VER. FIGHTING — 2026

THE AUTHORITATIVE RANKINGS FOR AI WEB CONTENT —

Artificial Intelligence Rankings

THE FILES AI READS BEFORE IT READS YOU

LLM Web Indexing Files

Researched from primary sources 13 Aug 2026 · every claim badged by evidence tier

Seven surfaces decide how AI systems discover, read, cite, and train on your site. This page breaks down each one: what it is, what it *provably* does, how to implement it, and — the part most guides skip — which benefits are ● VERIFIED and which are still ○ THEORY. Every file below ships with a copy-paste starter. AIR's own domain reports (https://airanks.net/d/airanks.net) audit all seven on any site, free.

WHAT IS IN THIS GUIDE

Ordered by measured impact. Every claim carries a badge: VERIFIED means a primary source confirms it, COMMUNITY-REPORTED means practitioners observe it, THEORY means plausible but unproven.

All seven, ranked by what they actually do

Ordered by measured impact, not by chronology. If you only have an afternoon, do row 01 and stop.

#	SURFACE	PATH	READ BY	EFFORT	VERDICT
01	your HTML	(the page itself)	every AI crawler	hours–days	Do this before anything else
02	llms.txt	/llms.txt	coding agents, IDE tools	30 min	Worth the half hour
03	robots.txt	/robots.txt	every declared crawler	15 min	Decide it deliberately, once
04	sitemap.xml	/sitemap.xml	AI search crawlers	automate it	Generate it, keep lastmod honest
05	ai.txt	/.well-known/ai.txt	nobody yet	5 min	Cheap insurance, zero effect today
06	JSON-LD	in your <head>	search engines; AI unclear	half a day	No measured citation lift
07	ai-plugin.json	/.well-known/ai-plugin.json	a deprecated program	skip	Historical. Build an MCP server instead

How to read the badges

● VERIFIED a primary source confirms it: the spec, the vendor's own documentation, or a study with methodology.

◐ COMMUNITY-REPORTED practitioners observe it; no vendor confirmation.

○ THEORY plausible, widely repeated, unproven. Not a reason to skip it — a reason to know what you're buying.

01 Render your content server-side

> **The one that outranks the other six.** Every file below is advice *about* your content. None of it matters if AI systems cannot read the content itself — and most of them cannot run your JavaScript.

PATH	ENFORCEABLE
`the page itself`	n/a — it is your own HTML

WATCH OUT FOR

`<noscript>` will not save you

EVIDENCE

> • VERIFIED GPTBot, OAI-SearchBot, ClaudeBot, PerplexityBot and Meta's crawler fetch HTML but *do not execute JavaScript* (Vercel/MERJ crawler telemetry). Only Gemini's Google infrastructure and Applebot render JS. A client-rendered SPA reads as an empty shell to most of the AI ecosystem. `Vercel & MERJ, 17 Dec 2024`

> • VERIFIED `<noscript>` does not save you. AI crawlers ignore it.

> • VERIFIED We ate this one ourselves: our own summarizer judged airanks.net "essentially no substantive content", because this site is a client-rendered app. The fix — real server-rendered content in the initial HTML — shipped the same day, and the follow-up summary is on our own domain report (https://airanks.net/d/airanks.net). The audit tool must pass its own audit.

DO THIS

1. Server-render or statically generate anything you want cited. If that is a rewrite you cannot fund this quarter, ship a server-rendered *floor*: the headline, the summary, and the substance in the initial HTML.
2. Verify the way a crawler would — fetch without JavaScript and read what actually comes back.

VERIFY LIKE A CRAWLER

```
# Read your own page the way an AI crawler does: no JavaScript, no mercy.
curl -sL -A "OAI-SearchBot" https://example.com/ | sed 's/<[^>]*>//g' | tr -s '[:space:]' ' '

# If that prints your nav and a footer and nothing else, the AI ecosystem sees
# nothing else either. Count the words you actually shipped:
curl -sL https://example.com/ | sed 's/<script[^>]*>.*<\/script>//g; s/<[^>]*>//g' | wc -w
```

02 llms.txt

> A curated markdown index of your best pages. Where robots.txt tells crawlers what they *may not* read, llms.txt tells AI readers what they *should* read.

PATH
`/llms.txt`

ENFORCEABLE
n/a — it is an invitation

WATCH OUT FOR
Google Search ignores it entirely

Proposed by Jeremy Howard (Answer.AI) in September 2024. The shape is fixed: an H1 title, a blockquote summary, then H2 sections of annotated links, with an `## Optional` section for the skippable tail.

EVIDENCE

> ● VERIFIED The format and its specification are public and stable — `llmstxt.org`, proposed by Jeremy Howard (Answer.AI) in September 2024. Major AI vendors publish one for their own API documentation.

> ● VERIFIED (NEGATIVE) Google's John Mueller, on Bluesky, **17 June 2025**: *"FWIW no AI system currently uses llms.txt."* He added that it is obvious from server logs — the consumer chatbots fetch your pages, but none of them fetch the file. `Search Engine Roundtable, 18 Jun 2025`

> ◐ COMMUNITY-REPORTED That coding agents and IDE tools now fetch it at inference time, and that adoption sits somewhere around 6–9% of major sites. Widely repeated by practitioners; we have not found a primary source for either number, so we are not going to badge it as though we had.

> ○ THEORY That consumer chat assistants (ChatGPT, Claude, Perplexity) consult it when answering. No vendor has confirmed this. Publish it for the agent ecosystem that provably does.

DO THIS

1. Hand-write it. It is a curation, not an export — a generated dump of every URL defeats the point.
2. Keep it to about a screenful. Describe each link's *payoff*, not its title. Put your API and docs first.
3. Optionally add a larger `llms-full.txt` carrying the whole story (~1% adoption, cheap to add).

/LLMS.TXT

```
# Example Corp

> Payments infrastructure for marketplaces. This file points AI readers at the
> pages worth reading; everything else is navigation.

## Docs
- [API reference](https://example.com/docs/api): every endpoint, with request
  and response examples. Start here if you are writing code.
- [Quickstart](https://example.com/docs/quickstart): first successful charge in
  about ten minutes.
- [Webhooks](https://example.com/docs/webhooks): event list, retry semantics,
  signature verification.

## Policy
- [Pricing](https://example.com/pricing): per-transaction rates by region.
- [Status](https://status.example.com): current and historical uptime.

## Optional
- [Changelog](https://example.com/changelog): dated release notes, 2019-present.
- [Engineering blog](https://example.com/blog): background reading, not reference.
```

Tooling: Mintlify auto-generates for docs sites; Yoast and AIOSEO ship WordPress plugins; validators exist at llmstxt.org. For most sites a text editor is the right tool.

03 robots.txt — the AI agent roster

Training and search are separate decisions. The major vendors run different crawlers for different purposes, and blocking the wrong one costs you visibility without protecting anything.

PATH
/robots.txt

ENFORCEABLE
no — a request, not a wall

WATCH OUT FOR
Training and search are separate gates

THE ROSTER

USER-AGENT	VENDOR	PURPOSE	BLOCKING IT COSTS YOU	HONORS ROBOTS?
GPTBot	OpenAI	Training	Nothing in search. This is the training opt-out.	• VERIFIED Yes
OAI-SearchBot	OpenAI	Search index	Your listing in ChatGPT search.	• VERIFIED Yes
ChatGPT-User	OpenAI	Live fetch	Live answers when a user asks about your page.	• VERIFIED Yes
ClaudeBot	Anthropic	Training	Nothing in search. Training opt-out only.	• VERIFIED Yes
Claude-SearchBot	Anthropic	Search index	Your listing in Claude's search results.	• VERIFIED Yes
Claude-User	Anthropic	Live fetch	Live answers about your page.	• VERIFIED Yes
Google-Extended	Google	Training gate	Gemini training only. **Googlebot is untouched** — this is a token, not a crawler.	• VERIFIED Yes
Applebot-Extended	Apple	Training gate	Apple Intelligence training. Applebot search is untouched.	• VERIFIED Yes
CCBot	Common Crawl	Open dataset	Inclusion in the open dataset many models train from.	• VERIFIED Yes
Meta-ExternalAgent	Meta	Training / AI	Meta AI training.	• VERIFIED Yes

USER-AGENT	VENDOR	PURPOSE	BLOCKING IT COSTS YOU	HONORS ROBOTS?
PerplexityBot	Perplexity	Search + live	**Your listing in Perplexity answers.**	◐ COMMUNITY-REPORTED Disputed
Bytespider	ByteDance	Training	—	◐ COMMUNITY-REPORTED No

EVIDENCE

> ● VERIFIED OpenAI's three agents have independent gates. Blocking `GPTBot` opts you out of training *without* removing you from ChatGPT search — and vice versa. Published IP ranges let you verify impostors.
> `OpenAI's crawler documentation`

> ● VERIFIED Anthropic splits the same way — `ClaudeBot`, `Claude-SearchBot`, `Claude-User` — and `Google-Extended` is a robots token rather than a crawler, so blocking it never touches Googlebot. `Anthropic·Google's crawler list`

> ◐ COMMUNITY-REPORTED ByteDance's `Bytespider` is repeatedly observed ignoring robots.txt despite vendor claims, and Cloudflare documented Perplexity fetching through undeclared headless browsers (2024). If opt-out matters to you, robots.txt alone is a request — WAF rules are the enforcement.

DO THIS

1. Decide training and search visibility *separately*. Write down which business you are in before you write a single `Disallow`.
2. Never block what you have not named. A bare `Disallow: /` under `User-agent: *` takes you out of every AI answer at once.
3. If opting out actually matters commercially, back it with WAF rules. The file is a request; the firewall is the enforcement.

/ROBOTS.TXT

```
# Training and search are SEPARATE decisions. Decide them separately.

# --- OpenAI ---
User-agent: GPTBot                # trains models
Disallow: /

User-agent: OAI-SearchBot         # the index behind ChatGPT search
Allow: /

User-agent: ChatGPT-User          # fetches your page when a user asks about it
Allow: /

# --- Anthropic ---
User-agent: ClaudeBot             # trains models
Disallow: /

User-agent: Claude-SearchBot      # search
Allow: /

User-agent: Claude-User           # live fetch
Allow: /

# --- Robots-only training gates (these are tokens, not crawlers) ---
User-agent: Google-Extended       # gates Gemini training; Googlebot is untouched
Disallow: /

User-agent: Applebot-Extended
Disallow: /

# --- Everyone else ---
User-agent: *
Allow: /

Sitemap: https://example.com/sitemap.xml
```

This example opts out of training and stays in search. That is a choice, not a recommendation — invert the `Disallow` lines if your business runs the other way.

04 sitemap.xml

> The oldest discovery mechanism still works on the newest crawlers. Generate it mechanically and keep `lastmod` honest.

PATH
/sitemap.xml

ENFORCEABLE
n/a

WATCH OUT FOR
An honest `lastmod`, or none at all

EVIDENCE

● VERIFIED AI search crawlers discover through sitemaps referenced in robots.txt. The classic limits hold: 50,000 URLs / 50MB per file, index files above that.

● VERIFIED — DATED, AND MOVING FAST Cloudflare's crawl-economics data is the eye-opener. In **July 2025** Anthropic's crawler fetched roughly **38,000 pages per referred visitor**, OpenAI's about 1,100:1, against Googlebot's ~5:1. `Cloudflare, Jul 2025`
Do not quote that figure as current. The ratios have fallen by more than an order of magnitude since, as the vendors split search crawling from training. The durable finding is the *direction* — AI crawlers still read far more than they send back — not the number. Cloudflare publishes it live on Radar. `How the ratio is measured`

○ THEORY That an honest `lastmod` speeds AI citation refresh. Plausible, unmeasured. Keep it truthful anyway — lying to crawlers is how you train them to ignore you.

DO THIS

1. Generate from your router or CMS, never by hand.
2. Reference it from robots.txt with a `Sitemap:` line. The format is unchanged since 2005 — `sitemaps.org/protocol`.
3. Set `lastmod` to the date the content changed. Stamping today's date on every URL nightly is the fastest way to make the field worthless.

/SITEMAP.XML

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>https://example.com/docs/api</loc>
    <!-- The date this page's CONTENT changed. Not today's date. -->
    <lastmod>2026-07-14</lastmod>
  </url>
  <url>
    <loc>https://example.com/pricing</loc>
    <lastmod>2026-08-02</lastmod>
  </url>
</urlset>
```

05 ai.txt

> **Nobody reads it yet.** Publish it anyway — it costs five minutes and it is the legal breadcrumb you will wish you had.

PATH
`/.well-known/ai.txt`

ENFORCEABLE
not technically — possibly legally

WATCH OUT FOR
No major vendor reads it today

Spawning's machine-readable *usage policy*, now an IETF draft (`draft-car-ai-txt-wellknown-00` , June 2026). Where robots.txt is binary access control, ai.txt expresses nuance: crawl me, cite me, don't train on me — per-agent rules, license terms, attribution requirements.

EVIDENCE

> ● VERIFIED The format and its IETF draft status. The EU AI Act requires machine-readable opt-outs be honored, which is the regulatory tailwind behind it.

> ○ THEORY *No major vendor has announced reading it.* Compliance today is approximately zero. This is cheap future-proofing, not an active control. We would rather say that plainly than sell you a file.

DO THIS

1. Serve the same short policy at `/.well-known/ai.txt` and `/ai.txt` — the draft location and the one older fetchers try.
2. State the license in one line a human can also read. Ours (https://airanks.net/ai.txt) permits training and reproduction with attribution.

/.WELL-KNOWN/AI.TXT

```
# https://example.com/.well-known/ai.txt
# Machine-readable usage policy. Also served at /ai.txt for older fetchers.

User-Agent: *
Allow: /
Disallow: /account/
Disallow: /checkout/

# Rights, in the order a lawyer would ask about them.
Training: n
Inference: y
Citation: required
License: https://example.com/terms#ai
Contact: legal@example.com

# In English, for the human who ends up reading this:
# Read us, quote us, cite us by name and link. Do not train on us.
```

06 JSON-LD structured data

> The most theory-encrusted item on this list. The best measured study found **no citation lift at all**. Ship it for search engines, not because someone told you it feeds the AI.

PATH
`in your <head>`

ENFORCEABLE
n/a

WATCH OUT FOR
JavaScript-injected markup is invisible

EVIDENCE

> **• VERIFIED (AND SOBERING)** The best measured study — Ahrefs, difference-in-differences across 1,885 pages that added JSON-LD against ~4,000 matched controls, Aug 2025 to Mar 2026 — found *no citation lift*: −4.6% in Google AI Overviews, +2.4% in AI Mode and +2.2% in ChatGPT, none of it statistically significant. No AI vendor documents JSON-LD as a citation signal. `Linehan & Guan, Ahrefs, 11 May 2026`

> **• VERIFIED** 53% of AI-cited pages carry JSON-LD — but that is correlation traveling with better technical quality generally. Google's July 2026 policy change also gates self-serving review and rating markup: publishing stars about yourself now risks eligibility rather than earning it.

> **• VERIFIED (THE TRAP)** JSON-LD injected by JavaScript is invisible to AI crawlers. Tests across ChatGPT, Claude, Perplexity and Gemini fetches found none of them extracted JS-injected markup. If your schema comes from a tag manager, AI systems have never seen it.

DO THIS

1. Emit `Organization` and `WebSite` site-wide, and `Article` or `FAQPage` where they are actually true.
2. Render it from the *server*. If a tag manager injects it, delete it and start again.
3. Skip anything Google's policy now treats as self-serving — ratings about yourself most of all.
4. Validate what you ship: `validator.schema.org`. Tooling: spatie/schema-org (PHP), schema-dts (TypeScript).

ORGANIZATION — SERVER-RENDERED

```
<!-- Server-rendered. If a tag manager injects this, AI crawlers never see it. -->
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Organization",
  "@id": "https://example.com/#org",
  "name": "Example Corp",
  "url": "https://example.com",
  "logo": "https://example.com/logo.png",
  "description": "Payments infrastructure for marketplaces.",
  "foundingDate": "2019-04-02",
  "sameAs": [
    "https://en.wikipedia.org/wiki/Example_Corp",
    "https://www.linkedin.com/company/example-corp"
  ]
}
</script>
```

07 ai-plugin.json

⚠ DEPRECATED APRIL 2024 — KEPT HERE SO YOU CAN RECOGNISE IT, NOT ADOPT IT

The OpenAI plugin manifest at `/.well-known/ai-plugin.json` pointed agents at your OpenAPI spec. The plugin program it served was replaced by GPT Actions and, increasingly, MCP (Model Context Protocol). **Its primary consumer no longer exists.**

◐ COMMUNITY-REPORTED Some 2026 agent runtimes still probe it as a fallback discovery path.

If you are building for agents in 2026, build an MCP server. That is where the ten minutes should go. The manifest below is a signpost to your OpenAPI spec, and the spec is the artifact that actually matters.

▶ SHOW THE HISTORICAL MANIFEST

Audit any site against all seven

The AIR domain report checks every surface on this page for any domain — presence, robots.txt AI-agent verdicts, server-rendered JSON-LD types, and live file contents — alongside the domain's AIR score and the AI's own summary of the site. Free, no account.

Want to see one first? Read our own report (https://airanks.net/d/airanks.net) — including the section where we failed our own server-rendering test. The build log (https://airanks.net/blog) documents how we fixed our files, before and after.

BEFORE YOU CLOSE THIS: THE SEVEN CHECKS

01	☐	**your HTML**	Do this before anything else
02	☐	**llms.txt**	Worth the half hour
03	☐	**robots.txt**	Decide it deliberately, once
04	☐	**sitemap.xml**	Generate it, keep lastmod honest
05	☐	**ai.txt**	Cheap insurance, zero effect today
06	☐	**JSON-LD**	No measured citation lift
07	☐	**ai-plugin.json**	Historical. Build an MCP server instead

Run all seven automatically on any domain at **airanks.net/d/<your-domain>** — free, no account. Audited by _____ on ____ / ____ / _____