FOR IMMEDIATE RELEASE

Nearly Eight in Ten U.S. Websites That Refuse OpenAI's Training Crawler Allow China's DeepSeek Crawler

The Department of Homeland Security excluded China's DeepSeek products from federal civilian agencies on October 20, 2026, and The Answer Gap finds that 14,659 of 15,101 U.S. government websites with a clear result let the training crawler of China's DeepSeek in.

NEW YORK, November 25, 2026. The United States Department of Homeland Security published an order on October 20, 2026 (91 Federal Register 66346, https://www.govinfo.gov/content/pkg/FR-2026-10-20/pdf/2026-21447.pdf) that excludes China's DeepSeek products from procurement by federal civilian executive branch agencies and requires their removal from those agencies' information technology systems. The order answers whether federal civilian agencies may buy and use DeepSeek products. This leaves a second question open for an agency or organization as the publisher of a public website: when DeepSeek's training crawler visits that website to collect material, do the agency's own robots.txt rules let it in or keep it out?

The Answer Gap, a global study by Invictus Communications, checks each robots.txt file for 31 AI crawlers behind 16 AI assistants, models and data sets, made by companies and organizations in the United States, China, France and Canada. Of the 15,101 United States government websites with a clear result, at federal, state and local level, 14,659 (97.1 percent) let DeepSeekBot in, either because a rule allows it or because no rule covers it. DeepSeekBot is the training crawler of DeepSeek, a Chinese AI company that published the crawler's name in September 2026. Of the 1,347 of these government websites whose rules refuse GPTBot, the training crawler of United States-based OpenAI, 954 (70.8 percent) also refuse Bytespider, the training crawler of ByteDance, the Chinese company that makes the Doubao assistant, and allow DeepSeekBot.

A robots.txt file is a public file in which a website tells crawlers, the automated software that collects pages for AI companies and search engines, which pages they may access. A rule can name one crawler, or apply to all crawlers at once.

The free AI-Readable Sources Index shows which AI crawlers a website allows and compares the results by sector and country. The Answer Standard is a method for checking which pages assistants open, what they cite and how accurately they answer. Together, they help leadership, communications teams and industry bodies base decisions about access, representation and protection on evidence.

What the new findings show Across all United States websites in the research, 7,751 of the 55,868 with a clear result have rules that refuse GPTBot, the crawler that OpenAI, a United States company, uses to collect text for training the models behind ChatGPT.

Of those 7,751 websites, the rules of 9.9 percent allow Bytespider, the training crawler of ByteDance, a Chinese company.

Of the same 7,751 websites, the rules of 79.7 percent allow DeepSeekBot, the training crawler of DeepSeek, a Chinese company.

Many of the same files also object to AI training in general. Of the 7,751 U.S. websites whose rules refuse GPTBot, 5,420 also refuse Bytespider and allow DeepSeekBot, and 2,813 of those carry the line “ai-train=no”, a signal that their content may not be used to train AI. In Canada, 336 of the 525 websites in the same group carry it.

The robots.txt file of The New York Times refuses the crawlers of OpenAI, Anthropic, Perplexity, Common Crawl and ByteDance, and names neither DeepSeekBot nor GrokBot. Its front page is protected by DataDome, a security service against automated visitors. A robots.txt file states a request, and a security service such as DataDome decides separately which visitors to admit. Worldwide, the rules of 30,143 of the 240,870 websites with a clear result refuse GPTBot. Of those, 87.2 percent also refuse Bytespider and 76.2 percent allow DeepSeekBot. Sixteen assistants, one look-up The free AI-Readable Sources Index shows, for any website in the research, what its robots.txt file tells ChatGPT, Claude, Gemini, Grok, Perplexity, DeepSeek, Copilot, Meta AI, Apple Intelligence, Alexa, Le Chat, Doubao, Command, OLMo, DuckAssist and Common Crawl: https://invictuscomm.com/index/lookup.php

The Answer Gap examines digital inclusion through access to information, digital identity, representation and informed choice. Digital inclusion concerns both obtaining information and making it available to others, and it applies to people with limited access to technology and to some of the world's largest institutions. An organization's digital inclusion can also affect the people and communities it serves, because customers, employees, investors, students and citizens may use AI answers in their search for information. Under the report's guiding principle, each organization decides what to open to AI, what to protect and whether its content may train AI.

The report summary is at https://invictuscomm.com/research/the-answer-gap/ Journalists, researchers and regulators can request the full report free of charge.

Images free to publish with credit, including eight United States infographics: https://invictuscomm.com/research/the-answer-gap/#media

Correction, November 26, 2026: an earlier version of this release gave two figures for GrokBot, and both have been withdrawn. It stated that the rules of 85.2 percent of the 7,751 United States websites that refuse GPTBot allow GrokBot. It gave 80.2 percent for the 30,143 such websites worldwide. SpaceX Super Intelligence, the company that makes Grok, publishes no name for the crawler Grok uses, so a rule for GrokBot may not apply to it.

About Invictus Communications Invictus Communications is a communications and research consultancy.

Media contact Lindsay McDonald, Invictus Communications, lindsay@invictuscomm.com