The Technical Signals AI Search Uses That Most SEOs Still Aren’t Optimizing

For the last couple of years, the topic of AI visibility has dominated SEO discussions. SEO teams have cleaned code and chunked content, all the easier for AI bots to access and ingest. If your brand has started earning regular citations in AI-generated answers or showing up in AI Overviews, you might be forgiven for thinking the battle for AI visibility is nearly won.

None of that is wrong, and none of it is wasted effort. It’s just not the whole story.

However, our audit of 50 major websites found that, while most have made it easier for AI to *find* them, almost none have made it possible for AI to truly *understand* them. And nearly two-thirds leave the question of which AI bots can access which content entirely to luck.

3 Layers Of AI Visibility

Ask a marketing team to define “AI visibility,” and the answer will likely focus on mentions and citations. Should someone ask any of the leading AI platforms a question relevant to your category, you want your brand and/or product to have a better than average chance of appearing in the response.

Defining AI visibility purely in terms of citations and mentions is old SEO thinking – get ranked, get found, get clicked – carried over to a new and very different form of search. But visibility in AI isn’t simply about getting your

brand, product and relevant links in front of the right eyeballs. It’s also about how AI understands your information and interacts with your website.

If you’ve parented a child through primary school, you already know that teaching them to read isn’t the same as teaching them to understand. A six-year-old learning phonics might read the words in their early reading book reasonably fluently. Tick, get yourself a cookie from the jar. But ask them about what they’ve just read, and you may find they absorbed or understood less than you thought.

Once a child has mastered both reading and comprehension, they can put what they learn into action – follow a recipe to cook a meal, write their own stories or essays, carry out further research to find answers to questions arising from the initial text, and so on.

Reading is only one step towards comprehension. And comprehension is only one step towards agency; what you decide to do with that information once you have it.

This is why we developed an audit framework for AI readiness with three distinct layers:

1. Retrievability

Can AI fetch and parse your content without stumbling? This is the reading bit, the foundational layer most SEO teams already optimize for, directly related to AI visibility in the sense of citations and brand mentions.

But unless you also optimize for the second layer, you’re basically crossing your fingers and hoping AI interprets your content correctly.

2. Attribution And Meaning

Can AI determine what your pages are about and who owns them?

This is how AI knows, without simply guessing, which number on the page is the product price, and which is the discount price available exclusively to loyalty card holders. It's how AI determines who the brand and/or author really is, and whether to consider them a trusted authority on the subject. It's how AI tells whether an article is about Jaguar the car brand, or one of the many Jaguar football clubs.

Giving AI greater confidence that it has correctly interpreted your content does two things: It increases the likelihood of your content being cited in relevant AI responses, while also reducing the risk of those responses misrepresenting your brand, your products or your claims.

3. Agent Transaction And Discovery

Can AI agents access your website's capabilities to carry out tasks, such as completing transactions? This final layer is about agency, and it's the difference between AI merely parroting information back to the user and being able to interact with your business on their behalf.

Under each of these layers, we grouped the likely protocols SEO teams might implement to achieve those outcomes. For example, implementing ARIA labeling or including AI user-agent directives in the robots.txt would largely relate to Retrievability, while the presence of an entity map or JSON-LD schema would relate to Attribution and Meaning.

Altogether, our framework includes 27 audit elements – 11 in Layer One, three in Layer Two, and 13 in Layer Three – rated and scored according to their current maturity as industry standards:

- **Established:** Production-ready standards in active use today.
- **Emerging:** Real protocols gaining traction with early movers, and growing fast.

- **Frontier:** Standards still being debated with no settled implementation so far.

Armed with this framework, we audited 50 major websites across retail, SaaS, travel, publishing and finance to see where the most common gaps in AI readiness might be.

I'll admit it's a tad unfair of me to apply the same framework and scoring system across the entire cohort, regardless of industry and business model. When using our AI readiness framework with clients, we weight the scoring according to their industry. But using the framework as a benchmarking tool across a varied cohort wouldn't work if we measured different things or applied different scoring to each website.

The purpose was to highlight where the biggest gaps are overall, rather than guessing at the nuances of every individual business case or commercial decision. A low score doesn't necessarily mean a site is underprepared for AI if other clues suggest they've deliberately adopted this approach.

See also: Machine-First Architecture: How To Build Websites Machines Can Identify, Read, Cite & Use

What The Data Shows

Using an instrumented browser, we were able to capture live HTTP responses, the rendered DOM, raw server HTML, and machine-discovery endpoints. All data was captured on the same day: June 12, 2026.

We then scored 12 Established signals against our agentic readiness framework (0/1/2), with the final score rendered as a percentage of the possible maximum. While we also tracked Emerging and Frontier protocols, we excluded these from the scoring.

- The highest score went to Airbnb.com with 79.2%.

- On average, sites have implemented just over half of the established protocols (mean 56.6%, median 58.3%).
- Only 10 sites scored below 50%.

While that might sound like a pass mark, it isn't. When you pull the three layers apart, it becomes clear where there's still work to be done.

1. Retrievability (Average 74.4%)

- **Established Audit Elements (8/11)**

1. txt & AI User-Agent Directives.
2. Accessibility Tree Integrity.
3. ARIA Labeling & Descriptive Names.
4. Semantic HTML & Document Hierarchy.
5. Token-Efficient DOM Density.
6. Server-Rendered / Clean HTML Delivery.
7. Form & Input Machine Usability.
8. Sitemap Declaration.

This is the layer that overlaps most heavily with conventional technical SEO, including elements that may have already been in place, making any new tweaks to optimize for AI retrievability easier to implement. Unsurprisingly, most sites perform reasonably well, with only three scoring below 50%. And there could be legitimate reasons for those low scores, as you'll see.

2. Attribution And Meaning (Average 38.5%)

- **Established Audit Elements (2/3)**

1. JSON-LD Schema & Semantic Richness.
2. Content Signals Policy.

Scores drop away sharply between the first and second layers.

The good news is we detected JSON-LD structured data on the homepage of 35 out of 50 websites (70%), with all but three scoring the 2-point maximum. The bad news is that still leaves nearly a third without any at all.

Schema is a vital part of how AI understands what your content actually means: This is a product, this is the price, and this is the brand that sells it. Poor or non-existent schema probably won't undermine your website's AI retrievability, and the LLM will still make its best guess at what everything means. But that best guess can also be wrong, leading to inaccurate AI responses or even outright hallucinations.

Perhaps more telling is that only five of the 50 audited sites have implemented Cloudflare's Content Signals Policy. This is a set of directives within the robots.txt file that spell out exactly what crawlers are permitted to do with your content in relation to search indexing, live AI query responses, and model training. A Content Signals Policy means your AI strategy isn't limited to a simplistic binary choice – "block all AI" or "allow all AI" – allowing you to take a more strategic approach.

3. Agent Transaction And Discovery (Average 2.1%)

- **Established Audit Elements (2/13)**

1. OAuth Discovery (Authorization Server).
2. OAuth Protected Resource Metadata.

Agentic browsers and agentic commerce have only been around since late 2025. As a result, of the 13 protocols we identified as directly related to Layer 3, we could only class two as Established.

Both ebay.com and wikipedia.org block same-origin fetch via CSP. But of the 48 sites where endpoint testing was possible,

46 scored zero. And while airbnb.com and vercel.com have both implemented OAuth authorization server metadata, neither has implemented OAuth protected resource metadata, meaning they only scored 50% each for Layer 3.

OAuth discovery and OAuth protected metadata make it possible for AI to interact with your site. For example, the OAuth discovery metadata tells a client application what it's allowed to access, how to identify itself, and where to get the information it needs to securely access an OAuth authorization server or protected API.

However, our team is closely monitoring nine other Emerging elements in Layer 3 that could see the space evolve quickly.

Some relate to the Model Context Protocol (MCP), which allows AI to query your servers directly. For example, instead of piecing together information from your website product pages, AI gets what it needs straight from your real-time product database, drastically reducing processing time and the risk of AI hallucination.

And then there are the relatively new ecommerce protocols that allow AI to complete transactions within AI conversations, like Google's Universal Commerce Protocol (UCP) and OpenAI's Agentic Commerce Protocol (ACP), which are likely to become established in the near future.

Layer 3 is very much the area to watch, with the potential to confer first-mover advantage on any brands willing to experiment.

A Low Score Isn't Always A Bad Score

I'm not suggesting every business should embrace AI in the same way, following our framework of established protocols like a checklist to be completed. While it's technically

possible for a website to score 100%, that doesn’t mean it should.

Some publishers within the cohort – including the BBC, CNN and The Guardian – specifically block most, if not all, AI bots. This is understandable. Their business models depend on people visiting their sites to get the latest news and read their particular brand of reporting.

Meanwhile, industries such as retail, SaaS, travel and financial services are increasingly dependent on reaching customers and influencing decisions in these AI spaces. And as more customers start interacting with these businesses via AI agents instead of through their websites, all three layers will become increasingly important.

Then again, one of the lowest overall scores (29.2%) belongs to amazon.com, which was also one of the three sites to score below 50% in Layer 1. Now, I don’t think for one minute that Amazon has simply overlooked AI search. Like the BBC, Amazon has specific AI crawler directives in place to block almost all AI bots from accessing the site, so this is clearly by design. And depending on their reasons for blocking the bots, there might not be much point in implementing some of the other measures.

Conversely, some sites, like airbnb.com and cloudflare.com, have no formal blocks in place but have implemented specific access rules for most of the major AI bots. Others, including eBay.com and tripadvisor.com, have been even more selective, implementing rules for the AI bots they do want and blocking those they don’t.

Each of the above sites has made deliberate, bot-specific decisions and updated their robots.txt with the appropriate AI directives. They’ve begun to implement a plan for AI even if, in some cases, that plan is to block it entirely – in which case an AI readiness score is kind of moot.

But nearly two-thirds of the audited sites (29 of 50) appear to have made no deliberate decision about AI agent access one way or the other. Nothing is blocked, and nothing is explicitly allowed either, with no rules for crawlers to follow.

What AI bots can access, how they interpret what they find, and how they use that information in AI responses is largely left to chance.

Signals Are Not Guarantees

At this point, we should also mention llms.txt, which goes beyond the AI directives and content signals within robots.txt to provide AI systems with a curated, human-readable guide to a site's content and structure.

Our framework classifies llms.txt as a Frontier element, excluded from scoring, because it is currently an unratified standard with no agreed specification body behind it. However, we found 11 of the 50 sites have published one, suggesting there's some enthusiasm for the idea of providing AI with a handy "CliffsNotes" guide to a brand and its content.

But before you instruct your team to update your site's robots.txt or add an llms.txt, it's worth clarifying what these files can – and can't – do.

Any AI directives documented in your robots.txt are preference statements, rather than hard-coded rules. Compliance is voluntary. The major AI crawlers – GPTBot, ClaudeBot, Google-Extended and Applebot-Extended – broadly respect them, but beyond that group, things get patchier. Some bots may crawl your content regardless.

Content signals carry even less enforcement weight. Well-behaved bots can choose to follow them, but there's no technical mechanism to force compliance. And, for now at

least, llms.txt is still little more than a signal of intent.

For example, Expedia has published an llms.txt file that reads like a love letter to AI, setting out the brand's canonical identity and explaining in neatly chunked copy what the brand is about, its capabilities, and so on.

Yet when it came to the rest of the site's AI architecture, Expedia.com scored just 33.3% overall – one of the lowest in the entire cohort. No JSON-LD structured data, no sitemap declaration, and only a quarter of the content is delivered server-side. Publishing such a well-written llms.txt file while leaving so much else unaddressed is a little like putting a sign in your window saying "open for business" but forgetting to unlock the door.

This doesn't mean such measures are wasted, just that results are far from guaranteed. But when has *anything* been guaranteed to work in SEO? It's why optimization has always been about strengthening as many signals as possible.

Updating your robots.txt with detailed instructions or creating a comprehensive llms.txt aren't replacements for doing the rest of the work to improve your site's AI architecture.

Block, Allow, Or Do Nothing: The Choice Is Yours

Mentions and citations were never the whole story; they're only the first layer of true AI visibility. Your brand could be cited regularly and still lose customers because of inaccuracies or hallucinations, or because the AI didn't have enough confidence in your information to recommend your products in the most relevant conversations.

Airbnb, eBay, Amazon and others aren't chasing mentions (in the latter's case, definitely not). Each one has made

deliberate, specific choices about which AI systems can train on their content, and which can act on a customer's behalf in the moment.

The decisions that will define your brand's AI presence won't all be made by your marketing team. Some of them may have already been made for you, hidden within default settings, legacy robots.txt files, and security policies written for a different era.

Whether you've planned and optimized for it or not, AI is almost certainly reading your content. But this is still your content. You can still exert some control over how AI accesses, interprets, and interacts with your brand.

These are all solvable problems – but only if you decide to solve them.

More Resources:

Featured Image: SvetaZi/Shutterstock

https://www.searchenginejournal.com/the-technical-signals-ai-search-uses-that-most-seos-still-arent-optimizing/586381/