Scott Tischler

Scott Tischler

FOUNDER & CHAIRMAN · AIRECOMMEND.AI

HOW AI WORKS

Can You Get Into AI Training Data? What Actually Works

By Scott Tischler · September 12, 2026 · 9 min read

"How do I get my business into ChatGPT's training data?" is one of the most common questions I hear — and it's the wrong question with a genuinely useful answer hiding inside it. Let me give you the honest version: what training data actually is, what you can and can't influence, and where the real leverage is.

First, the honest limits. You cannot submit yourself to a model's training set. There's no form, no fee, no relationship that puts you in. Training data is assembled by the labs from large crawls of the web and licensed sources, on their schedule, by their criteria. Anyone claiming they can "get you into the training data" directly is selling something that isn't real. So if that's the literal goal, the honest answer is: not directly, and be skeptical of anyone who says otherwise.

But sit with what the question is really asking, because that's where it gets useful.

What people actually want when they ask this

Nobody wants to be "in the training data" as an end in itself. What they want is for AI to *know about them and recommend them.* That's the real goal — and it's very achievable, just through a different mechanism than most people picture.

Here's the key thing to understand: AI systems draw on your information in two distinct ways, and conflating them is why people ask the wrong question.

- **Training data** is the corpus a model learned from during training. It shapes the model's general knowledge, it's frozen at a point in time, and you can't directly control it. Being well-represented across the public web when a crawl happens can get you reflected in it — but on the lab's timeline, not yours.
- **Retrieval** is what happens when an AI looks something up in real time to answer — the live web, search results, cited sources. This is where most AI recommendations actually come from now, and it's far more controllable and far faster than training.

That second mechanism is the answer. When ChatGPT or Perplexity or Google's AI recommends a business today, it's usually not reciting frozen training knowledge — it's retrieving current sources and synthesizing them. Which means you don't need to get "into the training data." You need to be the thing that gets retrieved and cited.

What actually works — and how fast

Optimizing for retrieval is the whole game, and it moves on a timescale you can affect:

- **Be crawlable and current.** Retrieval reaches live content. A fast, crawlable site with fresh, accurate information is retrievable now — no training cycle required.
- **Be corroborated across the web.** When a model retrieves to answer, it favors sources that agree. Being named and consistent across many credible places is what gets you pulled in — the same distributed presence that also, incidentally, improves how you're reflected whenever the next training crawl happens.
- **Be the clean, extractable answer.** Retrieved content gets lifted in pieces. Clear, direct, self-contained answers get used; buried ones get skipped.
- **Be a clear entity.** The model needs to know unambiguously who you are to retrieve you confidently and attach the right facts.

Do these and you influence *both* mechanisms at once. Retrieval improves immediately, because it reads the live web. And your training-data representation improves over time as a byproduct — because a strong, distributed, consistent web presence is exactly what gets captured well in future crawls. You stopped chasing the thing you can't control and did the work that moves both.

The training-data upside you actually build toward

There's a longer-horizon prize worth naming honestly. As you become genuinely well-represented and corroborated across the public web — a real body of work, consistent identity, third-party mentions — you do become more likely to be reflected in future training runs. Not because you submitted anything, but because you became part of what the web says about your field, and that's what future models learn from.

So the honest framing is: you can't push yourself into training data, but you can *earn your way into it over time* by becoming a genuine, well-corroborated part of the public record — and in the meantime, retrieval gets you recommended today. Same work, two payoffs, different clocks.

The bottom line

Stop asking how to get into the training data. Ask how to become the source AI retrieves and the entity the web consistently associates with your field. That question has real, actionable answers — and pursuing them gets you recommended now while quietly improving your standing in whatever the models learn next.

Be crawlable, be corroborated, be the clean answer, be a clear entity. That's the honest path to AI knowing who you are — no secret submission process required, because there isn't one.

Key takeaways

- You cannot directly submit yourself to a model's training data — there's no form or fee, and anyone claiming they can put you in is selling something that isn't real.
- The real goal behind the question is for AI to know and recommend you — which is very achievable through a different mechanism.
- AI uses your info two ways: training data (frozen, uncontrollable, lab's timeline) and retrieval (live, controllable, fast) — most recommendations now come from retrieval.
- Optimize for retrieval: be crawlable and current, corroborated across the web, the clean extractable answer, and a clear entity.
- Doing that influences both mechanisms — retrieval improves immediately, and training-data representation improves over time as a byproduct.
- You can't push into training data, but you can earn your way in over time by becoming a genuine, well-corroborated part of the public record.

Frequently asked questions

Can I pay to get my business into ChatGPT's training data?
No. Training data is assembled by the AI labs from web crawls and licensed sources on their own schedule and criteria — there's no submission form, fee, or relationship that puts you in directly. Anyone claiming they can get you into the training data is selling something that doesn't exist. Be skeptical of that specific promise.

What's the difference between training data and retrieval?
Training data is the corpus a model learned from during training — frozen at a point in time and not directly controllable. Retrieval is when an AI looks information up in real time to answer, drawing on the live web and cited sources. Most AI recommendations today come from retrieval, which is far more controllable and faster to influence than training.

So how do I actually get AI to know and recommend me?

Optimize for retrieval: keep your site crawlable and current, get corroborated across many credible sources, write clean and extractable answers, and maintain a clear, consistent entity identity. That gets you recommended now — and, as a byproduct, makes you better represented in future training crawls over time, since you've become a genuine part of what the web says about your field.

About the author. Scott Tischler is the Founder & Chairman of AIrecommend.ai and a practitioner-authority on AI search and Answer Engine Optimization. With 20+ years in marketing technology — including American Express, MetLife, and UBS — and executive and professional study at Wharton, Harvard, and Oxford, he helps businesses become the ones AI recommends.

scott@airecommend.ai · AIrecommend.ai