

members/api/comments/counts/ Disallow: /r/ Disallow:
webmentions/receive/ # Allow rules User-agent: Google
Disallow rules # robots.txt generated by Robots.txt Help
Ai2Bot Disallow: / User-agent: Ai2Bot-Dolma Disallow: / U
iHitBot agent: Amazonbot Disallow: / Us
andibot agent: anthropic-ai Disallow: / U
applebo -agent: Applebot-Extended Disa
User-agent: bedrockbot Disallow: / User-agent: Brightbo
User-agent: Bytespider Disallow: / User-agent: CCBot D
User-agent: ChatGPT-User Disallow: / User-agent: Claud
Disallow: / User-agent: Claude-User Disallow: / User-age
Disallow: / User-agent: **THE PROTOCOL GAP:** ClaudeBot
User-agent: cohere-ai Disallow: / User-agent GoogleOth
Disallow: / User-agent:cohere-training-data-crawler Dis
User-agent: Cotoyogi Disallow: / User-agent: Crawlspac
User-agent: **BRAZIL** Disallow: / User-agent: DuckAssistB
User-agent: EchoboxBot Disallow: / User-agent: Facebo
Disallow: / User-agent: Factset_spyderbot Disallow: / Use
irecrawlAgent Disallow: / User-agent: FriendlyCrawler I
User-agent: Google-CloudVertexBot Disallow: / User-age
Google-Extended Disallow: / User-agent: GoogleOther D
User-agent: GoogleOther- **March 2026** : / User-agent:
GoogleOther-Video Disallow: / User-agent: GPTBot Disal
User-agent: iaskspider/2.0 Disallow: / User-agent: ICC-C
Disallow: / User-agent: ImagesiftBot Disallow: / User-age
ng2dataset Disallow: / User-agent: ISSCyberRiskCrawl
User-agent: Kangaroo Bot Disallow: / User-agent: meta-e

This report is a joint publication of Journalism Relay Project, Momentum - Journalism and Tech Task Force, and the International Fund for Public Interest Media (IFPIM).

Sérgio Spagnuolo (Journalism Relay Project) developed the methodology and led on data analysis. Ester Borges, Bruno Fiaschetti, and Paula Miraglia (Momentum) provided contextual analysis for Brazil. Jessica White and Irene Jay Liu (IFPIM) provided overall editorial and research support.



Contents

Key Findings	4
Introduction	4
About this project	5
Robots.txt: An Underutilized Tool in Brazil	6
Robots.txt Among Brazilian Publishers: What the Data Tells Us	6
Blocked agents	7
Implications for Publishers in Brazil's Evolving AI Landscape	8
Conclusion and Next Steps	10
Methodology	11

The Protocol Gap: Brazil

Key Findings

- An overwhelming majority of news websites in Brazil have either no AI-scraping policy at all or a very permissive approach. As of November 2025, only 7.2% of Brazilian news websites disallow at least one AI crawler through the use of a robots.txt file, even though most news sites (75%) have a robots.txt file in their website.
- Websites in Brazil that implement robots.txt directives focus their efforts on what appear to be the most well-known AI crawlers from OpenAI, Common Crawl, Google, ByteDance, Amazon, Apple, Meta and Huawei.
- This data indicates that many websites do not use robots.txt as a mechanism to signal their preferences regarding AI companies scraping their content.
- Keeping an updated robots.txt file is not a foolproof solution, but it is a readily available tool for newsrooms to state the preferences regarding the use of their content, in line with the organization's content strategy and business model. Whatever their stance, organizations need to be able to have a clear public direction as to how their content is used to train new AI models.

Introduction

In recent years, news sites have been receiving an influx of traffic from a new kind of visitor: AI crawlers. From Googlebot to OpenAI's "GPTbot," Anthropic's "Claudebot" and Meta's "ExternalAgent," these crawlers all have one thing in common: a persistent need for data to train Large Language Models (LLMs), develop AI assistants, and feed AI-powered search engines. These new visitors are noteworthy especially when we consider that around 30% of the global web traffic today comes from bots. News sites are attractive destinations as they produce timely, reliable information that can help to improve the quality of their AI models and outputs.

This surging demand for AI data presents new opportunities—but many challenges—for media organizations. It raises important legal and ethical issues around unauthorized use, copyright infringement, and privacy violations. But the very survival of journalism hangs in the balance, too: without mechanisms in place to signal and enforce permissions on AI crawlers, media organizations are unable to effectively protect and monetize the value of their work as it gets scraped, repackaged, and redistributed by AI systems. This gap risks accelerating the sustainability crisis facing media outlets that are already grappling with declining visibility and traffic from digital platforms—what many publishers have already called a "traffic apocalypse"—and the collapse of traditional business models based on advertising.

News sites are becoming increasingly aware of this growing challenge, but there are currently few ways to manage AI crawlers and none of them are completely

foolproof. Some sites are implementing hard paywalls, while others are launching lengthy lawsuits claiming copyright infringement. Technical responses are also emerging: in July 2025, the web security and infrastructure provider Cloudflare announced that its protections would block AI crawlers by default and help to enforce a permission-based model. Content marketplaces—like the one piloted by Microsoft or startups such as ProRata.ai and Tollbit—and new data solutions developed by the likes of Flexolmo and OpenMined are also emerging as ways to give content producers more control over how their content gets accessed, used, and compensated by AI companies. But many more still resort to managing content scraping by AI crawlers by implementing the Robots Exclusion Protocol (known as robots.txt).



Robots.txt is a text file added to the root domain of a website that contains instructions to search engines and web crawlers about which pages or files they are allowed to access. Despite its limitations, robots.txt is one of the few free tools available to content creators to signal whether they want their data to be used by AI companies. Robots.txt has already been used as evidence in legal complaints against unauthorized crawling and scraping in Canada, the United States, and the United Kingdom. The file can also serve as leverage in licensing negotiations, especially in contexts where AI models rely heavily on large volumes of unconsented, uncompensated data.

```
# robots.txt for http://www.wikipedia.org/ and friends
#
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#
# Observed spanning large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /
# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /
# Wikipedia work bots:
User-agent: IsraBot
Disallow: /
User-agent: Orthogaffe
Disallow: /
# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /
User-agent: DGC
Disallow: /
User-agent: Zao
Disallow: /
# Some bots are known to be trouble, particularly those designed to copy
# entire sites. Please obey robots.txt.
User-agent: sitecheck.internetseer.com
Disallow: /
User-agent: Zealbot
Disallow: /
User-agent: MSIECrawler
Disallow: /
User-agent: SiteSnagger
Disallow: /
User-agent: WebStripper
Disallow: /
User-agent: WebCopier
Disallow: /
User-agent: Fetch
Disallow: /
```

Screenshot of robots.txt file used by Wikipedia.org

Despite growing awareness about the unauthorized scraping and crawling for AI, media responses to protect their content online remain patchy at best. Small and medium-size media outlets remain at a higher disadvantage because they do not necessarily have the resources to invest in sophisticated infrastructure to detect and block AI crawlers nor do they have the leverage to pursue individual licensing deals with large AI companies—Reddit, for example, agreed to license its content to Google for AI training in February 2024.

While Robots.txt remains the most widely available mechanism, this research begins to uncover critical gaps in newsroom use of robots.txt to manage AI crawlers, with a lens on Global South markets. In this first report mapping over 4,000 news websites in Brazil, we find that the vast majority are not giving any directives regarding AI crawlers, even though many media organizations already have a robots.txt file attached to their site.¹

As AI companies continue to fight for market dominance and harvest vast troves of data to feed their systems, the need for more informed and coordinated responses is especially crucial for media organizations in the Global South. Coordinated robots.txt adoption across news organizations can help to establish shared participation in standards that benefits the entire journalism community—from large outlets to independent publishers—by setting clear directives towards AI companies. These simple text documents continue to be

a relevant tool to empower media organizations in the age of AI, transforming what might otherwise be passive data extraction into an active editorial decision.

About this project

This report is the result of a collaboration between Journalism Relay Project, Momentum - Journalism and Tech Task Force, and the International Fund for Public Interest Media. In this first country report focusing on Brazil, we sought to map how many news sites are using robots.txt to manage AI crawlers. It is part of a broader research project that involves partners in Brazil, Indonesia, and South Africa and aims to strengthen our global understanding of evolving norms and practices around AI access to journalistic content, with a particular focus on key markets in the Global South. Through technical and qualitative research, we aim to increase awareness among media organizations on how AI systems access their content and help to inform strategies and policies to manage AI crawlers in ways that best serve newsroom interests. We have developed a detailed methodology (see annex to this report) for all those who may want to replicate this research on robots.txt in other countries and regions.

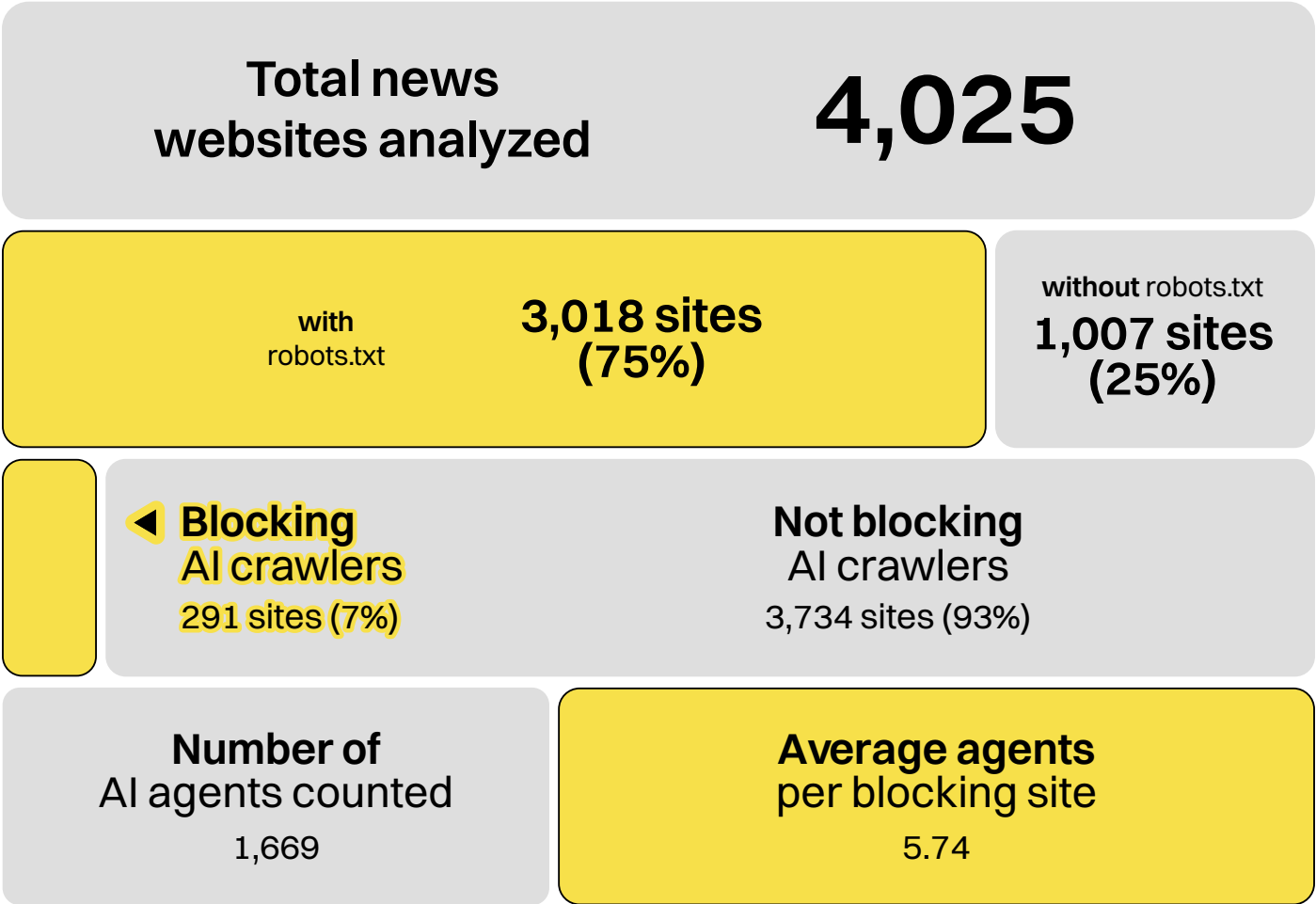
¹Website data from Brazil was based on the most recent national research from Atlas da Notícia (atlas.jor.br), from June 2025.

Robots.txt: An Underutilized Tool in Brazil

Our analysis found that an overwhelming majority of news websites in Brazil have either no AI-scraping policy at all or a very permissive approach, with several websites displaying only the base robots.txt files from their content management systems, if any.

As of November 2025, only about 7.2% of Brazilian news websites had at least one AI directive, although most (75%) were identified as having a robots.txt file attached to their website.

Robots.txt Among Brazilian Publishers: What the Data Tells Us



Source: The Protocol Gap

This is a strong indication that the vast majority of websites are unaware of the capabilities of setting their public statements regarding the permissiveness of AI scraping their content, instead relying on the default robots.txt file that comes in their content management systems.

Websites in Brazil that implement AI blocking take a much more targeted approach, focusing their efforts on what appear to be the most prominent AI crawlers rather than casting a wide net. Brazilian sites block an average of just about six AI agents per blocking site.

This selective blocking strategy could reflect several factors: these countries may have less awareness of the full spectrum of AI crawlers operating online, their websites might be primarily concerned with blocking only the most well-known AI systems, or there could be cultural and regulatory differences that make comprehensive AI blocking less of a priority compared to the more cautious approach adopted by media organizations in countries like the United States. Further study is required to understand prevailing factors.

Blocked agents

Among the websites that do block at least one AI agent, the top 10 most blocked bots are:

AI agent	Description	% of blocks*
gptbot	From OpenAI. Data is used to train current and future models, removing paywalled data, PII and data that violates the company's policies.	10.2
ccbot	From <u>Common Crawl Foundation</u> . Web archive going back to 2008. <u>Cited in thousands of research papers per year</u> .	9.77
google-extended	From Google. Used to train Gemini and Vertex AI generative APIs. Does not impact a site's inclusion or ranking in Google Search.	9.53
claudefbot	From Anthropic. Scrapes data to train LLMs and AI products offered by Anthropic.	9.17
bytespider	From ByteDance. Downloads data to train LLMs, including ChatGPT competitors.	8.81
amazonbot	From Amazon, Includes references to crawled website when surfacing answers via Alexa; does not clearly outline other uses.	8.51
applebot-extended	From Apple. Apple has a secondary user agent, Applebot-Extended ... [that is] used to train Apple's foundation models powering generative AI features across Apple products, including Apple Intelligence, Services, and Developer Tools.	8.39
meta-externalagent	From Meta. Meta-ExternalAgent crawls the web for use cases such as training AI models or improving products by indexing content directly.	8.21
petalbot	From Huawei. Operated by Huawei to provide search and AI assistant services.	3.82
universal_block	Uses the directive to block all external agents. User-agent: * Disallow: /	1.92

For a list of authorship and descriptions of bots, see: <https://github.com/ai-robots-txt/ai-robots.txt/blob/main/table-of-bot-metrics.md>

* Percentage calculated from all observed blocks (not percentage of news websites blocking).

Implications for Publishers in Brazil's Evolving AI Landscape

The limited use of robots.txt to disallow AI crawlers highlights a gap among Brazilian news publishers as they contend with the impact of emerging technologies. Newsrooms in Brazil are increasingly adopting AI tools for a range of purposes—from translating articles to producing videos. But the rise of AI is also exacerbating sustainability challenges. In a media market that has long been concentrated among a few large conglomerates, the lack of proactive and coordinated responses risks exacerbating challenges linked to the unequal relationships with digital platforms and difficulties in negotiating collectively with Big Tech companies.

Many Brazilian publishers have expressed concern over unauthorized use of their content, ethical risks, and falling website traffic—and express a desire for fair regulation and agreements—but collective strategies and concrete agreements have yet to emerge. To date, no licensing agreements between media organizations and technology platforms have been made public in Brazil. In August 2025, Folha de São Paulo, one of Brazil's largest newspapers, filed a lawsuit accusing OpenAI of unfair competition and copyright infringement, claiming the company used its content without permission or payment to train its AI models. AI Overviews is also exacerbating financial pressures, with evidence showing its impact on traffic to new sites and potential sources of revenue.

Failing to protect content may represent a missed opportunity to secure both resources and legitimacy for journalism. From a sustainability perspective, remuneration for the use of journalistic content could be a vital source of funding to at least partially offset the loss of audience and advertising revenue resulting from the digital shift in information consumption. Facing ongoing

financial strain, the media ecosystem could find in this new funding source a way to strengthen its democratic role.

From a legitimacy standpoint, the extraction of journalistic content without recognition of its social value may further erode journalism's position as a primary source of information. In Brazil, recent studies reveal that—particularly among younger populations—social media platforms have become the main tools for accessing information, including news and political content.

Compounding this issue is the fact that most of these companies and their technologies are based in the United States and China—meaning that, in both design and operation, these tools are not only ill-suited to meet the needs and specificities of the Global South, but often reinforce and deepen North-South power asymmetries.

Despite the many challenges, Brazil is in a pivotal moment to act. All three branches of government—Executive, Legislative, and Judiciary—are advancing measures that could reshape journalism's future.

Brazil's federal government, echoing broader aspirations of other countries in the Global South, has called for greater influence in global debates on digital sovereignty and AI regulation. Recent examples include the G20 Digital Economy Working Group's declaration on information integrity in 2024, as well as the more recent BRICS declaration delivered at the BRICS summit in Rio de Janeiro in July 2025, which placed AI high on the bloc's agenda. Executive engagement could accelerate regulatory progress and open space for civil society and the journalism sector to shape these discussions.

Where Robots.txt Falls Short, Newsrooms Pay

Because Google does not distinguish between its traditional search engine and its AI Overviews search feature, publishers cannot easily use robots.txt to prevent their content from being used for AI-generated summaries without disappearing from search altogether. This tradeoff carries significant economic consequences. Many publishers—including in Brazil—have reported steep declines in traffic to their websites since AI Overviews was rolled out, which is impacting their ability to earn revenue. This is linked to the fact that users are less likely to click through to news websites if they access their information through an AI-generated summary. In September 2025, Cloudflare announced a Content Signals Policy that would integrate into a robots.txt file and could give site owners stronger controls to block crawlers for AI training while still allowing them to be visible in traditional search. However, it remains to be seen whether this will stem the loss in website traffic.

Meanwhile, the legislative branch is advancing key bills aimed at regulating platforms and their relationship with journalism:

- Bill No. 2630 (2020) seeks to strike a balance between freedom of expression and the fight against disinformation, mandating greater algorithmic transparency and requiring platforms that benefit from journalistic content to compensate those who produce it.
- AI Bill - No. 2338 (2023) sets general guidelines for AI development and use in Brazil. It aims to ensure responsible AI practices that protect rights and democratic systems. With regards to journalistic activity specifically, the bill states that automated use of works—such as extraction, reproduction, storage, and transformation, including data and text mining in AI systems—by research institutions, media organizations, museums, archives, and libraries does not constitute copyright infringement. This exemption is contingent upon the tools being used only as necessary, without harming the economic interests of rights holders, not competing with the normal exploitation of the works, and not being used for commercial purposes without fair compensation. The discussion about the obligation to remunerate for the use of content protected by copyright is, however, the most sensitive point of the bill. Big Tech companies are putting pressure on deputies to avoid payment, making lawmakers cautious about advancing stricter regulations given the potential economic and geopolitical implications of such measures.

Brazil's Supreme Court has also shaped this debate through initiatives like the "Fake News Inquiry" and the recent ruling that redefined social media liability for user-generated content. While the Court's stance has raised concerns about the risk of over-censorship and its impact on freedom of expression, it has influenced the national debate on AI governance and regulation.

Amid these policy shifts, the news industry remains largely reactive to technological upheaval. This study could serve as a wake up call. With only 7.2% of Brazilian news outlets currently using robots.txt to block AI crawlers, the findings highlight the urgency of deepening discussions on content appropriation and developing viable avenues for recognition and remuneration. Strengthening industry coordination around these issues could set the first step toward a more sustainable digital future for Brazilian journalism.

Conclusion and Next Steps

This study serves as a point of departure for deeper reflections on the complex relationship between journalism and digital platforms in the Global South—particularly when it comes to the unauthorized use of journalistic content by generative AI tools and the development of strategies to fairly compensate media outlets.

Despite its limitations—most notably, the technical constraints of robots.txt as a content protection tool and the reliance on exclusively public data—the findings signal a broader sense of unpreparedness, or even hopelessness, among Brazilian publishers confronted with the capture of their content by AI companies. Awareness of how few Brazilian outlets actively protect their content via robots.txt could also mark a turning point away from fragmentation and lack of engagement towards a collective discussion about AI regulation, and could inspire a new wave of strategic reflection within the media industry.

This study represents the first chapter of a partnership between Journalism Relay Project, Momentum – Journalism and Tech Task Force, and the International Fund for Public Interest Media (IFPIM), aimed at building a critical body of knowledge and contributing to discussions on pathways to journalism sustainability from the perspective of the Global South.

Next steps include:

1. **Further Research:** Based on the data and conclusions presented here, we plan to move forward with a qualitative analysis of the Brazilian landscape through interviews with local publishers. Our goal is to gather further insights that will help us more clearly and deeply understand the dilemmas faced by journalism in Brazil regarding content capture and the pursuit of fair regulatory frameworks for the use of their material.
2. **Publisher Tools:** The decision whether or not to block crawlers is an individual decision that must fit with a publisher's strategy.
3. **Cross-Border Partnership:** The partnership also aims to replicate the methodology used in this study to examine the media ecosystems of Indonesia and South Africa—two other Global South countries that, like Brazil, share similar characteristics: historically concentrated media markets, unequal relationships with digital platforms, and difficulties in negotiating collectively with Big Tech companies.

We encourage all researchers who may want to replicate this study in other markets and encourage them to consult the full methodology that follows.

Methodology

1.1. Data origins

This project used programmatic file verification to ascertain the existence of blocking parameters for AI agents in robots.txt files of 4,025 news organizations from Brazil. Website data from Brazil was based on the most recent national research from Atlas da Notícia (atlas.jor.br), from June 2025. The main variables considered were the name of the news organizations, their website addresses and their locations.

The methodology centered on examining these files for specific patterns and its relations to AI crawlers, based on the standard parameter **Disallow:/** in robots.txt files. By consulting an open source database of known AI crawler identifiers, including those from major technology companies like Google, Meta, Amazon, Apple, Anthropic and OpenAI, the analysis identified when websites specifically attempted to signal restrictions to AI crawlers towards their content.

This aspect of the methodology was particularly relevant given the recurring concerns about AI systems scraping web content without explicit permission. It is notable to mention that robots.txt files are non-binding instruments that cannot technically prevent data collection from any bots, but merely indicate a site's policy towards automated content aggregation.

When examining a website, the system first ensured proper URL formatting before attempting to access its robots.txt file. The system then implemented error handling and timeout protocols to manage various real-world scenarios such as server timeouts, missing files, connection issues or DNS blocking, creating a comprehensive logging system.

To account for slow-responding servers that might need time to “wake up,” the system implemented a retry mechanism with a 3-second delay when robots.txt files were not found on the first attempt. This approach significantly improved detection reliability while maintaining efficient processing times for responsive servers.

The parsing logic distinguished between restrictive directives (**Disallow:/** or **Disallow:***) and permissive ones (**Allow:** directives or empty **Disallow:** statements), ensuring accurate classification of AI blocking policies. Sites were categorized along a spectrum from permissive to super restrictive based on the number and type of AI agents they blocked.

Websites with malformed or incomplete robots.txt files were generally classified as permissive toward AI crawlers. This included sites with empty **Disallow:** directives (which technically allow access) and those using only partial blocking rules that restricted specific URLs or directories while leaving the majority of their content unrestricted. In the absence of comprehensive blocking directives, these sites were effectively treated as allowing AI access to their content.

We manually checked 90 random results and were able to check that most of the results were indeed accurate. Despite these robust methodological safeguards, out of extra caution we estimate that a 3% error rate persists due to technical factors including DNS resolution issues, server timeouts, and temporary network connectivity problems that can result in false positives or, in some cases, false negatives in robots.txt detection.

1.2 Data collection

Once the robots.txt content was successfully retrieved, the analysis moved to pattern recognition. The system processed the file's content line by line, identifying User-agent directives and examining them against the known AI crawler database (looking for the **/Disallow** parameter).

This pattern matching was implemented with case-insensitive comparison to catch variations in how crawler names might be written, making the analysis less prone to redundancies. The results of this analysis were structured to provide insights about each website's stance on AI crawlers.

For each analyzed website, the system determined whether a robots.txt file existed, whether it contained any blocking directives, and specifically which known AI crawlers were being blocked under the **/Disallow** parameter.

This structured approach to data collection and analysis made it possible to conduct fast (average of 90 minutes of code run) large-scale examinations of how different websites approached AI crawler access.

Collection and analysis were based in R programming language.

²Atlas da Notícia, site - <https://atlas.jor.br>

³List of known AI crawlers, 2025, Github repository - [ai-robots.txt](https://github.com/ai-robots-txt)

⁴OECD (2025), “Intellectual property issues in artificial intelligence trained on scraped data”, OECD Artificial Intelligence Papers, No. 33, OECD Publishing, Paris, <https://doi.org/10.1787/d5241a23-en>.

1.3 Data classification

As data was collected from the robots.txt files, the algorithm automatically made classifications based on the level of permissiveness discretely assigned to each site.

Rather than using arbitrary absolute thresholds, the system first calculates the baseline average number of AI agents blocked by sites that actually implement AI blocking policies. This average serves as the reference point for determining what constitutes normal, moderate, or excessive blocking behavior within the specific dataset being analyzed.

The classification system then creates five distinct tiers using multiples of this calculated average.

Classification	AI Agents Blocked	Description
Permissive AI agent policy	0 AI agents	Sites that impose no restrictions on AI crawlers
Somewhat restrictive	1x to 2x average	Sites with typical or slightly above-average blocking behavior
Restrictive	2x to 5x average	Sites blocking two to five times the average
Very restrictive	5x to 10x average	Sites blocking five to ten times the average
Super restrictive	10x+ average	Sites with exceptionally comprehensive blocking policies

This methodology ensures that the classification remains meaningful regardless of the absolute numbers in any given dataset, as it adapts to the actual distribution of blocking behaviors observed. For instance, if the average site blocks three AI agents, a site blocking thirty agents would be classified as super restrictive, but if the dataset average were ten agents, that same thirty-agent site would fall into the restrictive category.

1.4 Data aggregations

Finally, the data was aggregated in a set of indicators that could be representative of the whole dataset.

metric	value / unit	date_observation
total_observations	data / websites	data
block_ai	data / websites	data
not_block_ai	data / websites	data
pct_not_blocked	data / percent	data
pct_blocked	data / percent	data
yes_robots_txt	data / websites	data
no_robots_txt	data / websites	data
pct_robots_txt	data / percent	data
pct_no_robots_txt	data / percent	data
ai_agents_count	data / ai agents	data
avg_agents_per_blocking_site	data / ai agents	data

members/api/comments/counts/ Disallow: /r/ Disallow:
/webmentions/receive/ # Allow rules User-agent: Google
Disallow rules # robots.txt generated by Robots.txt Helper
AI2Bot Disallow: / User-agent: Ai2Bot-Dolma Disallow: / U
aiHitBot Disallow: / User-agent: Amazonbot Disallow: / Use
Andibot Disallow: / User-agent: anthropic-ai Disallow: / Us
Applebot Disallow: / Use -agent: Applebot-Extended Disal
User-agent: bedrockbot Disallow: / User-agent: Brightbot
User-agent: Bytespider Disallow: / User-agent: CCBot Disa
User-agent: ChatGPT-User Disallow: / User-agent: Claude
Disallow: / User-agent: Claude-User Disallow: / User-age
Disallow: / User-agent: THE PROTOCOL GAP: ClaudeBot D
User-agent: cohere-ai Disallow: / User-agent GoogleOthe
Disallow: / User-agent:cohere-training-data-crawler Disa
User-agent: Cotoyogi Disallow: / User-agent: Crawlspac
User-agent: BRAZIL Disallow: / User-agent: DuckAssistBo
User-agent: EchoboxBot Disallow: / User-agent: Facebook
/ User-agent: Factset_spyderbot Disallow: / User-agent: F
Disallow: / User-agent: FriendlyCrawler Disallow: / User-a
Google-CloudVertexBot Disallow: / User-agent: Google-Ex
Disallow: / User-agent: GoogleOther Disallow: / User-ager
GoogleOther- March 2026 : / User-agent: GoogleOther-Vi
/ User-agent: GPTBot Disallow: / User-agent: iaskspider/2
User-agent: ICC-Crawler Disallow: / User-agent: Imagesif
/ User-agent: img2dataset Disallow: / User-agent: ISSCyb
Disallow: / User-agent: Kangaroo Bot Disallow: / User-age
meta-externalagent Disallow: / User-agent: Meta-Externa