

Google adds a switch for publishers to opt out of becoming AI training data

Google just announced it's giving website publishers a way to opt out of having their data used to train the company's AI models while remaining accessible through Google Search. The new tool, called Google-Extended, allows sites to continue to get scraped and indexed by crawlers like the Googlebot while avoiding having their data used to train AI models as they develop over time.

The company says Google-Extended will let publishers "manage whether their sites help improve Bard and Vertex AI generative APIs," adding that web publishers can use the toggle to "control access to content on a site." Google confirmed in July that it's training its AI chatbot, Bard, on publicly available data scraped from the web.

Google-Extended is available through robots.txt, also known as the text file that informs web crawlers whether they can access certain sites. Google notes that "as AI applications expand," it will continue to explore "additional machine-readable approaches to choice and control for web publishers" and that it will have more to share soon.

Already, many sites have moved to block the web crawler that OpenAI uses to scrape data and train ChatGPT, including *The New York Times*, CNN, Reuters, and Medium. However, there have been concerns over how to block out Google. After all, websites can't close off Google's crawlers completely, or else they won't get indexed in search. This has led some sites, such as *The New York Times*, to legally block Google instead by

updating their terms of service to ban companies from using their content to train AI.

<https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex>