NEWS MEDIA
ALLIANCE

Wholesaling of Search Indexes:
Exploiting Publisher Content Through the Backdoor

August 13, 2025

High-quality newspapers, magazines, and digital media publishers play an important role in their communities and society as a whole by fostering an informed public and a healthy democracy. These publishers invest considerable time and resources to produce journalism and original creative content that keeps their readers informed, engaged, and entertained. However, the proliferation of generative artificial intelligence (AI) models, applications, and developers over the last few years presents not only benefits but considerable challenges to publishers and their public service mission. Now, a new challenge emerges: traditional search engines have begun exploiting their special privilege to copy websites for indexing purposes, turning around to resell this indexed content for non-search uses.

Far too many AI developers simply scrape and use publisher online content without authorization or compensation to train AI models or to provide them with real-time content in response to user queries (grounding or Retrieval Augmented Generation, RAG). These harms are exacerbated by the actions of leading search engines – including Google, Microsoft Bing, and Brave – who make their search indexes available to third parties, including for exploitation by AI companies for the purposes of operating their "answer engines." Publishers by and large allow the crawling of their websites for search indexing purposes to ensure visibility in search engine results and to bring in monetizable traffic. But they have not consented to the downstream exploitation of this indexed content, especially for purposes that, through this act, deepen significant risks to publishers and for which licensing opportunities exist.

These uses are particularly concerning considering the significant monetization effects AI applications already have on publishers, with a recent study by TollBit estimating that AI answer engines deliver 91 percent fewer referrals to news websites than traditional search engines.[1] And this trend is only getting worse – recently, every two scrapes by Google used to lead to one visitor but today that ratio is 18:1, while OpenAI's crawl to visitor ratio has increased from 250:1 to 1,500:1 and Anthropic's from 6,000:1 to 60,000:1 in just six months.[2]

While the direct misappropriation of publisher content by AI developers is well-reported, less is known about the downstream wholesaling of search indexes by platforms to AI developers, which equally reduces publisher control over the use and distribution of their content and impacts emerging and

[1] TollBit, TollBit Artificial Intelligence User Agent Index - Q1 2025 at 15 (Feb. 24, 2025), https://tollbit.com/bots/24q4/.

[2] Rob Thubron, *Cloudflare Tests "Pay-for-Crawl" System to Charge AI Firms for Scraping Website Content*, TechSpot (Jul. 1, 2025), https://www.techspot.com/news/108521-cloudflare-tests-pay-crawl-system-charges-ai-firms.html; Ethan Hays (@ethanhays), X (Jun. 27, 2025, 6:32 PM), https://x.com/ethanhays/status/1938651733976310151.

existing licensing and partnership markets. This document outlines the issue, discusses some of the challenges, and highlights outstanding questions publishers have about these practices.

How Search Indexing Works and How Indexes are Made Available to Third Parties

For decades, search engines and publishers operated under a mutual understanding that inclusion in search results helps both parties – publishers benefit from visibility that drives monetizable traffic while search engines can access a comprehensive catalogue of web content to deliver better results. Accordingly, publishers generally allow search engines to crawl their sites and include them in an index that powers the search engine. To crawl the internet, search providers' bots, such as Googlebot or Bingbot, identify themselves and can be verified by the website host. Indexing websites involves scraping the content published on those sites, including potentially making full copies of copyright protected works. According to Google, Googlebot renders the entire page and stores "key signals" like keywords and website freshness in an index containing "hundreds of billions of webpages",[3] while others note that when "Google indexes a page, it saves a copy of the page in its search results database, or index—which is not publicly accessible. It also saved another copy of the page in its cache—which was publicly accessible."[4] Similarly, Bing stores copies of crawled pages in Bing Cache, and Bingbot "crawls the page at regular intervals to update the content in our index and stores a copy in Bing cache."[5] Search engine results then link directly back to the content, providing publishers with valuable user traffic.

Recently, however, in addition to using their search indexes to provide search results, indexing companies have started providing access to their indexes to third-party AI developers, often via Application Programming Interfaces (APIs). Google's offerings include Custom Search JSON API,[6] which allows developers "to retrieve and display search results from Programmable Search Engine programmatically" in JSON format, and Grounded Generation API which enables the creation of

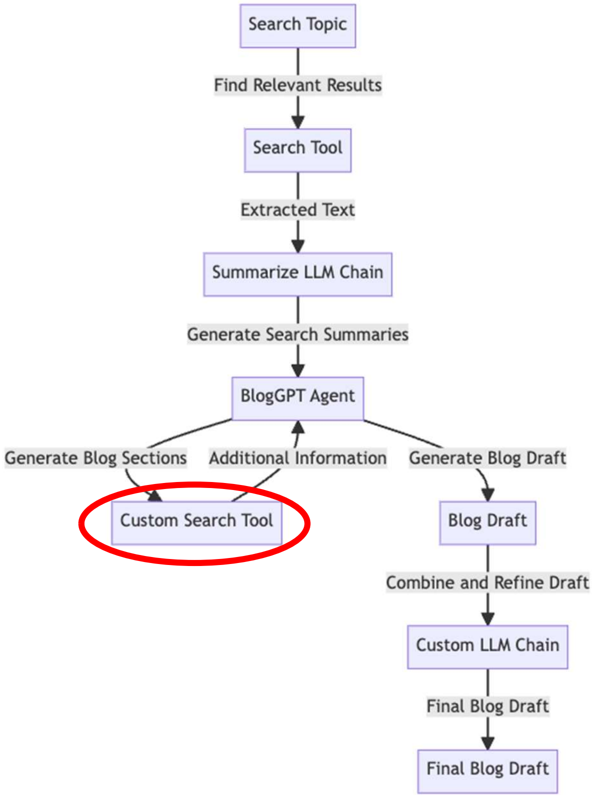
[3] *How Google Search Organizes Information*, GOOGLE SEARCH, https://www.google.com/search/howsearchworks/how-search-works/organizing-information/ (last visited Aug. 8, 2025).

[4] Tan Siew Ann, *Removal of Google Cached Pages: Impact and Alternatives*, SEMRUSH BLOG, (Mar 20, 2024) https://www.semrush.com/blog/google-cached-pages/; Tr. of Bench Trial at 6303, *United States v. Google LLC* (Oct. 18, 2023) (20-cv-3010), avail. at https://thecapitolforum.com/wp-content/uploads/2023/10/20231018-APM-BT24-AM-Google.pdf (Google's VP, Search: "We go out using a process of crawling the web and other mechanisms for acquiring the content, and we create an index of hundreds of billions of documents that we hope is comprehensive."); *Id.* at 6305*;* Cyrus Shepard, *Google's Index Size Revealed: 400 Billion Docs (& Changing)*, ZYPPY LIST (Jun. 24, 2025), https://zyppy.com/seo/google-index-size/ ("Not only does Google store each document, it creates a tremendous amount of data about each document, including all the words and concepts related to each document."); John Lanchester, *Googled: The End of the World as We Know It by Ken Auletta*, THE GUARDIAN (Feb. 21, 2010), https://www.theguardian.com/books/2010/feb/21/googled-ken-auletta ("[W]hat Google does instead is make a copy of the entire internet – everything they can get access to – store it on their own servers, and then index it").

[5] *Content Removal: Report Broken Links or Outdated Cache Pages*, MICROSOFT BING, https://www.bing.com/webmasters/help/bing-content-removal-tool-cb6c294d (last accessed Aug. 8, 2025).

[6] *Custom Search JSON API*, GOOGLE, https://developers.google.com/custom-search/v1/overview (last accessed Aug. 8, 2025).

"generative answers to your prompts using information on Google Search or your own data." [7] "[I]nformation on Google Search" includes copyrighted publisher and other web content that publishers spend considerable time and resources to produce, even perhaps encompassing content copied for indexing purposes that does not appear on Google's search engine results page (SERP). It is unclear whether Custom Search JSON API can be used for AI purposes, including to ground Large Language Models (LLMs), but sources indicate this may be possible.[8] The graph below shows how Custom Search JSON API would assumedly fit into the architecture of an AI service designed for drafting new blog posts.



Graphics Source: Minhajul Hoque, *BlogGPT: Your Personal AI Blog Writer*, MEDIUM (Jul. 12, 2023), https://medium.com/@minh.hoque/bloggpt-your-personal-ai-blog-writer-2e58119199f0 (*highlight added*)

Microsoft and Brave offer similar services. Microsoft's Bing Search API v7[9] allows developers to retrieve "web documents indexed by Bing," and integrate functionalities such as searching for news to get

[7] *AI Applications Pricing*, GOOGLE CLOUD, https://cloud.google.com/generative-ai-app-builder/pricing#grounded_generation_api_pricing (last accessed Aug. 8, 2025).
[8] Minhajul Hoque, *BlogGPT: Your Personal AI Blog Writer*, MEDIUM (Jul. 12, 2023), https://medium.com/@minh.hoque/bloggpt-your-personal-ai-blog-writer-2e58119199f0; Yuriks, *Google Search API: Best Way to Work with JSON Response*, OPENAI DEVELOPER COMMUNITY (Aug. 3, 2023), https://community.openai.com/t/google-search-api-best-way-to-work-with-json-response/315525.
[9] Scheduled to be discontinued shortly, although some developers will reportedly retain access. *See* Tom Warren, *Microsoft Shuts Off Bing Search APIs and Recommends Switching to AI,* THE VERGE (May 15, 2025), https://www.theverge.com/news/667517/microsoft-bing-search-api-end-of-support-ai-replacement.

"comprehensive results," including an "authoritative image of the news article, related news and categories, provider info, article URL, and date added." [10] Brave Search API provides "third-party chatbots with a valuable, affordable way to access search functionality."[11] Like Google's Grounded Generation API, this is clearly aimed at AI developers, with Brave even emphasizing in a lawsuit that the low cost of its API is "why companies, like Cohere, Mistral AI, Perplexity, and You.com, are coming to Brave."[12]

In addition to API-based sharing of search indexes, Google provides access to its search through Vertex AI, which allows users to ground supported models using Google Search (Grounding with Google Search).[13] Grounding with Google Search seemingly returns "the relevant web pages and snippets" and "a detailed set of information that includes not only the final text answer but also the sources it used to generate that answer."[14] As News/Media Alliance has discussed elsewhere, even snippets can be substantial and "collectively provide ample information on any news story to satisfy the casual reader skimming the news,"[15] in potential violation of copyright law. Such extracts are highly substitutional when used to power generative AI chatbots and "answer engines". In addition to extracting creative content, Vertex AI Search – Google's "out-of-the-box RAG system" which seems to also pull from publicly available online resources if combined with Grounding with Google Search – seemingly also "identifies structural and content elements, including titles, section headings, paragraphs, and tables,"[16] used to provide high-quality and more relevant AI generated output to users.

Recently, Anthropic also announced functionality for its Anthropic API that allows Claude to answer questions beyond its knowledge cut-off point. The API appears to generate a targeted search query, retrieve and analyze relevant results, "and provides a comprehensive answer with citations back to the source material."[17] It is unclear which search index powers Anthropic's new answer engine functionality and exactly what content it provides to developers using it. Mistral also recently announced a similar functionality, allowing developers to "combine Mistral models with diverse, up-to-date information from web search, reputable news, and other sources" using Mistral Agents API.[18]

[10] *Bing Search API Pricing*, MICROSOFT, https://www.microsoft.com/en-us/bing/apis/pricing (last accessed Aug. 8, 2025).

[11] Complaint at 9, *Brave Software Inc. v. News Corp.*, 3:25-cv-02503 (N.D. Cal. Mar. 12, 2025).

[12] *Id.*

[13] *Grounding Overview*, GOOGLE CLOUD, https://cloud.google.com/vertex-ai/generative-ai/docs/grounding/overview (last accessed Aug. 8, 2025).

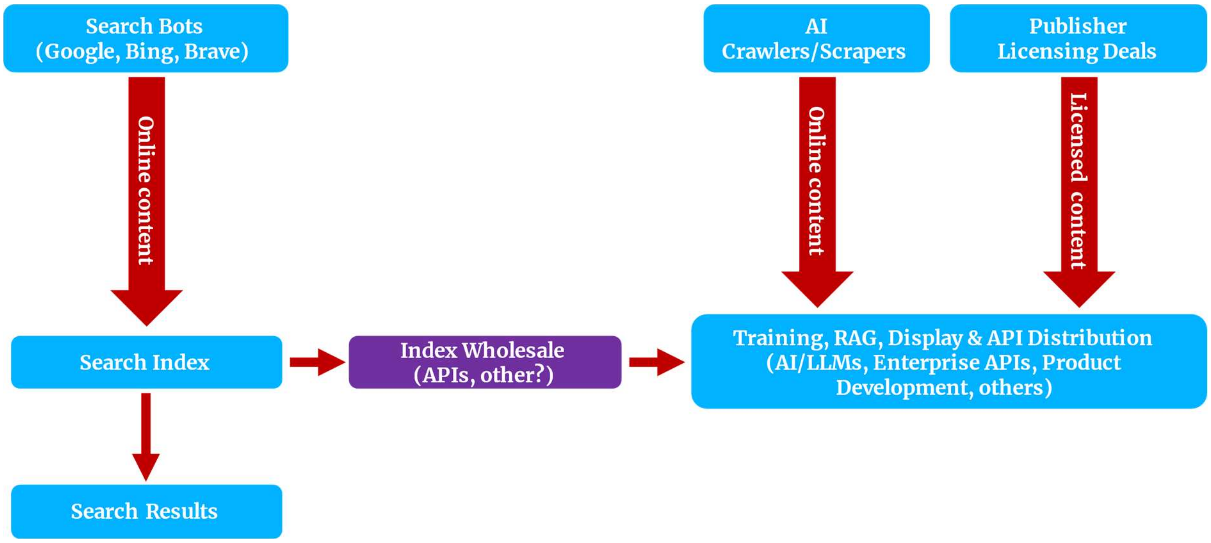
[14] *Understanding Google Search Grounding*, GOOGLE, https://google.github.io/adk-docs/grounding/google_search_grounding/#detailed-description (last accessed Aug. 8, 2025).

[15] News/Media Alliance, *How Google Abuses Its Position as a Market Dominant Platform to Strong-Arm News Publishers and Hurt Journalism* (Sep. 6, 2022), https://www.newsmediaalliance.org/wp-content/uploads/2022/09/NMA-White-Paper_REVISED-Sept-2022.pdf

[16] Kaz Sato and Guangsha Shi, *Your RAGs Powered by Google Search Technology, Part 2*, GOOGLE CLOUD BLOG (Feb. 14, 2024), https://cloud.google.com/blog/products/ai-machine-learning/rags-powered-by-google-search-technology-part-2.

[17] *Introducing Web Search on the Anthropic API*, ANTHROPIC (May 7, 2025), https://www.anthropic.com/news/web-search-api.

[18] *Build AI Agents with the Mistral Agents API*, MISTRAL (May 27, 2025), https://mistral.ai/news/agents-api.


Search Bots (Google, Bing, Brave)
Online content
Search Index
Search Results
Index Wholesale (APIs, other?)
AI Crawlers/Scrapers
Online content
Publisher Licensing Deals
Licensed content
Training, RAG, Display & API Distribution (AI/LLMs, Enterprise APIs, Product Development, others)

<u>Wholesaling of Search Indexes Has the Potential to Significantly Undermine Publishers' Business Models and Revenue Streams</u>

The unauthorized sub-licensing of and making available publisher content scraped for search purposes can have significant effects on publishers, undermining economic models and ultimately threatening the availability of high-quality content and journalism to local communities. Licensing, advertising, and subscription revenues – all of which rely on publishers' ability to attract and retain readers on their websites – keep publishers in business and enable their continued investments in new content. The effects of unfettered search index resale are in many ways analogous to companies licensing publisher content for limited internal-use purposes suddenly and unexpectedly starting to supply publisher content via APIs to a broad pool of third-party corporates without permission. These disruptions of expectations can have broad ripple-effects and would clearly violate the original agreement as a breach of the rights included in such agreement, typically subject to negotiated compensation.

With the proliferation of generative AI models and services and the resultant changes in consumer habits, content licensing is becoming an increasingly important avenue for publishers to offset decreases in website traffic and any resultant loss of advertising and subscription revenue. In the past two years, over 145 licensing deals between AI developers and rightsholders have been reported, including multiple with publishers such as Axel Springer, The Atlantic, AP, The Guardian, The Financial Times, News Corp., and others.[19] Some deals are reportedly worth millions of dollars per year and grant the

[19] *See, e.g.*, Helen Coster, *Global News Publisher Axel Springer Partners with OpenAI in Landmark Deal*, REUTERS (Dec. 13, 2023), https://www.reuters.com/business/media-telecom/global-news-publisher-axel-springer-partners-with-openai-landmark-deal-2023-12-13/; Sara Fischer, *Exclusive: AP Strikes News-Sharing and Tech Deal with*

right to use content for various purposes, including real-time information retrieval. The sale of access to search indexes may provide these same services to AI developers, bypassing the publisher and its IP rights, and directly threatening these licensing markets.

The sharing of access to search indexes and results between the big tech companies is particularly concerning to publishers. In May 2023, Google noted that to drive "Search growth & diversification … we must introduce new user access points, meeting users where they are," and sought approval – presumably granted – for the syndication of the search index to "pilot embedded Google Search with up to 15 partners in 2023 to validate hypotheses and refine offering."[20] This decision to syndicate search results to AI and non-AI third parties appears to have marked a significant shift in Google's approach to search. The use of the Google search index to provide responses in AI powered search engines such as Perplexity, caused significant concern for publishers due to the lack of onward traffic from those services to original publisher websites.[21] A 2024 email from OpenAI to Google confirmed that "Google works with Meta to provide search features and link grounding in Meta's AI chatbots and AI products."[22] The email suggested that this was a "very low-cost partnership." Recent reports indicate that Llama 3, which powers Meta AI,[23] also uses Brave Search "to answer questions about recent events that go beyond its knowledge cutoff or that require retrieving a particular piece of information from the web."[24] Llama 3 is also the "only assistant that has Google and Bing" providing real-time information.[25]

Documents from indexers suggest several of them license or otherwise share copyrighted content scraped from publisher sites to AI developers, overstepping their bounds and granting rights they do not

OpenAI, Axios (Jul. 13, 2023), https://www.axios.com/2023/07/13/ap-openai-news-sharing-tech-deal; Sara Fischer, *Exclusive: The Atlantic, Vox Media ink Licensing, Product Deals with OpenAI*, Axios (May 29, 2024), https://www.axios.com/2024/05/29/atlantic-vox-media-openai-licensing-deal; *Major UK Media Organizations Join ProRata's Movement to Credit and Compensate Content Owners in the Age of Generative AI*, BusinessWire (Nov. 20, 2024), https://www.businesswire.com/news/home/20241119176784/en/Major-UK-Media-Organizations-Join-ProRata%E2%80%99s-Movement-to-Credit-and-Compensate-Content-Owners-in-the-Age-of-Generative-AI; Kyle Wiggers, *Microsoft Starts Paying Publishers for Content Surfaces by Copilot,* TechCrunch (Oct. 1, 2024), https://techcrunch.com/2024/10/01/microsoft-starts-paying-publishers-for-content-in-copilot/; Alexandra Bruell, Sam Schechner & Deepa Seetharaman, *OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over $250 Million*, The Wall Street Journal (May 22, 2024), https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba.

[20] Google Presentation: Search Entry Points (May 2023), Exhibit no. PXR0113, *United States v. Google*, 1:20-cv-03010-APM (D.D.C. 2025), available at https://www.justice.gov/atr/media/1399411/dl?inline.

[21] This also has the effect of nullifying any Perplexity bot controls adopted by publishers as Perplexity can get the content via backdoors, while also reducing the need for developers like Perplexity to develop or operate proprietary crawlers, and bear the accompanying overheads, to conduct their business.

[22] Email from Michelle Fradin (OpenAI) to Philipp Schindler (Google), OpenAI and Google Search API (Aug. 15, 2024), Exhibit no. PXR0181, *United States v. Google*, 1:20-cv-03010-APM (D.D.C. 2025), available at https://www.justice.gov/atr/media/1397566/dl.

[23] *Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date,* Meta (Apr. 18, 2024), https://ai.meta.com/blog/meta-llama-3/.

[24]Jennifer D'Souza, *A Catalog of Transformer Models*, ORKG (Mar. 2025), https://orkg.org/comparison/R1364660.

[25] Alex Heath, *Q&A: Mark Zuckerberg on Winning the AI Race*, The Verge (Apr. 19, 2024), https://www.theverge.com/2024/4/18/24134370/mark-zuckerberg-meta-interview-llama-3-ai-assistant-race.

possess:

- **Bing:** Bing states that its Web Search API – reported to be discontinued in August 2025 – results include the title, URL, and "snippet for the top ten webpage results."[26] The terms of service for Bing Search API, specifically mentioned in the PubHub Content License Agreement (see below), grants third-party LLM developers the right to "display Content received from the Bing Search APIs" and to use the web results for grounding LLMs – i.e., providing AI services that give users real-time responses based on publisher content.[27]
- **Google**: Google notes that its Custom Search JSON API results "include the URL, title and text snippets that describe the results," in addition to rich snippet information where available.[28] As noted above, Grounding with Google Search provides at least "the relevant web pages and snippets," in addition to other detailed information.[29]
- **Brave**: Brave has explicitly stated that with its Search APIs, it "allows third-party chatbots to essentially step into the shoes of a Brave search engine user and leverage Brave's search technology at the time of 'inference'" and that its API provides the chatbot with "real-time Brave search results, which include links to third-party webpages and snippets from those webpages, allowing the chatbot to provide accurate, up-to-date responses to user queries."[30] In addition, Brave also provides "alternative snippets for AI," amounting to "up to five 'snippets' per website" for use by AI developers.[31] Brave's Search APIs also provide "an enriched set of structured data about a webpage that better represents and describes the page."[32]

When publishers have tried to challenge these uses, the response is not always constructive. Following reports of Brave's scraping and resale of publisher content to third-party AI developers, News Corp. contacted and sent a cease-and-desist letter to Brave, calling on them to stop their copyright infringing practices. Instead of engaging in productive dialogue, Brave sued News Corp., asking for a declaratory judgement of no copyright infringement, misuse, or breach of contract. [33] Brave decided to voluntarily dismiss the lawsuit in June 2025.[34]

[26] *Use and Display Requirements of Bing Search APIs, with Your LLM*, MICROSOFT, https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/use-display-requirements-llm (last accessed Aug. 8, 2025).
[27] *Legal Terms for Bing Search APIs, with Your LLM*, MICROSOFT (Feb. 2025), https://www.microsoft.com/en-us/bing/apis/llm-legal.
[28] *Use REST to Invoke the API*, GOOGLE, https://developers.google.com/custom-search/v1/using_rest (last accessed Aug. 8, 2025).
[29] *Understanding Google Search Grounding*, GOOGLE, https://google.github.io/adk-docs/grounding/google_search_grounding/#detailed-description (last accessed Aug. 8, 2025).
[30] Complaint at 15, *Brave Software Inc. v. News Corp.*, 3:25-cv-02503 (N.D. Cal. Mar. 12, 2025).
[31] *Id*.
[32] *Id*. at 16.
[33] *Id*.
[34] *See Brave Software, News Corp. Voluntarily Dismiss Case*, CHAT GPT IS EATING THE WORLD (Jun. 11, 2025), https://chatgptiseatingtheworld.com/2025/06/11/brave-software-news-corp-voluntarily-dismiss-case/.

Search Engines Make Money by Facilitating Access to Their Search Indexes

These companies commercialize third-party access to publisher content, charging for the use of their APIs, with some limited exceptions for small-case users and lower transaction speeds. No monies are shared with publishers. While Google Custom Search JSON API provides 100 free search queries per day and charges $5 for every additional 1,000 queries, limited to 10,000 queries per day, Grounding with Google Search costs $35 per 1,000 requests.[35] Bing Search API, meanwhile, varies between $10 and $25 per 1,000 transactions.[36] Brave allows 1 query per second for free and goes up to $9 per month for 50 queries per second, noting that all of these offerings grant the right to "use data for AI inference" and enable access to news.[37]

Opaque Behavior Leave Publishers with Little Choice and Fail to Provide Clarity

For the owners of websites, large and small, the take it or leave it approach by search engines provide little to no visibility or clarity about the purposes for which their data is used post-scraping/indexing. The practice of wholesaling is not specifically hidden, but neither is it publicized in any manner that might offer publishers an understanding of what it entails, and whether they would like to allow their content to be included or not.

Bing. Bing's relationship with publishers who wish to be included in Bing's news search is governed by the Bing PubHub Content License Agreement, which extracts an excessive grant of rights beyond what is necessary to power Bing's search index. The Content License Agreement grants Microsoft broad licenses, including to copy, reproduce, and distribute publisher content as part of "any Microsoft offering," including via APIs, and to distribute it without limitation to third parties.[38] Licensed content consists of "all images, video and text (so long as a text snippet from each individual article does not exceed 250 words, excluding the headline) contained in the website domain(s) specified by Licensor on the Bing PubHub portal."[39] The Agreement does not outline limitations Microsoft places on downstream users of its APIs, how and when publisher content is made available through the APIs, or how publishers may terminate the Agreement and enforce deprecation of their content from historic and current versions of the API and downstream databases. Publishers are simply required to grant Microsoft extremely broad rights to reuse and repurpose their content for purposes beyond search, without control or compensation. Such terms are particularly problematic in light of Bing's API functionalities.

[35] *Custom Search JSON API*, GOOGLE, https://developers.google.com/custom-search/v1/overview (last accessed Aug. 8, 2025); *Cost of Building and Deploying AI Models in Vertex AI*, GOOGLE, https://cloud.google.com/vertex-ai/generative-ai/pricing (last accessed Aug. 8, 2025).

[36] *Bing Search API Pricing*, MICROSOFT, https://www.microsoft.com/en-us/bing/apis/pricing (last accessed Aug. 8, 2025).

[37] *Brave Search API*, BRAVE, https://brave.com/search/api/ (last accessed Aug. 8, 2025); *See* Complaint at 15, *Brave Software Inc. v. News Corp.*, 3:25-cv-02503 (N.D. Cal. Mar. 12, 2025) ("Brave earns revenue [] through its Brave Search API offering.").

[38] *Bing PubHub Content License Agreement*, MICROSOFT, https://www.bing.com/webmasters/help/pubhub-content-licence-agreement-9e7ed342 (last accessed Aug. 8, 2025).

[39] *Id.*

While Bing provides webmasters with the opportunity to opt out and control uses of their content in Bing Chat and Microsoft's own generative AI products,[40] provisions governing its Search API do not make clear whether any limitations or opt-outs are communicated to downstream users, thereby significantly undermining any such controls and creating a loophole for Microsoft to continue profiting from publisher content. Microsoft announced in May 2025 that it would be removing access to its Search API in August 2025 for most – if not all – users, encouraging them to start using "Grounding with Bing Search as part of Azure AI Agents" instead.[41] While Bing clarifies that "Grounding with Bing Search" does not provide "access to raw content returned" in response to queries and that the model response "includes citations with links to the websites used to generate the response, and a link to the Bing query used for the search,"[42] it is unclear how this change will affect publishers.

Brave. Unlike Google and Bing, Brave explicitly and intentionally ignores publishers' robots.txt instructions and resells its index to AI developers without offering any reasonable ways for publishers to prevent such uses. This is particularly egregious as many AI developers, including Anthropic which competes directly with publishers, license Brave's services. Brave, which has created its own search index and makes it available to AI developers, notes that its web crawler "does not advertise a differentiated user-agent because we must avoid discrimination from websites that allow only Google to crawl them."[43] It only respects "no crawling" and NOINDEX requests if they apply to all search engines and/or Google, noting that "otherwise Google benefits from its dominant position."[44] At the same time, all of Brave's paid API subscriptions provide users with "rights to use data for AI inference,"[45] which is particularly disconcerting considering that Brave also provides "extra alternate snippets" for users of its Data for AI product, which can be up to 260 words long – far surpassing Google's Featured Snippets' 50-word limit.[46] This can amount to an almost full-length news article, and significantly complicates and takes away publishers' control over how and by whom their content is used. As one observer noted, "[b]y claiming they want to avoid discrimination, Brave is actually saying that they get to decide what they crawl and index, not the publisher… Brave crawls the web in a stealthy way, without an obvious way to control or block it, and goes on to resell the crawled content for AI training."[47]

[40] *Announcing New Options for Webmasters to Control Usage of Their Content in Bing Chat*, MICROSOFT BING BLOGS (Sep. 22, 2023), https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat.

[41] Paresh Dave, *Microsoft Cuts Off Access to Bing Search Data as It Shifts Focus to Chatbots*, WIRED (May 14, 2025), https://www.wired.com/story/bing-microsoft-api-support-ending/.

[42] *Grounding with Bing Search*, MICROSOFT (Aug. 7, 2025), https://learn.microsoft.com/en-us/azure/ai-services/agents/how-to/tools/bing-grounding?view=azure-python-preview&tabs=python&pivots=overview.

[43] *Brave Search API*, BRAVE, https://brave.com/search/api/ (last accessed Aug. 8, 2025).

[44] steeven, *Stop Website Being Shown in Brave Search*, BRAVE COMMUNITY (Aug. 4, 2023), https://community.brave.com/t/stop-website-being-shown-in-brave-search/.

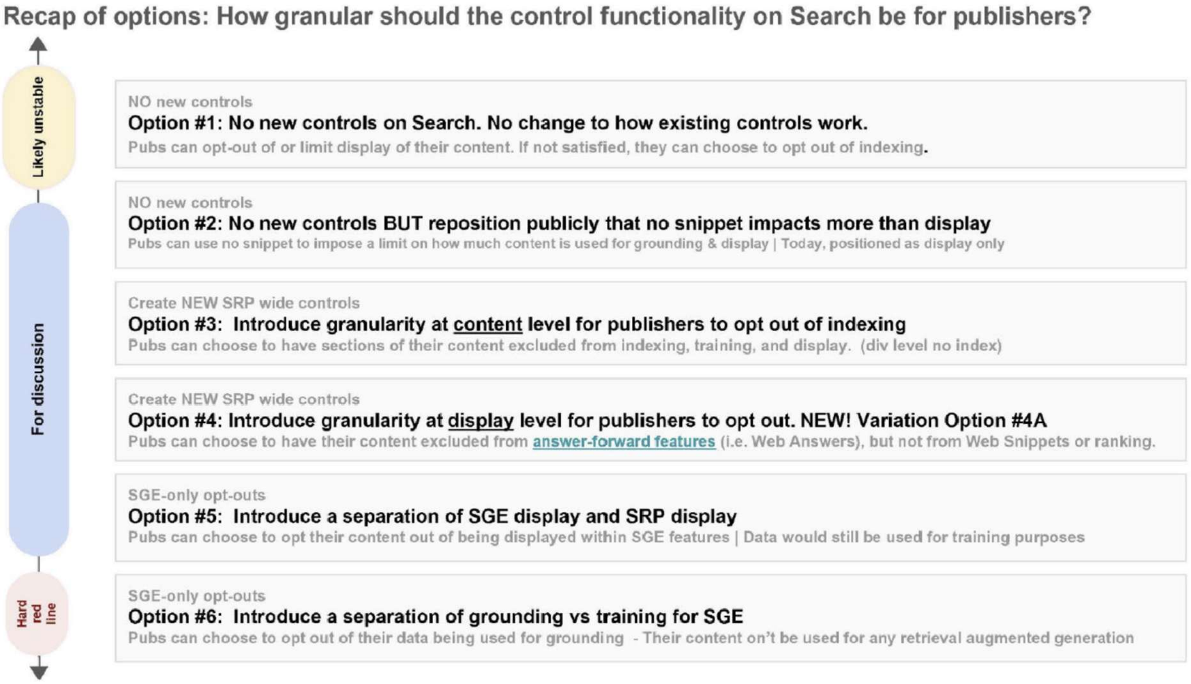
[45] *Brave Search API*, BRAVE, https://brave.com/search/api/ (last accessed Aug. 8, 2025).

[46] Alex Ivanovs, *The Shady World of Brave Selling Copyrighted Data for AI Training*, STACKDIARY (Jul. 16, 2023), https://stackdiary.com/brave-selling-copyrighted-data-for-ai-training/.

[47] Pierre Far, *Crawlers, Search Engines and the Sleaze of Generative AI Companies*, SEARCH ENGINE LAND (Jul. 13, 2023), https://searchengineland.com/crawlers-search-engines-generative-ai-companies-429389.

Google. It is unclear to what extent Google's controls limit Google's own (or their clients') use of publisher content. Similar to Bing, Google's terms of service grants it a broad license to use publisher content without authorization, including displaying, distributing, and modifying such content to develop "new technologies and services for Google."[48] Google does not specify what these uses may entail or how publishers may limit or prohibit the use of their content for unexpected use purposes. The controls that Google does offer are ambiguous and offer publishers only minimal control.

This strategy appears to be intentional, at least with regard to the tying of Google's AI products and services and Google Search. In 2024 an internal Google deck called "Search (incl SGE) Publisher Controls," discussing transparency measures and publisher control options, stated that Google should have a "hard red line" when it comes to the "separation of grounding vs training for SGE" that would allow publishers to opt-out of their data being used for grounding/RAG.[49] Google seemingly decided not to provide new controls, "BUT reposition publicly that [NOSNIPPET tag] impacts more than display," while cautioning against saying publicly that it "opts [publishers] out of training" or "grounding, as this is evolving into a space for monetization."[50]


Likely unstable
For discussion
Hard red line

Recap of options: How granular should the control functionality on Search be for publishers?

NO new controls
Option #1: No new controls on Search. No change to how existing controls work.
Pubs can opt-out of or limit display of their content. If not satisfied, they can choose to opt out of indexing.

NO new controls
Option #2: No new controls BUT reposition publicly that no snippet impacts more than display
Pubs can use no snippet to impose a limit on how much content is used for grounding & display | Today, positioned as display only

Create NEW SRP wide controls
Option #3: Introduce granularity at content level for publishers to opt out of indexing
Pubs can choose to have sections of their content excluded from indexing, training, and display. (div level no index)

Create NEW SRP wide controls
Option #4: Introduce granularity at display level for publishers to opt out. NEW! Variation Option #4A
Pubs can choose to have their content excluded from answer-forward features (i.e. Web Answers), but not from Web Snippets or ranking.

SGE-only opt-outs
Option #5: Introduce a separation of SGE display and SRP display
Pubs can choose to opt their content out of being displayed within SGE features | Data would still be used for training purposes

SGE-only opt-outs
Option #6: Introduce a separation of grounding vs training for SGE
Pubs can choose to opt out of their data being used for grounding - Their content on't be used for any retrieval augmented generation

Source: https://www.justice.gov/atr/media/1399381/dl?inline.

[48] *Terms of Service*, GOOGLE, https://policies.google.com/terms (last accessed Aug. 8, 2025).
[49] Search (incl SGE) Publisher Controls, Exhibit no. PXR0026, *United States v. Google*, 1:20-cv-03010-APM (D.D.C. 2025), available at https://www.justice.gov/atr/media/1399381/dl?inline.
[50] *Id.*

Such limited controls provide publishers with a false sense of control. Google's President of Global Affairs previously warned that by displaying only the title, URL, and thumbnails, NOSNIPPET would lead to a 45 percent reduction in clickthrough traffic, making this an unacceptable option for publishers.[51] Similarly, while publishers can opt-out of being included in Google's Gemini apps and Vertex AI through a protocol called Google-Extended, it does not offer meaningful protection or choice to publishers, and leaves questions as to its scope and how it is effectuated.[52] Based on statements during the Google Search trial, as a prime example of unexpected, unauthorized uses of search indexed publisher content, Google's search index is used to power and supplement Google's Gemini AI model to generate AI Overview summaries, even when a publisher employs Google-Extended to opt-out of Google's AI uses.[53]

Google-Extended offers insufficient level of granular control and opting-out using the Google-Extended is not a true opt-out from other Google models. Further, regardless of the efficacy of Google-Extended or other controls, Google can potentially use index sharing through an API to undermine such limitations – even if publishers can seemingly opt out of Vertex AI and Gemini App's use of their content, assumedly leaving Vertex AI clients with access to a redacted version of its search index, Google's other APIs may still provide access to full indexes for AI grounding purposes. Google's documentation falls woefully short of clarifying such uses or providing publishers with sufficient control over them.

Anthropic. Meanwhile, while Anthropic identifies the three bots it uses for AI purposes – ClaudeBot for AI training, Claude-User for information retrieval, and Claude-SearchBot for search indexing – and provides instructions on how to block each specific bot,[54] this advice is buried amongst a document otherwise advising rightsholders to either completely remove content they don't want scraped from their sites, block all scraping – including beneficial search indexing – using robots.txt, or using the "NOINDEX" tag to tell Anthropic's partners "not to index your content so that they don't send it to us in

[51] Kent Walker, *Now Is the Time to Fix the EU Copyright Directive*, THE KEYWORD BLOG (Feb. 7, 2019), https://blog.google/around-the-globe/google-europe/now-time-fix-eu-copyright-directive/.

[52] Danielle Romain, *An Update on Web Publisher Controls*, THE KEYWORD BLOG (Sep. 28, 2023), https://blog.google/technology/ai/an-update-on-web-publisher-controls/ ("Today we're announcing Google-Extended, a new control that web publishers can use to manage whether their sites help improve Bard and Vertex AI generative APIs").

[53] Davey Alba, *Google Can Train Search AI with Web Content After AI Opt-Out*, BLOOMBERG (May 3, 2025), https://www.bloomberg.com/news/articles/2025-05-03/google-can-train-search-ai-with-web-content-even-after-opt-out ("'Once you take the Gemini' AI model 'and put it inside the search org, the search org has the ability to train on the data that publishers had opted out of training, correct?' asked Diana Aguilar, a Department of Justice lawyer. 'Correct — for use in search,' Collins responded."); Khushita Vasant, *Google Executives Worried About 'Ceding a New Ecosystem' to AI Search, US Judge Hears*, MLEX (May 7, 2025), https://www.mlex.com/mlex/articles/2336512/google-executives-worried-about-ceding-a-new-ecosystem-to-ai-search-us-judge-hears ("[Google witness] agreed that AI Overviews use a customized Gemini model, which works in tandem with existing search systems. It incorporates Google search quality, Google search ranking and Google search features such as the Knowledge Graph. Google uses search signals to help train the Gemini model that is underlying the AI Overview, she said.").

[54] *Does Anthropic Crawl Data from the Web, and How Can Site Owners Block the Crawler?*, ANTHROPIC, https://support.anthropic.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler (last accessed Aug. 8, 2025).

response to your web search query."[55] It is unclear exactly as to how the latter option works, which partners it refers to, and to which use cases it is applicable. More disconcertingly, though, Anthropic was recently reported to use Brave to power its AI search products, effectively rendering its own controls for ClaudeBot and Claude-SearchBot meaningless as it can still gather the information through Brave, which doesn't respect robots.txt controls if they do not also apply to Google.[56] This is an example of what may happen in case the wholesaling of search indexes is not addressed – a race to the bottom where legitimate publisher controls and robots.txt measures are bypassed by using actors who do not respect such measures in an effort to challenge Google.

Meta. There have recently been reports that Meta is also building its own search index in an effort to reduce its reliance on third-party services, such as Google.[57] Meta's decision to create its own search index exemplifies the substantial benefits of such an endeavor to technology companies – it is a convenient and relatively easy way for technology companies to gather vast amounts of user data, bypassing AI controls and opt-outs adopted by publishers. For example, in response to a question about why Google would allow the outsourcing of its search results to competitors, Meta CEO Mark Zuckerberg noted: "I guess I wouldn't have been surprised if they didn't want to do it. But it seems like they are building up a whole model around this, so it makes sense. It's good for Google. It shows Google prominently and links to Google. They pay Apple a ton of money for distribution. They're not paying us. So, I think it's good for them on that … There's not a ton of money flowing either way."[58] There may not be a lot – if any – of monetary consideration exchanged between Meta and Google but the real advantage for both companies is the control over data and distribution channels, which hold the promise of future riches. Increased access to user data and distribution channels is particularly enticing especially when threatened with increased competition that challenges existing business models, including traditional search. Data acquired through search index monetization can support a variety of alternative revenue streams, including online advertising technology. So while there are considerable benefits to the creator of the search index, publishers lose control over their content and access to data about their readers which form the basis of their business models.

Significant Questions Remain Regarding the Amount of Data Copied and Provided to Third Parties, Use Limitations, and Clients

[55] *Reporting, Blocking, and Removing Content from Claude*, ANTHROPIC, https://support.anthropic.com/en/articles/10684638-blocking-and-removing-content-from-claude (last accessed Aug. 8, 2025).

[56] Kyle Wiggers, *Anthropic Appears to Be Using Brave to Power Web Search for Its Claude Chatbot*, TECHCRUNCH (Mar. 21, 2025), https://techcrunch.com/2025/03/21/anthropic-appears-to-be-using-brave-to-power-web-searches-for-its-claude-chatbot/.

[57] Roger Montti, *Meta Takes Step to Replace Google Index in AI Search*, SEARCH ENGINE JOURNAL (Oct. 30, 2024), https://www.searchenginejournal.com/meta-takes-step-to-remove-google-from-ai-search/531200/; Adam Clark, *Meta Is Working on an AI Search Engine. What's at Stake for Google.*, BARRON'S (Oct. 29, 2024), https://www.barrons.com/articles/meta-platforms-stock-price-google-ai-9c5c3d65.

[58] Alex Heath, *Q&A: Mark Zuckerberg on Winning the AI Race*, THE VERGE (Apr. 19, 2024), https://www.theverge.com/2024/4/18/24134370/mark-zuckerberg-meta-interview-llama-3-ai-assistant-race.

1. **To What Extent Is Indexed Content Made Available to Third Parties?** Publishers should know what content is being shared. Is access to search indexes limited to links and short extracts provided by APIs on a query-by-query basis, or does it include full copies of sites or other more detailed content to licensees?

2. **Which Uses Are Allowed Under Terms of Service with Third Parties?** Indexers must clearly disclose how they are allowing indexed content to be used in AI search, grounding, RAG, or other AI uses so that downstream uses can be monitored. Currently, search providers impose ambiguous restrictions on how the information fetched through their APIs and other offerings may be used. For example, Bing Search API's terms forbid users from using the API results for AI training purposes, but explicitly allow LLM and grounding uses.[59] Microsoft, however, reportedly informed two Bing-powered search engines that their access would be limited if they continued to use Bing's search index for AI chatbot purposes.[60] For Grounding with Google Search, Google states the customer shall "not, and will not allow its End Users or any third party to, cache, copy, frame, syndicate, resell, analyze, train on, or otherwise learn from Grounded Results or Search Suggestions," yet it is unclear whether Google actually does any due diligence or enforcement of its terms and services.[61] It is also unclear as to what restrictions apply to Google's API offerings, including Custom Search JSON API. Brave, meanwhile, remains seemingly quiet on allowable downstream uses when it comes to the training or operation of AI systems or models.[62]

3. **Who Are the Clients?** There is no reliable information available regarding the search index client base. Some may be actors, including AI developers, who do not have search engines at all. Publishers need access to client information in order to effectively control the use and users of their content and to protect their brands.

4. **How Much Money Do the Search Companies Make from Selling Access to Their Indexes?** It is unclear how much search engines profit from making their indexes available to third parties. Publishers must know how much large tech companies profit from, directly or indirectly, from search index revenue and business models to accurately establish their own revenue losses.

5. **Are Opt-Outs – If They Are Available to Publishers – Communicated to Downstream Users?** There are also significant questions regarding whether and how opt-outs are communicated to downstream users, e.g., if a publisher opts-out of Google's own AI services using Google-

[59] *Use and Display Requirements of Bing Search APIs, with Your LLM*, MICROSOFT, https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/use-display-requirements-llm (last accessed Aug. 8, 2025) (prohibiting the use of "any responses … to train, evaluate or improve any LLM, including your proprietary LLM"); *Legal Terms for Bing Search APIs, with Your LLM*, MICROSOFT, https://www.microsoft.com/en-us/bing/apis/llm-legal (last accessed Aug. 8, 2025) ("You are granted a non-exclusive, non-assignable, non-transferable, revocable license to … use Web Results only for Grounding an LLM, including your own, proprietary LLM and display respective source attribution as set forth in the Bing Search APIs, with your LLM Use and Display Requirements ("LLM Use").").

[60] Emma Roth, *Microsoft Reportedly Orders AI Chatbot Rivals to Stop Using Bing's Search Data*, THE VERGE (Mar. 25, 2023), https://www.theverge.com/2023/3/25/23656336/microsoft-chatbot-rivals-stop-using-bing-search-index.

[61] *Service Specific Terms*, GOOGLE, https://cloud.google.com/terms/service-terms (last accessed Aug. 8, 2025).

[62] *Brave Search API Terms of Use Agreement*, BRAVE, https://api-dashboard.search.brave.com/app/documentation/general/terms-of-service (last accessed Aug. 8, 2025).

Extended, does Google communicate this opt-out to AI developers who may be using its Custom Search API to ground their LLMs.

6. **What Data Is Generated Through Grounding APIs?** It is unclear as to what data is being generated through the use of grounding APIs regarding how often a publisher's content appears in the outputs generated by an LLM using that service. This is vital to allow publishers to understand the difference between the use of their IP, brand, and journalism in AI models, and any clickthrough from that article back to the originating source.

Conclusion

The wholesale of search index access by the leading search engine providers poses significant risks to publishers. It takes away publishers' control over the use and distribution of their content, diverts revenues away from publishers while benefiting the search engine providers, and disincentivizes AI developers from seeking fair licensing arrangements directly with the publishers. Legitimate business expectations are disrupted when content crawled for an authorized purpose can be repurposed and resold for other, unauthorized purposes. Combined with the lack of information regarding the exact amount of content made available, the clients, the monetization, and opt-out communications, these uses seriously hamper publishers' ability to effectively enforce their rights, control the use of their content, and to continue investments in the production of high-quality journalism.