

THE PROTOCOL GAP:

INDONESIA

OCTOBER 2026

Asosiasi Media Siber Indonesia

MONASH University

pr2media
PEMANTAU REGULASI & REGULATOR MEDIA
MEDIA REGULATION AND REGULATOR WATCH

if International Fund for Public Interest Media

Journalism Relay Project

This report is a joint publication of the Journalism Relay Project, Asosiasi Media Siber Indonesia (Indonesian Cyber Media Association, AMSI), Monash University Indonesia, PR2Media, and the International Fund for Public Interest Media (IFPIM).

Sérgio Spagnuolo (Journalism Relay Project) developed the methodology and led the data analysis. Ika Karlina Idris (Monash University Indonesia) and PR2Media provided contextual analysis for Indonesia. 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 Indonesia, but Growing 5

- Robots.txt Among Indonesia Publishers: What the Data Tells Us 6
- Blocked agents 6

Implications for Publishers in Indonesia's Evolving AI Landscape 7

Conclusion and Next Steps 9

Methodology 10

The Protocol Gap: Indonesia

Key Findings

- Most news websites in Indonesia are not using robots.txt as a mechanism to signal their preferences regarding AI companies scraping their content. As of September 2026, only 10.1% of news websites in Indonesia disallow at least one AI crawler through the use of a robots.txt file, even though most (74.6%) have a robots.txt file on their website. The percentage of news websites restricting AI crawlers has nearly doubled since the first analysis in July 2025, suggesting awareness is slowly growing.

- Sites that do disallow AI crawlers target an average of 11.3 bots, indicating that publishers that take action tend to adopt broad restrictions rather than singling out individual crawlers. Frequently restricted bots include several well-known crawlers from Common Crawl Foundation, OpenAI, ByteDance, Anthropic, Meta, and Amazon.

- Implementation of proactive AI crawling policies varies across publishers in Indonesia, attributable to gaps in institutional resources, technical capacity, and incentive structures. Closing these gaps will require further research, collective tools, and legal reform to give publishers stronger grounds to decide how their content is used by AI platforms.

Introduction

In recent years, news sites have been receiving an influx of traffic from a new kind of visitor: AI crawlers. Operated by companies like Google, OpenAI, Anthropic, and Meta, these automated bots 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 visitors are noteworthy especially when we consider that global web traffic coming from bots of various kinds is now surpassing human traffic. News sites are particularly attractive targets because they produce timely, reliable content that can improve the quality of AI outputs.

This surging demand from AI bots raises a number of challenges for media organizations. Not only do they inflict additional strain on servers, but they also raise legal and ethical questions around unauthorized use and copyright. The very survival of journalism hangs in the balance, too: without mechanisms to signal and enforce permissions, publishers cannot effectively protect or earn from 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 and the collapse of traditional business models based on advertising.

News sites are becoming increasingly aware of this growing challenge. 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, and in July 2026 it introduced a default block on "multi-purpose" crawlers and more options for websites to manage AI crawlers by function (search, agent, training). 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, OpenMined and Mozilla Data Collective 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 website owners to signal whether they want their data to be used by AI companies. While not legally binding, 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. As such, robots.txt functions as a mechanism for managing crawler access but also as a strategic tool through which publishers can signal consent, refusal, or conditional use. However, its effectiveness depends on voluntary compliance by AI developers, and its role within broader governance frameworks remains contested.

Despite growing awareness about the unauthorized scraping and crawling for AI, media responses to protect their content online remain patchy at best. While robots.txt remains a 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.

Following the first two reports focused on Brazil and South Africa, this third report analyzes 1,565 news websites in Indonesia. We find that as of September 2026, only 10.1% of Indonesian news websites have configured their robots.txt file to disallow at least one AI crawler, despite 74.6% having a robots.txt file in place. The figure has doubled since our first analysis in July 2025, suggesting awareness is slowly growing.

These findings point to significant technological and institutional gaps. In an environment marked by platform dependency, financial precarity, and regulatory uncertainty, Indonesian outlets may lack the resources, technical capacity, or incentive structures to implement AI protection strategies proactively. Indonesia's copyright law does not currently provide publishers with clear legal grounds to seek compensation when AI companies profit from their content, but proposed reforms may provide an opportunity to shift this.

Traffic loss, server strain, and bargaining asymmetry are shared challenges across many markets, but for outlets that operate with small teams and limited resources, they can be especially hard to absorb. Many publishers in Indonesia are feeling direct pressures from AI disruption: according to a recent study by Monash University Indonesia, many publishers have seen a downward trend in traffic, page views, and click-through rates after AI Overviews was rolled out. As AI-powered search features increasingly serve users summaries rather than sending them to news sites, traffic losses are contributing to weakening advertising revenue, since many outlets still depend on page views to generate income.

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 benefit 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 the Journalism Relay Project, the International Fund for Public Interest Media, Asosiasi Media Siber Indonesia (Indonesian Cyber Media Association, AMSI), Monash University Indonesia, and PR2Media. In this country report focusing on Indonesia, 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.

Robots.txt: An Underutilized Tool in Indonesia, but Growing

Our analysis found that an overwhelming majority of news websites in Indonesia either do not have an AI-scraping policy or adopt a very permissive approach, with several websites displaying only the base robots.txt files from their content management systems, if any.

In September 2026, 10.1% of Indonesian news websites had at least one AI directive, although most (74.6%) were identified as having a robots.txt file attached to their website. This figure is an increase from 5.8% websites based on the analysis conducted in July 2025.

This increase contrasts with Brazil's rate of 7% of websites that block at least one AI agent in January 2026 (against 5% in July 2025), and South Africa's rate of 30.4% in April 2026.

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

Total news websites analyzed: **1,565***

News websites **with** robots.txt: **1,168 websites, 74.63%**

News website without robots.txt: **397 websites, 25.37%**

News website Blocking AI crawlers: 158 websites, 10.1%

News website Not blocking AI crawlers: 1,407 websites, 89.9%

Number of AI agents counted: **1,679**

Average agents per blocking site: 11.34

*In the analysis conducted on September 30, 2026, 1,565 sites had readings from an initial sample of 1,862 URLs. 297 sites were undetermined (could not reach a robots.txt file).

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*
ccbot	From Common Crawl Foundation. Web archive going back to 2008.	6.45%
gptbot	From OpenAI. Data is used to train current and future models, removing paywalled data, PII and data that violates the company's policies.	6.39%
chatgpt-user	From OpenAI. Visits websites when ChatGPT users request information and enables ChatGPT to include links in its responses.	4.91%
bytespider	From ByteDance. Downloads data to train LLMs, including ChatGPT competitors	4.38%
openai	Part of OpenAI's suite of crawlers.	4.32%
claudebot	From Anthropic. Scrapes data to train LLMs and AI products offered by Anthropic.	2.96%
meta-externalagent	From Meta. Meta-ExternalAgent crawls the web for use cases such as training AI models or improving products by indexing content directly.	2.72%
amazonbot	From Amazon, Includes references to crawled website when surfacing answers via Alexa; does not clearly outline other uses.	2.54%
anthropic-ai	From Anthropic. Scrapes data to train Anthropic's AI products.	2.37%
petalbot	From Huawei. Used to provide recommendations in Huawei assistant and AI search services.	2.25%

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 Indonesia's Evolving AI Landscape

The fact that only 10.1% of Indonesian news websites block at least one AI crawler, despite 74.6% having a robots.txt file on their website, shows that most publishers are not using robots.txt as a mechanism to signal their preferences regarding AI companies scraping their content. A range of factors may influence publishers' approaches, including resource constraints, business interests, or technical capacity to constantly monitor bot activity and navigate what content to protect or leave open. As AI restructures the relationship between publishers, platforms and audiences, these decisions also carry broader implications for visibility and audience reach.

One of the key challenges facing publishers in Indonesia is the significant loss in traffic referrals and weakening programmatic advertising revenues driven by the emergence of AI-powered search and summaries. According to a recent study by Monash University Indonesia and AMSI in Indonesia, traffic to news sites has declined significantly since the launch of Google's AI Overviews: among the top 5 news sites, average visitors fell from roughly 40% from September 2024 to November 2025. The hardest drop was observed by the top 16–20 news sites, about 59% over the same period. Across 44 outlets involved in the qualitative study, the average visitor trend dropped 34.07%, with a median of 45.10% traffic declining. These traffic losses translate into falling advertising revenue, as many publishers still depend on page views and programmatic advertising[1].

While blocking AI crawlers may offer some level of protection by limiting how freely AI can take and reuse their content, it may also impact discoverability if the publisher's content is not cited in AI answers. This poses a dilemma for publishers: stay open for visibility and referrals, or restrict bots to protect content and strengthen bargaining power.

A Strategic Dilemma: To Protect Content or Remain Visible?

Publishers face a difficult trade-off between protecting their content and staying visible to audiences on dominant search platforms.

Until mid-2026, Google did not allow content producers to opt-out of AI Overviews for content published online without impacting their visibility on Google Search. (Publishers were able to disallow google-extended in robots.txt to block some AI training use of their content for Gemini-based AI models).

Since August 2026, Google has rolled out distinct opt-out controls globally, giving publishers more control over whether they choose their content to be used by AI features such as AI Overviews and AI Mode. This move followed a decision in June by the UK's Competition and Markets Authority, requiring fairer and more transparent treatment from Google Search, notably giving publishers the ability to opt out of their content being used to power AI features.

However, this opt-out control continues to raise a strategic dilemma for publishers.

While opting out may not hurt a site's position in regular search results, it does mean losing visibility and visits from Google's AI features, which more and more people use to find information. By opting out, news sites may also drop out of Top Stories when those appear inside AI Overviews, since the setting treats everything shown in an AI section the same way.

The controls also do not apply to the Gemini app, so opted-out content can still appear there. On the other hand, allowing full access carries its own risk, because content owners do not have editorial control over how their content is used or cited. In addition, users are more likely to stay on AI summaries without clicking through to the original source, reducing traffic to the original site. This is why publisher associations such as AMSI in Indonesia note that a simple yes-or-no choice is not enough: publishers need more say over how their content is used, attributed, measured, licensed and compensated.

1 "AI Blockers and Generative AI Affordances: Traffic Risks and Protection Strategies For Indonesian News Organisations," Asosiasi Media Siber Indonesia (AMSI) & Monash University, June 2026.

Another challenge relates to significant gaps in Indonesia's legal and regulatory framework governing the unauthorized and unremunerated extraction of journalistic content by AI companies. Under Indonesia's current 2014 Copyright Law, the economic rights of journalistic works are not recognized in a way that obliges digital platforms or AI companies to share revenue when they profit from news content. There is no publishers' right or neighboring right for news, which is the legal mechanism that would allow media organizations to claim compensation for the commercial reuse of their content. This gap has allowed AI crawling to operate largely outside the scope of copyright enforcement.

That may be set to change. As of May 2026, the Legislative Board within the Indonesian House of Representatives (DPR RI) has circulated a draft revision that would explicitly classify journalistic works as a protected economic asset. Under this revision, AI companies would be required to obtain a license or pay royalties through a Collective Management Organization (CMO) for a wide range of uses, including aggregation, algorithmic recommendations, and AI model training. This would represent a significant departure from the EU's Copyright Directive, which allows AI companies to train on copyrighted works unless rightsholders explicitly opt out. The Indonesian draft flips this default: no automatic safe harbor, and mandatory licensing as a condition of use.

As an association representing digital media organizations across regions, Asosiasi Media Siber Indonesia (AMSI) has formally engaged with this legislative process, advocating for the revision on the grounds that technical measures like robots.txt are necessary but not sufficient. Blocking crawlers may reduce unauthorized extraction, but a legal framework is needed to convert that defensive action into enforceable rights and compensation mechanisms. Several implementation questions remain. The effectiveness of the CMO mechanism will depend on whether a credible and representative collective body can be established for the news sector, and whether the Ministry of Communication and Digital (Komdigi) has the institutional capacity to oversee compliance — a persistent challenge in Indonesia's regulatory environment. These efforts have also faced criticism from Google, as they continue to actively lobby the government against such measures. The final text, including its specific provisions on AI training data, remains subject to legislative revision.

In addition to the intellectual property approach, AMSI, the Press Council, and other media stakeholders in Indonesia are actively engaging with the Indonesian Commission for the Supervision of Business Competition. They aim to compel global platforms to compensate media companies through competition law. This approach is driven by similar initiatives in other countries, such as Australia through the News Media Bargaining Code—which is currently strengthening its leverage through the News Bargaining Incentive; Canada through the Online News Act, which requires global platforms to negotiate fair compensation for domestic news media and involves the Competition Bureau Canada if negotiations stall; and South Africa, carried out by its Competition Commission. As a precedent, in 2025, the Indonesian Competition Commission fined Google LLC 202.5 billion IDR (around $11.27 million USD) for abusing its dominant position regarding the Google Play Billing System on its app store. Google appealed the decision to the Supreme Court, but this appeal was rejected by the Supreme Court in May 2026.

Not all news organizations are equally positioned to navigate this changing technological and legal landscape, including negotiating copyright or content-use arrangements with digital platforms. Many small outlets operate with only three to ten journalists and function largely as a news division without dedicated legal, business, or technical teams. This makes them especially vulnerable, because they often lack the capacity to measure their losses, review the impact of AI crawling on their traffic and revenue, or engage platforms in meaningful negotiations. Some large media companies, by contrast, have more leverage and may pursue direct negotiations with digital platforms for their own institutional benefit rather than through collective action. In this context, AMSI's role is especially important: as the largest association of news organizations in Indonesia, its strategies, advocacy, and capacity to build a collective front can play a decisive role in shaping the future of journalism in Indonesia.

Conclusion and Next Steps

Even as the global debate over how AI systems acquire and use data intensifies, this study suggests that many Indonesian news publishers have not yet taken steps to manage how AI systems access their content. While the share of websites blocking at least one AI crawler has doubled since July 2025, it remains a small minority of the ecosystem. The gap between having a robots.txt file (74.6% of sites) and actually using it to set AI-specific directives (10.1%) points to an opportunity: many publishers already have the technical infrastructure in place but have not yet configured it for this purpose.

Next steps include:

1. **Further Research:** Based on the data and conclusions presented in this report, further research in Indonesia will focus on four key areas : longitudinal analysis of AI's impact on traffic and revenue, comparative evaluation of blocking and non-blocking strategies, deeper study of the vulnerability of small and local media, and audience research on how AI summaries are changing news discovery, attribution, and trust. Together, these insights can help clarify not only the scale of the harm, but also which technical and collective responses are most viable for sustaining journalism in Indonesia.

2. **Publisher Tools and Collective Action:** Whether to restrict AI crawlers is an individual decision that must align with each publisher's strategic priorities. However, individual technical measures alone are not sufficient to address the broader challenges facing news publishers. AMSI is now actively pursuing complementary efforts to strengthen publishers' rights, explore sustainable models for content use, and support collective approaches to engagement with AI platforms. This includes:

    - On the legal front, a legal framework is needed so that publishers' preferences on how their content may be accessed and used become enforceable rights that AI platforms are obliged to respect, rather than voluntary signals.

    - On the technical and commercial front, AMSI is exploring practical and scalable approaches to data sharing, content licensing, and valuation, helping to identify sustainable ways for publishers to engage with AI systems and participate in emerging commercial opportunities.

    - In parallel, partnerships among publishers, technology providers, and data initiatives can help advance these efforts by exploring shared infrastructure, new models for responsible data access, and fairer arrangements for the use of journalistic content.

3. **Cross-Border Partnership:** This report contributes to a rich, comparative dataset spanning Brazil, South Africa and Indonesia. The tricontinental CTRL+J partnership will continue to develop shared frameworks and policy recommendations grounded in Global South media conditions that 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 news organizations' websites from Indonesia. The analysis conducted on September 30, 2026 (and presented in this report) is based on an initial sample of 1,862 URLs. Of these, 1,771 corresponded to unique sites and 1,565 robots.txt files were successfully verified. The sample was compiled from multiple sources, including AMSI's member outlets, Press Council-verified online media, and human-vetted databases from 2022 and 2023 (altogether matching 1,224 unique sites in the sample). The remaining URLs were supplemented by other media databases. The sample does not include every domain in AMSI's datasets. Inclusion in AMSI's member list reflects membership status at the time the list was compiled, while inclusion in the Press Council–verified media list reflects the verification status recorded at that time. Neither constitutes an endorsement by AMSI of an outlet's editorial practices. 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[2], 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[3] 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.

2 List of known AI crawlers, 2025, Github repository - ai.robots.txt
3 OECD (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.

1.2 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.

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 behaviour
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: / Us
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 Dis
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: Crawlspace
User-agent: INDONESIA Disallow: / User-agent: DuckAssis
User-agent: EchoboxBot Disallow: / User-agent: Facebool
Disallow: / User-agent: Factset_spyderbot Disallow: / User
FirecrawlAgent Disallow: / User-agent: FriendlyCrawler D
User-agent: Google-CloudVertexBot Disallow: / User-agen
Google-Extended Disallow: / User-agent: GoogleOther Dis
User-agent: GoogleOther- October 2026 : / User-agent:
GoogleOther-Video Disallow: / User-agent: GPTBot Disallo
User-agent: iaskspider/2.0 Disallow: / User-agent: ICC-Cra
Disallow: / User-agent: ImagesiftBot Disallow: / User-agen
img2dataset Disallow: / User-agent: ISSCyberRiskCrawler
User-agent: Kangaroo Bot Disallow: / User-agent: meta-ex

