

ESTUDIO DE INVESTIGACIÓN DE PPN WORLD N.º 1 · EDICIÓN 1

El 94 % de los comunicados de prensa está abierto a los rastreadores de IA

Un censo prerregistrado de las reglas para rastreadores de IA en 1503 sitios de comunicados



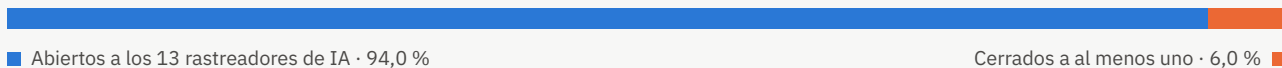
Por **Émile Bolduc** · PPN World
Publicado el 8 de octubre de 2026

Leímos el archivo robots.txt detrás de cada uno de los 500.129 comunicados de prensa. El 94,0 % está abierto a todos los rastreadores de IA que probamos. Cuando un editor rechaza, rechaza entrenar a la IA, no que esta lo cite.

Edición 1 · Archivos leídos el 7 de octubre de 2026 · Publicada el 8 de octubre de 2026 · Prerregistrada
<https://www.ppnworld.com/es/research/press-releases-ai-crawlers> · Textos, figuras y datos: CC BY 4.0

94,0 %

de los 500.129 comunicados de prensa está abierto a todos los rastreadores de IA probados



500.129

COMUNICADOS VERIFICADOS

1503

SITIOS

13

RASTREADORES DE IA PROBADOS



PRERREGISTRADO

RESUMEN

Los asistentes y los motores de respuesta ya repiten comunicados de prensa, lo que plantea una pregunta práctica a quienes los publican: ¿están esos comunicados abiertos a los rastreadores de IA? Antes de leer un solo archivo, registramos tres hipótesis. Después leímos el archivo robots.txt de cada sitio detrás de los 500.129 comunicados del archivo de PPN World del 1 de julio de 2026 al 6 de octubre de 2026 —1503 sitios de empresas, distribuidores, gobiernos, ciudades, universidades y ligas— y aplicamos cada archivo, como especifica la RFC 9309, a la dirección de cada comunicado, para 13 rastreadores de IA y para los de Google y Bing. El 94,0 % de los comunicados está en sitios que dejan entrar a todos esos rastreadores de IA, y el 99,0 % está abierto a los tres rastreadores que alimentan las respuestas de búsqueda con IA. Los rechazos son escasos y apuntan al entrenamiento: el 3,3 % de los sitios aparta al rastreador de entrenamiento de OpenAI, frente al 0,6 % para Googlebot (H1), y 28 sitios rechazan ese rastreador de entrenamiento mientras admiten el de búsqueda de OpenAI, y ninguno hace lo contrario (H2). Organismos públicos y editores comerciales no difieren (3,8 % frente a 4,2 %; H3 no confirmada), y las copias archivadas muestran el mismo panorama doce meses antes. A finales de 2023, casi la mitad de los sitios de noticias más leídos ya bloqueaba a los rastreadores de OpenAI (Fletcher, 2024); los comunicados han seguido abiertos.

CONTENIDO

- 01 Hallazgos principales
- 02 Por qué medirlo
- 03 Cómo lo medimos
- 04 Casi todos los comunicados están abiertos
- 05 Entrenamiento rechazado, respuesta admitida
- 06 Quién rechaza
- 07 Un año antes
- 08 La otra barrera: el cortafuegos
- 09 Qué significa para los comunicadores
- 10 Notas de método
- 11 Datos, prerregistro y cita

01 Hallazgos principales

01

94,0 %

El 94 % de los comunicados está abierto a todos los rastreadores de IA

73 de los 1443 sitios que respondieron (5,1 %) rechazan al menos uno de los 13 rastreadores de IA probados; ponderado por comunicados, el 6,0 %. Los cinco sitios más grandes que respondieron —con el 42,1 % de esos comunicados— no rechazan ninguno.

02

99,0 %

La búsqueda con IA está aún más abierta

Solo 34 sitios (2,4 %) apartan siquiera a uno de los rastreadores que alimentan las respuestas de búsqueda con IA, los de OpenAI, Anthropic o Perplexity. El 99,0 % de los comunicados está abierto a los tres.

03

3,3 % frente a 0,6 %

Los rastreadores de IA se rechazan más que Google (H1)

48 sitios (3,3 %) rechazan a GPTBot; 9 (0,6 %) rechazan a Googlebot. 39 rechazan a GPTBot y admiten a Googlebot; ninguno hace lo contrario ($p < 0,001$). Esperado, y ahora replicado a escala completa.

04

28 frente a 0

Los editores rechazan el entrenamiento, no la respuesta (H2)

28 sitios rechazan a GPTBot, el rastreador de entrenamiento de OpenAI, y admiten a OAI-SearchBot, que alimenta la búsqueda de ChatGPT. Ninguno hace lo contrario ($p < 0,001$). Todo sitio que aparta a algún rastreador de IA aparta a uno de entrenamiento.

05

3,8 % frente a 4,2 %

Gobiernos y empresas se comportan igual (H3, no confirmada)

El 3,8 % de los sitios de organismos públicos y el 4,2 % de los comerciales rechazan a GPTBot (diferencia de $-0,3$ puntos, IC del 95 % $-3,4$ a $+2,0$ puntos). Los distribuidores lo rechazan más que las salas de prensa de las empresas (8,3 % frente a 2,0 %), pero en ambos casos es poco común.

06

4,3 % → 4,7 %

Un año no cambió nada

Frente a copias archivadas de los mismos archivos del otoño de 2025 (1152 sitios), los bloqueos dirigidos a la IA pasaron del 4,3 % al 4,7 % de los sitios: 21 empezaron y 17 dejaron de hacerlo (p 0,63).

07

75 sitios

El cortafuegos es la otra barrera

75 sitios (5,0 %) rechazaron nuestro cliente automatizado en la propia página del comunicado, incluso cuando se presentaba como navegador: aproximadamente tantos como los que rechazan a un rastreador de IA en el robots.txt. Si un rastreador de IA real atraviesa esos cortafuegos no puede verse desde fuera.

02 Por qué medirlo

Un comunicado de prensa se escribe para que se repita. En 2026, esa repetición la hace cada vez más la IA: asistentes que responden preguntas, buscadores que resumen y los modelos detrás de ambos. Que un comunicado pueda llegar a ellos depende primero de un simple archivo de texto, el robots.txt, con el que un sitio le indica a cada rastreador adónde puede ir. El archivo distingue a los rastreadores por su nombre: un editor puede rechazar al que recopila datos de entrenamiento y admitir al que busca páginas para responder preguntas.

Los medios de noticias respondieron pronto y en voz alta: a finales de 2023, el 48 % de los sitios de noticias más usados en diez países bloqueaba a los rastreadores de OpenAI (Fletcher, 2024), y las restricciones a los rastreadores de IA se extendieron con rapidez entre las fuentes más usadas de la web (Longpre et al., 2024). Sobre los comunicados se ha discutido más que medido: a los comunicadores se les dice a la vez que las respuestas de la IA ya citan comunicados y que los editores le cierran la puerta a la IA. Este estudio lo mide, para una población completa de sitios de comunicados, con el análisis registrado antes de leer un solo archivo.

03 Cómo lo medimos

La población es cada comunicado del archivo de PPN World con fecha de publicación del 1 de julio de 2026 al 6 de octubre de 2026, de las fuentes que el archivo sirve: 500.129 comunicados con un enlace utilizable, en 1503 sitios. Un sitio es aquí un origen —esquema, host y puerto—, porque eso es lo que rige un archivo robots.txt. Cada sitio toma el tipo de editor de las fuentes que enlazan a él: 781 pertenecen a gobiernos y organismos públicos (nacionales, regionales y locales) y 346 a empresas y distribuidores comerciales.

El 7 de octubre de 2026 pedimos el robots.txt de cada sitio identificándonos como cliente de investigación y, si se nos rechazaba, otra vez como navegador. Cada archivo se aplicó exactamente como especifica la RFC 9309 (Koster et al., 2022) —el grupo que nombra al rastreador o, si no, el grupo *; gana la regla coincidente más larga— a la dirección de cada comunicado del sitio, no solo a su página de inicio, porque un sitio puede cerrar un directorio y dejar abierto el resto. Un sitio rechaza a un rastreador cuando más de la mitad de sus comunicados le están cerrados. Los 60 sitios cuyo archivo no pudo obtenerse quedan fuera de las tasas y se cuentan como rechazos en una prueba de robustez.

Probamos 13 rastreadores de IA en tres familias fijadas de antemano según la documentación de cada operador (Anthropic, s. f.; Apple, s. f.; Common Crawl, s. f.; Google, s. f.; Meta, s. f.; OpenAI, s. f.; Perplexity, s. f.): los que recopilan datos de entrenamiento (GPTBot, ClaudeBot, CCBot, Meta-ExternalAgent, Bytespider y los controles de entrenamiento Google-Extended y Applebot-Extended); los que indexan páginas para las respuestas de búsqueda con IA (OAI-SearchBot, Claude-SearchBot, PerplexityBot); y los que abren una página cuando un usuario lo pide (ChatGPT-User, Claude-User, Perplexity-User). Googlebot y Bingbot son la referencia. Registramos tres hipótesis, con corrección de Holm: los rastreadores de IA se rechazan más que Googlebot (H1); el rastreador de entrenamiento de OpenAI más que su rastreador de búsqueda (H2); y los organismos públicos difieren de los editores comerciales (H3).

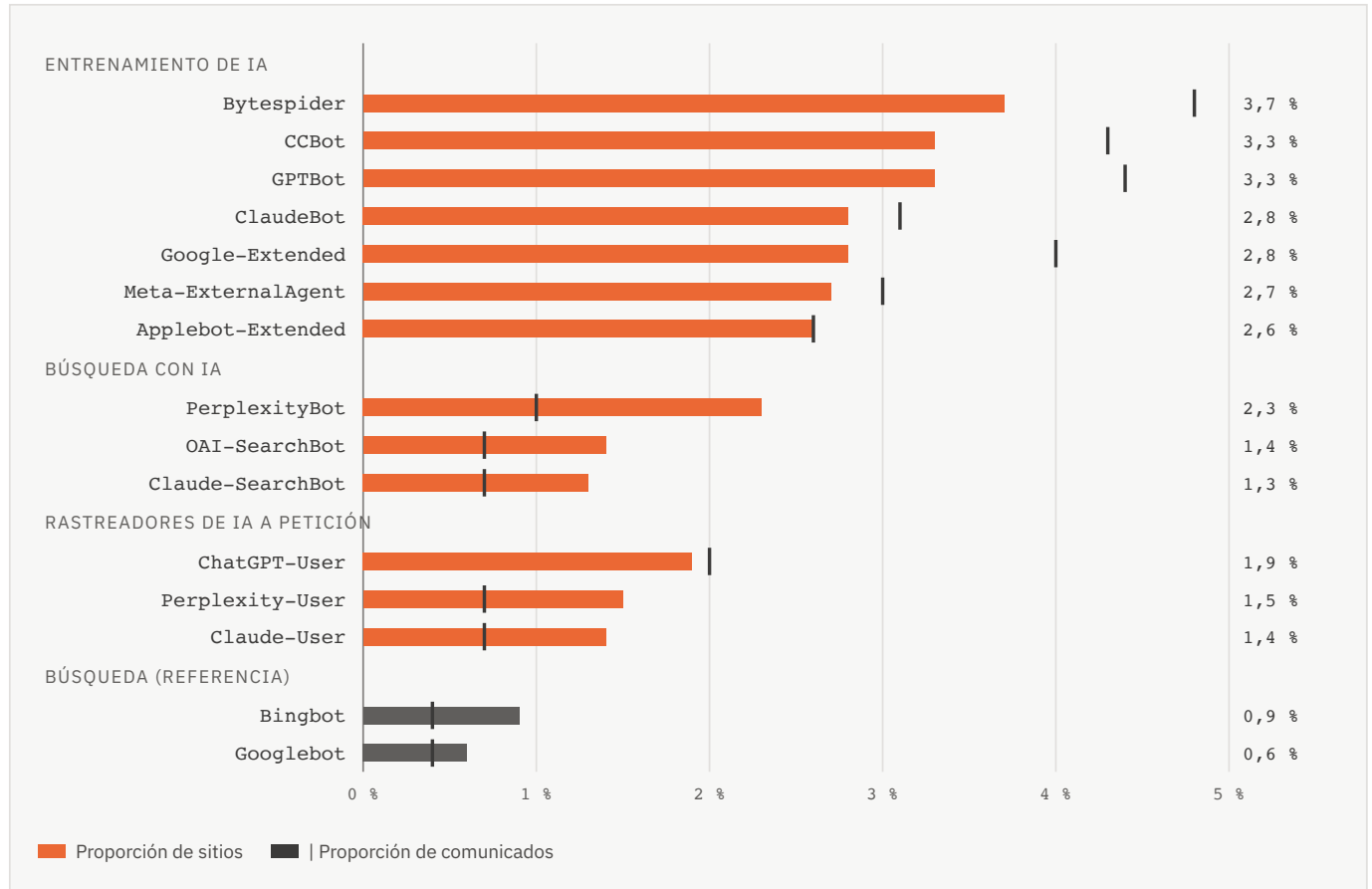
04 Casi todos los comunicados están abiertos

Casi ningún archivo cierra la puerta. 73 de los 1443 sitios que respondieron (5,1 %, IC del 95 % 4,0 % a 6,3 %) rechazan al menos a un rastreador de IA, y 34 rechazan a uno de búsqueda con IA. Ponderado por los comunicados de cada sitio, el 94,0 % de los comunicados está abierto a todos los rastreadores de IA y el 99,0 % a los tres de búsqueda con IA. Los cinco sitios más grandes que respondieron —con el 42,1 % de esos comunicados— no rechazan ninguno.

FIGURA 1

Los rastreadores que rechazan los sitios de comunicados

Proporción de sitios cuyo robots.txt cierra la mayoría de sus direcciones de comunicados a cada rastreador (barras), y proporción de comunicados cerrados a él (marcas). Sitios que respondieron por su robots.txt.



Los rastreadores más rechazados son los que recopilan datos de entrenamiento:

Bytespider, GPTBot y CCBot encabezan la lista. Los menos rechazados son los de búsqueda con IA y los que abren una página para un usuario. Googlebot es rechazado por 9 sitios, 8 de ellos porque cierran sus comunicados a todos los rastreadores de la lista. H1 se confirma: 39 sitios rechazan a GPTBot y admiten a Googlebot, y ninguno hace lo contrario (prueba exacta de McNemar, p ajustada por Holm $< 0,001$).

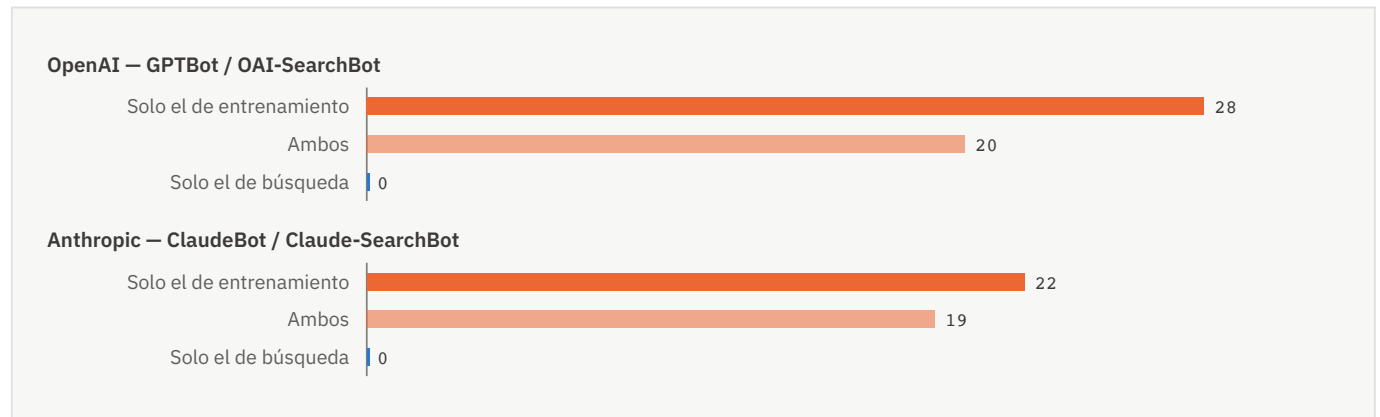
05 Entrenamiento rechazado, respuesta admitida

H2 preguntaba si los editores distinguen entre entrenamiento y respuesta. Lo hacen, en un solo sentido. De los sitios que respondieron, 20 rechazan a los dos rastreadores de OpenAI; 28 rechazan a GPTBot pero admiten a OAI-SearchBot, el rastreador con el que, según OpenAI, aparecen sitios en la búsqueda de ChatGPT (OpenAI, s. f.); y ninguno hace lo contrario (prueba exacta de McNemar, p ajustada por Holm $< 0,001$; H2 confirmada). El par de Anthropic se reparte igual: 22 sitios rechazan solo a ClaudeBot y ninguno solo a Claude-SearchBot.

FIGURA 2

Entrenamiento rechazado, respuesta admitida

Sitios que rechazan el rastreador de entrenamiento de cada empresa, su rastreador de búsqueda o ambos. Ningún sitio rechaza solo el de búsqueda.



Parte de la diferencia viene de cómo se escriben los archivos. 35 de los 48 sitios que rechazan a GPTBot lo nombran en su archivo; los rastreadores de búsqueda, más recientes, rara vez se nombran, así que caen bajo el grupo *, que suele admitirlos. Algunos editores eligen la separación de forma explícita: gov.ie, en Irlanda, nombra a GPTBot, ClaudeBot, Google-Extended y Meta-ExternalAgent y deja en paz a los rastreadores de búsqueda. En cualquier caso, un comunicado cerrado al entrenamiento sigue, casi siempre, abierto a ser citado.

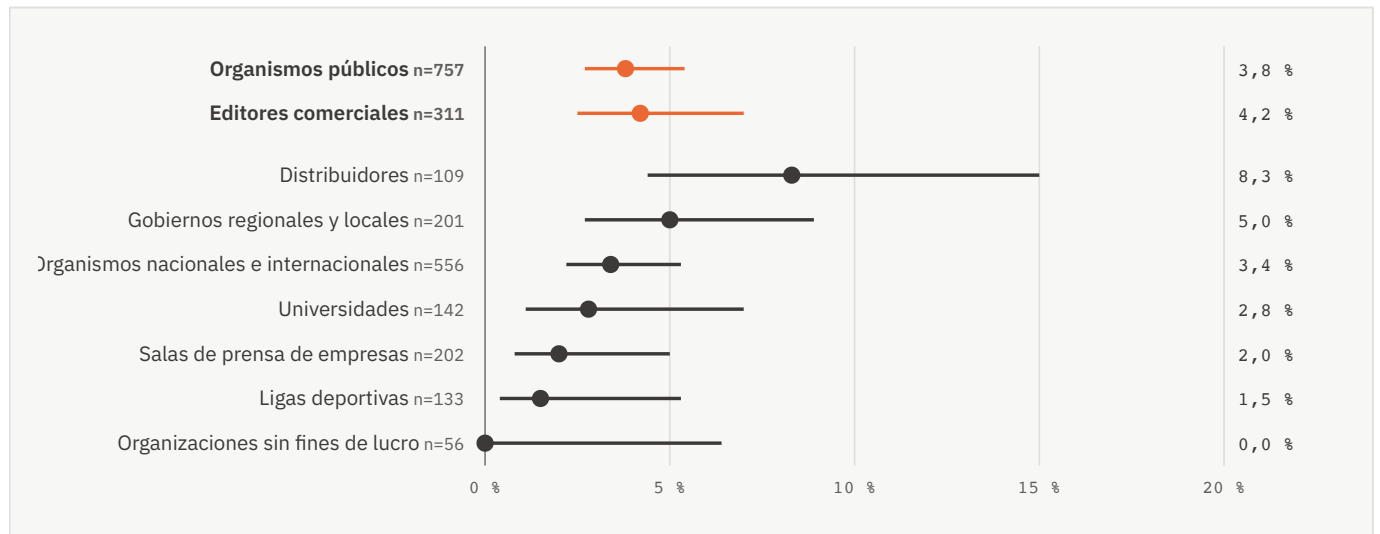
06 Quién rechaza

H3 no encontró diferencias entre organismos públicos y editores comerciales. El 3,8 % de los sitios de organismos públicos (29 de 757) y el 4,2 % de los comerciales (13 de 311) rechazan a GPTBot; la diferencia, $-0,3$ puntos (IC del 95 % $-3,4$ a $+2,0$ puntos), no se distingue de cero (p ajustada por Holm 0,86; H3 no confirmada). Dentro del grupo comercial, los sitios de distribución rechazan a GPTBot más a menudo que las salas de prensa de las empresas: 9 de 109 (8,3 %) frente a 4 de 202 (2,0 %).

FIGURA 3

Quién rechaza a GPTBot

Proporción de sitios que rechazan a GPTBot, por tipo de editor, con intervalos del 95 %. Las dos filas agrupadas son la comparación de H3. Sitios que respondieron por su robots.txt.



Los organismos públicos son de registro público, así que los que rechazan pueden nombrarse. Los que más comunicados publican son el gobierno de la ciudad de Moscú, la Generalitat Valenciana, el Ayuntamiento de Barcelona, gov.ie en Irlanda, la autoridad española de la competencia (CNMC), la oficina del gobernador de Texas, el Ministerio del Interior de Austria y la Organización Mundial del Comercio. El servicio regional de noticias de la Toscana va más allá y cierra sus páginas de comunicados a todos los rastreadores, Google incluido.

Entre los países con al menos 20 sitios, rechazar a un rastreador de IA es más habitual en parte de Europa —Países Bajos (16,7 %), España (16,1 %), Alemania (13,0 %)— que en Estados Unidos (4,2 %), Canadá (4,4 %) o Japón (1,9 %). Los grupos son pequeños y las cifras son descriptivas.

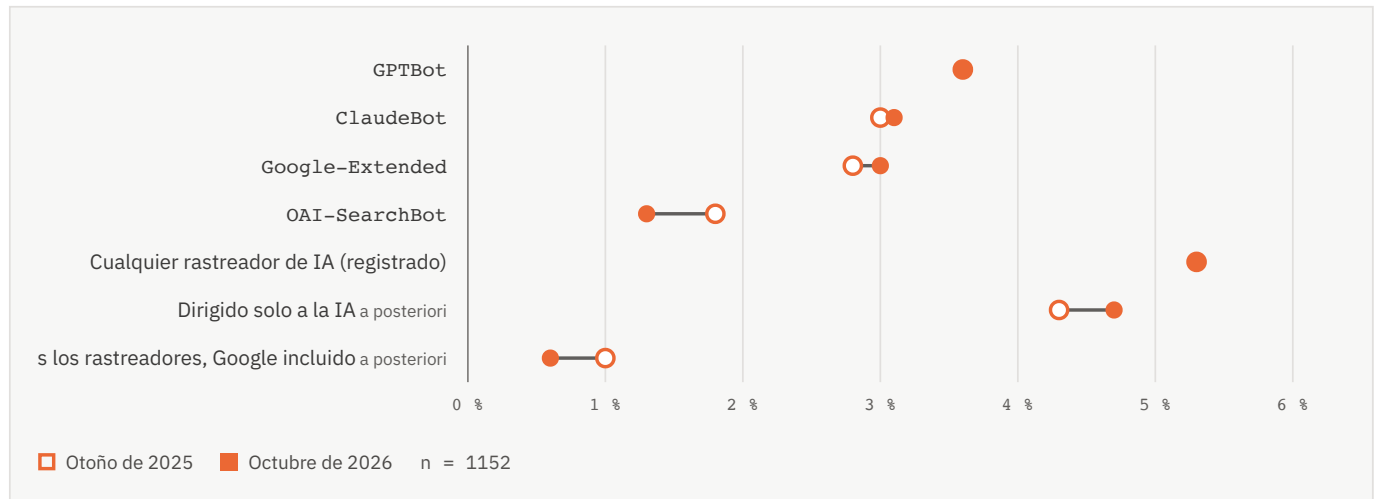
07 Un año antes

Para ver si esto cambia, leímos en la Wayback Machine de Internet Archive (Internet Archive, s. f.) la copia archivada del robots.txt de cada sitio más cercana al 7 de octubre de 2025, con dos meses de margen a cada lado, y la aplicamos a las mismas direcciones de comunicados. 1152 sitios (el 80 % de los que respondieron) tenían una copia utilizable; la copia mediana estaba a 6 días de la fecha buscada.

FIGURA 4

Doce meses después, lo mismo

Proporción de sitios que rechazan cada rastreador en la copia archivada de su robots.txt más cercana al 7 de octubre de 2025 (hueco) y hoy (relleno), en los sitios con una copia utilizable. Las filas marcadas a posteriori se separaron después de ver los datos.



La proporción que rechazaba a GPTBot era del 3,6 % entonces y del 3,6 % ahora. Con la medida registrada —cualquier rastreador de IA rechazado—, 61 sitios rechazaban entonces y 61 ahora, con 23 que empezaron y 23 que dejaron de hacerlo. Ese vaivén mezcla dos cosas, que separamos después de verlo: los bloqueos generales que cierran un sitio a todos los rastreadores, Google incluido (11 sitios entonces, 7 ahora), y los bloqueos dirigidos solo a la IA, que pasaron del 4,3 % al 4,7 % de los sitios (21 empezaron, 17 dejaron de hacerlo; p 0,63). Los editores modifican sus archivos en los dos sentidos; el total no se ha movido.

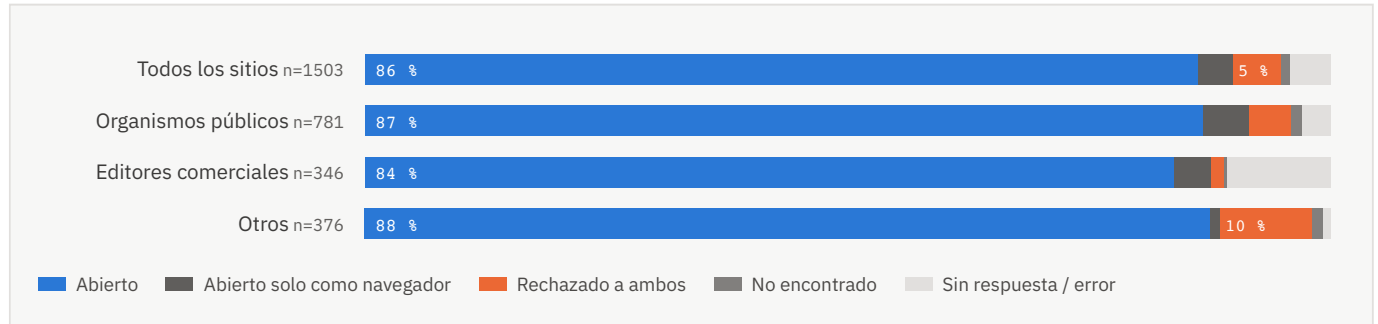
08 La otra barrera: el cortafuegos

El robots.txt es una petición, y está escrita. Un cortafuegos delante de un sitio se aplica, y no está escrito en ningún lugar que un rastreador pueda leer. Al abrir el comunicado más reciente de cada sitio, 75 sitios (5,0 %) rechazaron de entrada nuestro cliente, tanto si se identificaba honestamente como si lo hacía como navegador, y otros 54 rechazaron la identidad honesta pero abrieron a la de navegador. 60 sitios no dieron ninguna respuesta utilizable por su robots.txt (tiempos agotados, errores de servidor, certificados inválidos); un segundo cliente HTTP, añadido después de la recogida, mostró que 30 de ellos —todos sitios de empresas, la mayoría páginas de relación con inversores— dejan sin respuesta a un cliente de investigación identificado y responden con un 403 a una identidad de navegador.

FIGURA 5

Lo que encontró un cliente automatizado en la página del comunicado

Resultado de abrir el comunicado más reciente de cada sitio: abierto a un cliente identificado honestamente; abierto solo a una identidad de navegador; rechazado a ambos; no encontrado; sin respuesta o error del servidor.



Un rastreador de IA real no es nuestro cliente. OpenAI y Google publican las direcciones de red que usan sus rastreadores, y un cortafuegos puede admitirlos y rechazar a todos los demás. Lo que sí puede decirse es que aproximadamente tantos sitios de comunicados ponen un cortafuegos contra clientes automatizados delante de sus comunicados como los que rechazan a un rastreador de IA en el robots.txt, y que lo segundo es una línea de texto que escribió el editor, mientras que lo primero suele ser un ajuste por defecto que el editor nunca vio.

Las demás señales son aún más raras. 29 de las 1351 páginas de comunicados que abrimos piden a los buscadores que no las indexen; 1 lleva la etiqueta no oficial noai; 10 archivos robots.txt incluyen Content Signals (Cloudflare, 2025), 5 de ellos para rechazar el entrenamiento de IA. En cambio, 81 sitios (5,4 %) publican un archivo llms.txt (Howard, 2024) —una guía escrita para sistemas de IA—, más de los que rechazan a un rastreador de IA.

09 Qué significa para los comunicadores

¿Bloquean los sitios de comunicados a los rastreadores de IA?

Rara vez. De los 1443 sitios que respondieron, detrás de 500.129 comunicados, el 5,1 % rechaza al menos a un rastreador de IA en el robots.txt. El 94,0 % de los comunicados está abierto a todos los rastreadores de IA probados, y el 99,0 % a los que alimentan las respuestas de búsqueda con IA.

¿Bloquear a GPTBot deja un comunicado fuera de la búsqueda de ChatGPT?

No. OpenAI documenta rastreadores distintos (OpenAI, s. f.): GPTBot recopila datos de entrenamiento, OAI-SearchBot hace aparecer sitios en la búsqueda de ChatGPT, y cada uno sigue sus propias reglas del robots.txt. 28 sitios de comunicados rechazan al primero y admiten al segundo; ninguno hace lo contrario. Del mismo modo, Google-Extended rige el entrenamiento y la fundamentación de Gemini, no la Búsqueda de Google (Google, s. f.).

Si el robots.txt permite los rastreadores de IA, ¿pueden leer el comunicado?

No necesariamente. Un cortafuegos puede rechazar a los clientes automatizados diga lo que diga el robots.txt: 75 de los 1503 sitios rechazaron de entrada nuestro cliente, y 30 sitios de empresas, la mayoría de relación con inversores, no dieron ninguna respuesta a un cliente identificado. Pruebe la sala de prensa con un cliente que no sea un navegador, no solo el archivo robots.txt.

¿Qué copia de un comunicado debe revisar un comunicador?

Cada sitio donde vive el comunicado —la sala de prensa, el distribuidor, cualquier sindicación—, porque cada uno tiene su propio archivo y su propio cortafuegos. Los cinco sitios más grandes de nuestra población que respondieron (el 42,1 % de esos comunicados) no rechazan a ningún rastreador de IA; la sala de prensa de la empresa es la copia que ella controla.

10 Notas de método

Estadística. H1 y H2 usan pruebas exactas de McNemar (McNemar, 1947) sobre el resultado pareado de cada sitio; H3 usa la prueba exacta de Fisher, con un intervalo híbrido de Newcombe (Newcombe, 1998) para la diferencia de proporciones. La corrección de Holm (Holm, 1979) cubre las tres pruebas. Los intervalos sobre proporciones de sitios son intervalos de Wilson del 95 % (Wilson, 1927). La población es un censo de los sitios del archivo, por lo que las proporciones de comunicados se dan sin intervalo.

Robustez, publicada íntegramente en los datos. Contar los archivos inaccesibles como rechazos eleva la proporción de sitios que rechazan a algún rastreador de IA al 8,8 % (GPTBot 7,2 %). Excluir los cinco sitios más grandes eleva la proporción ponderada por comunicados al 10,3 %. Usar solo la respuesta al cliente identificado da un 3,1 % para GPTBot. Contar un sitio como rechazo si uno solo de sus comunicados está cerrado da un 3,7 % para GPTBot y un 1,2 % para Googlebot. Cada archivo se leyó una segunda vez al menos 6 horas después de la primera: de los 1438 sitios que respondieron las dos veces, ninguno cambió de veredicto para GPTBot (5 no respondieron la segunda vez). Ninguna conclusión cambia.

Desviaciones respecto al registro

1. Parámetros de seguimiento. El registro aplicaba cada archivo a la dirección completa del comunicado, con sus parámetros. Algunas fuentes añaden etiquetas de seguimiento de campaña (utm_ y similares) a cada enlace, y una regla general contra los parámetros se lee entonces como el rechazo de comunicados que en realidad están abiertos: toda la producción de un distribuidor cambiaba así. Esas etiquetas (5660 direcciones de comunicados) se eliminan antes de aplicar las reglas. Si se conservan, como se registró, GPTBot es rechazado por el 3,5 % de los sitios y algún rastreador de IA por el 5,2 %; cambian 2 sitios, ninguna conclusión.

2. Copias archivadas. El registro pedía la captura con estado 200 más cercana al 7 de octubre de 2025. El índice de búsqueda de la Wayback Machine tardaba unos 40 segundos por consulta, así que cada captura se localizó con su redirección a la captura más cercana y se conservó solo si se había archivado con estado 200. Un sitio cuya captura más cercana era un error cuenta como no cubierto.
3. Las señales a nivel de página solo se leen en las páginas que se abrieron. Una página de rechazo lleva su propia etiqueta noindex, que no dice nada del comunicado que hay detrás.
4. Añadido después de ver los datos, y señalado como tal: la separación del cambio interanual entre bloqueos generales y bloqueos dirigidos a la IA; la misma comparación sin los sitios cuyo archivo actual se niega a cualquier cliente automatizado (48 entonces, 54 ahora); y la verificación de los sitios inaccesibles con un segundo cliente.

Limitaciones

- El robots.txt es una petición. Este estudio informa de lo que pide cada archivo, no de si cada rastreador lo cumple; OpenAI y Perplexity indican que sus rastreadores activados por el usuario pueden no seguir el robots.txt (OpenAI, s. f.; Perplexity, s. f.).
- Nuestro cliente no es un rastreador de IA. Los cortafuegos que admiten rastreadores verificados por su dirección son invisibles desde fuera, así que las cifras sobre cortafuegos miden la exposición a clientes automatizados en general, no a la IA.
- Los archivos son los servidos un día concreto. Un archivo puede cambiar a la mañana siguiente.
- La población es el archivo de PPN World —unos 1.500 sitios, con predominio del inglés y de organismos públicos—, no todos los editores de comunicados del mundo.
- Un comunicado puede estar en varios sitios. Probamos la copia a la que enlaza el archivo; la sala de prensa del emisor o una copia sindicada pueden responder de otra forma.
- Algunos comunicados de organismos públicos enlazan a plataformas de terceros (Google Noticias, X, Instagram, Mailchimp), cuyos archivos son de la plataforma, no del organismo.
- Las copias archivadas cubren cuatro de cada cinco sitios y las capturó el propio rastreador de Internet Archive, al que quizá no se le sirvió lo mismo que a otros.

Referencias

- Anthropic. (s. f.). Does Anthropic crawl data from the web, and how can site owners block the crawler? Claude Help Center. Recuperado el 8 de octubre de 2026, de <https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler>
- Apple. (s. f.). About Applebot. Apple Support. Recuperado el 8 de octubre de 2026, de <https://support.apple.com/en-us/119829>
- Cloudflare. (2025, September 24). Content Signals Policy. <https://contentsignals.org/>
- Common Crawl. (s. f.). CCBot. Recuperado el 8 de octubre de 2026, de <https://commoncrawl.org/ccbot>
- Fletcher, R. (2024, February 22). How many news websites block AI crawlers? Reuters Institute for the Study of Journalism. <https://doi.org/10.60625/risj-xm9g-ws87>

- Google. (s. f.). Google's common crawlers. Google for Developers. Recuperado el 8 de octubre de 2026, de <https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- Howard, J. (2024, September 3). The /llms.txt file. llms-txt. <https://llmstxt.org/>
- Internet Archive. (s. f.). Wayback Machine. Recuperado el 8 de octubre de 2026, de <https://web.archive.org/>
- Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022, September). Robots Exclusion Protocol (RFC 9309). Internet Engineering Task Force. <https://doi.org/10.17487/RFC9309>
- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A. M., Dinh, A., Chitongo, C., Yin, D., . . . Pentland, S. (2024). Consent in crisis: The rapid decline of the AI data commons. *Advances in Neural Information Processing Systems*, 37, 108042–108087. <https://doi.org/10.52202/079017-3431>
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>
- Meta. (s. f.). Meta web crawlers. Meta for Developers. Recuperado el 8 de octubre de 2026, de <https://developers.facebook.com/documentation/sharing/webmasters/web-crawlers>
- Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890. [https://doi.org/10.1002/\(SICI\)1097-0258\(19980430\)17:8%3C873::AID-SIM779%3E3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C873::AID-SIM779%3E3.0.CO;2-I)
- OpenAI. (s. f.). Overview of OpenAI crawlers. OpenAI Developers. Recuperado el 8 de octubre de 2026, de <https://developers.openai.com/api/docs/bots>
- Perplexity. (s. f.). Perplexity crawlers. Perplexity Docs. Recuperado el 8 de octubre de 2026, de <https://docs.perplexity.ai/docs/resources/perplexity-crawlers>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>

11 Datos, prerregistro y cita

Cada resultado por sitio detrás de las figuras se publica en CSV con licencia CC BY 4.0, con la fecha en que se leyó cada archivo. Los organismos públicos figuran por su dirección; las empresas y los distribuidores, por tipo y rango, porque PPN World no nombra a fuentes comerciales con las que no tiene acuerdo. El prerregistro se publica sin cambios.

↓ Descargar los datos (CSV)

Una fila por sitio: tipo, país, comunicados, estado del robots.txt, rastreadores rechazados ahora y doce meses antes.

Leer el prerregistro

Registrado el 7 de octubre de 2026, antes de leer ningún archivo robots.txt, y servido tal como se guardó: por eso llama a este

estudio por su número de trabajo, n.º 03, y menciona un estudio piloto no publicado.

Citar este estudio

Libre de citar, graficar y republicar con atribución. Citas sugeridas:

APA

Bolduc, É. (2026). El 94 % de los comunicados de prensa está abierto a los rastreadores de IA: un censo prerregistrado de las reglas para rastreadores de IA en 1503 sitios de comunicados (PPN World Research Paper No. 1, Edition 1). PPN World.
<https://www.ppnworld.com/es/research/press-releases-ai-crawlers>

MLA

Bolduc, Émile. El 94 % de los comunicados de prensa está abierto a los rastreadores de IA: un censo prerregistrado de las reglas para rastreadores de IA en 1503 sitios de comunicados. PPN World, 8 Oct. 2026, www.ppnworld.com/es/research/press-releases-ai-crawlers.

CHICAGO

Bolduc, Émile. 2026. El 94 % de los comunicados de prensa está abierto a los rastreadores de IA: un censo prerregistrado de las reglas para rastreadores de IA en 1503 sitios de comunicados. PPN World Research Paper 1. PPN World.
<https://www.ppnworld.com/es/research/press-releases-ai-crawlers>.

Licencia

Textos, figuras y datos se publican con licencia [CC BY 4.0](#)

Correcciones

Ninguna por ahora. Cualquier corrección se publicará aquí con su fecha; una cifra citada nunca se cambia en silencio.

SOBRE EL AUTOR



Émile Bolduc

Centro de investigación de PPN World

Émile Bolduc escribió este estudio para PPN World, que sigue los comunicados de empresas, distribuidores y organismos públicos de todo el mundo. Para una pregunta de método o de datos, o una extracción que no hemos publicado, escriba al Centro de investigación.

[Perfil de LinkedIn](#)

[Escribir al Centro de investigación](#)