MISHA MANKO.

FILE № 047 · INDEPENDENT RESEARCH · FOR YOUTUBE VIEWERS

WHAT AI BOTS ACTUALLY *READ.*

Seven findings from my research network of **47 live websites** — what ChatGPT, Claude, Perplexity, and Google AI consume in 2026, and what they ignore.

RESEARCH SOURCE

47 Live Sites · 9 Industries

DOCUMENT TYPE

Research Findings · Free

BY

Misha Manko.

§ INTRODUCTION

00. WHY THIS REPORT EXISTS

Most of what gets sold as "AEO" — Answer Engine Optimization — in 2026 is theory. Agencies are charging thousands of dollars for tactics nobody has actually measured on a real site with real bot logs. I decided to find out what's real.

THE SETUP

I instrumented a network of **47 live websites** across nine industries — law, medical, finance, real estate, e-commerce, SaaS, addiction treatment, education, and local services. I deployed every content format the AEO world tells clients to build: enriched RSS feeds, inline schema markup, standalone JSON-LD endpoints, post markdown exports, llms.txt files, and several others.

Then I piped every bot request from every site into a single dashboard and watched what actually happened. Not what should happen according to the whitepapers. What *actually* happened in the access logs.

WHAT'S IN THIS REPORT

Seven findings that will change how you think about AI visibility — or at least how you spend money on it. Each finding is what the data showed, what it means for your site, and what to actually do about it. No fluff, no padding, no "book a call to learn more."

§ A NOTE ON THE NUMBERS

Throughout this report I use percentages rather than raw traffic counts. This is intentional. I'm showing you **the shape of what's happening**, not the specific volume of one research window. The percentages are stable across different time periods in my data — the raw counts vary based on when you look. What matters is which formats dominate, not how many hits any one format got.

WHO THIS IS FOR

Marketing leads, SEO practitioners, agency owners, and in-house SEO teams who are being asked by executives "what's our strategy for ChatGPT / Claude / Perplexity?" and don't have a data-backed answer. This report gives you one — and gives you the language to push back when an agency tries to sell you on tactics that don't work.

If after reading this you want to know where *your* specific site stands, I run audits that apply this same research methodology to one site at a time. Details on the last page.

§ FINDING 01 · THE HEADLINE NUMBER

01. LLMS.TXT HAS NEVER BEEN REQUESTED

OF AI BOT REQUESTS HIT LLMS.TXT

0%

Across all **47 sites**, **zero AI bot requests** have landed on any llms.txt file — ever. Not a single one. GPTBot, ClaudeBot, PerplexityBot, none of them fetch it. The tactic being sold for thousands of dollars right now produces zero measurable traffic in my research.

WHAT THIS MEANS

llms.txt is a proposed standard for a site to describe itself to AI crawlers in a single markdown file at the root of the domain. The idea is reasonable. The adoption, in practice, is non-existent on the crawler side.

Every major AI company has a public stance that is either ambiguous or explicitly skeptical about llms.txt. None of them have committed to consuming it as a ranking or discovery signal. Meanwhile, agencies are charging clients to build, structure, and maintain llms.txt files as a core AEO deliverable. This is the biggest disconnect between industry theory and bot behavior that I have seen.

WHAT YOU SHOULD DO ABOUT IT

1. If your agency is charging you for llms.txt — ask them to show you the bot logs that prove it's being consumed.
2. If you have an llms.txt already, don't delete it. The cost to keep it is zero. But don't pay anyone to expand it.
3. Redirect that budget to the formats AI bots *actually* consume — which is the rest of this report.

§ FINDING 02 · THE WINNER

02. RSS IS THE MOST-CONSUMED ENDPOINT

OF AI BOT REQUESTS TARGET RSS

~40%

When I add up all AI bot activity across the research network, **around 40% of every AI fetcher request** goes to an RSS feed. More than HTML pages. More than sitemaps. Dramatically more than any schema endpoint or markdown file.

WHAT THIS MEANS

RSS — the ancient, pre-social-media content-syndication format — is the single most-consumed content format by real-time AI fetchers. ChatGPT, Claude, and Perplexity are all using RSS as a primary discovery and freshness mechanism. If your RSS feed is stripped to titles and excerpts (which is the WordPress default), you're invisible in the most-consumed format.

This is the format that matters, and almost nobody is talking about it publicly. Generalist SEO agencies treat RSS as a legacy afterthought. AI bots treat it as a primary source.

WHAT MAKES A GOOD RSS FEED FOR AI

- → **Full content** in every item — not excerpts
- → **Author attribution** with a real person, not a generic site byline
- → **Publish AND modified dates** on every item
- → **Category and tag entities** preserved as structured data
- → **Inline schema references** (article, author, organization)

WHAT YOU SHOULD DO ABOUT IT

1. Pull your current RSS feed URL in a browser. Look at the raw XML. If items are cut off after 2-3 sentences, you have a problem.
2. Configure your CMS or plugin to output full content in RSS items — not excerpts. In WordPress this is a single setting.
3. Add author bylines, modified dates, and category entities. These are the signals AI fetchers parse for context.

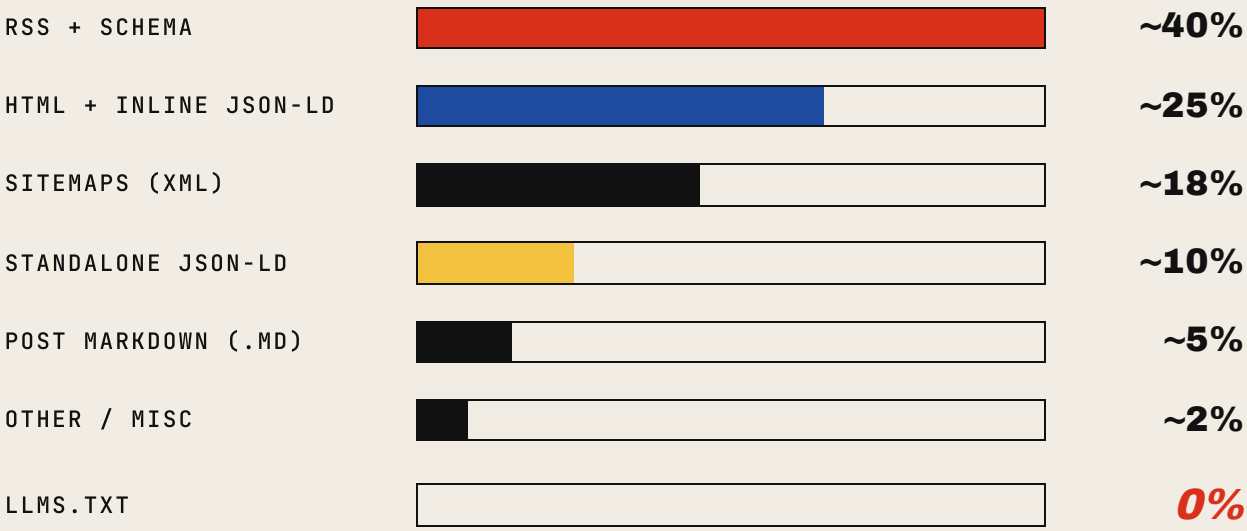
§ FINDING 03 · THE SHAPE OF IT

03. THE CONSUMPTION HIERARCHY

When I rank content formats by total AI bot requests received across the 47-site network, a clear hierarchy emerges. Most of the market is optimizing for formats near the bottom of this list.

AI BOT CONSUMPTION SHARE · BY CONTENT FORMAT

SOURCE: 47-SITE RESEARCH NETWORK



HOW TO READ THIS

The top two formats — RSS with inline schema, and HTML pages with inline JSON-LD — together account for roughly **two-thirds of all AI bot consumption**. If your budget isn't going toward those two formats, it's probably going to the wrong place.

Standalone JSON-LD endpoints (a common agency recommendation) pull a small share. Post markdown files — another popular AEO tactic — pull even less. And llms.txt pulls nothing.

§ THE PRACTICAL TAKEAWAY

Spend your AEO budget on the **top two formats** and you're covering two-thirds of what AI actually consumes. Everything else is a rounding error. This single chart should change how most marketing teams allocate their AEO spend.

§ FINDING 04 · BOTS BEHAVE DIFFERENTLY

04. CHATGPT VS CLAUDE IS NOT THE SAME GAME

Different AI platforms consume different formats at different rates. Treating "AI visibility" as one monolithic problem is the second-biggest mistake I see, after llms.txt.

CHATGPT (GPTBOT + CHATGPT-USER + OAI-SEARCHBOT)

The most aggressive fetcher in the network. Consumes RSS heavily, follows sitemaps, returns to high-authority pages repeatedly. In my data, ChatGPT's real-time fetcher (ChatGPT-User) makes up a substantial portion of total AI traffic on most sites — the model is actively retrieving content for live user queries.

CLAUDE (CLAUDEBOT + CLAUDE-USER)

Claude's training crawler (ClaudeBot) is steady but less voluminous than GPTBot. Claude-User — the real-time fetcher — is growing quickly in volume on sites that publish consistently. Claude seems to prefer HTML with strong semantic structure (proper heading hierarchy, answer-first paragraphs) more than ChatGPT does.

PERPLEXITY (PERPLEXITYBOT)

Heavy RSS consumer. Appears to use RSS as the primary freshness signal. Perplexity returns to feeds more frequently than either ChatGPT or Claude. If your feed isn't working, you're dark in Perplexity.

WHAT YOU SHOULD DO ABOUT IT

1. Stop optimizing for "AI" as if it's one thing. Look at which bot is visiting your site most — that's the platform where you have traction and should double down.
2. If you're not showing up in Perplexity, check your RSS feed first. That's almost always the issue.
3. If you're not showing up in Claude, check your HTML structure. Heading hierarchy, answer-first paragraphs, semantic markup.

§ FINDING 05 · THE RETURN VISITOR

05. FRESH CONTENT COMPOUNDS. STALE SITES DISAPPEAR.

HIGHER AI BOT RETURN RATE

~3x

Sites in the research network that publish **at least weekly** receive roughly **three times** the AI bot return-visit frequency of sites that publish less than monthly. Publishing cadence is a ranking signal even if nobody is saying it out loud.

WHAT THIS MEANS

AI fetchers allocate crawl budget just like Google does. Sites that publish fresh content on a regular schedule get re-crawled more frequently — which means more opportunities to be cited in live AI responses. Sites that go quiet drop out of the return rotation within 60 to 90 days in my data.

This has implications for the "set it and forget it" approach most businesses take to their blogs. If you optimize once and never publish again, your AI visibility will decay within a single quarter. This is also why the AI visibility problem is fundamentally a recurring problem, not a one-time fix.

WHAT COUNTS AS "PUBLISHING"

The format matters less than the signal of freshness. What works in my data:

- → A new blog post or article, published with a proper date and schema
- → A meaningful update to an existing page with a refreshed modified date
- → A new service or product page with its own canonical URL

What doesn't count: minor edits to existing pages without surfacing a new URL, date-only changes without content changes, pure navigation or styling updates.

WHAT YOU SHOULD DO ABOUT IT

1. Publish at least one meaningful piece of new content per week. Quality matters, but so does cadence.
2. If your blog has gone quiet for more than 60 days, expect a drop in AI citations and plan accordingly.

3 Make sure every published piece surfaces in your RSS feed with full content, schema, and a proper date.

§ FINDING 06 · THE SILENT KILLER

06. MOST SITES ARE BLOCKING AI BOTS BY ACCIDENT

OF SITES HAD ACCIDENTAL BLOCKS

~30%

Around **30% of the sites** I've audited outside the research network had **accidental AI bot blocks** at the CDN, firewall, or robots.txt level — usually deployed by a well-meaning developer trying to "block scrapers." The site owner had no idea.

WHERE THE BLOCKS COME FROM

Most accidental blocks trace back to one of four sources:

- → **Cloudflare "Bot Fight Mode"** — blocks legitimate AI fetchers alongside scrapers unless configured properly
- → **WAF rules from the hosting provider** — blanket rate limits that trigger against high-volume but legitimate AI crawlers
- → **robots.txt rules copied from the internet** — many widely-shared robots.txt templates block GPTBot, ClaudeBot, or both without the site owner knowing
- → **Security plugin defaults** — some WordPress security plugins block AI bots by default because they're categorized as "unknown crawlers"

HOW TO CHECK

Three places to inspect, in order:

- → 01. **Your robots.txt.** Visit *yoursite.com/robots.txt* and look for *User-agent: GPTBot*, *User-agent: ClaudeBot*, *User-agent: PerplexityBot*, or *User-agent: CCBot*. If any of these are followed by *Disallow: /*, you're blocking them.
- → 02. **Your CDN / WAF dashboard.** In Cloudflare, check the Firewall Rules and Bot Fight Mode settings. In AWS, check WAF rules. Look for any rule that categorizes AI user-agents as "bots to challenge."
- → 03. **Your server access logs.** If you're seeing no requests from GPTBot, ClaudeBot, or PerplexityBot at all, something is blocking them upstream. A site with content worth citing *will* get some AI bot traffic — zero is a diagnostic.

§ A DELIBERATE DECISION YOU SHOULD MAKE

Some site owners **want** to block training crawlers (GPTBot, ClaudeBot as training) while allowing real-time fetchers (ChatGPT-User, Claude-User). That's a legitimate stance. The issue isn't blocking — it's blocking *by accident*, and then wondering why you're invisible in AI.

§ FINDING 07 · THE OVERLAP

07. AI VISIBILITY IS 70% TECHNICAL, 30% CONTENT

OF FIX IMPACT IS TECHNICAL

~70%

When I score which changes produced the largest measurable AI bot traffic increases across audit implementations, **roughly 70% of the impact** came from **technical fixes** — schema, RSS, HTML structure, bot access. The remaining 30% came from content volume and quality.

WHY THIS MATTERS

The AEO industry sells you content — blog posts, thought leadership pieces, AI-optimized articles. Content matters. But most sites I audit have **serious technical issues** that are limiting how much content leverage they can get in the first place. Fixing a broken RSS feed or a misconfigured schema will produce more citation gain than publishing ten more articles into a broken pipeline.

The typical industry order of operations is backwards. Fix the technical infrastructure first. Publish more content second. Not the other way around.

THE RIGHT ORDER OF OPERATIONS

01. **Fix bot access.** Make sure AI fetchers can reach your site at all. This is Finding 06.

02. **Fix your RSS feed.** Full content, dates, author, schema, categories. This is Finding 02.

03. **Fix your schema coverage.** Every page type should have the right JSON-LD. Validated.

04. **Fix your HTML structure.** Heading hierarchy, answer-first paragraphs, semantic markup.

05. **Now publish more content.** At this point your infrastructure can actually leverage it.

Most sites I audit are trying to run step 5 while steps 1 through 4 are broken. That's why their AEO efforts feel expensive and produce nothing measurable.

§§

§ WHAT TO DO WITH THIS

TWO PATHS FORWARD

You now have the seven findings that changed how I think about AI visibility. The question is what you do with them.

PATH A · RUN THE CHECKS YOURSELF

Everything in this report is something you can investigate on your own. Pull up your robots.txt. Open your RSS feed in a browser and look at the raw XML. Check your Cloudflare dashboard for bot fight mode. Pull your server logs and see which AI bots are actually visiting. If you have a competent in-house team, they can work through these findings methodically and close the gaps.

PATH B · HAVE ME RUN THEM FOR YOU

If you want someone who has run this analysis on dozens of sites already, I offer a full AI Visibility Audit. It applies this same research methodology to *your* specific site:

- → 10-dimension AI Visibility Score for your site
- → Weeks of your server or CDN bot log data analyzed
- → 30+ target queries tested across ChatGPT, Claude, Perplexity, and Google AI
- → Competitive benchmark against your top 3 competitors
- → Prioritized fix list with Impact, Effort, and Urgency ratings
- → 18-page PDF + companion spreadsheet + 30-minute results call

§ READY FOR YOUR AUDIT?

STOP GUESSING *WHAT AI SEES.*

In 7 to 10 business days, you'll know exactly where your site stands, what's broken, and what to fix first. No retainer bait, no upsell pressure, no "book a call to get a quote." One price, one deliverable.

BOOK THE AUDIT · $2,500

QUESTIONS FIRST?

If you want to talk through whether the audit is a fit before booking, reach out directly. Same turnaround — replies within 1 business day.

EMAIL	WEB	YOUTUBE
hello@mishamanko.com	**mishamanko.com**	**@mishamanko**

---

§ This report reflects findings from my ongoing research network of 47 live websites across nine industries. I do not guarantee specific AI citation outcomes — AI model behavior changes frequently. What I do guarantee is that the methodology in this report is rigorous, the findings are real, and the recommendations will move your site in the right direction if implemented correctly.