative, you are automatically opted out from AI model training.

HubSpot

HubSpot, a popular marketing software, automatically uses data from customers to improve its machine-learning model. Unfortunately, there's not a button to press to turn off the use of data for AI training. You have to **send an email** to privacy@hubspot.com with a message requesting that the data associated with your account be opted out.

OpenAI: ChatGPT and Dall-E

People reveal all sorts of personal information while using a chatbot. OpenAI provides some options for what happens to what you say to ChatGPT—including allowing its future AI models not to be trained on the content. "We give users a number of easily accessible ways to control their data, including self-service tools to access, export, and delete personal information through ChatGPT. That includes easily accessible options to opt out from the use of their content to train models," says Taya Christianson, an OpenAI spokesperson. (The options vary slightly depending on your account type, and data from enterprise customers is not used to train models).

On its help pages, OpenAI says ChatGPT web users without accounts should navigate to **Settings** and then uncheck **Improve the model for everyone**. If you have an account and are logged in through a web browser, select **ChatGPT, Settings, Data Controls,** and then turn off **Chat History & Training**. If you're using ChatGPT's mobile apps, go to **Settings**, pick **Data Controls,** and turn off **Chat History & Training**. Changing these settings, OpenAI's support pages say, won't sync across different browsers or devices, so you need to make the change everywhere you use ChatGPT.

OpenAI is about a lot more than ChatGPT. For its Dall-E 3 image generator, the startup has a **form that allows you to send images** to be removed from "future training datasets." It asks for your name, email, whether you own the image rights or are getting in touch on behalf of a company, details of the image, and any uploads of the image(s). OpenAI also says if you have a "high volume" of images hosted online that you want removed from training data, then it may be "more efficient" to add GPTBot to the robots.txt file of the website where the images are hosted.

Traditionally a website's robots.txt file—a simple text file that usually sits at websitename.com/robots.txt—has been used to tell search engines, and others, whether they can include your pages in their results. It can now also be used to tell AI crawlers not to scrape what you have published—and AI companies have said they'll honor this arrangement.

Perplexity

Perplexity is a startup that uses AI to help you search the web and find answers to questions. Like all of the other software on this list, you are automatically opted in to having your interactions and data used to train Perplexity's AI further. Turn this off by clicking on your **account name**, scrolling down to the **Account** section, and turning off the **AI Data Retention** toggle.

Quora

Quora says it "currently" doesn't use answers to people's questions, posts, or comments for training AI. It also hasn't sold any user data for AI training, a spokesperson says. However, it does offer opt-outs in case this changes in the future. To do this, visit its **Settings** page, click to **Privacy,** and turn off the "**Allow large language models to be trained on your content**" option. Despite this choice, there are some Quora posts that may be used for training LLMs. If you reply to a machine-generated answer, the company's help pages say, then those answers may be used for AI training. It points out that third parties may just scrape its content anyway.

Rev

Rev, a voice transcription service that uses both human freelancers and AI to transcribe audio, says it uses data "perpetually" and "anonymously" to train its AI systems. Even if you delete your account, it will still train its AI on that information.

Kendell Kelton, head of brand and corporate communications at Rev, says it has the "largest and most diverse data set of voices," made up of more than 6.5 million hours of voice recording. Kelton says Rev does not sell user data to any third parties. The firm's terms of service say data will be used for training, and that customers are able to opt out. People can opt out of their data being used by **sending an email** to support@rev.com, its help pages say.

Slack

All of those random Slack messages at work might be used by the company to train its models as well. "Slack has used machine learning in its product for many years. This includes platform-level machine-learning models for things like channel and emoji recommendations," says Jackie Rocca, a vice president of product at Slack who's focused on AI.

Even though the company does not use customer data to train a large language model for its Slack AI product, Slack may use your interactions to improve the software's machine-learning capabilities. "To develop AI/ML models, our systems analyze Customer Data (e.g. messages, content, and files) submitted to Slack," says Slack's privacy page. Similar to Adobe, there's not much you can do on an individual level to opt out if you're using an enterprise account.

The only real way to opt out is to have your administrator email Slack at feedback@slack.com. The message must have the subject line "Slack Global model opt-out request" and include your organization's URL. Slack doesn't provide a timeline for how long the opt-out process takes, but it should send you a confirmation email after it's complete.

Squarespace

Website-building tool Squarespace has built in a toggle to stop AI crawlers from scraping websites it hosts. This works by updating your website's robots.txt file to tell AI companies the content is off limits. To block the AI bots, open **Settings** within your account, find **Crawlers**, and turn off **Artificial Intelligence Crawlers**. It points out this should work for the following crawlers: Anthropic, OpenAI's GPTBot and ChatGPT-User, Google Extended, and CCBot.

Substack

If you use Substack for blog posts, newsletters, or more, the company also has an easy option to apply the robots.txt opt-out. Within your **Settings** page, scroll to the **Publication** section and turn on the toggle to **Block AI training**. Its help page points out: "This will only apply to AI tools that respect this setting."

Tumblr

Blogging and publishing platform Tumblr—owned by Automattic, which also owns WordPress—says it is "working with" AI companies that are "interested in the very large and unique set of publicly published content" on the wider company's platforms. This doesn't include user emails or private content, an Automattic spokesperson says.

Tumblr has a "prevent third-party sharing" option to stop what you publish being used for AI training, as well as being shared with other third parties such as researchers. If you're using the Tumblr app, go to account S**ettings**, select your blog, click on the **gear icon**, select **Visibility**, and toggle the "**Prevent third-party sharing**" option. Explicit posts, deleted blogs, and those that are password-protected or private, are not shared with third-party companies in any case, Tumblr's support pages say.

WordPress

Like Tumblr, WordPress has a "prevent third-party sharing" option. To turn this on, visit your website's dashboard, click on **Settings**, **General**, and then through to **Privacy**, select the **Prevent third-party sharing** box. "We are also trying to work with crawlers (like https://commoncrawl.org/) to prevent content from being scraped and sold without giving our users choice or control over how their content is used," an Automattic spokesperson says.

Your Website

If you are hosting your own website, you can update your robots.txt file to tell AI bots not to scrape the pages. Most news websites don't allow their articles to be crawled by AI bots. WIRED's robots.txt file, for example, doesn't allow bots from OpenAI, Google, Amazon, Facebook, Anthropic, or Perplexity, among others. This opt-out isn't just for publishers though: Any website, big or small, can alter its robots file to exclude AI crawlers. All you need to do is add a disallow command; working examples can be found here.