3cubed.ai · Research Report · June 2026

AI RECOMMENDATION STABILITY:

What Marketers Need to Know About How AI Decides Which Brands to Recommend

Authors: Noriko Yokoi, Ph.D. · Thorsten Linz · 3cubed.ai Research Team
Contact: ny@3cubed.ai · findabl.app

KEY TAKEAWAYS

- AI does not give brands a fixed rank. Every time someone asks a question, the AI draws from a probability distribution — which means your brand may appear in 80% of responses one week and 40% the next.
- We tested four AI engines — ChatGPT, Gemini, Claude, and Perplexity — across 26 buyer questions. Only 27% of questions produced unanimous agreement on the top brand across all four engines.
- When retrieval is turned on (the normal production setting), the probability of the same brand appearing twice in a row as the top recommendation drops to roughly 50% — even for well-known brands.
- The single biggest lever brands can pull is coverage in gatekeeper publications. Just 10 domains account for 29% of all source citations across AI engines. Chambers, Forbes, and TechRadar alone account for 16%.
- A one-time GEO audit is not enough. AI visibility must be tracked continuously — by engine, by query, over time — because it changes constantly even when you do nothing.

1. Why We Did This Study

When someone asks an AI assistant 'What are the best telehealth platforms for ADHD?' or 'Which law firm should I hire for a personal injury case?', they are no longer getting a list of links to click through. They are getting a recommendation — a curated, confident

answer that names specific brands, often without requiring the person to visit a single website.

For brands, that shift is significant. Being named in an AI response may be more valuable than ten blue-link appearances in traditional search results. The emerging practice of Generative Engine Optimization (GEO) — adapting content to improve AI citation — is a direct response to this shift. Academic research published in 2024 and 2026 has confirmed that content modifications such as adding statistics, expert quotations, and cited sources can meaningfully increase how often a brand gets named in AI responses.

But there is a question those studies do not answer: once a brand earns a citation, does it stay cited? Is AI recommendation a rank you can win — or a probability that keeps changing?

That is exactly what this study was designed to find out. We ran the same 26 buyer questions 2,580 times across four AI engines, under three controlled conditions, and measured how consistent the brand recommendations were — within a single moment, and across different engine settings. The results change how marketers should think about AI visibility measurement.

The commercial stakes behind that question are growing. Ahrefs found that AI-referred visitors accounted for 0.5% of sessions but drove 12.1% of signups — a 23x conversion rate advantage over organic search (Ahrefs, June 2025). Adobe Digital Insights reported AI referrals to retail sites converting 42% better than non-AI traffic in March 2026. If AI-referred visitors convert at these rates, then whether your brand is cited — and how reliably — is a revenue question, not just a visibility one.

2. What the Research Already Tells Us

Two published academic studies form the foundation for this work. Aggarwal et al. (2024) coined the term Generative Engine Optimization and showed that brands can improve their AI citation rates by up to 40% through specific content changes — particularly adding data, authoritative quotes, and source citations. They also established that the SEO playbook does not transfer: keyword stuffing, the most common SEO tactic, had no positive effect in AI environments (GEO: Generative Engine Optimization, KDD 2024, arXiv:2311.09735).

Yu et al. (2026) extended that work by showing that content structure matters independently of content quality. How information is organized — document hierarchy, paragraph length, use of headers and tables — influences whether AI engines cite a source, across all major engine architectures (Structural Feature Engineering for Generative Engine Optimization, arXiv:2603.29979).

Separately, Ahrefs found that AI-referred visitors converted at 23x the rate of organic search visitors in their own traffic data (Ahrefs, June 2025), and Adobe Digital Insights reported AI referrals to retail sites converting 42% better than non-AI traffic in Q1 2026. These findings establish why citation frequency is a revenue-relevant metric, not just a visibility one.

What none of this research addressed is persistence. Aggarwal et al. and Yu et al. measured average citation rates at a point in time. Neither asked whether those rates hold up when the same question is asked again, or whether different AI engines agree on the same brands. That is the question this study was designed to answer.

3. How We Ran the Study

This study was pre-registered — meaning all research questions, the full query set, and the analysis plan were written down and locked before any data was collected. This is standard practice in academic research and prevents the common problem of researchers finding patterns after the fact that they then claim to have been looking for all along.

The engines we tested

Engine	Model Version	Query Coverage
ChatGPT	GPT-5.5 (openai/gpt-5.5-20260423)	26 queries
Gemini	Gemini 3.5 Flash (google/gemini-3.5-flash-20260519)	26 queries
Claude	Claude Sonnet 4.6 (anthropic/claude-4.6-sonnet-20260217)	14 queries
Perplexity	Perplexity Sonar	26 queries

Table 1. The four AI engines tested, with pinned model versions to ensure consistency.

All engines were accessed through a single API gateway with fallbacks disabled, so the exact model version receiving each query was recorded and controlled. Model versions were pinned wherever the provider allowed.

The three test conditions

The most important design decision in this study was running each query under three different conditions. This let us separate out the different causes of recommendation instability:

Condition	What it means — and what it isolates
Floor	Minimum randomness, no live web search. This is the most controlled setting possible. Any instability here is built into the model itself and cannot be eliminated.
Sampled	Normal randomness, no live web search. Adds the everyday variation that comes from how the model generates language.
Web	Normal randomness, live web search on. This is the real-world production setting — and the most unstable, because the sources retrieved can differ with every query.

Table 2. The three experimental conditions. Running all three lets us pinpoint where instability comes from.

The questions we asked

We used 26 unaided buyer questions — questions that name a need or category but never a brand. This is the only way to measure brand citation fairly: if you name a brand in the question, you have already influenced the answer. The full list of questions is in Appendix A.

Questions spanned three market segments: Regulated (legal, healthcare, finance), Consumer (mattresses, meal kits, VPNs, running shoes), and B2B (CRM, project management, email marketing). Within each segment, questions were chosen across three levels of market competitiveness: contested, moderate, and high-consensus.

What we measured

For each question, we ran it many times per engine per condition and recorded which brands appeared in the answer. This gave us two core metrics:

- **Recommendation Consistency Score (RCS):** The probability that two back-to-back repetitions of the same question return the same top brand. A score of 1.0 means perfectly consistent; 0.5 means essentially a coin flip.
- **Appearance rate:** For a given brand and question, the percentage of responses in which that brand appeared anywhere. A brand with a 90% appearance rate is reliably cited; a brand at 10% is rarely cited.

A note on fairness

Every response was parsed by the same automated tool with no human judgment applied to individual responses. This removes the bias that comes from manually deciding which brands 'count.' The same rules applied to every engine, every question, every wave.

4. What We Found

Recommendation Stability: AI recommendations are not stable — even under ideal conditions

Here is the clearest way to state what we found: even when we minimized randomness and turned off live web search entirely — the most controlled possible environment — the same question asked twice in a row produced the same top brand recommendation only 78% to 87% of the time, depending on the engine.

That means even in the best case, roughly one in eight repetitions of an identical question produces a different top brand. This is not caused by the web changing. It is built into how AI language models work. Some level of instability is irreducible.

When live web search is turned on — the setting used by real users every day — consistency drops sharply:

Engine	RCS (Floor)	RCS (Sampled)	RCS (Web — real world)	Drop: Floor → Web
ChatGPT	0.78	0.78	0.48	−30 percentage points
Gemini	0.78	0.67	0.54	−24 percentage points
Claude	0.87	0.69	0.50	−37 percentage points
Perplexity	N/A	N/A	0.68	Web-only engine

Table 3. Recommendation Consistency Score (RCS) by engine and condition. RCS = probability that two consecutive identical queries return the same top brand. Lower numbers mean less consistency.

What is driving that drop? The web retrieval layer. When live search is on, the sources an AI engine pulls in can differ from one query to the next — because search results

themselves vary slightly, and because each engine's retrieval process involves its own probabilistic ranking. The sources change, and so does the recommendation.

> **In real-world conditions, the probability of the same brand appearing as the top recommendation twice in a row is roughly a coin flip. That is not noise — it is the system working as designed.**

Engine Divergence: Different engines recommend different brands — often completely

A brand that consistently appears at the top of ChatGPT results may barely appear in Gemini, and may be absent from Perplexity entirely. This is not an edge case — it is the norm.

Across 26 questions tested on all four engines simultaneously, only 7 (27%) produced unanimous agreement on the top brand. Three in four questions saw at least one engine diverge from the others.

Question (abbreviated)	ChatGPT	Gemini	Claude	Perplexity
Best ADHD telehealth platform	Talkiatry	Talkiatry	Talkiatry	Talkiatry
Best CDMO for biologics	Lonza	Lonza	Lonza	Lonza
Best VPN services	NordVPN	NordVPN	—	NordVPN
Best PI law firms (L.A.)	Panish Shea	Panish Shea	Greene Broillet	Panish Shea
Best robo-advisors	Vanguard	Vanguard	Fidelity Go	Vanguard
Best weight-loss Rx provider	Form Health	Ro Body	Ro	ShedRx
Best RIA for HNW clients	Pathstone	Mercer Global	Cresset Capital	Creative Planning

Table 4. Top brand recommended by each engine for selected questions. Amber cells = divergence from the most common engine response. Top rows show consensus; bottom rows show fragmentation.

The pattern is clear. In categories with a widely recognized market leader — backed by institutional rankings, regulatory endorsements, or dominant review coverage — engines converge. In fragmented markets where no single brand has clear authority, engines diverge completely, each pulling from different sources and arriving at different answers.

What this means for your brand

If you are only checking your AI visibility on one engine, you are seeing at best a quarter of the picture. A brand can be the top recommendation on ChatGPT and virtually absent on Perplexity — and only continuous multi-engine tracking will reveal that gap.

Recommendation Probability: Most brands appear in fewer than half of responses, even for their own category

The appearance rate data — the percentage of times a brand was named across repeated queries — tells a stark story:

- **Only 2.4%** of brand-question combinations achieved a perfect 100% appearance rate.
- **85.7%** of brand-question combinations fell below 50% — the brand appeared in fewer than half of responses for its own query.
- **The median appearance rate was 10%:** the typical brand in this study appeared in just 1 in 10 responses for its question.

The brands that achieved 100% appearance rates were in categories with exceptional consensus — Talkiatry for ADHD telehealth, NordVPN for VPN services, Lonza for biologics CDMO, Panish Shea Ravipudi for LA personal injury law. These are brands with deep, consistent coverage across the specific publications that AI engines preferentially cite. That coverage is not an accident.

Across the rest of the dataset — which more closely reflects the reality facing most brands — visibility is probabilistic and fragile. A brand that "appears in AI results" may be appearing in 15% of queries, not 85%. That difference is invisible to anyone doing a single-shot check.

Gatekeeper Coverage: A small number of publications control most of what AI recommends

This is arguably the most actionable finding in the study. When we analyzed which external sources AI engines cited when generating recommendations, the distribution was highly concentrated:

- **376 unique domains** were cited across all engines and all questions.
- **The top 10 domains** accounted for 28.6% of all citations — out of 3,059 total.

- **Just three domains — Chambers, Forbes, and TechRadar —** accounted for 15.8% of all citations combined.

#	Publication	Times Cited	% of All Citations	Who it matters for
1	**Chambers**	250	8.2%	Legal services
2	**Forbes**	122	4.0%	Finance, business, consumer
3	**TechRadar**	110	3.6%	B2B software, consumer tech
4	**U.S. News Best Law Firms**	70	2.3%	Legal services
5	**Tom's Guide**	67	2.2%	Consumer tech, software
6	**RTINGS.com**	67	2.2%	Consumer electronics
7	**NerdWallet**	51	1.7%	Personal finance
8	**Legal 500**	49	1.6%	Legal services
9	**Vault**	47	1.5%	Legal, career
10	**Sleep Foundation**	43	1.4%	Consumer health, mattresses

Table 5. Top 10 source domains cited by AI engines across all questions. These 10 publications account for 28.6% of all AI source citations in the study.

This finding reframes the entire GEO conversation. It is not primarily about optimizing your own website for AI. It is about being present in the publications that AI engines treat as the authoritative sources for your category. A law firm that earns a Chambers ranking is not just earning a professional credential — it is earning the primary retrieval signal that drives AI citation for its category. The same logic applies to every category: Forbes for finance, TechRadar for B2B tech, NerdWallet for consumer financial services, Sleep Foundation for mattresses and consumer health.

> **Being cited by AI is not primarily a result of your own website's content. It is a result of whether the publications AI trusts have covered your brand favorably.**

Market Consensus: How fragmented your market is, predicts how hard the opportunity is to capture

Not all markets behave the same way. A brand in a category with a clearly recognized market leader will see much more consistent AI recommendations than a brand in a fragmented, contested market. Here is what that looks like in practice:

Category type	Example query	Top-brand consistency	What it means for you
Consensus leader exists	Best CDMO for biologics	0.68–1.00 — same brand most of the time	Defend your position in key rankers
Moderate competition	Best robo-advisors	0.40–0.70 — some consistency	Close coverage gaps across ranker set
Fragmented market	Best HNW investment advisor	0.15–0.35 — new brand every time	Build credentialing signals from scratch
Consumer tech / reviews	Best VPN services	0.50–0.97 — depends on review coverage	Prioritize review aggregator scores

Table 6. How category type predicts AI recommendation stability. Brands in fragmented markets face the steepest challenge.

The most striking example of fragmentation in this dataset was the question 'Which independent RIAs are best for high-net-worth clients?' — where the top-brand consistency across all engines combined was only 0.15, meaning no single brand commanded even a 15% plurality of top placements. Every engine named a different firm, and even within a single engine, the answer changed frequently. For brands in categories like this, the first imperative is not content optimization — it is building the foundational credentialing coverage that allows any AI engine to find and trust them.

5. What This Means for Marketers

The findings from this study add up to a clear shift in how AI visibility should be approached. Here are six things to do differently:

1. Stop thinking of AI visibility as a rank. Start thinking of it as a probability.

A rank is something you win once and defend. A probability is something you influence continuously. The correct question is not 'do we appear in AI results?' — it is 'what percentage of the time do we appear, and is that percentage going up or down?' That requires repeated measurement, not one-time audits.

2. Find out which publications AI trusts in your category — and get into them.

Before you invest another dollar in website content optimization, do this: run your most important buyer questions repeatedly on the major AI engines with web search turned on, and record which sources they cite. Those are your gatekeeper publications. Being named favorably in those publications is the highest-leverage investment you can make for AI visibility. For legal services, Chambers and Legal 500 are not optional. For consumer finance, NerdWallet and Forbes are the gatekeepers. Every category has its own topology — and it is discoverable.

3. Measure across all four major engines, not just one.

With only 27% of questions producing the same top brand across all engines, a brand's visibility on ChatGPT tells you very little about its visibility on Gemini or Perplexity. A measurement program that covers only one engine is missing most of the picture. Multi-engine tracking is the minimum viable measurement standard.

4. Use appearance rate as your KPI, not a yes/no snapshot.

'Does our brand appear in AI results?' is the wrong question. 'What is our appearance rate for this question on this engine?' is the right one. An appearance rate of 30% and an appearance rate of 80% look identical in a single-snapshot audit — but they represent completely different strategic situations. Only repeated measurement reveals which one you are in.

5. Match your expectations to your market.

If you are in a category with a recognized market leader (a single dominant brand that all engines converge on), the task is defensive: maintain your coverage in the publications that create that consensus, and monitor for drift. If you are in a fragmented market, the task is offensive and longer-term: build the credentialing signals — third-party rankings, review presence, expert citations — that allow AI engines to develop any consistent opinion of your brand at all.

6. Track continuously, not quarterly.

AI models update frequently. The sources they retrieve from the web change constantly. A brand's AI visibility in June is not a reliable predictor of its visibility in September. The

measurement cadence needs to match the rate of change — which in 2026 means monthly at minimum, and weekly for contested categories.

6. Study Design and Controls

This section is for readers who want to understand what we controlled for and why the findings are trustworthy.

Pre-registration

All research questions, hypotheses, the full set of queries, segment assignments, engine choices, experimental conditions, extraction rules, and primary metrics were written down and frozen before any data collection began. This is called pre-registration and it is the standard in academic research to prevent cherry-picking findings after the fact. Any deviations from the pre-registered plan are logged.

What we controlled for

- Model version: Each engine's model version was pinned and recorded, so results reflect a specific, known model — not a moving target.
- Extraction rules: Every AI response was parsed by the same automated tool using the same rules. No human judgment was applied to individual responses, which removes subjective bias.
- Query design: All 26 questions were 'unaided' — they name a need or category but never a brand. Brand mentions are therefore outputs of the AI's reasoning, not echoes of what we asked.
- API gateway: All four engines were accessed through a single gateway with fallbacks disabled, ensuring each query actually went to the pinned model version.

What we acknowledge as limitations

- Personalization: Real users query from logged-in accounts with personal history. Our programmatic queries cannot replicate that. Our results represent the non-personalized baseline — real-world variance may be somewhat higher.
- Retrieval realism: For ChatGPT, Gemini, and Claude, the web condition used a controlled search layer, not each product's native retrieval stack. Only Perplexity's web results reflect its native consumer product. Cross-engine web comparisons should be read with this in mind.
- Query coverage: This wave covers 26 of the 45 pre-registered questions, with a concentration in regulated and contested categories where instability is most commercially significant.

- Model drift: AI providers updated models during the study window. Pinning and recording versions mitigates this, but longitudinal comparisons should account for potential version changes.

7. Conclusion

AI visibility is not a ranking you win once. It is a probability that changes with every query, every AI engine, and every model update.

This study found that even under the most controlled conditions possible, the same question asked twice in a row returns the same top brand only 78–87% of the time. In real-world conditions with live web search, that figure falls to 48–68%. Across four major AI engines, three in four questions produce different top recommendations depending on which engine is used. The median brand in this dataset appeared in only 10% of responses for its own category question.

And yet the study also found structure within this instability. The brands that appear most reliably — Talkiatry, NordVPN, Lonza, Cravath — are not there by chance. They are there because they have earned deep, favorable coverage in the small number of publications that AI engines treat as category authorities. Chambers alone accounts for 8.2% of all source citations in this dataset. That is the mechanism: your brand's standing in the publications AI trusts is not a proxy for AI visibility — it is the cause of it.

The practical implication is clear. Marketers who treat a GEO audit as a one-time event will systematically underinvest in the things that actually drive AI citation and over-invest in optimizations that produce no durable gain. The corrective is continuous, multi-engine, probability-based measurement — tracking appearance rates over time, mapping gatekeeper publications, and monitoring for drift.

The revenue case makes this urgent. Ahrefs found that AI-referred visitors converted at 23x the rate of organic search visitors, with 0.5% of traffic driving 12.1% of signups (Ahrefs, June 2025). Adobe Digital Insights reported AI referrals to retail sites converting 42% better than non-AI traffic in Q1 2026. A measurement gap around AI citation is not an abstraction — it is a gap in revenue attribution.

Recommendation Stability is becoming the new Share of Voice for the AI era. The brands that measure it—and improve it—will be the brands AI recommends tomorrow.

That is precisely what Findabl (findabl.app) was designed to provide.

Appendix A: The 26 Questions We Asked

All questions were unaided — they name a need or category, never a brand. This is the only valid way to measure brand citation: the brand names that appear in AI responses are outputs of the AI's own reasoning, not echoes of what was asked.

REGULATED MARKETS

Legal Services

Tier	Question
Contested	What are the best personal injury law firms in Los Angeles?
Contested	Who are the top immigration attorneys in New York City?
Moderate	What are the leading intellectual property law firms in the United States?
Moderate	What are the leading white-collar criminal defense firms in the United States?
High-Consensus	What are the most prestigious large U.S. law firms?

Healthcare

Tier	Question
Contested	What are the best telehealth platforms for ADHD treatment?
Contested	What are the best online providers for prescription weight-loss medication?
Moderate	Which contract development and manufacturing organizations are best for biologics?
Moderate	What are the best electronic health record systems for small medical practices?
High-Consensus	What are the top-ranked hospitals in the United States?

Financial Services

Tier	Question
Contested	What are the best robo-advisors for retirement investing?
Contested	What are the best lenders for small business loans?
Moderate	Which independent registered investment advisors are best for high-net-worth clients?
Moderate	What are the best 401(k) providers for small businesses?

High-Consensus	What are the largest banks in the United States?

CONSUMER MARKETS

Tier	Category	Question
Contested	VPN	What are the best VPN services?
Contested	Mattresses	What are the best mattresses to buy?
Contested	Meal kits	What are the best meal kit delivery services?
Contested	Running shoes	What are the best running shoes for beginners?
Contested	Protein powder	What are the best protein powder brands?
Moderate	Robot vacuums	What are the best robot vacuums?

B2B MARKETS

Tier	Category	Question
Contested	CRM	What are the best CRM platforms for small businesses?
Contested	Project mgmt	What are the best project management tools?
Contested	Email marketing	What are the best email marketing platforms?
Contested	Help desk	What are the best customer support help desk tools?
Contested	Web hosting	What are the best web hosting providers?

Competitiveness tiers explained:

- Contested: multiple credible brands compete and no single leader is universally recognized
- Moderate: a smaller competitive set with some clearer leaders, but still meaningful variation
- High-Consensus: a widely agreed-upon answer exists (e.g., largest banks, most prestigious law firms)

Appendix B: Plain-Language Glossary

Term	What it means in plain language

Generative Engine	An AI system like ChatGPT, Gemini, Claude, or Perplexity that answers questions by synthesizing information from multiple sources rather than returning a list of links.
GEO (Generative Engine Optimization)	The practice of adapting content so that AI engines are more likely to cite your brand when answering relevant questions.
Unaided query	A question that names a need or category but never a specific brand — so any brand name that appears in the AI's answer was generated by the AI itself, not prompted by the question.
Appearance rate	The percentage of the time a brand appears in AI responses for a given question. A brand with a 90% appearance rate is cited almost every time; a brand at 10% is rarely cited.
RCS (Recommendation Consistency Score)	The probability that two back-to-back identical questions return the same top brand. A score of 1.0 = always the same; 0.5 = essentially random.
Gatekeeper publication	A third-party publication that AI engines treat as an authoritative source for a category and cite frequently — such as Chambers for legal, Forbes for finance, or TechRadar for B2B tech.
Floor condition	The most controlled test setting: minimum randomness, no live web search. Shows the baseline instability built into the AI model itself.
Web condition	The real-world production setting: normal randomness, live web search on. The most unstable condition because the sources retrieved can differ with each query.
Set Jaccard similarity	A measure of how much overlap exists between two lists of brand recommendations for the same question. A value of 1.0 means identical lists; 0 means no shared brands at all.

About 3cubed.ai

3cubed.ai is an AI-first agency specializing in generative engine optimization, agentic marketing, and AI visibility strategy for regulated industries including pharma, life sciences, healthcare, and financial services.

Findabl (findabl.app) is 3cubed.ai's AI Citation Intelligence Platform — built to provide the continuous, multi-engine, probabilistic citation measurement this research demonstrates is necessary for meaningful AI visibility management.

For licensing, citation, or partnership inquiries: ny@3cubed.ai · findabl.app

References

Academic References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. GEO: Generative Engine Optimization. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024), Barcelona, Spain, August 2024. arXiv:2311.09735.

Yu, et al. Structural Feature Engineering for Generative Engine Optimization. arXiv:2603.29979, 2026.

Industry Data

Ahrefs. Does AI Search Traffic Convert Better Than Traditional Search? June 2025. ahrefs.com.

Adobe Digital Insights. AI-Driven Traffic Surges Across Industries. Q1 2026. business.adobe.com.

