Sites scramble to block ChatGPT web crawler after instructions emerge

reader comments

109 with

Without announcement, OpenAI recently added details about its web crawler, GPTBot, to its online documentation site. GPTBot is the name of the user agent that the company uses to retrieve webpages to train the AI models behind ChatGPT, such as GPT-4. Earlier this week, some sites quickly announced their intention to block GPTBot's access to their content.

In the new documentation, OpenAI says that webpages crawled with GPTBot "may potentially be used to improve future models," and that allowing GPTBot to access your site "can help AI models become more accurate and improve their general capabilities and safety."

OpenAI claims it has implemented filters ensuring that sources behind paywalls, those collecting personally identifiable information, or any content violating OpenAI's policies will not be accessed by GPTBot.

News of being able to potentially block OpenAI's training scrapes (if they honor them) comes too late to affect ChatGPT or GPT-4's current training data, which was scraped without announcement years ago. OpenAI collected the data ending in September 2021, which is the current "knowledge" cutoff for OpenAI's language models.

It’s worth noting that the new instructions may not prevent web-browsing versions of ChatGPT or ChatGPT plugins from accessing current websites to relay up-to-date information to the user. That point was not spelled out in the documentation, and we reached out to OpenAI for clarification.

The answer lies with robots.txt

According to OpenAI’s documentation, GPTBot will be identifiable by the user agent token “GPTBot,” with its full string being “Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; GPTBot/1.0; +https://openai.com/gptbot)”.

The OpenAI docs also give instructions about how to block GPTBot from crawling websites using the industry-standard robots.txt file, which is a text file that sits at the root directory of a website and instructs web crawlers (such as those used by search engines) not to index the site.

Advertisement
It’s as easy as adding these two lines to a site’s robots.txt file:

```
User-agent: GPTBot
Disallow: /
```

OpenAI also says that admins can restrict GPTBot from certain parts of the site in robots.txt with different tokens:

```
User-agent: GPTBot
Allow: /directory-1/
Disallow: /directory-2/
```

Additionally, OpenAI has provided the specific IP address blocks from which the GPTBot will be operating, which could be blocked by firewalls as well.

Despite this option, blocking GPTBot will not guarantee that a site’s data does not end up training all AI models of the future. Aside from issues of scrapers ignoring robots.txt

files, there are other large data sets of scraped websites (such as The Pile) that are not affiliated with OpenAI. These data sets are commonly used to train open source (or source-available) LLMs such as Meta's Llama 2.

Some sites react with haste

While wildly successful from a tech point of view, ChatGPT has also been controversial by how it scraped copyrighted data without permission and concentrated that value into a commercial product that circumvents the typical online publication model. OpenAI has been accused of (and sued for) plagiarism along these lines.

Accordingly, it's not surprising to see some people react to the news of being able to potentially block their content from future GPT models with a kind of pent-up relish. For example, on Tuesday, VentureBeat noted that The Verge, Substack writer Casey Newton, and Neil Clarke of Clarkesworld, all said they would block GPTBot soon after news of the bot broke.

But for large website operators, the choice to block large language model (LLM) crawlers isn't as easy as it may seem. Making some LLMs blind to certain website data will leave gaps of knowledge that could serve some sites very well (such as sites that don't want to lose visitors if ChatGPT supplies their information for them), but it may also hurt others. For example, blocking content from future AI models could decrease a site's or a brand's cultural footprint if AI chatbots become a primary user interface in the future. As a thought experiment, imagine an online business declaring that it didn't want its website indexed by Google in the year 2002—a self-defeating move when that was the most popular on-ramp for finding information online.

It's still early in the generative AI game, and no matter which way technology goes—or which individual sites attempt to opt out of AI model training—at least OpenAI is providing the

option.

https://arstechnica.com/?p=1960108