

WUNDERLANDMEDIA

# Blocking Google's AI Crawler

The control reference: which directive does what, what it does not do, and which side of the line your business sits on.

---

By Kemal | [wunderlandmedia.com](https://wunderlandmedia.com)

**68%**

Of local business queries  
show an AI Overview

**45%**

Of US consumers use AI  
to find local businesses

**60%**

Of AI citations go to  
aggregators, not businesses

## 1 Which Side of the Line Are You On?

### One Question Decides Everything

Does your site make money **from the visit**, or **from what happens after it**? Answer that honestly before you touch a single directive. Every control in this document is downstream of that answer.

### THE DECISION TREE

**Start:** Is revenue a function of pageviews, ad impressions, subscription conversions on article pages, or affiliate clicks?

**YES** → You are a publisher-shaped business. A summary that satisfies the reader is a direct loss. Blocking is a coherent, if risky, negotiating position. **Verdict: blocking is defensible. Proceed with the Search Console setting, not sitewide nosnippet, and accept that nobody can tell you the traffic cost yet.**

**NO** → Is revenue a function of contacts, calls, bookings, quotes or purchases? Service businesses, local trades, B2B firms, agencies, SaaS, ecommerce, clinics. **Verdict: do not block. Being described accurately inside an AI answer is the goal, not the injury. Blocking is self-harm.**

**BOTH** → A blog funded by display ads that also sells a service is genuinely split. The property-level toggle is a bad fit for you. **Verdict: go selective. Use data-nosnippet at block level (Section 4).**

### PUBLISHERS AND PLUMBERS SOLVE DIFFERENT EQUATIONS

#### PUBLISHER ARITHMETIC

Sells attention. A visitor who reads a summary instead of the article is revenue that did not happen, millions of times a month. Their content is what the summary is made of, so withholding it is genuine leverage. At their scale, absence is noticeable.

#### SERVICE BUSINESS ARITHMETIC

Sells jobs. Needs eleven phone calls, not eleven thousand visits. A prospect who reads "Company X handles heat pump replacement, 4.8 stars from 200 reviews" and then calls is not a lost pageview. That is the outcome you were paying for.

### WHAT THE BLOCKING WAVE ACTUALLY BLOCKS

- ☐ **79% of major UK and US news sites block at least one AI training crawler**  
January 2026 Buzzstream analysis of roughly 100 of the biggest news sites. 71% block retrieval bots; 34% of the top 50 block every bot tested.
- ☐ **Blocking everything: BBC, NYT, WSJ, Telegraph, Daily Mail, NBC News, NY Post**  
Allowing everything: Fox News, Politico, the Independent, Substack. The wave is real but it is not uniform.
- ☐ **They are blocking OpenAI, Anthropic and Common Crawl, not Google**  
Those publishers remain wide open to Googlebot. This is the detail that gets dropped from every headline about it.
- ☐ **Blocking Google itself is still talk, as of July 2026 reporting**  
Reddit has discussed it internally. USA Today, Politico, Reuters, the Economist and People Inc are reported to be reassessing. No major publisher had pulled the plug on Google's crawlers.

### Do Not Copy a Business Model You Do Not Have

The client who prompted this ran a five-truck HVAC company: 22 pages, eleven inbound calls a month from organic. He wanted to copy the strategy of an organisation whose entire revenue model is advertising against pageviews. He is not that. Almost nobody is.

## 2 The Control Reference

### The Single Most Misreported Fact in This Topic

**Google-Extended governs Gemini model training and grounding only. It does not remove you from AI Overviews.** AI Overviews and AI Mode are surfaces inside Google Search, built on the regular Search index and crawled by regular Googlebot. Add User-agent: Google-Extended / Disallow: / and you have opted out of Gemini training. You are still fully eligible to be summarised in an AI Overview tomorrow morning. Google confirmed this in October 2023, shortly after introducing the token, and it has not changed.

| Control                                     | What it controls                                                                                                               | What it does NOT control                                                                                      | Organic ranking                                              | AI Overviews / AI Mode                                                     |
|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|----------------------------------------------------------------------------|
| <b>Googlebot</b>                            | Crawling for the Search index. Everything Google Search shows is built on this.                                                | Nothing is exempt. Block it and you leave Search entirely.                                                    | <b>Removes you</b>                                           | <b>Removes you (via removal from Search)</b>                               |
| <b>Google-Extended</b>                      | Whether crawled content may be used for training future Gemini models, and for grounding in Gemini Apps and the Vertex AI API. | <b>AI Overviews. AI Mode.</b> Search inclusion. Ranking. This is the misreported one.                         | <b>None (Google states it is not a ranking signal)</b>       | <b>No effect whatsoever</b>                                                |
| <b>Google-CloudVertexBot</b>                | Crawling requested by site owners themselves for building Vertex AI Agents.                                                    | Google Search and other products. Nobody crawls you with it unless a site owner asked.                        | <b>None</b>                                                  | <b>None. Not part of this debate.</b>                                      |
| <b>Search Console generative AI setting</b> | Property-level opt-out from AI Overviews, AI Mode and generative AI features in Discover.                                      | AI training (that is Google-Extended). The Gemini app. Page-level granularity, at least for now.              | <b>Google says none, but see the caveat on page 4</b>        | <b>Removes you, honoured within a day or two</b>                           |
| <b>nosnippet</b> (meta robots)              | Whether any text from the page can be shown as a preview, including in AI features.                                            | Indexing. You still rank, you just have nothing under the link.                                               | <b>Destroys your organic snippet on every query, forever</b> | <b>Removes your content from AI Overviews</b>                              |
| <b>max-snippet:[n]</b>                      | Caps preview text length at n characters. max-snippet:0 behaves exactly like nosnippet.                                        | Indexing. Which parts get used, only how much.                                                                | <b>The cap applies to your normal snippet too</b>            | <b>Softer version of the same trade</b>                                    |
| <b>data-nosnippet</b> (HTML attribute)      | One specific element. Google will not use that block of text in any snippet or AI feature.                                     | The rest of the page, which stays fully available. It is not a sitewide directive.                            | <b>None outside the wrapped block</b>                        | <b>Excludes only the wrapped block</b>                                     |
| <b>robots.txt</b>                           | Which user agents may crawl which paths. The place Google-Extended and third-party bots are declared.                          | What Google does with content it has already been allowed to crawl. It is a crawl control, not a use control. | <b>Depends entirely on which agent you disallow</b>          | <b>Cannot remove you from AI Overviews without also removing Googlebot</b> |
| <b>noindex</b>                              | Whether the page appears in Search at all.                                                                                     | Needs no explanation. You are gone.                                                                           | <b>Removes you</b>                                           | <b>Removes you</b>                                                         |

### Read the Table This Way

There is exactly one control that removes you from AI Overviews without damaging your ordinary search listing: the Search Console setting. Everything else either does nothing to AI Overviews (Google-Extended, CloudVertexBot) or takes your organic snippet down with it (nosnippet, max-snippet:0, noindex). data-nosnippet is the only surgical instrument in the set.

3

The Search Console Generative AI Setting

What It Is

A property-level setting in Search Console, under Settings, covering AI Overviews, AI Mode and the generative AI features in Discover. The default for every property is "include," so nothing happens unless you deliberately switch it. For most of the last three years the honest answer to "can I leave AI Overviews without leaving Search?" was no. That changed in June 2026, and it changed because a regulator forced it.

THE REGULATORY CHAIN THAT PRODUCED IT

| Date               | What happened                                                                                                                                                                               |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 30 September 2025  | The UK's Competition and Markets Authority designates Google with Strategic Market Status in general search, having found Google handles more than 90 percent of UK general search queries. |
| Designation effect | Unlocks a power under the Digital Markets, Competition and Consumers Act 2024: the CMA can impose binding conduct requirements without first proving wrongdoing in court.                   |
| 3 June 2026        | The CMA uses it, ordering Google to give publishers a genuine opt-out from AI Overviews, AI Mode and related AI features that does not damage their ordinary search performance.            |
| 17 June 2026       | Google begins honouring the choice. Content drops out within a day or two of switching the setting.                                                                                         |
| July 2026          | The setting starts appearing outside the UK. Rollout is partial and uneven.                                                                                                                 |
| By March 2027      | Google has nine months from the June order to comply fully, so finer-grained (page-level) control could arrive as late as this.                                                             |

THREE THINGS IT DOES NOT DO

- ☐ It does not cover AI training  
That is still Google-Extended, a completely separate control. Switching one does nothing to the other.
- ☐ It does not cover the Gemini app  
Search surfaces only. The standalone assistant is outside its scope.
- ☐ It is property-level, not page-level, at least for now  
All or nothing across the property. If you need to protect a few blocks of text and keep the rest, this is the wrong tool. See Section 4.

AVAILABILITY: WHERE YOU CAN AND CANNOT FIND IT

|                                                                                                                                  |                                                                                                                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>AVAILABLE</b><br>Launched UK-first in June 2026. Started appearing outside the UK from July 2026. Some US properties have it. | <b>NOT AVAILABLE (EARLY SEPTEMBER 2026)</b><br>Germany, Austria and Switzerland. Some US properties still do not have it. If you cannot find the setting, that is the rollout, not you. |
|----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

The Caveat Nobody Wants to Print

Google says the setting does not affect regular results, but AI Overviews and organic results are not cleanly separated in the live SERP. Top Stories, for instance, now appears inside the AI Overview on a meaningful share of trending news queries. What opting out costs you in practice is **genuinely unmeasured**. There is no public traffic dataset yet. Anyone telling you the number is making it up.

*"The Search Console opt-out is roughly three months old, the rollout is uneven, nobody has published credible before-and-after traffic data on using it, and the regulatory picture is still moving with Google's full compliance deadline running into 2027. If you are a publisher weighing this seriously, you are making a judgement call with incomplete information and you should say so out loud rather than pretend otherwise."*  
Kemal, Wunderlandmedia

## 4 Surgical Controls vs Blunt Instruments

### Why This Section Exists

Before the Search Console setting existed, the only lever pointed at AI features was the preview controls. Google's own AI features documentation still points site owners at exactly these to "limit the information shown from your pages in Search." They work. They also cut both ways, which is the entire problem.

### DATA-NOSNIPPET IN PRACTICE

An HTML attribute you put on a specific element, not a sitewide directive. It tells Google not to use that particular block of text in any snippet or AI feature while the rest of the page stays fully available.

```
<!-- A price table you renegotiate quarterly -->
<div data-nosnippet>
  <table class="pricing">...</table>
</div>

<!-- A process breakdown that is your actual differentiator -->
<section data-nosnippet>
  <h3>Our seven-step commissioning process</h3>
  <p>Step one...</p>
</section>

<!-- Inline: lead-time text that goes stale -->
<p>Book online today. <span data-nosnippet>Current lead time: 3 weeks.</span></p>
```

### WHAT TO WRAP AND WHAT TO LEAVE OPEN

#### LEAVE FULLY OPEN

Service descriptions. Service area. Credentials. Reviews. Contact details. These are what get you *named* in the answer, so make them as machine-readable as you can.

#### REASONABLE TO WRAP

An exact price table you renegotiate quarterly. A detailed process breakdown that is your actual differentiator. Availability or lead-time text that goes stale.

### TWO RULES I HOLD TO ON CLIENT SITES

- ☐ **The money pages stay completely unrestricted**  
Service pages, location pages, contact, and the pricing page if it is a real sales asset. No exceptions.
- ☐ **Never use a sitewide directive when a block-level one exists**  
Blanket rules are how people delete their own snippets and then spend six weeks wondering why click-through fell off a cliff.

### THE BLUNT INSTRUMENTS

- ☐ **nosnippet removes your regular organic snippet too**  
Your listing becomes a blue link with no description underneath, on every query, forever. For a business whose search listing is a sales pitch, that is a self-inflicted conversion problem far bigger than anything an AI Overview was doing to you.
- ☐ **max-snippet:0 behaves identically to nosnippet**  
There is no hidden difference. Zero characters is zero characters.
- ☐ **max-snippet:[number] is a softer version of the same trade**  
You cap the text length, and the cap applies to your normal snippet too. You cannot cap AI usage alone.
- ☐ **noindex needs no explanation**  
You are gone.

#### The Short Version

If the Search Console setting is available on your property, use that instead. Rolling out sitewide nosnippet in 2026 costs you more in ordinary search visibility than it saves in AI exposure.

5 Third-Party AI Crawlers: A Separate Argument

**Worth Saying Clearly, Because the Two Conversations Get Mashed Together**

GPTBot, ClaudeBot, PerplexityBot, CCBot, Applebot-Extended and the rest have **nothing to do with your Google rankings**. They are other companies' crawlers. Disallowing every one of them tomorrow changes nothing about how Google sees you.

THE DISTINCTION THAT IS THE WHOLE DECISION

| Crawler type       | Examples                                                     | Cost of blocking it                                                                   |
|--------------------|--------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Training crawlers  | GPTBot, ClaudeBot, CCBot, Applebot-Extended, Google-Extended | Very little visibility. This is the low-stakes half of the decision.                  |
| Retrieval crawlers | OAI-SearchBot, ChatGPT-User, Perplexity-User, PerplexityBot  | Makes you uncitable in the products where a growing share of buying research happens. |
| Googlebot          | Googlebot, Googlebot-Image, Googlebot-News                   | Leaves Search entirely. Not comparable to either of the above.                        |

**The Accident People Keep Having**

Plenty of sites have blocked retrieval bots while intending to block only training bots. The product names look similar and the robots.txt lines sit next to each other. The result is a site that is invisible inside ChatGPT and Perplexity for reasons its owner never chose.

SAMPLE ROBOTS.TXT: ALLOW RETRIEVAL, DECLINE TRAINING

```
# Google Search: always allow. This is not the AI decision.
User-agent: Googlebot
Allow: /

# Gemini model training and grounding. Does NOT affect AI Overviews.
User-agent: Google-Extended
Disallow: /

# Third-party TRAINING crawlers: low cost to decline.
User-agent: GPTBot
User-agent: ClaudeBot
User-agent: CCBot
User-agent: Applebot-Extended
Disallow: /

# Third-party RETRIEVAL crawlers: block these and you become uncitable.
User-agent: OAI-SearchBot
User-agent: ChatGPT-User
User-agent: PerplexityBot
User-agent: Perplexity-User
Allow: /
```

**DO**

Decide training and retrieval separately, per product. Check what you already have in robots.txt before adding lines, because inherited rules from a theme or a plugin are common.

**DON'T**

Paste a "block all AI bots" list from a forum post. That is how the retrieval accident happens, and it is also how people convince themselves they have opted out of AI Overviews when they have not.

**A Tidier Way to State Preferences**

An `llms.txt` file is worth a look if you want to declare your preferences more legibly, with the honest caveat that adoption is still thin.

6

The Local Business Case

45%

Of US consumers now use AI tools to find local businesses, up from 6 percent the year before. BrightLocal's 2026 Local Consumer Review Survey, a representative panel of 1,002 US adults. That makes AI the third-largest discovery channel behind Google and Facebook.

83→71%

Google's own share of local discovery over the same period. Removing yourself from the channel that grew seven-fold in a year is a strange move for a business that needs to be found.

WHITESPARK: 540 QUERIES, THREE US CITIES, SIX SERVICE INDUSTRIES

| Query type                                | AI Overview shown | Example                                                                    |
|-------------------------------------------|-------------------|----------------------------------------------------------------------------|
| Average across all local business queries | 68%               | The headline number, and the one that gets quoted without the split below. |
| Pure local intent                         | 15%               | "plumbers near me"                                                         |
| Informational                             | 92%               | "how much do plumbers charge"                                              |
| Hybrid                                    | 97%               | Queries mixing a question with a place                                     |

The Split Matters More Than the Average

If your traffic is mostly "near me" queries, AI Overviews barely touch you at 15 percent. If you win customers by answering questions, you are in the 92 to 97 percent band and the AI Overview is the SERP. Work out which of those you actually are before you decide anything.

THE AGGREGATOR ARGUMENT

60 / 40

In Whitespark's Houston plumber sample, 60 percent of AI citations went to third-party publishers like Yelp, Thumbtack, Reddit and HomeGuide. Only 40 percent went to actual local businesses.

Read That Number Again

On the queries where AI Overviews dominate, aggregators are already taking most of the citations. **If you opt out, you do not remove the AI Overview. You just hand your 40 percent to Thumbtack.** The answer still gets written. Your name is simply not in it.

WHAT TO DO INSTEAD OF BLOCKING

- ☐ **Make service and location pages easier for a machine to summarise correctly**  
This is the thing that mattered for the HVAC client, and it is what I did instead of touching a single directive.
- ☐ **Name your service area explicitly, in text**  
"Company X in Chandler handles heat pump replacement" is the sentence you want the AI Overview to write. Give it the words.
- ☐ **Put credentials and reviews where a crawler can read them**  
Star ratings and review counts are the details that get carried into a summary and turn a citation into a call.
- ☐ **Go after the informational queries the aggregators currently own**  
The 92 percent band is where Yelp and Thumbtack are winning. That is a content problem, not a robots.txt problem.

The Real Risk

The risk for a business the size of a five-truck HVAC company is not being used by AI search. It is being described badly, or not found at all while an aggregator takes the slot.

7 Bandwidth Versus Visibility

One Legitimate Reason to Block Has Nothing to Do With Content Policy

A crawler is hammering your server. That is an infrastructure decision and it needs no philosophical justification, only a log file. It is a completely different argument from opting out of AI features, and conflating the two is how people end up making a visibility decision they meant to make a hosting decision.

**50%+** Cloudflare's own data suggests more than half of AI crawl traffic is spent re-fetching pages that have not changed. I have watched small sites on modest hosting get knocked over by crawlers that pull the same pages dozens of times an hour.

TELLING THE TWO DECISIONS APART

|                      | Bandwidth blocking                                                 | AI feature opt-out                                               |
|----------------------|--------------------------------------------------------------------|------------------------------------------------------------------|
| What triggers it     | A log file showing repeated re-fetches and load spikes             | A judgement about how your business makes money                  |
| Which bots           | Whichever specific agent is misbehaving, regardless of who owns it | A category of features, across all your content                  |
| Right instrument     | Rate limiting first, then a targeted disallow or firewall rule     | Search Console setting, or block-level data-nosnippet            |
| Reversibility        | Easy. Lift it when behaviour improves.                             | Reversible, but the visibility cost while it is on is unmeasured |
| Justification needed | None beyond the logs                                               | A real business case, because you are trading away discovery     |

DOING IT PROPERLY

- ☐ **Rate-limit before you block**  
Most crawler load problems are a frequency problem, not a presence problem. Throttling keeps you citable.
- ☐ **Block the badly behaved ones specifically**  
Identify the agent in your logs and act on that one. A blanket ban punishes crawlers that were behaving.
- ☐ **Check your Cloudflare bot settings now if you are on a free plan**  
From 15 September 2026 Cloudflare blocks mixed-use crawlers by default on ad-bearing pages for new customers, new sites on existing accounts, and all free-tier accounts.
- ☐ **Understand the new bot categories**  
Cloudflare splits bots into Search, Agent and Training. A default policy applied to your account is still a policy you are living with.

You May Find a Policy Applied to You That You Never Chose

The ground is shifting at the infrastructure layer as well as the search layer. If you have not looked at your bot settings recently, look now, before you conclude that your AI visibility problem is something Google did.

*"Rate-limit them. Block the badly behaved ones. That is an infrastructure decision and it needs no philosophical justification, only a log file."*  
Kemal, Wunderlandmedia

Decision Summary: What to Set, and When

| Your situation                                       | Recommended action                     | What to set                                                                                                         |
|------------------------------------------------------|----------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Service business, local trade, clinic, B2B, agency   | Do not block anything                  | Leave Googlebot and the Search Console setting alone. Make service and location pages machine-readable.             |
| Ecommerce or SaaS                                    | Do not block. Go selective at most.    | Optional data-nosnippet on volatile pricing blocks. Nothing sitewide.                                               |
| Ad-funded publisher or large content site            | Defensible to opt out, with eyes open  | Search Console generative AI setting, if available on your property. Never sitewide nosnippet.                      |
| Comparison or review affiliate                       | Same equation as a publisher           | Search Console setting. Watch that you are not also blocking retrieval bots by accident.                            |
| Mixed: ad-funded blog that also sells a service      | The property-level toggle is a bad fit | Block-level data-nosnippet on the specific text you want withheld. Money pages stay unrestricted.                   |
| You want out of Gemini model training                | Safe, low-cost, does what it says      | User-agent: Google-Extended / Disallow: /. Understand it does nothing to AI Overviews.                              |
| You want out of AI Overviews and AI Mode             | Only one control does this cleanly     | Search Console setting. Not Google-Extended. Not robots.txt.                                                        |
| You want out of ChatGPT, Claude, Perplexity training | Separate decision, no Google effect    | Disallow GPTBot, ClaudeBot, CCBot, Applebot-Extended. Keep OAI-SearchBot, ChatGPT-User and Perplexity-User allowed. |
| A crawler is knocking your server over               | Legitimate, no justification needed    | Rate-limit first, then a targeted disallow on the specific misbehaving agent. Check your Cloudflare bot settings.   |
| You are in DE, AT or CH and cannot find the setting  | Not available yet                      | That is the rollout, not you. Do not reach for nosnippet as a substitute.                                           |

"I told him not to block Google's AI crawler, and I told him why in about four sentences: the organisations he read about sell advertising, he sells heat pump installations, and an AI answer that names his company and lists his service area does his job for him rather than instead of him."

Kemal, Wunderlandmedia

For the Twenty-Page Service Business, the Uncertainty Does Not Really Bite

The downside of staying visible is that some people read a summary instead of your homepage. The downside of blocking is that you disappear from the fastest-growing way your customers find businesses like yours, in exchange for leverage you do not have.

Read the Full Post

wunderlandmedia.com/should-you-block-google-ai-crawler

Want a second pair of eyes on your robots.txt and your snippet directives before you change anything? That is the kind of thing I do. Bring your log files. wunderlandmedia.com