Cloudflare response to the CMA consultation on the Publisher Conduct Requirement relating to Google's general search services

February 25, 2026

Cloudflare appreciates the opportunity to respond to the public consultation by the Competition and Markets Authority on its proposed Publisher Conduct Requirement (CR) relating to Google's general search services.

1. Summary

Cloudflare believes the CMA's proposed Publisher Conduct Requirement (CR) represents a critical step toward rebalancing the digital publishing ecosystem, yet it currently stops short of addressing the structural root of the problem: Google's dual-purpose use of its search crawler. While the CR is a welcome acknowledgment of publisher rights, the mechanism the CMA has proposed ultimately delegates the definition and implementation of those rights back to Google, a company with a vested interest in maintaining its data pipeline.

To ensure a fair and competitive future for digital publishing, Cloudflare recommends three core pillars for the final CR:

A) **Mandatory crawler separation:** This remains the only viable technical solution to restore true autonomy and choice to publishers. By requiring Google to separate traditional search indexing from AI training and grounding crawlers, the CMA would enable publishers to protect their intellectual property without sacrificing vital search traffic or having to navigate a maze of Google content controls.

B) **Granular, unified controls:** If the CMA proceeds with its proposed model, it must require a single, comprehensive set of controls that allow site-wide and page-level distinctions between search indexing, AI training/fine-tuning, and AI grounding/summarization. Publishers must be able to exercise distinct, granular controls over those economically different uses of their content. Crucially, there must be an option to only opt into search indexing ('traditional' search), without being subject to down- or misranking in general search results.

C) **Independent enforcement and proactive transparency:** Oversight cannot be left to Google's self-assessment alone. The CMA should mandate regular, independent

third-party audits—including adversarial testing and on-site inspections—to verify that opt-out choices are respected and do not result in search ranking penalties. Furthermore, Google must be required to proactively notify publishers when it extends controls to new generative AI services or otherwise changes its data use or control parameters.

Without these robust safeguards, the power imbalance between dominant search platforms and content creators will continue to stifle innovation and undermine the viability of the open Internet. We urge the CMA to adopt an approach that places technical control back in the hands of the publishers who create the very content upon which the AI ecosystem depends.

2. About Cloudflare

Cloudflare provides cybersecurity, reliability and web performance services to millions of websites and applications. With millions of customers relying on the security and integrity of our network, Cloudflare leverages predictive AI to continuously detect and address threats – both known and new vulnerabilities – before they can cause harm. Our AI technologies learn from each encounter, improving their ability to recognise and counteract new and emerging threats and ensuring that our defences remain robust and up-to-date. Cloudflare's bot management service can identify and block bots (including AI bots) from scraping content without consent and/or engaging in other abusive behaviour.

3. Google's search dominance creates an unfair advantage in AI, undermining competition and fair compensation

Publishers are currently trapped in a dilemma: they cannot afford to block Google's search crawler without losing vital traffic, yet this same access allows Google to ingest their content for generative AI search features like AI Overviews. This creates a paradox where Google reproduces publisher content—often without attribution or compensation—effectively replacing the very sources it relies on and depriving publishers of ad revenues because of lower click-through rates.

This privileged access also gives Google an unfair advantage over competing AI companies and undermines a fair marketplace for content that is used for training and grounding Google's generative AI models. While competitors must often negotiate for access to content, Google bypasses this negotiation by leveraging its search dominance. Cloudflare data confirms this structural advantage: Googlebot accesses significantly more unique web pages than its peers, reaching almost twice as many pages as GPTBot

and ClaudeBot, and 167 times more than PerplexityBot.[1] Because publishers fear the consequences of blocking Googlebot for their search ranking, Google maintains a near-exclusive pipeline of free data that its rivals simply cannot match.

4. The CMA's proposed controls are a less effective remedy than crawler separation

Based on the above, we share the CMA's conclusion that publishers currently "are unable to exercise sufficient choice over how their Search Content is used by Google across its generative AI services and features."[2] This lack of control leads to adverse impacts for publishers, limiting their ability to restrict how Google uses their content and making it more challenging for them to monetise it.

However, we are concerned that the proposed CR does not actually solve the underlying issue of promoting effective and transparent control for publishers over their content, and thus ultimately undermines the intended purpose of these measures. In particular, we disagree with the CMA's assessment that requiring Google to develop new controls for publishers to opt out of the use of data for AI purposes would constitute an "equally effective remedy"[3] compared to imposing a remedy that would require crawler separation.

We consider the new proposed controls to be less effective than crawler separation for the following reasons:

- **The proposed controls fail to shift the underlying power dynamic, as publishers remain dependent on Google's proprietary systems:** Even with the new controls contemplated in the proposed Publisher CR, it would still be effectively impossible for publishers to block Googlebot from accessing their content with a technical crawler blocking tool of their choice, such as Cloudflare's Crawl Control, without jeopardizing their appearance in search results. Content creators will have to manage the proposed controls as specified by Google. Ultimately, this leaves publishers with less autonomy than if they were able to enforce their preferences by only allowing access upfront to Google's crawler for traditional search purposes and at the same time being able to block it outright for undesired purposes, such as generative AI training/fine-tuning and grounding.
- **The proposed solution provides less effective transparency for publishers on Google's use of their content**: Without the ability to control the technical

[1] See Appendix for full data set
[2] CMA Publisher Conduct Requirement, paras. 1.15-1.16, pages 8-10
[3] CMA Publisher Conduct Requirement, para 5.12, page 51

mechanisms that allow or disallow Google's access to their content for specific purposes, publishers will not be able to depend on their own data to view access by Google. Instead, they will need to interface with Google and its platform, relying on information provided by Google in a level of detail, format and timeline that Google will largely define and set the parameters for.

These shortcomings would remain true even if the CMA enhances its current proposed CR to include additional specificity and more tailored controls and transparency requirements, which we discuss in more detail below. Ultimately, behavioral tweaks to existing systems are insufficient to address the deep-seated power imbalance. A more fundamental, structural shift is required to ensure genuine transparency and market fairness.

5. Crawler separation: a prerequisite for effective control and innovation

Instead, as we have previously submitted to the CMA, the most effective way to give publishers control over the use of their content is to require Googlebot to be split up into separate crawlers.[4] Such an approach appropriately puts the burden on Google to designate the intended purpose of its crawling, rather than putting the burden on publishers to decode Google's proposed controls. With split crawlers, publishers could allow crawling for traditional search indexing, which they need to attract traffic to their sites, but also block access for unwanted use of their content in generative AI services and features.

The technical and structural benefits of separating search indexing from AI training and grounding are fundamental to fostering a competitive and innovative digital ecosystem, and would by far outweigh any costs that may be incurred by Google or third parties compared to the proposed controls. Maintaining distinct crawlers is already an established industry standard for leading AI firms, including Anthropic and OpenAI, which operate separate bots for different purposes.

Allowing a dominant provider to maintain a multi-purpose bot grants an unfair structural advantage that stifles competition. **Crawler separation would instead establish a level playing field**, ensuring that competition in the AI market is based on model performance rather than the ability to leverage a legacy search monopoly to bypass publisher controls. By mandating separation, the CMA would signal that market power cannot be used to

[4] More broadly, as reflected in our proposed responsible AI bot principles, we believe that all AI bots should have one distinct purpose and declare it, so that website owners can make clear decisions over who can access their content and why. Unlike its leading competitors, such as OpenAI and Anthropic, Google does not comply with this principle for Googlebot, which is used for multiple purposes (search indexing, AI training, and inference/grounding).

dictate the terms of data access, thereby encouraging fair competition between established players and emerging AI startups.

Separating crawlers allows for superior technical optimization. A crawler limited to traditional search functions must behave differently—scraping more broadly and frequently—than a crawler dedicated to generative AI training or grounding (RAG). Splitting these functions enables more efficient data retrieval and reduces technical complexity over the long term.

More importantly, innovation in AI is entirely dependent on the continued viability of high-quality, original online content. Crawler separation enables content creators to monetize and control their output through explicit, granular authorizations (e.g., enabling access to crawlers used for search indexing while blocking those used for AI training). **This ensures the long-term sustainability of the data pipeline upon which all AI innovation depends**.

Implementing crawler separation is technically straightforward, primarily requiring the use of distinct User-Agents (such as 'Googlebot-search' and 'Googlebot-train') to declare their specific end-purpose. Rather than requiring a retrospective revision of existing models, the focus should remain on future training and real-time grounding. This approach provides a clear technical safeguard that achieves the core objective of ensuring content is only accessed for explicitly permitted purposes. It also ensures minimal technical burdens compared to the complex behavioral controls currently proposed. Such a structural shift would effectively mitigate economic harm to publishers while driving broader industry progress overall.

6. Recommendations to improve the proposed Publisher CR

In the above sections, we have set out extensively why we still firmly believe that crawler separation is the only path to achieve the CMA's stated goals with the Publisher CR. If however, the CMA chooses not to proceed with crawler separation, we urge it to increase the prescribed granularity of publisher controls and strengthen the transparency and compliance requirements that are currently proposed.

A) The proposed controls lack granularity and leave important decisions to Google

Under its proposal, the CMA distinguishes between two use-cases by Google for Search Content:

- Training and grounding of broader generative AI services (e.g. Gemini), which should be covered by a further developed Google-Extended control, and

- grounding of search generative AI features, which should be covered by a separate, new tool that is still to be developed.

This distinction is technically confusing, as the two controls cover different services (broader generative AI services vs. search generative AI features) and use categories (training and grounding vs. grounding only). This is not intuitive and not every publisher will be able to tell the difference. The proposal also lacks the granularity that is both desired by many publishers and necessary to provide them with meaningful choice over Google's use of their content across its AI services and features. We urge the CMA to be more specific and prescriptive in defining those categories. **Publishers must be able to specify how their data is utilized—whether for search indexing, grounding, or training/fine-tuning—at both the site and individual page levels**. To counter the effect of Google's strategic market status in search, publishers should have an easy way to opt out of Google using their data for anything other than traditional search, without affecting their general search ranking. Leaving so much discretion to Google on defining these important parameters when it has a vested interest in disincentivizing publishers from opting out of any AI crawling, especially with a view to strategic monetization and licensing of content, exposes the CR to a high risk of failure.

As the CMA correctly recognizes, the scope of Google-Extended and how it develops with different products and services is unclear. We welcome the CMA's clarification that the scope of Google-Extended should be clearly defined in a product-agnostic way. But this definition should not be left to Google. In particular, we consider it **problematic that Google-Extended currently offers a single control covering both training and grounding use cases**. Separate controls allow publishers to treat their data as two different types of assets: IP for intelligence (training/fine-tuning) and links for traffic (grounding). For example, a publisher could require a license for the use of their content for data for training or fine-tuning, but might be willing to allow grounding, even without a license fee—if it reliably sends high-intent traffic back to their site via citations.

In addition, **the new control to be developed by Google should also apply to the fine-tuning of models underlying generative AI search features**. This would make it much more likely for publishers to adopt the control. Google's claim that opting-out of fine-tuning risks downranking or mis-ranking publisher content in organic search results proves it is leveraging search dominance to secure an unfair advantage for its generative AI products. This is quite alarming and must be challenged. **Crucially, there must be an option to only opt into search indexing ('traditional' search), without being subject to down- or misranking in general search results.** If necessary to prevent any negative effects on search ranking, Google should be obliged to distinguish between models

underlying generative AI search features and those underlying ranking, even if this means creating duplicative models.

B) The CMA should mandate a comprehensive set of granular controls, clearly defining primary AI use-cases

Based on our own conversations with publishers about their needs, we believe there should be one comprehensive set of controls for the use of Search Content from both broader generative AI services and search generative AI features at both site and page level, *at least* distinguishing between the following primary use categories, which should be clearly defined in the CR:

(a) search indexing,
(b) AI training/fine-tuning, and
(c) AI grounding or summarisation.

These are the same categories we have proposed for crawler separation, which are also part of our Content Signals Policy and have been integrated in the recently adopted RSL (Really Simple Licensing) standard.

C) Strengthening proactive and granular transparency

We appreciate the CMA's strong focus on imposing robust transparency requirements throughout the proposed CR. Regarding transparency over use of publisher content and in relation to Google's controls, Cloudflare believes the CMA has identified the right themes. However, it will be challenging for publishers and third parties to keep track of any future changes, even if the information is made available in a publicly accessible manner. For example, we appreciate that Google would be required to publicly disclose when and how their controls are extended to any new generative AI services as they are released. However, given the rapid pace of innovation, and the imperative for Google to be transparent about the scope of its controls, we strongly recommend that the CMA require Google to **provide publishers and third parties with a proactive notification mechanism for relevant updates**. This is preferable to burdening publishers with the resource-intensive task of independently tracking and interpreting Google's internal changes.

In order to increase transparency for publishers, we also agree with the CMA's provisional view that, in its information for publishers, Google must provide, *at a minimum*, performance data such as 'Clicks, 'Impressions' and 'CTR' in aggregate for search generative AI features as separate metrics from those provided in relation to

non-generative AI search results. We think **the CMA should go even further and require per-feature metrics**, which will enable publishers to better assess the commercial impact of Google's use of their content and support their bargaining position.

D) Ensuring independent and ongoing enforcement

With regard to compliance mechanisms, we appreciate that Google is asked to publish a non-confidential version of its compliance reports. We recommend adding a requirement that Google provide **specific information in these non-confidential reports on ways Google implements, monitors, and confirms compliance with the obligations not to circumvent publishers' opt-out choices and to ensure that publishers' opt-out choices do not impact their search rankings.**

We also agree with the CMA's consideration that there is an intrinsic value in having the initial baseline compliance audit carried out by an **independent third party** to increase publishers' trust in the measures. We further recommend that this independent audit should also include mechanisms like **random on-site inspections for additional verification**, in addition to adversarial testing. Furthermore, we recommend the CMA require **follow-up audits by an independent third party at regular intervals to ensure ongoing compliance**. Those would complement Google's own six-monthly compliance reports, which cannot be independently verified.

Given Google's wide discretion over the implementation of the controls under the current proposal, we consider it very likely that the controls will fail to be effective. We therefore urge the CMA to **clearly commit in the CR to rapidly deploying more stringent, structural remedies—such as mandatory structural crawler separation—should non-compliance or ineffective implementation occur**.

7. Conclusion

The rapid evolution of generative AI has fundamentally shifted the Internet's value exchange. The UK has a unique opportunity to lead globally by establishing an online content marketplace defined by technical autonomy rather than platform dependency.

Cloudflare remains convinced that mandatory crawler separation is a prerequisite for a truly innovative digital ecosystem. By requiring Google to designate its crawling purposes upfront—aligning with standards already set by other leading AI companies—the CMA can establish a level playing field. This structural shift moves beyond tweaks to existing systems controlled by Google, ensuring that competition is based on model performance rather than search dominance. Crucially, this will protect the economic viability of

publishers, securing the long-term sustainability of the high-quality data pipeline that fuels all AI progress.

If the CMA proceeds with behavioral controls, they must be granular, subject to proactive notification, and verified by independent third-party audits to ensure publisher choices are not circumvented. **In particular, there must be an option to only allow search indexing without being subject to down- or misranking in general search results.** In addition, disaggregated performance metrics are essential to monitor for any punitive impacts on site visibility in general search results or user engagement.

We look forward to collaborating with the CMA to ensure the UK becomes a world-class hub for a fair, open marketplace for content that drives the next generation of AI innovation.

APPENDIX: Cloudflare data sets on Google's privileged access to online content

Over a period of two months (December 1st, 2025 to January 31st, 2026), Googlebot saw (in rounded multiple terms):

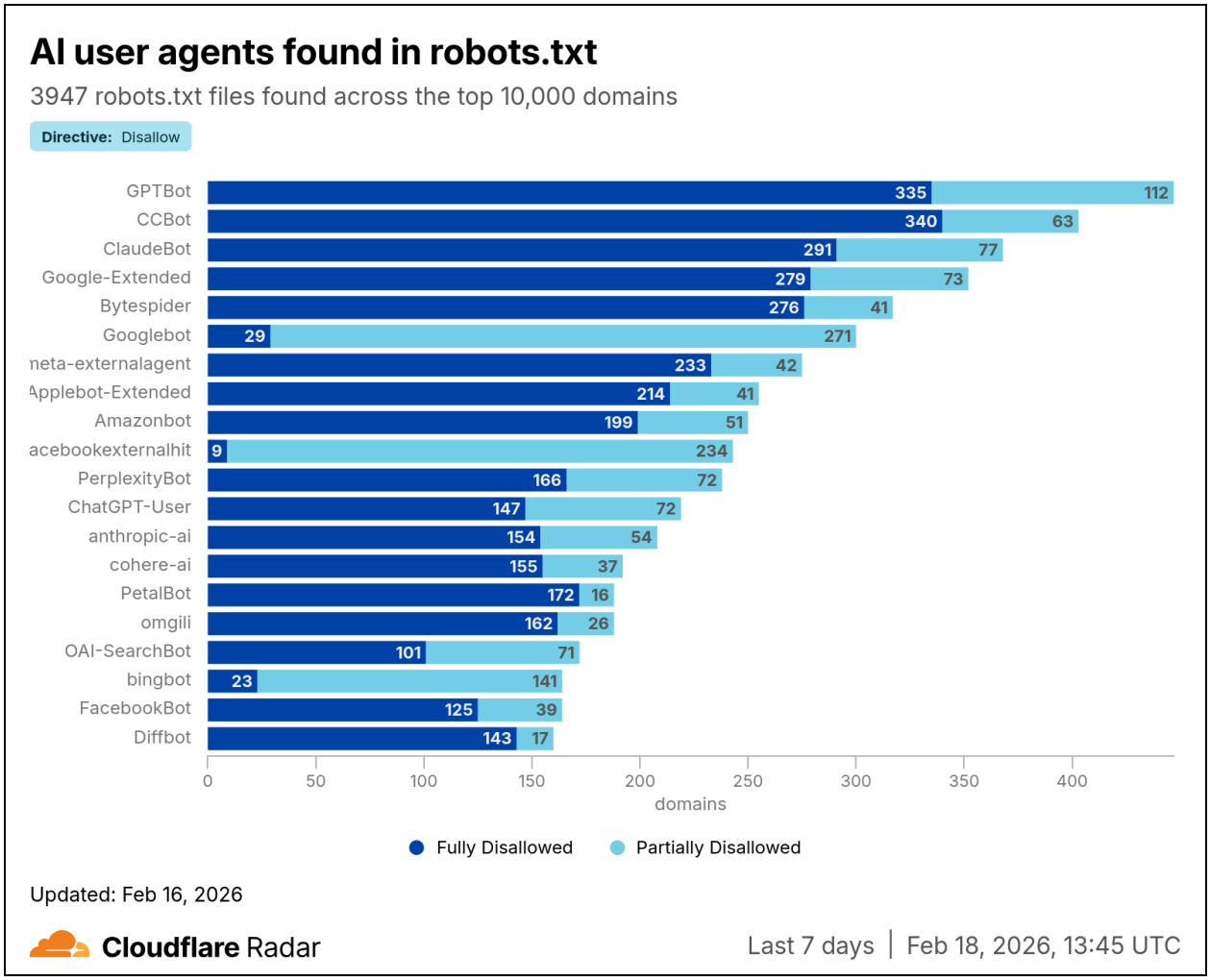
- vs. ~1.70x the amount of unique URLs seen by ClaudeBot;
- vs. ~1.76x the amount of unique URLs seen by GPTBot;
- vs. ~2.99x the amount of unique URLs by Meta-ExternalAgent;
- vs. ~3.26x the amount of unique URLs seen by Bingbot;
- vs. ~5.09x the amount of unique URLs seen by Amazonbot;
- vs. ~14.87x the amount of unique URLs seen by Applebot;
- vs. ~23.73x the amount of unique URLs seen by Bytespider;
- vs. ~166.98x the amount of unique URLs seen by PerplexityBot;
- vs. ~714.48x the amount of unique URLs seen by CCBot; and
- vs: ~1801.97x the amount of unique URLs seen by archive.org_bot.

Googlebot also stands out in other Cloudflare datasets:

Even though it ranks as the most active bot by overall traffic, publishers are far less likely to disallow Googlebot in their robots.txt file compared to other crawlers. This is likely due to its importance in driving human traffic to their content—and, as a result, ad revenue—through search.

As shown below, almost no website explicitly disallows the dual-purpose Googlebot in full, reflecting how important this bot is to driving traffic via search referrals. (Note that

partial disallows often impact certain parts of a website that are irrelevant for search engine optimization, or SEO, such as login endpoints.)



Robots.txt merely allows the expression of crawling preferences; it is not an enforcement mechanism. Publishers rely on “good bots” to comply. To manage crawler access to their sites more effectively—and independently of a given bot’s compliance—publishers can set up a Web Application Firewall (WAF) with specific rules, technically preventing undesired crawlers from accessing their sites. Following the same logic as with robots.txt above, we would expect websites to block mostly other AI crawlers but not Googlebot.

Indeed, when comparing the numbers for customers using AI Crawl Control, Cloudflare’s own AI crawler blocking tool that is integrated in our Application Security suite, between July 2025 and January 2026, one can see that the number of websites actively blocking

other popular AI crawlers (e.g., GPTBot, Claudebot), was nearly seven times as high as the number of websites that blocked Googlebot and Bingbot. (Like Googlebot, Bingbot combines search and AI crawling and drives traffic to these sites, but given its small market share in search, its impact is less significant.)