

SEO, AEO, GEO and AI Search Readiness Benchmark — India 2026

What 1,948 Wikidata-listed Indian business websites tell an AI engine when it comes to cite them

SEOIndia original research · census collected 3 September 2026 · methodology, instrument and dataset published with this report

The answer first

Almost no Wikidata-listed Indian business website in our census blocks AI search crawlers: 4 of 1,948 reachable home pages disallow OpenAI's search bot at the root. But being allowed in is not the same as being citable. Of the 1,944 home pages that permit AI search, 817 serve no JSON-LD structured data at all, 1,061 expose no Organization entity an engine could attribute an answer to, and 1,455 show neither an author nor a date on the home page. Only 445 of 1,944 (22.9%) clear all four preconditions we measured for being retrieved, parsed and attributed. The gap in India's AI-search readiness is not access. It is legibility.

1. Why this study exists

Most claims about "AI search readiness" are opinions dressed as numbers: a vendor's client base, a handful of screenshots, a survey of people who already care. We wanted a measurement that anyone could repeat without trusting us. So we built a census with a public frame, a fixed instrument and one rule: every headline carries its own denominator, and the wording never says more than the measurement can.

The question is practical. When a generative engine such as ChatGPT, Perplexity, Gemini or Google's AI Overviews assembles an answer to a commercial question, it needs three things from a website before it can cite it: permission to fetch, a page it can retrieve, and a page it can parse into an entity with a name, an author and a date it can trust. A business can be excellent at classic SEO and still fail the third step, because nothing in a ranking report tells you whether your home page describes itself in a form a machine can read.

This benchmark measures those three steps, in that order, across a population defined by someone other than us.

2. How we measured

Population. Every Indian business with an official website recorded in Wikidata, collected 3 September 2026: 4,116 Wikidata rows, collapsed to 2,596 distinct hosts after removing duplicates across entities. We call this "Wikidata-listed Indian business websites in the measured census" throughout, and we do not generalise it to "Indian businesses": the Wikidata frame over-represents companies large or notable enough to have an entry.

Instrument. One request to each host's home page plus its robots.txt, with the user agent `SEOIndia-AISearchReadinessStudy/1.0 (+https://seoindia.co/; research; one page per host)` and a 1.5 second delay between hosts. Redirects were followed and the final host recorded. Hosts whose robots.txt disallowed all user agents were not fetched, by design, and are counted as unreachable in the run. Server-delivered HTML only: nothing injected by JavaScript is seen, which makes every "has no schema" figure an upper bound on the true share.

Reachability. 1,948 of 2,596 hosts (75%) answered; 648 (25%) did not. The unreachable group is not one thing: 181 names did not resolve, 127 answered HTTP 403 to our agent, 79 refused the connection, 77 failed TLS trust, 46 answered a permanent redirect our client does not follow, 38 timed out, 21 disallowed all agents in robots.txt, 11 answered 404, and the rest were assorted connection failures. Every "permitting" and "parseable" share below is computed on reachable hosts only, and says so.

Reproducibility. We re-probed a seeded sample of 60 hosts the next day. 59 of 60 were reachable again; the one failure was a timeout. Eight of twelve measured fields agreed on all 60 hosts, and every disagreement traced to that single transient host. The frame file, the run CSV, the sample list and the verification script ship with this report so a reader can reproduce any row.

What is measured	Numerator / denominator	Collection	Exclusions	Limitation
Reachable home pages	1,948 / 2,596 hosts in the frame	2026-09-03 16:26–21:47 UTC	none	one attempt per host; some slow hosts are counted as unreachable

3. Findings

3.1 AI crawler posture: search access is open, training is where sites draw the line

Among the 1,948 reachable home pages, 4 disallow OpenAI's search crawler (OAI-SearchBot) at the root of robots.txt. That is 0.2%. Blocking of Perplexity's crawler is similarly rare: 7 hosts. Training crawlers are a different matter: 102 hosts block GPTBot (5.2%), 115 block CCBot

(5.9%), and 121 (6.2%) block at least one of the four agents we checked. 98 hosts block a training agent while permitting search agents; 4 block both; none block search only.

The pattern is coherent. Where an Indian business has taken a position on AI crawlers, it has chosen to keep its content out of training corpora while leaving the door open to being found and cited. That is a defensible commercial choice. What is striking is how few have taken any position: 1,846 of 1,948 (94.8%) block none of the four agents, and 272 have no robots.txt at all.

Measure	Numerator / denominator	Value	Limitation
Blocks OAI-SearchBot	4 / 1,948 reachable	0.2%	root-level Disallow only; path-level rules not counted
Blocks GPTBot · CCBot · any of four	102 · 115 · 121 / 1,948	5.2% · 5.9% · 6.2%	under-count for the same reason
No robots.txt	272 / 1,948	14.0%	absence of a file, not a policy

Because 99.8% of reachable hosts permit AI search, the rest of this report is computed on the 1,944 permitting hosts. The four that block are excluded from every downstream share so that "permitted but unreadable" is measured on sites that actually asked to be read.

3.2 Retrieval eligibility: the basics are mostly in place

1,941 of the 1,948 reachable home pages are served over HTTPS. 23 carry a noindex directive on the home page, which in most cases is a misconfiguration rather than a decision. 1,288 declare a sitemap in robots.txt. 701 carry a self-referencing canonical; 141 resolve to a different host than the one in the frame, usually a www or brand-domain redirect that was recorded but not penalised.

Retrieval, in other words, is rarely the barrier. A generative engine that wants to fetch an Indian business home page can almost always do so. The barrier comes after the fetch.

Measure	Numerator / denominator	Value	Limitation
HTTPS	1,941 / 1,948	99.6%	final URL after redirects
Home page noindex	23 / 1,948	1.2%	meta robots on the home page only
Sitemap declared	1,288 / 1,948	66.1%	robots.txt declaration, not sitemap validity
Self-referencing canonical	701 / 1,948	36.0%	exact-match test; equivalent variants counted as non-self

3.3 Parse surface: what a machine can read once it has the page

1,131 of 1,948 reachable home pages carry at least one JSON-LD block. 886 declare an Organization entity. 343 carry a BreadcrumbList, 242 an FAQPage, and 662 publish an llms.txt file, the emerging convention for telling language models what a site is. Only 208 home pages expose an author surface and 415 a date surface.

The most common JSON-LD types tell you what the ecosystem's plugins emit by default rather than what businesses have decided to say: Organization (843), WebSite (815), SearchAction (606), PostalAddress (547), ContactPoint (521), ImageObject (501), WebPage (449). FAQPage (242) and Person (299) are rarer because they require someone to write something.

Two numbers deserve a second look. 245 home pages carry JSON-LD without an Organization node, which usually means a theme emitted WebSite and WebPage and nobody added who the site belongs to. And 77 hosts publish an llms.txt while serving no JSON-LD at all: they adopted the newest convention before the oldest one.

Measure	Numerator / denominator	Value	Limitation
Any JSON-LD	1,131 / 1,948	58.1%	server HTML only; JS-injected schema missed
Organization entity	886 / 1,948	45.5%	JSON-LD only; Microdata/RDFa not counted
Author surface · date surface	208 · 415 / 1,948	10.7% · 21.3%	presence on the home page, not correctness
llms.txt present	662 / 1,948	34.0%	file exists; content not evaluated

3.4 The cross-tab: permitted but unparseable

This is the headline that matters, computed on the 1,944 hosts that permit AI search.

817 of them (42.0%) serve no JSON-LD structured data. 1,061 (54.6%) expose no Organization entity. 1,455 (74.8%) show neither an author nor a date on the home page; 74 show an author only, 282 a date only, 133 both. 438 declare an Organization but not a self-referencing canonical, so the entity is there but the engine has to guess which URL it belongs to.

Put the four preconditions together (JSON-LD present, an Organization entity, a self-referencing canonical, and no noindex directive) and 445 of the 1,944 permitting home pages (22.9%) are what we call fully parseable. That composite is not a quality score, and a site can be fully parseable and still say nothing useful. It is the floor: below it, an engine that wants to cite the business has to reconstruct who it is from prose.

The commercial consequence is direct. When two suppliers are equally relevant to a question and one describes itself as an entity with a name, a location and a canonical address while the

other is a WebPage with a logo, the citation goes to the one the engine can name. Being open to AI search and illegible to it is the worst of both positions: you pay the crawl cost and forgo the credit.

Measure	Numerator / denominator	Value	Limitation
No JSON-LD	817 / 1,944 permitting	42.0%	upper bound (JS-injected schema missed)
No Organization entity	1,061 / 1,944	54.6%	upper bound (Microdata/RDFa missed)
Neither author nor date	1,455 / 1,944	74.8%	home pages often carry no byline legitimately; this is a surface count, not a content judgement
Fully parseable (4 preconditions)	445 / 1,944	22.9%	composite; not a quality score

3.5 Answerability and media: the page itself

931 of 1,948 home pages show a visible FAQ block; 242 of those back it with FAQPage schema and 689 do not. Heading discipline is uneven: 593 home pages have no H1, 985 exactly one, 370 two or more. Title tags: 64 are empty, 424 are under 30 characters, 970 sit in the 30–65 range, 490 run longer. Meta descriptions: 524 are missing, 163 are shorter than 70 characters, 676 sit in the 70–160 range, 585 run longer. The median home page carries 900 words, 9 lists and no tables.

Media and performance follow the same shape. The median home page carries 30 images; 753 home pages have at least one image with no alt text. Median time to first byte was 754 ms from our vantage point; 682 hosts answered within 500 ms, 529 within a second, 737 took longer than a second and 152 longer than three. The median page weighed 152 KB of HTML.

None of these are AI-specific. They are the classic SEO hygiene items, and they matter to generative engines for the same reason they matter to Google: a page with one H1, a readable title, a real description and labelled images is easier to summarise correctly. A home page with no H1 (593 of them) leaves the engine to infer the subject from the logo.

Measure	Numerator / denominator	Value	Limitation
Visible FAQ · with FAQPage schema	931 · 242 / 1,948	47.8% · 12.4%	visible-block detection by heading pattern
No H1	593 / 1,948	30.4%	server HTML only
Missing meta description	524 / 1,948	26.9%	tag absent, not tag weak
Any image missing alt	753 / 1,948	38.7%	at least one image; not a ratio of images
TTFB over 1 s	737 / 1,948	37.8%	one request from one vantage point; not a Core Web Vitals field value

3.6 By industry: direction only

981 of the 1,948 reachable hosts carry an industry label in Wikidata; 967 do not. Among the seven labelled sectors with at least 20 hosts, software (41 hosts) and artificial intelligence (31) lead on Organization schema at 78% and 77%, followed by e-commerce (71%, 28 hosts) and information technology (70%, 27). Automotive sits at 56% (27 hosts), retail at 50% (24), and financial services trails at 38% (21 hosts). On JSON-LD of any kind, software (90%) and e-commerce (89%) lead; financial services (62%) and retail (58%) trail.

These are observed shares in small groups, not rankings. A sector with 21 hosts moves nine points when two sites change. What the direction says is plausible and worth testing at scale: the sectors that sell software describe themselves to machines; the sectors that sell trust in person have not started.

Measure	Denominator	Value	Limitation
Organization schema, software · AI · financial services	41 · 31 · 21 hosts	78% · 77% · 38%	small n; direction only; 967 hosts have no sector label

4. What a site owner should do first: five zero-cost fixes, in order

- 1. Declare who you are.** Add an Organization node in JSON-LD on the home page with name, url, logo, sameAs to your official profiles, and a ContactPoint. This is the single change that moves a site from the 1,061 to the 883. Most CMS platforms can emit it from settings already filled in.
- 2. Point the entity at one URL.** Set a self-referencing canonical on the home page and make the Organization's url match it. 438 sites have the entity but not the address.
- 3. Give the home page a byline and a date where it is honest to do so.** A "last reviewed" date and a named responsible person or team are surfaces engines look for

when they judge freshness and accountability. Do not invent a journalist; name the business owner or the team.

4. **Fix the basics the engine reads first.** One H1. A title between 30 and 65 characters that names the business and what it does. A meta description that a human would accept as a summary. Alt text on every image that carries meaning.
5. **Decide your crawler policy on purpose.** If you want to be cited but not trained on, say so in robots.txt: allow OAI-SearchBot and PerplexityBot, disallow GPTBot and CCBot. 98 sites in the census have already made exactly that choice. If you have no robots.txt at all, you have made no choice, and the default is that everything is permitted.

None of these require a budget. All of them require someone to own the answer to "what does our home page say about us to a machine", which in most Indian businesses is nobody's job yet.

5. Limitations, in the body where they belong

- **The frame is Wikidata.** It over-represents notable companies and under-represents the small businesses that make up most of the economy. Every share here is a share of this census, not of Indian business.
- **One page per host.** We measured the home page. A company whose product pages carry rich schema while its home page carries none is scored as having none. This is deliberate: the home page is where an engine starts, and it is the page a business controls most completely.
- **Server-delivered HTML only.** Schema injected by tag managers or client-side frameworks is invisible to this instrument, as it is to some crawlers. Every "has no schema" share is therefore an upper bound.
- **Root-level robots rules only.** A site that blocks an agent from a subdirectory is counted as permitting. The blocking shares are under-counts.
- **One request, one vantage point, one day.** Time to first byte is a single sample, not a field metric. Reachability includes some hosts that were merely slow on 3 September.
- **Sector labels are sparse.** 967 of 1,948 reachable hosts have no industry label in Wikidata, so the sector cut describes half the census and should be read as direction only.

6. Reproduce it

The frame (`sample-frame-wikidata-20260903.txt` , 2,596 hosts), the census output (`run-20260903/aisr-india-2026-20260903-214748.csv`), the 60-host reproducibility sample and its repro (`repro-20260904/`), the instrument and the verification script that recomputes every number in this report from the raw rows (`publication-20260904/verify.js` , output `VERIFY-20260904.md`) are published with this report. Run the verifier and every figure above should reappear; if one does not, the report is wrong and we want to know.

7. What this means for an Indian business in 2026

The visibility contest has moved. Ranking on a results page still matters, but a growing share of commercial questions are now answered inside an assistant, with two or three sources cited and everyone else invisible. Those citations go to pages an engine can fetch, parse and attribute. In this census, three in four of the Indian business websites that permit AI search have not done the third part.

That is bad news and good news at once. Bad, because most Indian businesses are open to being cited and are not being cited. Good, because the fixes are cheap, the field is uncrowded, and the businesses that describe themselves properly this year will be the ones the engines have learned to name by the time the rest catch up.

If you want to know where your own site sits against these 1,944 home pages, SEOIndia runs the same instrument against a single domain as a free readiness check, and our [AEO services](#) and [GEO services](#) exist to close the gap the check finds. The umbrella offer, [AI Search SEO](#), covers both.

Frequently asked questions

Does this study show that Indian businesses block AI? No. It shows the opposite for this census: 4 of 1,948 reachable Wikidata-listed Indian business home pages block OpenAI's search crawler at the root. Blocking is concentrated on training crawlers, and even there it is under 7%.

Why is "no structured data" called an upper bound? Because the instrument reads server-delivered HTML only. A site that injects JSON-LD with JavaScript is counted as having none. The true share without schema can only be lower than 42.0% of permitting hosts, never higher.

Is 22.9% "fully parseable" a quality score? No. It is the share of permitting home pages that clear four preconditions: JSON-LD present, an Organization entity, a self-referencing

canonical, and no noindex directive. A fully parseable page can still be thin. An unparseable page cannot be attributed no matter how good it is.

Can I apply these numbers to my sector? Only as direction. Sector groups in the census are small (21 to 41 hosts), and half the reachable hosts carry no sector label in Wikidata. Use the sector cut to decide what to test on your own site, not to rank industries.

How do I check my own site the same way? Fetch your home page and robots.txt as a crawler would, then ask four questions: is an AI search agent allowed, is there JSON-LD, does it contain an Organization with a url that matches your canonical, and does the page show who is responsible and when it was last reviewed. The published instrument does exactly this for one host, and SEOIndia will run it for you on request.