Rillor

EXPLAINER

Robots.txt and model training crawlers: what an opt-out covers

What a robots.txt opt-out from model training reaches: which crawler, which use, which time and which copy, from the protocol, crawler documentation, compliance studies and the law.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A robots.txt opt-out reaches the crawlers it names, from the time they next read the file, on the host that serves it, and the protocol itself states no rule about how content is used.

- RFC 9309, the IETF standard of September 2022, describes rules "that crawlers are requested to honor when accessing URIs" and states that they "are not a form of access authorization" [1]. Each crawler finds its rules by its own name, called a product token, and the standard has no field for what fetched content may be used for [1, 6].
- Three operators document a separate product token for each purpose: crawling for training, crawling for search and fetching a page because a user asked [20, 21, 29]. The pages of four operators, read on 8 November 2026, say their fetchers acting on a user's request may not follow robots.txt rules, or generally ignore them [20, 27, 28, 29].
- Two training tokens crawl nothing. One search engine operator's documentation says its token "is used in a control capacity" and "does not impact a site's inclusion in Google Search", and a device maker's says its own "does not crawl webpages" [22, 24].
- An opt-out looks forward. One model developer's page says a block "signals that the site's future materials should be excluded" from its training datasets, and an IETF draft on attaching usage preferences says robots.txt updates "do not retroactively apply to resources" [6, 21].
- A robots.txt file covers one host. The European Union Intellectual Property Office (EUIPO) wrote in May 2025 that a reservation through robots.txt on one website "does not guarantee a valid reservation of rights for the licensed copies of the same content hosted on other websites" [16].
- Measured compliance depends on the bot and the rule. Over 40 days on 36 sites at one institution, a 2025 study found bots "less likely to comply with stricter robots.txt directives", and fewer than 40% of assistant and model-search bots re-checked the file within a 168-hour window [33].
- Vocabularies that state a use, such as training, are drafts or industry specifications. On 8 November 2026 the two Internet-Drafts of the IETF AI Preferences (aipref) working group, one defining the vocabulary and one attaching it to robots.txt and HTTP headers, showed the datatracker state "I-D Exists", after a charter milestone of August 2026 for sending them to the Internet Engineering Steering Group (IESG), and the vocabulary draft says its contents "DO NOT REFLECT CONSENSUS" of the working group [4, 5, 6].
- EU law lets rightholders reserve text and data mining of online content by "machine-readable means", and the EU AI Act asks providers of "general-purpose AI models" to identify and comply with those reservations [11, 12]. Germany's Federal Court of Justice has set 17 December 2026 for its judgment in a case about building an image-text dataset for training [19].

Background

Robots.txt began as a convention. A document on robotstxt.org records "a consensus on 30 June 1994 on the robots mailing list" and states that "It is not enforced by anybody" [2]. The IETF published the protocol as RFC 9309, on the Standards Track, in September 2022 [1]. One search engine operator's guide says the file is "used mainly to avoid overloading your site with requests" [3].

Model training brought new names into the file. The Reuters Institute dates one developer's release of its web crawlers, with instructions for blocking them, to 7 August 2023, and a second developer's opt-out to 28 September 2023 [34]. That opt-out is a training token that crawls nothing (Finding 3) [22, 34]. EU law had provided for machine-readable reservations since Directive (EU) 2019/790, published on 17 May 2019, and the EU AI Act, published on 12 July 2024, tied them to providers of "general-purpose AI models" [11, 12]. Robots.txt became one candidate means among several [13, 15].

This report asks what a robots.txt opt-out from model training reaches, through four edges: which crawler, which use, which time and which copy. It is written for data and machine learning teams that collect or license training data, and for the rights and legal staff who review that work. It is research, never legal advice.

Method

What was examined. RFC 9309, the crawler documentation of eight operators, the vocabularies that state a use, five measurements of compliance and adoption, and the legal texts, reports and rulings that refer to machine-readable reservations of rights.

Sources. Thirty-five sources, each opened on 8 November 2026 and read for the passages cited: standards, drafts, specifications and one search operator's guide [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], EU texts [11, 12, 13, 14, 15], three government and agency reports [16, 17, 18], a German court notice [19], operator documentation [20, 21, 22, 23, 24, 25, 26, 27, 28, 29] and measurements and a vendor policy [30, 31, 32, 33, 34, 35].

How. Operator pages are quoted as they read on 8 November 2026, with each page's modified date where it gave one. Operators are described by kind, and product tokens appear only inside quotations. Study figures are given by category, with each study's scope and dates. Two pages refused plain downloads and were opened through a fetch tool, so quotes from them are short [7, 27]. The German notice is summarized in the author's translation. Laws, drafts and reports are summarized with their dates and status.

Findings

1. What RFC 9309 defines, and what it leaves out

RFC 9309 defines groups of rules and the names that select them. "Crawlers set their own name, which is called a product token, to find relevant groups" [1]. Matching is case-insensitive, several matching groups "MUST be combined into one group", a crawler with no matching group obeys the * group if there is one, and with neither "no rules apply" [1]. "The most specific match found MUST be used" [1]. The file sits "in the top-level path of the service" [1].

Failures and caching have their own rules. After a 400 to 499 answer, "the crawler MAY access any resources on the server" [1]. After a 500 to 599 error, "the crawler MUST assume complete disallow", and after about 30 days it may treat the file as unavailable or keep a cached copy [1]. "Crawlers SHOULD NOT use the cached version for more than 24 hours, unless the robots.txt file is unreachable" [1]. Crawlers "MAY interpret other records that are not part of the robots.txt protocol", such as sitemaps [1], and the usage lines in Finding 8 sit in the file beside these rules [6, 30].

What the RFC leaves out is use. Its rules concern access to URIs, with no field for what fetched content may be used for [1]. The attachment draft of the IETF AI Preferences (aipref) working group states the split: "Acquisition relies on Allow/Disallow rules", and "usage preference relies on Content-Usage rules" [6]. The RFC warns that it "is not a substitute for valid content security measures" [1], and one search operator's guide says a disallowed page "can still be indexed if linked to from other sites" [3].

2. How operators split training, search and user-triggered fetches

Table 1 summarizes the crawler documentation of eight operators as it read on 8 November 2026.

Table 1. Crawlers by purpose, from each operator's documentation

Operator, by kind	Training	Search	Fetches a user starts	Source
A model developer	Own crawler	Own crawler	"robots.txt rules may not apply"	[20]
A second model developer	Own crawler, for "future materials"	Own crawler	Own fetcher, which a site can disable	[21]
A search engine operator	Control token, no crawler	Own crawler	Not on the pages read	[22, 23]

Operator, by kind	Training	Search	Fetches a user starts	Source
A device maker	Control token that "does not crawl webpages"	Own crawler	Not on the page read	[24]
A social media company	Crawler for training "or improving products by indexing content directly"	Separate crawler, to "improve Meta AI search result quality"	"may bypass robots.txt rules"	[27]
An answer engine	None documented	Crawler "not used to crawl content for AI foundation models"	"generally ignores robots.txt rules"	[28]
An online retailer	Crawler whose data "may be used to train" models	Separate crawler	"may not follow all robots.txt directives"	[29]
A non-profit crawl archive	One crawler, for an archive that is "a sample of the web"	None on the pages read	None on the pages read	[25]

One model developer's page says "Each setting is independent of the others" and "If your site has allowed both bots, we may use the results from just one crawl for both use cases to avoid duplicative crawling" [20]. Finer rules differ. The archive says "We obey the Crawl-delay parameter for robots.txt" [25], while the device maker and the retailer say their crawlers do not follow it [24, 29].

Names can be faked. The archive says "we are aware of crawlers falsely identifying themselves as CCBot" [25], and the charter of an IETF working group says "Current solutions (such as IP allowlisting, User-Agent strings, and shared API keys) have significant limitations" [10].

Sites list names unevenly. In a 2024 audit's sample of domains from the C4 corpus, one developer's crawlers were restricted for 25.9% of the text, counted in tokens, and a search crawler for 1.0%, asymmetries that "tend to advantage less widely known AI developers" [31]. The audit also found that "4.5% of websites disallowed" user agents a developer does not recognize, "but not the documented agent" of that developer's crawler [31].

3. Control tokens and preview controls

One search engine operator's documentation says "Google-Extended doesn't have a separate HTTP request user agent string" and that "the robots.txt user-agent token is used in a control capacity" [22]. The token manages whether crawled content may be used for training future generations of the operator's model family and for grounding, and it "does not impact a site's inclusion in Google Search nor is it used as a ranking

signal in Google Search" [22]. The device maker's page says "Applebot-Extended is only used to determine how to use the data crawled by the Applebot user agent", and that pages disallowing it "can still be included in search results" [24].

Answers inside search follow page-level controls. One search operator points to "nosnippet, data-nosnippet, max-snippet, or noindex controls" to limit what its search results show [23]. The device maker names the nosnippet meta tag for its "broad world knowledge answers" [24]. The EU's General-Purpose AI Code of Practice encourages signatories that run a search engine to ensure that their compliance with a rights reservation "does not directly lead to adverse effects on the indexing of the content" [13].

Page-level tags can also carry a training preference. The retailer respects "noarchive (do not use the page for model training)" [29].

4. User-triggered fetchers

Four operators say robots.txt may not govern fetches a user starts. One model developer: "ChatGPT-User is not used for crawling the web in an automatic fashion. Because these actions are initiated by a user, robots.txt rules may not apply" [20]. The answer engine: "Since a user requested the fetch, this fetcher generally ignores robots.txt rules" [28]. The social media company's user fetcher "may bypass robots.txt rules", and another of its crawlers "might bypass robots.txt when performing security or integrity checks" [27]. The retailer's "may not follow all robots.txt directives" [29]. The second model developer says disabling its user fetcher "may reduce your site's visibility for user-directed web search" [21].

The vocabularies give this class its own category. The working group's vocabulary draft defines "AI Use" as "Using an asset as input to a generative AI model, where the asset is not directly provided by the user" [5]. The UK government reported in March 2026 that the IETF group "had not reached consensus on certain key issues such as how to treat RAG" [18].

In the 2025 study of 36 sites over 40 days, assistant and model-search bots "have the lowest re-check rates with less than 40% of bots checking robots.txt within a 168 hour window" [33]. Cloudflare measured user-action crawling on its network "up over 21x from January through early December" 2025 [35].

5. Timing: caching, re-checks and data already collected

A change reaches a crawler only when it fetches the file again. RFC 9309's cache limit is 24 hours [1]. One operator says its systems may take up to 24 hours to reflect changes [28], and another says its search results can take about 24 hours to adjust [20]. The retailer's crawlers "fetch host-level robots.txt files or use a cached copy from the last 30 days", and when a file cannot be fetched they behave as if it does not exist [29]. RFC 9309 asks a crawler to assume complete disallow after a 500 to 599 error [1].

The attachment draft says updates to robots.txt "do not retroactively apply to resources", and that changes "only apply after a crawler has retrieved the updated "robots.txt" and subsequently retrieved the affected resource again" [6]. One search operator says re-crawling a page "can take anywhere from several days to

several months" [23].

The operators' words look forward. One model developer says a block "signals that the site's future materials should be excluded from our AI model training datasets" [21]. The archive's FAQ says that after a disallow its crawler stops crawling the site and periodically checks the file for updates [25].

Material gathered earlier sits outside the file. The archive publishes "the full opt-out list for every legal request we have received", because "We have many downstream users who are not aware of the legal notices we have received" [26]. The page does not describe changes to past snapshots [26]. Under EU law, copies made under the general mining exception "may be retained for as long as is necessary for the purposes of text and data mining" [11]. Copyright owners told the U.S. Copyright Office that an opt-out would be "impossible to utilize after training occurs" [17]. For an image linked from a dataset of URLs, the EUIPO writes that a crawler "may skip the 'discovery phase' where robots.txt is typically checked" [16].

6. Location: copies, platforms and third-party datasets

A file covers the service whose top-level path serves it [1]. One model developer asks site owners to repeat the rules "for every subdomain that you wish to opt out from" [21], and the retailer says its crawlers follow the robots rules of each host of a domain [29].

Copies elsewhere carry their own files, or none. The EUIPO writes that reservations "may be copy/location-specific", and that a robots.txt reservation on one website "does not guarantee a valid reservation of rights for the licensed copies of the same content hosted on other websites" [16].

Many owners cannot edit the file. One search operator's guide says that on a hosted content management system "you might not need to (or be able to) edit your robots.txt file directly" [3]. Of 1,182 artist sites in a 2025 study, "over 78%" sat on one of eight hosting providers, most of which "do not allow for modification of robots.txt" [32]. For any opt-out Congress might consider, the U.S. Copyright Office wrote that it should not be "sufficient for AI companies to merely honor platform-level flags, like robots.txt, because in many cases copyright owners have no control over the platforms where their works appear" [17].

Embedded signals travel with copies. The attachment draft says embedding is "expected to be an effective means" because "metadata travels with the asset", though plain text has "no standardized and interoperable means of embedding metadata" [6]. The IPTC Data Mining property sits in image metadata, and TDMRep can sit in EPUB and PDF metadata [7, 8].

Third-party datasets sit outside a site's file. Under the Code, the crawler commitment "does not affect the application of Union law on copyright and related rights to protected content scraped or crawled from the internet by third parties and used by Signatories" [13]. The Commission's template asks providers for "the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained" [14].

7. Measured compliance and adoption

Table 2 lists the five measurements reviewed, each dated and scoped.

Table 2. Adoption and compliance measurements

Measurement	Scope and dates	Finding	Source
Reuters Institute	15 most widely used online news sources in ten countries, end of 2023	48% blocked one developer's crawlers, 24% the second developer's training token, which the Institute calls an "AI crawler"	[34]
Liu and others	40,455 sites (the "Stable Top 100k") in the Tranco top 100k every month and with a robots.txt file "in every snapshot of the Common Crawl data", October 2022 to October 2024	Sites that "fully disallow at least one AI crawler user agent": 12 to 14% of the 2,551 most popular (the "Stable Top 5k"), against 8 to 10% of the rest	[32]
Longpre and others	14,000 web domains, corpus samples of April 2024	Restriction rates, given any listed organization restricted, from 91.5% to 17.1%	[31]
Kim and others	36 sites at one institution, 40 days, 130 self-declared bots	Bots "less likely to comply with stricter robots.txt directives"	[33]
Cloudflare Radar	Cloudflare's top 10,000 domains, 1 January to 2 December 2025	"AI crawlers were the most frequently fully disallowed user agents"	[35]

The Reuters Institute's country range for the first developer ran "from 79% in the USA to just 20% in Mexico and Poland", and "During 2023, none of the websites we examined had reversed their decision after deciding to block" [34]. Of the sites blocking the second developer's training token, 97% also blocked the first developer's crawlers [34].

Liu and others found that "59% of artists have never heard about the term "robots.txt"" among 203 professional artists, and "only 17 sites having noai and 16 having noimageai tags" in the Tranco top 10,000 in October 2024 [32]. They found that reverse-proxy blockers "offer stronger protections" with "their own set of limitations" [32].

Kim and others report weighted compliance for crawl delay, endpoint access and disallow-all rules [33]. The category of bots gathering training data scored 0.560, 0.352 and 0.766, assistant bots 0.910, 0.131 and 1.000, model-search bots 0.895, 0.623 and 0.348, and headless browsers 0.036, 0.278 and 0.011 [33]. They also found "numerous instances of potential spoofing of well-known bots" [33].

Cloudflare found that two crawlers serving both search and model features "leaned heavily towards partial disallow" [35]. It measured crawling for model training "reaching as much as 7-8x search crawling and 32x user action crawling at peak" [35]. The EUIPO describes crawlers "masking their identities by not using known user-agent identifiers" [16], and respondents to the UK consultation described "stealth crawling" [18].

8. Usage-preference vocabularies and their status

Five vocabularies state what content may be used for, apart from whether it may be fetched (Table 3).

Table 3. Usage vocabularies on 8 November 2026

Vocabulary	Status	Carried in	Training term	Source
aipref drafts	IETF Internet-Drafts, datatracker state "I-D Exists"	A Content-Usage line in robots.txt or an HTTP header	`train-ai`	[4, 5, 6]
Content Signals Policy	Cloudflare policy of 24 September 2025, CC0	A Content-Signal line in robots.txt	`ai-train`	[30]
TDMRep	W3C community group final report of 10 May 2024, "not a W3C Standard"	`/.well-known/tdmrep.json`, HTTP headers, HTML, EPUB and PDF metadata	`tdm-reservation`	[7]
RSL 1.0	Recommendation of 10 December 2025	A License line in robots.txt, HTTP Link headers, HTML, RSS	`ai-train` in license terms	[9]
IPTC Data Mining	Photo Metadata Standard 2025.1 of 26 November 2025	Image metadata (XMP)	"Prohibited except for search engine indexing" among its values	[8]

The aipref vocabulary defines training as "Using an asset to modify the learned parameters of a generative AI model" [5]. Each category resolves to "allowed", "disallowed" or "unknown", unknown by default, and where statements differ "the most restrictive preference applies" [5]. "A recipient can only apply preferences it understands", and a recipient "has a choice whether to follow those preferences" [5]. The attachment draft's robots.txt form is:

```
User-Agent: *
Allow: /
Content-Usage: train-ai=n
```

Under it, "no preferences are implied for resources that are disallowed" [6]. The charter set August 2026 for sending both drafts to the IESG, and puts "Technical enforcement of preferences" out of scope [4].

Cloudflare's policy says "the absence of a signal conveys no meaning", that "content signals express preferences", and that "they are not technical countermeasures against scraping" [30]. Its robots.txt comment text asserts that restrictions "ARE EXPRESS RESERVATIONS OF RIGHTS UNDER ARTICLE 4" of Directive 2019/790, a claim the policy makes for itself [30]. In TDMRep, `tdm-reservation: 1` means "TDM rights are reserved" [7]. The IPTC property says "the absence of a prohibition does not indicate that the asset owner grants permission for data mining or any other use of an asset" [8].

Two operators name the IETF work. One model developer's page says that "As of October 30, 2026" it "has begun recognizing a machine-readable signal" on training, rolled out "over time and globally", and that "If your robots.txt file names ClaudeBot, we follow that rule, even if your general setting says something different" [21]. One search operator's page of 30 October 2026 says "Google intends to support the AIPREF shared machine-readable training preference signal (with "user-agent: __usage_ai-training") in early 2027" [22], a syntax that differs from the attachment draft's Content-Usage line [6]. The vocabulary draft says legal agreements that explicitly permit a use "likely apply despite the existence of machine-readable statements of preference" [5].

9. The legal frame (EU, Germany, UK, U.S.)

European Union. Article 4 of Directive (EU) 2019/790 gives an exception "for reproductions and extractions of lawfully accessible works", on condition that the use "has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online" [11]. Recital 18 names "metadata and terms and conditions of a website or a service" among those means [11]. Article 3's exception for scientific research by research organisations and cultural heritage institutions carries no reservation [11].

Article 53(1)(c) of the EU AI Act asks providers of "general-purpose AI models" to identify and comply with such reservations, "including through state-of-the-art technologies" [12]. Recital 106 applies this "regardless of the jurisdiction" of training, and the obligations apply from 2 August 2025, or from 2 August 2027 for models placed on the market before that date [12].

The General-Purpose AI Code of Practice of 10 July 2025 is "a voluntary tool" [13]. Under Measure 1.3, signatories commit "to employ web-crawlers that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt)", as RFC 9309 specifies, and to identify and comply with "other appropriate machine-readable protocols to express rights reservations" [13]. On 1 December 2025 the Commission opened a consultation to agree such protocols and said it would publish a list, reviewed "at least every two years" [15].

The EUIPO's May 2025 report calls the protocol "a voluntary mechanism which relies on the good faith of crawler deployers" [16]. In its analysis, where the protocol serves as a mining opt-out, the cause of action for ignoring it "may not be a breach of contract, but an infringement of copyright and related rights" [16].

Germany. The Federal Court of Justice heard case I ZR 281/25 on 3 September 2026 and has set 17 December 2026 for its judgment [19]. In the court's summary, a photographer sued a non-profit association that publishes a dataset of 5.85 billion image-text pairs, built from a U.S. dataset of image URLs [19]. The Higher Regional Court of Hamburg held on 10 December 2025 that the copying was covered by § 44b and § 60d of the German Copyright Act, and that the photographer had not shown that a reservation written in natural language was machine-readable in the second half of 2021 [19].

United Kingdom. The government's report of 18 March 2026 says "a broad copyright exception with opt-out is no longer the government's preferred way forward" [18]. It says "The REP is a voluntary standard but has been widely adopted by industry" [18].

United States. The Copyright Office released Part 3 of its report as a pre-publication version on 9 May 2025, and its status page expects no substantive changes in the final version [17]. One commenter group said robots.txt "has significant limitations because it is only effective to the extent it is recognized and respected" [17].

10. What a capture record can keep for each signal

The Commission's template asks for crawler names, "Purposes of the crawler(s)", a "General description of crawler behaviour" such as "respect of robot.txt and other protocols, while crawling", and the "Period of data collection" [14]. Table 4 lists fields a capture record can keep for those answers, from the sources reviewed.

Table 4. Signals a capture record can keep

Field	What it holds	Basis
Robots.txt as fetched	The file, its HTTP status and when it was fetched	[1, 6]
Matched group	The product token whose rules applied	[1]
Host	The host or subdomain read	[1, 21]
Usage lines	Content-Usage, Content-Signal or License lines	[6, 9, 30]
Headers and page tags	Content-Usage or TDMRep headers, and noindex, nosnippet or noarchive tags	[6, 7, 23, 29]
Embedded metadata	TDMRep in EPUB or PDF files, the IPTC Data Mining value	[7, 8]
Site terms	The site's terms and conditions, with their date	[11]

Field	What it holds	Basis
Origin	The crawl, or the third-party dataset and its stated opt-out practice	[14]

Because a preference change applies only after a resource is fetched again, the fetch time is part of each reading [6].

Implications

For teams collecting training data. Each signal is a reading of one host by one crawler at one time [1, 6]. A record that keeps the allow and disallow rules apart from any usage lines keeps acquisition apart from use [6].

For teams buying or licensing datasets. A collector's record can answer the template's questions on crawlers, purposes, behaviour and period [14]. Items taken from URL lists may not have passed a robots.txt check at download, and copies on other hosts carry their own signals [16]. In the vocabulary draft's words, written agreements that explicitly permit a use "likely apply" over machine-readable preferences [5].

For site owners. The documentation lists separate tokens by purpose at three operators [20, 21, 29], control tokens that crawl nothing [22, 24] and user fetchers that may not follow robots.txt [20, 27, 28, 29]. Page-level tags govern answers in search features [23, 24].

For rights and legal staff. The EU AI Act's obligations apply from 2 August 2025 [12], and the Code is "a voluntary tool" [13]. The aipref documents are Internet-Drafts, and the vocabulary draft does not reflect consensus [5, 6]. The U.S. report is a pre-publication version [17], and the German judgment is set for 17 December 2026 [19].

Rillor's data collection checks each source's terms before anything is captured and records where each record came from and when.

Limits of this analysis

- Operator pages change often. Four were modified between 26 October and 4 November 2026 [21, 22, 28, 29]. Quotes are as read on 8 November 2026.
- Eight operators were read, which is not a complete list of crawlers, and one audit found sites listing user agents a developer does not recognize [31].
- Two pages were opened through a fetch tool, and quotes from them are short [7, 27].
- Every adoption and compliance figure is dated and scoped (Table 2). None gives a current share of all websites.
- Cloudflare reports on its own network and its own policy [30, 35].
- No source reviewed says that an opt-out removes content from data already collected, from published datasets or from trained models.
- None of the operator pages read says whether it reads Content-Signal, TDMRep or RSL lines.
- No list of agreed opt-out protocols appeared on the Commission's consultation page on 8 November 2026 [15].
- The EU AI Act is cited from its Official Journal text of 12 July 2024 [12], without later amendments.
- The German case is summarized from the court's notice in the author's translation, and no outcome is predicted [19].
- This is research on what standards, operators, studies and laws say, never legal advice.

Sources

1. RFC 9309: Robots Exclusion Protocol. M. Koster, G. Illyes, H. Zeller and L. Sassman, IETF, Standards Track, September 2022, viewed 8 November 2026. https://www.rfc-editor.org/rfc/rfc9309.html

2. A Standard for Robot Exclusion. M. Koster, robotstxt.org, consensus of 30 June 1994, viewed 8 November 2026. https://www.robotstxt.org/orig.html

3. Introduction to robots.txt. Google Search Central, last updated 10 December 2025, viewed 8 November 2026. https://developers.google.com/search/docs/crawling-indexing/robots/intro

4. AI Preferences (aipref) working group charter, charter-ietf-aipref-01. IETF Datatracker, viewed 8 November 2026. https://datatracker.ietf.org/wg/aipref/about/

5. A Vocabulary For Expressing AI Usage Preferences, draft-ietf-aipref-vocab-08. P. Keller and M. Thomson, IETF Internet-Draft, 14 September 2026, viewed 8 November 2026. https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/

6. Associating AI Usage Preferences with Content in HTTP, draft-ietf-aipref-attach-05. G. Illyes and M. Thomson, IETF Internet-Draft, 19 August 2026, viewed 8 November 2026. https://datatracker.ietf.org/doc/draft-ietf-aipref-attach/

7. TDMRep: TDM Reservation Protocol, Final Community Group Report. W3C Text and Data Mining Reservation Protocol Community Group, 10 May 2024, viewed 8 November 2026. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

8. IPTC Photo Metadata Standard 2025.1, property 11.11 Data Mining. IPTC Photo Metadata Working Group, version 2025.1 rev 1, 26 November 2025, viewed 8 November 2026. https://iptc.org/std/photometadata/specification/IPTC-PhotoMetadata

9. Really Simple Licensing (RSL) 1.0 Specification, RSL-SPEC-1.0. RSL, Recommendation, 10 December 2025, viewed 8 November 2026. https://rslstandard.org/rsl

10. Web Bot Authentication (webbotauth) working group charter. IETF Datatracker, viewed 8 November 2026. https://datatracker.ietf.org/wg/webbotauth/about/

11. Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market. Official Journal of the European Union, L 130/92, 17 May 2019, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790

12. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 12 July 2024, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

13. General-Purpose AI Code of Practice, Copyright chapter, and the code's page. European Commission, Shaping Europe's digital future, published 10 July 2025, viewed 8 November 2026. https://ec.europa.eu/newsroom/dae/redirection/document/118115 and https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

14. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, 24 July 2025, viewed 8 November 2026. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models and https://ec.europa.eu/newsroom/dae/redirection/document/118480

15. Commission launches consultation on protocols for reserving rights from text and data mining under the AI Act and the GPAI Code of Practice. European Commission, 1 December 2025, last updated 16 December 2025, viewed 8 November 2026. https://digital-strategy.ec.europa.eu/en/consultations/commission-launches-consultation-protocols-reserving-rights-text-and-data-mining-under-ai-act-and

16. The Development of Generative Artificial Intelligence from a Copyright Perspective. European Union Intellectual Property Office, 12 May 2025, viewed 8 November 2026. https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025 and https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2025_GenAI_from_copyright_perspective/2025_GenAI_from_copyright_perspective_FullR_en.pdf

17. Copyright and Artificial Intelligence, Part 3: Generative AI Training, pre-publication version, and the Office's status page. U.S. Copyright Office, 9 May 2025, viewed 8 November 2026. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf and https://www.copyright.gov/ai/

18. Report on Copyright and Artificial Intelligence. Department for Science, Innovation and Technology, Department for Culture, Media and Sport and Intellectual Property Office, GOV.UK, 18 March 2026, viewed 8 November 2026. https://www.gov.uk/government/publications/report-and-impact-assessment-on-copyright-and-artificial-intelligence/report-on-copyright-and-artificial-intelligence

19. I ZR 281/25, Erstellen eines Datensatzes für KI-Training, hearing and judgment notice. Bundesgerichtshof, press office, hearing 3 September 2026, judgment set for 17 December 2026, viewed 8 November 2026. https://www.bundesgerichtshof.de/SharedDocs/Termine/DE/Termine/IZR281-25.html

20. Overview of OpenAI Crawlers. OpenAI, developer documentation, viewed 8 November 2026. https://developers.openai.com/api/docs/bots

21. Does Anthropic crawl data from the web, and how can site owners block the crawler? Anthropic, Claude Help Center, modified 4 November 2026, viewed 8 November 2026. https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler

22. Google's common crawlers. Google Crawling Infrastructure documentation, last updated 30 October 2026, viewed 8 November 2026. https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers

23. AI Features and Your Website. Google Search Central, last updated 10 December 2025, viewed 8 November 2026. https://developers.google.com/search/docs/appearance/ai-features

24. About Applebot. Apple Support, published 4 September 2026, viewed 8 November 2026. https://support.apple.com/en-us/119829

25. CCBot, and the FAQ. Common Crawl Foundation, viewed 8 November 2026. https://commoncrawl.org/ccbot and

https://commoncrawl.org/faq

26. Common Crawl Foundation Opt-Out Registry. Common Crawl Foundation blog, 17 September 2025, viewed 8 November 2026. https://commoncrawl.org/blog/common-crawl-foundation-opt-out-registry

27. Web crawlers. Meta for Developers, viewed 8 November 2026. https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/

28. Perplexity Crawlers. Perplexity documentation, modified 29 October 2026, viewed 8 November 2026. https://docs.perplexity.ai/docs/resources/perplexity-crawlers

29. Amazon Web Search crawlers. Amazon Developer, last updated 26 October 2026, viewed 8 November 2026. https://developer.amazon.com/docs/amazonbot/amazonbot.html

30. Giving users choice with Cloudflare's new Content Signals Policy. Will Allen, Cloudflare blog, 24 September 2025, viewed 8 November 2026. https://blog.cloudflare.com/content-signals-policy/

31. Consent in Crisis: The Rapid Decline of the AI Data Commons. S. Longpre, R. Mahari and 47 others, Data Provenance Initiative, arXiv 2407.14933, version 2, 24 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.14933

32. Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. E. Liu, E. Luo, S. Shan, G. M. Voelker, B. Y. Zhao and S. Savage, ACM Internet Measurement Conference 2025, arXiv 2411.15091, version 2, 7 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2411.15091

33. Scrapers selectively respect robots.txt directives: evidence from a large-scale empirical study. T. Kim, K. Bock, C. Luo, A. Liswood, C. Poroslay and E. Wenger, ACM Internet Measurement Conference 2025, arXiv 2505.21733, version 2, 23 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2505.21733

34. How many news websites block AI crawlers? Richard Fletcher, Reuters Institute for the Study of Journalism, University of Oxford, 22 February 2024, viewed 8 November 2026. https://reutersinstitute.politics.ox.ac.uk/how-many-news-websites-block-ai-crawlers

35. Cloudflare Radar 2025 Year in Review. Cloudflare blog, 15 December 2025, viewed 8 November 2026. https://blog.cloudflare.com/radar-2025-year-in-review/

About this report

Published by Rillor on 11 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Robots.txt and model training crawlers: what an opt-out covers. 11 November 2026. https://rillor.com/insights/robots-txt-and-model-training-crawlers

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data