Get Found by ChatGPT

By Joe Giler

Table of Contents

Chapter 1: What Actually Changed: How AI Assistants Pick Who to Mention

In 2023, if you wanted customers to find you, you optimized for a search results page. You fought for a blue link. You wrote a title tag that would earn a click. The whole game was: get on page one, get the click, land them on your site, convert them there.

That game still exists. It is not dead. Anyone who tells you "SEO is dead" is selling you something — usually a course about the thing that supposedly replaced it. But a second game now runs alongside it, and most small businesses are not playing it at all.

The second game is this: when somebody opens ChatGPT, Claude, Gemini, Perplexity, or Copilot and types *"who's the best commercial roofer in Tulsa"* or *"what's a good project management tool for a two-person design studio"*, the assistant answers in prose. It names three to six options. Sometimes it links them. Sometimes it doesn't. The user reads that answer and, very often, acts on it — they pick from the list, or they search the one name that sounded most credible.

If your business is not in that list of names, you did not lose a ranking. You were never in the room.

That is the change. Not "AI is coming for SEO." The change is that a meaningful slice of buying decisions now gets pre-filtered by a model that decided, in about four seconds, which handful of businesses deserved to be mentioned by name — and you have almost no direct control over that decision. You have indirect control. That's what this book is about.

The honest state of play in 2026

Let me give you the truth first, because the rest of the book is worthless if you start with the wrong picture.

AI assistants are a real traffic and referral channel, but they are not yet the dominant one for most businesses. If you run analytics on a typical small business site right now, you'll likely see referral traffic from `chatgpt.com`, `perplexity.ai`, `claude.ai`, `copilot.microsoft.com`, and `gemini.google.com` somewhere in the low single-digit percentages of total sessions — call it 1% to 6% for most sites, higher for B2B software and professional services, lower for local trades. That's an estimate based on the shape of the channel, not a study I'm citing. Check your own numbers; I'll show you how in a minute. Anyone quoting you a precise industry-wide percentage is guessing too, they're just not admitting it.

But the traffic quality is unusually good. This is the part that matters more than volume. A person who arrives from an AI assistant has usually already had a conversation about their problem, gotten a recommendation, and clicked through with intent to evaluate you specifically. They're not tire-kicking off a broad keyword. Many operators report noticeably higher conversion rates from AI referrals than from generic organic search. I believe that's directionally true and it matches what I see, but treat it as a pattern, not a law — measure your own.

And a lot of the value never shows up as a click at all. This is the frustrating part. Someone asks ChatGPT for a recommendation, ChatGPT names you, and the person types your business name into Google and arrives via "organic branded search." Your analytics credits Google. The assistant did the work. This is called the zero-click or dark-referral problem and it is real. It means your AI visibility is almost certainly larger than what you can measure, and it means you should not judge this channel purely on referral session counts.

So: real, growing, high-quality, hard to measure. That's the honest frame. Now let's take apart the machinery.

Three different machines, and why the difference matters

People say "the AI" as if it's one thing. It isn't. When an assistant names businesses, one of three very different processes is happening, and your strategy differs for each.

1. Pure parametric recall — the model answers from memory. The model was trained on an enormous pile of text scraped from the open web, books, code, forums, and licensed data, with a cutoff date. When you ask it "what are the major email marketing platforms," it doesn't look anything up. It generates the answer from statistical patterns baked into its weights during training. Mailchimp, Klaviyo, ConvertKit, Brevo come out because those names appeared thousands of times in that training data, in contexts that associate them with email marketing.

You cannot edit this. There is no form to fill in. The only lever is: *be written about, by other people, in public, a lot, before the next training run.* That's slow, it's indirect, and it's why Chapter 2 of this book is entirely about earning third-party mentions. For a business with no existing web footprint, parametric recall is essentially unwinnable in the short term. Accept that and focus on the other two machines.

2. Retrieval-augmented answers — the model searches, then reads, then answers. This is what happens when you ask ChatGPT with search enabled, Perplexity, Google's AI Overviews and AI Mode, Copilot, or Claude with web search. The system silently rewrites your question into one or more search queries, fires them at a search index (Bing powers a lot of this; Google powers its own; Perplexity blends multiple sources including its own crawler), pulls back a set of pages, reads chunks of them, and writes an answer grounded in what it just read — usually with citations.

This is the winnable game. It is winnable this quarter, not next year. If a page of yours is retrievable, readable, and clearly answers the question, it can be cited in an answer within days of publishing. There's no training-run lag. Most of the practical tactics in this book target this machine.

3. Tool and connector calls — the model queries a structured source. Increasingly, assistants don't just read web pages; they call APIs and structured data sources. Shopping queries hit product feeds. Local queries can hit map and business-listing data. Some assistants call partner APIs directly. When an assistant tells you a product's price and stock status, it usually got that from a feed, not from reading your HTML.

This machine rewards boring, unglamorous data hygiene: accurate Google Business Profile, clean product feeds, correct structured data, consistent name-address-phone across directories. Not exciting. Extremely effective for local and e-commerce.

Here's the practical takeaway: **ask yourself which machine is answering your customer's question.** "Best CRM for nonprofits" is mostly machine 1 and 2. "Plumber near me open now" is mostly machine 3. "Does the Sony WH-1000XM6 support multipoint" is machine 2. Your effort should follow the machine that serves your buyers.

What actually happens between the question and the answer

Let me walk the retrieval pipeline step by step, because every step is a place where you can win or lose, and most people only think about the last one.

Step one: query fan-out. The user types something conversational — "I need a bookkeeper who understands Shopify sellers, ideally in Texas." The system does not search for that string. It decomposes the intent into several tighter queries: *bookkeeper for Shopify sellers*, *ecommerce accountant Texas*, *Shopify bookkeeping services*, maybe *best accountants for online sellers 2026*. This is called query fan-out, and Google has described AI Mode as doing exactly this kind of decomposition.

Implication: the phrase you need to rank for is not the long conversational question. It's the two-to-five-word professional phrasing the system will actually search. Optimize for the sub-queries, not the sentence.

Step two: candidate retrieval. Each sub-query returns a set of results from an index. If your page isn't indexed in whatever index that assistant uses, you're out — permanently, for that answer. Note the plural: being in Google's index does not mean you're in Bing's. ChatGPT's search has historically leaned heavily on Bing. If you have never once checked whether Bing has your site indexed, go do it after this chapter. I've seen sites with perfect Google presence and thirty pages missing from Bing entirely.

Step three: fetch and chunk. The system pulls the candidate pages and breaks them into passages — a few hundred words each, roughly. It is not evaluating your page. It is evaluating *passages* from your page. A brilliant 4,000-word guide where the actual answer lives in paragraph 31, buried mid-sentence, may lose to a mediocre page that states the answer cleanly under a matching heading.

This single fact should change how you write. Write in retrievable units: a heading that states the question, then a paragraph that answers it completely, without requiring the reader to have read the previous 2,000 words. Self-contained passages win.

Step four: relevance and selection. The system ranks passages by semantic match to the sub-query, then selects a working set that fits in its context budget. It's looking for passages that directly address the question, that appear factual and specific, and — importantly — that come from sources it treats as credible. Credibility signals here overlap heavily with classic SEO authority signals, because in most cases the underlying retrieval is a search index that already computed them.

Step five: synthesis and hedging. The model writes the answer. Here's a behavior worth understanding: models are trained to hedge on claims they can't support and to prefer specifics they can support. If your page says "we are the leading provider of X," the model has nothing to work with — every competitor says that too, and it's an unverifiable brag. If your page says "we've completed 340 commercial roof replacements in the Tulsa metro since 2011 and we're certified for Carlisle and Firestone systems," the model has concrete, attributable material. Guess which one gets quoted.

Step six: citation. Not every source read gets cited. Systems cite selectively, often preferring sources that contributed a distinctive fact. Another argument for having distinctive facts.

The uncomfortable truth about who gets mentioned

Now let me say the hard part out loud, because you'll waste money if I don't.

In a large share of "best X" and "top X for Y" answers, the sources the model reads are not vendor websites. They are third-party listicles, review platforms, Reddit threads, YouTube reviews, trade publications, and comparison sites. The model reads *G2, Capterra, TrustRadius, Reddit, a trade magazine's roundup, and two independent blogs* — and then names the companies those sources name.

Sit with that. In those queries, **your website is not the input. Other people's pages about you are the input.** You can have the most beautifully optimized site on earth and be invisible, because the question was answered from sources you don't own.

This is why "AI SEO" services that only touch your own site are half a strategy at best. It's also why Chapter 2 exists and why it's the longest chapter in Part 1.

The flip side: for informational and comparison and how-to queries, your own content absolutely can be the cited source. And for branded queries — "what does Same Day Creative Solutions do," "is Acme Roofing legit," "how much does Acme charge" — your own pages plus third-party mentions are nearly the entire input. Different query types, different levers.

The four query types and what each one demands

Sort every query your buyers might ask into one of four buckets. Your plan falls out of the sort.

Type A — Branded. "Is [your company] any good?" "What does [your company] charge?" "Is [your company] legit?" These fire when someone has already heard of you — from an ad, a referral, or another AI answer. They are the highest-stakes queries you have, and almost nobody prepares for them.

What wins: an unambiguous About page, clear pricing information or at least an honest pricing-model explanation, real reviews on third-party platforms, and a presence in enough places that the model doesn't have to say "I don't have information about this company." That last outcome — the model shrugging — is a lost sale, and it happens constantly to businesses with thin footprints.

Type B — Category / "best of". "Best X for Y." Dominated by third-party sources. Highest competition, hardest to influence, biggest prize.

What wins: being present in the roundups, review platforms, and communities the models read. Plus, occasionally, having your own genuinely useful comparison content that's honest enough to be citable.

Type C — Problem / informational. "How do I stop condensation in a metal building?" "What's the difference between a 1099-NEC and a 1099-MISC?" "Why is my Shopify checkout abandoning at 70%?" These are where a small business can win fastest, because expertise beats budget and the field is less crowded.

What wins: specific, structured, genuinely expert content that answers the exact question in a self-contained passage. This is the single best use of your writing time in the first 90 days.

Type D — Transactional / local. "Emergency electrician open now near Round Rock." "Buy a 12x16 shed delivered in Ohio." Heavily influenced by structured data — maps, feeds, listings.

What wins: Google Business Profile accuracy, consistent NAP data, service-area pages with real detail, product feeds that don't lie about stock.

Most businesses should attack C first (fast wins, full control), D second if applicable (boring, mechanical, effective), A third (cheap insurance), and B as a long campaign (slow, compounding, the real prize).

Why models pick some sources over others

You can't read the ranking function. Nobody outside these companies can, and the people inside can't fully explain it either. But you can observe consistent behaviors, and these hold up across assistants:

- **Specificity beats fluency.** A passage with numbers, model names, dates, and constraints gets used more than a smooth paragraph of generalities. "Most policies exclude flood damage" is weaker than "standard HO-3 homeowners policies in the U.S. exclude flood damage; flood coverage requires a separate NFIP or private flood policy." The second is quotable. The first is filler. (That's a general example — check your own policy and talk to a licensed insurance agent.)
- **Structure beats prose.** Clear H2/H3 headings that state questions, short answer paragraphs, tables, and lists all chunk cleanly. Wall-of-text pages chunk badly.
- **Freshness matters, unevenly.** For pricing, versions, regulations, and "best of" queries, recent dates matter a lot. For evergreen how-to content, less so. A visible, accurate last-updated date helps. Faking one is both dishonest and eventually detectable — don't.
- **Consensus across sources beats one loud claim.** If five independent sources say a tool is good for solo consultants, the model says it. If only your site says it, the model usually doesn't. This is the single most important mechanic in the whole book.
- **Accessibility is binary.** If your content requires JavaScript to render, sits behind a login, is inside a PDF that's slow to fetch, or is blocked in robots.txt, it may as well not exist. More on this shortly.
- **Answering the actual question beats matching the keyword.** Retrieval is semantic. Stuffing "best CRM for nonprofits" fourteen times does nothing useful and may actively hurt by making the passage read as spam.

The crawler question — who's allowed to read you

This trips up more businesses than anything else in this book, and it's a five-minute fix. There are multiple distinct bots, and they do different jobs.

Roughly, as of 2026:

- **GPTBot** (OpenAI) — general crawling, historically associated with training data collection.
- **OAI-SearchBot** (OpenAI) — builds the index used for ChatGPT search results.
- **ChatGPT-User** (OpenAI) — fetches a page live when a user or the assistant follows a link during a conversation.
- **ClaudeBot** and related Anthropic agents — crawling and live user-initiated fetches.
- **PerplexityBot** and **Perplexity-User** — index building and live fetch.
- **Google-Extended** — a control token that governs use of your content for certain Google AI training purposes; it does *not* control whether you appear in Google Search or AI Overviews, which are governed by Googlebot.
- **Bingbot** — still enormously important, because several assistants lean on Bing's index.
- **Applebot** and **Applebot-Extended** — Apple's crawler and its AI-training control.

These names and behaviors change. Verify current bot names in each provider's published documentation before you edit anything — don't take a list from a book (including this one) as gospel a year from now.

The decision you have to make consciously: blocking training crawlers is a legitimate choice. Some publishers block them on principle or as leverage in licensing negotiations. That's a real business position and I won't argue you out of it.

But understand the trade. If you block the *search* and *live-fetch* bots — OAI-SearchBot, ChatGPT-User, PerplexityBot, Bingbot — you are opting out of being cited in AI answers. You've chosen invisibility. Many sites did this accidentally, by copying an aggressive "block all AI bots" robots.txt snippet off a forum post in 2024 and never revisiting it. If you're wondering why you get zero AI referrals, check this first.

Go look at `yoursite.com/robots.txt` right now. Actually go look. Then check your CDN and firewall — Cloudflare and similar services have one-click "block AI bots" toggles, and plenty of business owners have one enabled without knowing it. The robots.txt can be clean while the WAF quietly 403s every crawler.

A reasonable posture for most businesses that want AI visibility: allow the search and user-fetch bots, and make your own call on the training bots. If you want maximum long-term parametric presence, allow those too.

llms.txt — what it is and what it isn't

You'll see this recommended everywhere, usually with more confidence than the evidence supports. Here's the straight version.

What it is: a proposed convention — a markdown file at `/llms.txt` that gives an LLM a curated map of your site: what you do, and links to your most important pages in a clean, easily-parsed format. Think of it as a sitemap written for a reader rather than a crawler.

What it isn't: a ranking factor, a guaranteed intake channel, or a substitute for having crawlable content. As of 2026, major AI providers have not broadly committed to consuming llms.txt as a primary source, and there's no solid public evidence that having one lifts your citation rate on its own.

My recommendation: make one anyway, because it takes twenty minutes and it costs nothing. Keep it accurate. Do not spend a week on it, do not pay someone $500 to generate one, and do not skip real work to build it. It's a cheap lottery ticket, not a strategy. If you're a developer-tools company, the odds are somewhat better — that ecosystem has adopted it more.

A workable minimal version:

- An H1 with your business name.
- A one-paragraph blockquote saying exactly what you do, for whom, and where.
- Sections with links to: your core service/product pages, your pricing page, your best explanatory content, your about/credentials page, and your contact page — each link followed by a one-line description of what's on it.

Structured data: still worth it, for a different reason than you think

Schema.org markup — the JSON-LD blocks that describe your organization, products, articles, FAQs, and reviews in machine-readable form — is not read directly by a language model in the way people imagine. The model reads text.

But structured data feeds the *retrieval systems* the model queries, and it removes ambiguity for anything that does parse your page. It's how a system knows with certainty that "$149" is a price and not a phone extension, that your business is a plumber and not a plumbing blog, that a review score belongs to your product.

The types worth implementing, in priority order for most small businesses:

1. **Organization** — with `name`, `url`, `logo`, `description`, and critically `sameAs` pointing to every profile you control (LinkedIn, Crunchbase, Wikidata if you have an entry, YouTube, GitHub, industry directories). `sameAs` is how you tell machines "all these identities are one entity."
2. **LocalBusiness** (or the specific subtype, like `Plumber` or `Dentist`) — with exact address, geo coordinates, hours, service area, and phone. Must match your Google Business Profile character-for-character.
3. **Product** and **Offer** — for e-commerce. Accurate price, currency, availability. Lying here gets you filtered.
4. **Article** with a real `author` that links to a real author page — see the E-E-A-T section below.
5. **FAQPage** — only for genuine FAQs actually visible on the page. Google reduced FAQ rich results in search, but the markup still clarifies question-answer pairing for parsers, and question-answer structure chunks beautifully. Don't fabricate questions nobody asks.

Validate everything with Google's Rich Results Test and the Schema.org validator. Broken JSON-LD is worse than none — a syntax error can invalidate the whole block silently.

The E-E-A-T layer, translated into things you can actually do

Experience, Expertise, Authoritativeness, Trust. Google's quality framework, and a decent proxy for what makes a source look citable to an AI system. Translated into a to-do list:

- **Put a real human name on your content, with a real bio page.** Not "Admin." Not "The Team." A named person with credentials, a photo, a LinkedIn link, and a sentence about why they're qualified to write this. Link the article to the author page with schema. This is the single highest-leverage trust fix on most small sites and it takes an afternoon.
- **Show first-hand experience.** "We've installed 200 of these" or "in fourteen years of doing this, the failure I see most is X." Models are trained to value primary experience and it reads as distinctive, quotable material.
- **Make your business physically real on the page.** Street address, real phone number, business registration or license numbers where applicable, years in operation. Anonymous sites get treated as risky.
- **Cite your own sources.** When you reference a standard, a regulation, or a statistic, link to the primary source. Pages that cite well get treated as more trustworthy.
- **Date your content and update it honestly.** A visible "reviewed January 2026" on a page that was genuinely reviewed then.
- **Be findable elsewhere.** Which is Chapter 2's entire subject.

How to measure this without fooling yourself

This is where most people go wrong, so let's be rigorous. There are four measurement layers and you need at least the first two.

Layer 1: Referral traffic. In Google Analytics 4, build a report filtered to session source containing any of: `chatgpt.com`, `chat.openai.com`, `perplexity.ai`, `claude.ai`, `copilot.microsoft.com`, `gemini.google.com`, `you.com`, `bing.com` (partly). Set it up as a saved comparison or a custom channel group so you can watch it weekly. Also track conversions from that segment separately — that's where the real story is.

Caveat you must know: attribution here is leaky. Some assistants strip or don't send referrer headers, some open links in ways that lose the source, and mobile app contexts vary. Your reported number is a floor, not a total.

Layer 2: Server logs / bot hits. Look at raw access logs (or Cloudflare analytics) for user-agent strings containing GPTBot, OAI-SearchBot, ChatGPT-User, ClaudeBot, PerplexityBot, Google-Extended, Bingbot. This tells you two things referral data can't: whether the bots can reach you at all, and which specific URLs they're fetching. A spike in ChatGPT-User hits on one URL means that page is being cited live in conversations. That's your single best leading indicator, and it's free.

Layer 3: Manual answer auditing. Build a list of 20–40 questions your actual buyers ask. Run them monthly across ChatGPT, Claude, Gemini, Perplexity, and Copilot. Record, in a spreadsheet: were you mentioned, were you cited with a link, which competitors appeared, and what sources did the assistant cite. Use a logged-out or temporary-chat session so personalization and memory don't contaminate the result — this matters more than people realize; if you've been talking about your own company for a year, your account will name it back to you and you'll conclude you're winning when you aren't.

Also expect variance. Ask the same question twice and you may get different names. Run each question two or three times and record how often you appear, not whether you appeared once. Treat it as a rate, not a binary.

Layer 4: Paid monitoring tools. Several vendors now track AI answer visibility at scale — Profound, Peec AI, Otterly.AI, and AI-visibility modules inside Semrush and Ahrefs, among others. They're genuinely useful for tracking dozens of prompts across engines over time. They cost real money, typically somewhere from tens to several hundred dollars a month depending on tier and vendor — verify current pricing directly, it moves. My advice: do Layer 3 manually for two or three months first. You'll learn more about the mechanics doing it by hand, and you'll know whether the spend is justified.

Setting expectations you can actually hit

Realistic timeline for a small business starting from a normal, modest web presence:

- **Weeks 1–2:** Fix crawler access, verify Bing indexing, add Organization and LocalBusiness schema, put real author bios up, publish llms.txt, set up measurement. Nothing visible happens yet. Do it anyway — everything downstream depends on it.
- **Weeks 3–8:** Publish 6–12 genuinely expert, well-structured pages targeting Type C problem queries. Start seeing occasional citations in Perplexity and ChatGPT search for narrow, low-competition questions. This is the first real signal.
- **Months 3–6:** Third-party mention work starts landing — a directory listing here, a roundup inclusion there, a few reviews. Category-query appearances start, inconsistently. Referral traffic becomes a visible line in analytics rather than noise.
- **Months 6–12+:** If the third-party work has been consistent, you begin appearing in "best of" answers with some reliability. Parametric recall may start to include you after subsequent model updates — no guarantee, no schedule.

Where it will not work, and I'd rather tell you now: brand-new businesses with no track record competing in saturated categories against companies with a decade of press. You will not talk your way into "best CRM" answers this year. Pick narrower ground — "best CRM for independent insurance agencies in Texas" — and win something real.

Quick Quiz

1. Name the three distinct mechanisms by which an AI assistant can produce a business name in an answer, and say which one you can influence fastest.
2. Why does a 4,000-word guide sometimes lose to a shorter page, even when the long guide contains a better answer?
3. What is the practical difference between blocking GPTBot and blocking OAI-SearchBot, and which one costs you citations?
4. When you run a manual audit of whether an assistant mentions you, why must you use a logged-out or temporary chat session?

Answers: 1. Parametric recall from training data (slowest, essentially uncontrollable short-term); retrieval-augmented search of live web pages (fastest to influence — publish a retrievable page and it can be cited within days); and tool/connector calls to structured sources like feeds and business listings. 2. Because retrieval works on passages, not whole pages — the system chunks your content and evaluates individual few-hundred-word segments, so an answer buried mid-paragraph on page four loses to a self-contained answer sitting directly under a matching heading. 3. GPTBot is associated with training-data collection; OAI-SearchBot builds the index ChatGPT search draws from. Blocking OAI-SearchBot (and ChatGPT-User, PerplexityBot, Bingbot) is what removes you from citations — blocking GPTBot only affects longer-term training presence. 4. Because personalization and conversation memory contaminate the result: if your account has previously discussed your own company, the assistant is far more likely to name it back to you, and you'll mistake personalization for genuine visibility.

Chapter 2: Earning Third-Party Mentions the Models Already Read

Here's the sentence that should reorganize your marketing budget: **for the queries that produce the most valuable recommendations, your own website is often not what the model reads.**

Ask any assistant "best payroll software for a 12-person agency" and watch the citations. You'll typically see some mix of software review platforms, a couple of independent blogs or trade publications, a Reddit thread, maybe a YouTube comparison, and possibly one or two vendor pages. The vendor pages are usually there to confirm a detail — pricing, a feature — not to establish who belongs on the list. The list came from everywhere else.

That's the mechanic. The model synthesizes a consensus from independent sources. If independent sources don't mention you, you don't exist in that consensus, no matter how good your site is.

So the work is: get mentioned, by name, in the specific places these systems read, in contexts that match how buyers search. This chapter is how.

First: find out what they're actually reading

Do not guess. Do not assume "it's all Reddit" or "it's all G2." The source mix varies enormously by industry, and you can observe yours directly in about ninety minutes.

The citation audit. Here's the process:

1. Write down 25 to 40 questions a real buyer of your product or service would type. Mix all four query types from Chapter 1, weighted toward category and problem queries. Use their words, not your industry jargon.
2. Run every question through Perplexity first — it's the most transparent about citations and the fastest to audit. Then ChatGPT with search on, then Google AI Mode or AI Overviews, then Claude with search, then Copilot. Use logged-out or temporary sessions.
3. For every answer, log every cited domain into a spreadsheet. One row per citation: query, engine, cited domain, cited URL, and whether you or a competitor was named.
4. Count the domains. Sort descending.

What comes out is a ranked list of the sources that actually decide your category. I've never seen a business run this exercise and not be surprised by at least three entries. You'll find a trade publication you'd forgotten, a niche directory you didn't know existed, a subreddit you dismissed, a comparison site you assumed was irrelevant.

That list is your target list. Everything else in this chapter is about getting onto it. Work the list in order of frequency — the domain cited eleven times is worth ten times the effort of the one cited once.

Also log the *competitors* named. Then run the reverse exercise: for your three closest competitors, search their brand name plus terms like "review," "vs," "alternative," and "best" and catalog where they appear that you don't. That gap list is a concrete, prioritized backlog.

Source category 1: Review and comparison platforms

For B2B software and services, these are frequently the single heaviest-weighted source category. G2, Capterra, GetApp, TrustRadius, Software Advice, Gartner Peer Insights. For consumer services: Trustpilot, Yelp, Google reviews, BBB, Angi, Houzz, Thumbtack — depending on trade.

Why they carry so much weight: they're structured, they aggregate many independent opinions, they're updated constantly, and they publish exactly the "best X for Y" pages that match category queries word for word. A retrieval system loves them.

What to actually do: