

How Often Does AI Recommend You?

What a million AI answers say about measuring brand visibility

Manicule Research · Manicule, Inc. · <https://manicule.com>

Abstract. More and more buyers ask ChatGPT, Claude or Gemini which software to use, and companies want to know how often they are recommended. That is harder to answer than it sounds, because an AI assistant rarely gives the same answer twice. In October 2026 we collected and analysed 1,053,650 answers to 173,554 different buyer questions across 200 software categories, almost all of them from the three assistants Manicule tracks, and combined them with eight earlier, smaller studies. This paper reports what we found, in plain terms. A single answer is a weak signal: it misjudges a brand by a median of 14 points. Only about a third of answers name a single top choice, and when they do, two identical requests put a different brand first 37% of the time on Claude and 55% on ChatGPT. The assistants disagree with each other, and each has its own shortlist. Who is asking changes the answer a great deal, while the user's country and how much the assistant searches change it no more than chance. We explain how Manicule's tracking is built on these results, what each part can and cannot detect, and where we got our own first summaries wrong.

Nine findings

1. **One answer tells you little.** It misjudges a brand's rate by a median of 14 points. Thirteen answers cut that to 7 (§3.1).
2. **The same request rarely gets the same top pick.** Two identical requests name a different brand first 37% of the time on Claude and 55% on ChatGPT, and only about a third of answers name a single top choice at all (§3.2).
3. **Who is asking matters; where from mostly doesn't.** Telling the assistant about the user changed its top pick in 63–74% of answers. A different country, or more web searching, changed it no more than chance (§3.3).
4. **Each assistant has its own shortlist.** The three agreed on a category's leader in under half of categories (48%). Claude names four times as many brands per answer as ChatGPT (§3.4).
5. **Wording matters as much as repetition.** Rewording a question moves a brand's rate as much as asking again, and real buyers' questions vary more than written ones (§3.5).
6. **Questions that name you flatter you.** Mixing them into a visibility score hides real losses (§3.6).
7. **The buying stage matters.** Category leaders are named in 53% of answers when buyers shortlist, and in 24% just before they decide (§3.7).
8. **AI does not simply favour incumbents.** Young companies get a larger share of top recommendations than of mentions (§3.8).
9. **Being named is not being recommended.** A brand can appear in most answers and still rarely be the first choice or a cited source (§3.9).

1 Why one answer is not enough

Ask an AI assistant for “the best CRM for a ten-person sales team” and it will name a few products, often with a favourite. Ask again and the list may change. Reword the question and it may change more. Ask a

different assistant and you may get a different list altogether.

Search engines are different. Their results stay much the same from hour to hour and from person to person; in one audit, personalization changed 11.7% of Google results (Hannak et al., 2013). AI answers do not behave that way. In 2,961 runs of twelve recommendation prompts, ChatGPT and Google’s AI answers repeated the same list of brands less than 1% of the time (Fishkin and O’Donnell, 2026), and small changes in wording are known to move model outputs a great deal (Sclar et al., 2024).

So “Does ChatGPT recommend us?” has no yes-or-no answer. The honest answer is a percentage, *the share of answers that name you*, with a margin of error. Measuring it well means choosing which questions to ask, how many times, on which assistants and under what conditions, and deciding how big a change has to be before it counts. This paper sets out the evidence behind each of those choices.

2 How we studied it

1,053,650	173,554	200	3
answers analysed	different questions	software categories	AI assistants
Collected 25–26 October 2026, plus about 11,000 answers from eight earlier studies.			

The evidence comes from nine studies run in September and October 2026. The first eight (Q1–Q8) were small, targeted experiments, about 11,000 answers from 14 models, each testing one question: how many runs a question needs, whether memory changes recommendations, how much a brand’s own website shapes an answer, and so on (Manicule Research, 2026a). The ninth, Q9, is the large study this paper is built on. It repeated the earlier measurements on the assistants Manicule tracks today and added new ones.

Q9 was designed around a lesson from the earlier studies: to learn what buyers in general are told, it is better to ask many *different* questions than to ask the same few again and again. So most of the budget went on breadth, and a smaller part on repetition and controlled tests (Table 1).

Table 1. How Q9’s 1,053,650 answers were spent. The assistants are the three Manicule tracks in production: Claude Sonnet 5.5, GPT-6 Luna (ChatGPT) and Gemini 3.5 Flash-Lite. The settings tests also used two larger models.

Part	What it measured	Answers
Breadth	160,000 different buyer questions across 200 categories, each asked once with web search on and once with it off, on Claude and ChatGPT; Gemini answered questions in 145 of the categories. Plus 13,397 questions following buyers from first problem to final decision, on all three. This is the map of who gets recommended.	872,277
Repetition	457 questions asked again and again on Claude and ChatGPT: 357 of them 30 times, and a calibration set of 100 asked 100 times. This measures how much answers vary when nothing changes.	41,382
Settings tests	10,000 questions asked a second time with one thing changed: who the user is, their country, a larger model, or more web searching.	139,991
Total		1,053,650

Most questions were written by a model in the voice of realistic buyer personas (NVIDIA Corporation, 2025), and never named a product. The 157 real questions in the repetition set come from WildChat, a public collection of real ChatGPT conversations (Zhao et al., 2024). Another model (Claude Haiku 5.5) read every answer and listed the products it recommended, and which one it put first. Appendix A gives the details.

Throughout, “Claude”, “ChatGPT” and “Gemini” mean these specific models reached through each company’s API with its own web search and an approximate United States location, which is how Manicule tracks them.

3 What we found

3.1 One answer tells you little

To see how much a single answer can be trusted, we used the repetition set: 457 questions asked 30 to 100 times each on Claude and ChatGPT, 41,382 answers in all. For the 100 questions asked 100 times, runs 31–100 give a reliable estimate of how often the assistant really names each brand. We checked how close the first 1, 3, 5, 13 or 30 runs came to that figure, across 4,055 brand–question pairs (Figure 1).

A single answer was a median 14.3 points away from the brand’s real rate, and it put the brand in the wrong quarter (for example “named in under 1 in 4 answers” when the truth was “1–2 in 4”) 31% of the time. Thirteen answers were 6.8 points off and in the wrong quarter 12% of the time. Across all 457 questions (16,706 brand–question pairs), asking one question at a time until the answer settled into a quarter, half of the pairs needed 18 runs or more. That matches our earlier study (Q6) exactly, on newer assistants and nine times as many questions.

Most pairs do settle in the end: 77% (Claude) and 78% (ChatGPT) within 30 runs, against 77% before, and in simulation only 3% of settled results were wrong. Reporting the rest as “between two quarters” instead of forcing a guess kept the simulated error to 2.4–2.6%, as in our earlier study. Checked against the independent 100-run estimates, the rule did less well (7–9% wrong), so we treat it as a useful guide rather than a guarantee.

Across a whole set of questions the picture is better, because errors on different questions partly cancel out. For a brand’s overall score on one assistant, we split each question’s answers in half and compared the halves. Across the 40 brand–assistant pairs with the most questions (23 or more each), a typical brand’s score could be off by up to 11.4 points with one answer per question, 7.1 with three, 5.5 with five and 4.3 with thirteen (95th percentile).

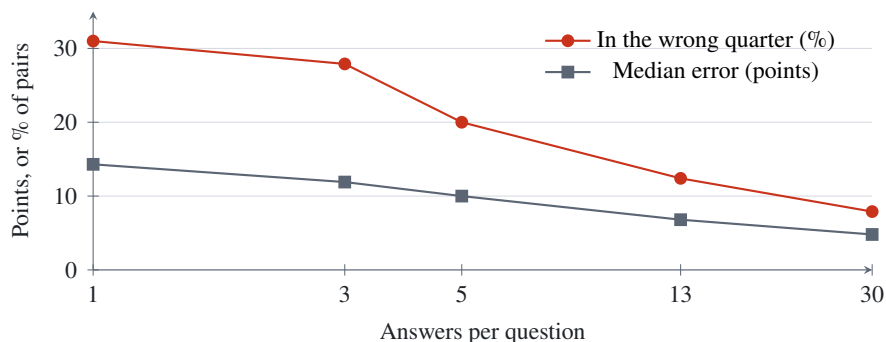


Figure 1. How far an estimate from the first few answers lands from a reliable estimate from 70 more (4,055 brand–question pairs on Claude and ChatGPT).

What this means. A dashboard that shows “100% visibility” from one answer, or a drop from one week’s single answer, is mostly showing chance. Visibility should be reported as a rate from many answers, with a margin of error. Manicule’s managed tracking asks each question 13 times per check for this reason.

3.2 The same request rarely gets the same top pick

First, a surprise: most answers don’t crown a winner. Only 29% of Claude’s answers, 39% of ChatGPT’s and 32% of Gemini’s (with search on) singled out one brand as the top choice; the rest offered several options side by side. Every “top pick” figure in this paper is about the answers that do pick one.

Before testing whether anything *changes* an assistant’s top pick, we measured how often it changes when nothing does. From the repetition set we took pairs of answers to exactly the same request, sent minutes apart, where both answers named a top pick, and checked whether it was the same brand.

It often was not. Two identical requests named a different brand first 37% of the time on Claude (95% confidence interval 31–43%, from 794 pairs) and 55% on ChatGPT (49–60%, from 1,479 pairs). On ChatGPT, in other words, the top pick changes more often than not between two identical requests. Our earlier memory study had measured 22%, but it used short questions without web search on older models; longer, more specific questions and live search both give an assistant more room to vary.

What this means. This is the yardstick for every claim that something “changed what AI recommends”. If a rewrite of your website, a new model or a different setting changed the top pick in 50% of answers, that may be no change at all. We now report every such result next to this baseline.

3.3 Who is asking matters; where from mostly doesn’t

We asked 10,000 questions a second time with one thing changed, and compared the top pick with the original answer, using about 1,800–3,700 pairs per test in which both answers named a top pick (Figure 2). A change counts as real only if it is clearly larger than the chance rate above, by a standard statistical test.

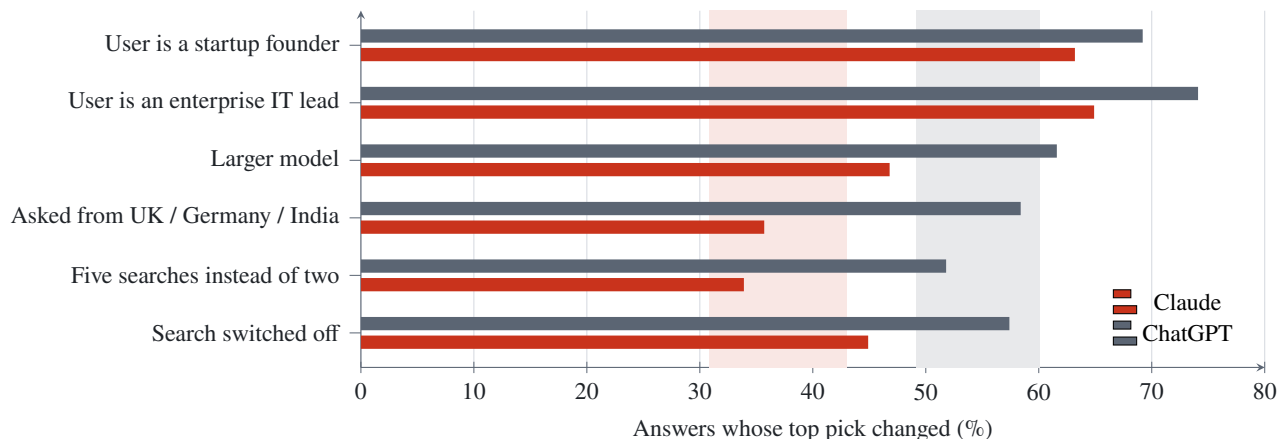


Figure 2. Changing one thing in the request, against chance. Shaded bands show how often two *identical* requests disagree (Claude 31–43%, ChatGPT 49–60%). Bars clearly beyond their band show a real effect. The country bar averages the three countries; the original answers were asked from the United States.

Who is asking. Adding a line about the user, written the way an assistant remembers earlier chats (“a technical founder of a six-person startup on a tight budget”, or “IT lead at a 5,000-person regulated company”), changed the top pick in 63–65% of Claude’s answers and 69–74% of ChatGPT’s, well above chance. On Claude the whole list changed too: the overlap between the two lists fell to 0.23–0.28, against 0.42 between identical requests (on a scale where 1 means the same brands). This matches our earlier study, where saved memories changed the top recommendation in 69% of answers.

A larger model. Swapping in Claude Opus 5.5 and GPT-6.1 Sol changed the top pick in 47% and 62% of answers, about 10 and 7 points above chance: a real but moderate effect.

Country and searching. Asking from the United Kingdom, Germany or India changed the top pick no more than chance. It nudged answers toward local companies, but only slightly: the share of named brands

based in the country rose from 2.0% to 2.3% for the United Kingdom, 2.1% to 2.5% for Germany and 1.3% to 2.0% for India. Allowing five web searches instead of two also made no measurable difference, partly because assistants often don't search even when they can: with search available, 54% of Claude's answers cited a web source, against 80% of ChatGPT's and 87% of Gemini's.

Web search on or off. Switching search off changed Claude's top pick in 45% of cases, modestly beyond chance (37%), and ChatGPT's in 57%, no more than chance (55%). Gemini's changed in 63%, but we have no chance figure for Gemini. What search did change was the length of the list. Without it, Claude named 12.5 brands per answer instead of 9.0, ChatGPT 2.9 instead of 2.2, and Gemini 8.5 instead of 7.1. Search narrows the field. Our earlier work found that it also changes facts such as prices and versions ([Manicule Research, 2026a](#)).

What this means. What an assistant knows about the person asking moves its recommendation more than any setting. Tracking should therefore use a clean session with no memory or instructions, so that results describe a neutral buyer and stay comparable week to week. Worries that results depend heavily on the user's country or on how much the assistant searches are not borne out for English-language business software.

3.4 Each assistant has its own shortlist

On the same question, Claude and ChatGPT named a given brand at rates 31 points apart on average, and for almost half of brand-question pairs (47%) the gap was more than 25 points. In 82% of the cases where a brand came up regularly for a question on one of them, the other almost never named it.

Much of that comes from how long their answers are. Claude named a median of 8.6 brands per answer, Gemini 7.2 and ChatGPT 2.1. Nearly nine in ten of the brands only one assistant named were in Claude's longer lists and never reached ChatGPT's one or two picks. Even at the top the assistants differ. Taking each assistant's most-named brand in a category as its leader, all three agreed in fewer than half of the 145 categories that all three answered (48%). On the same question, two assistants put a different brand first 60–72% of the time, most often Gemini and ChatGPT (72%). That is more than the 37–55% by which one assistant disagrees with itself, so the assistants really do have different favourites, not just different luck. They also cite different sources. In a typical category, links to the websites of the brands an answer named made up 69% of ChatGPT's citations, against 24% of Gemini's and 20% of Claude's.

What this means. An average across assistants describes none of them. Manicule reports each assistant separately. ChatGPT's short lists make it the hardest place to be named, and its heavy use of vendor websites makes a company's own pages matter most there.

3.5 Wording matters as much as repetition

Our earlier study (Q5) asked three wordings of the same need, up to 100 times each. Rewording moved a brand's mention rate about as much as simply asking again: both varied by about 30 points. Single cases were extreme. One payments brand appeared in 1 of 99 answers to "recommend" wording and in all 100 answers to "compare" wording of the same need, and in 43% of cases a brand's best and worst wordings were 30 points or more apart. In Q9, a brand's rate varied by 34 points from one real question to the next, and by 30 points across written ones. Real questions also drew more brands into each answer (9.3 against 5.6), and almost all of them drew at least one (99.7%, against 79% of written questions). People write messier, more varied questions than models do.

What this means. To describe what buyers in general hear, ask many different questions rather than the same few many times. Sets built only from model-written questions understate how much wording matters, so real buyer phrasing, such as questions from sales calls, is worth including.

3.6 Questions that name you flatter you

A question like “Does Zentrio integrate with NetSuite?” almost always gets an answer that mentions Zentrio. It says nothing about whether the assistant would bring Zentrio up unprompted. When we replayed Manicule’s drop alert on real answers (Manicule Research, 2026a), adding 8 such questions to 9 general ones meant a real 20-point loss was caught 13% of the time instead of 37%; adding 16 meant it was caught 2% of the time. Even with 25 general questions, adding 8 brand-named ones cut the catch rate from 77% to 54%. The questions that keep naming the brand dilute the drop.

The same effect explains a widely shared October 2026 claim that Supabase is AI’s top database recommendation “by far”. On 60 neutral questions Supabase was indeed the most mentioned provider, at 62%. On a set shaped like Supabase’s own tracking, half of whose questions named it, the figure was 93%.

What this means. Manicule keeps questions that name the brand out of every visibility figure and caps them at three per workspace. They are still useful for seeing how an assistant describes you.

3.7 The buying stage matters

We labelled every question with the buyer’s stage: describing a problem, exploring kinds of tools, shortlisting, comparing, or deciding. The category leader was named most when buyers shortlisted and least right before they decided (Figure 3). Late-stage answers name fewer brands (3.1 per answer when deciding, against 7.2 when exploring), but when they name the leader they put it first more often.



Figure 3. How often answers name the category’s leading brand, by the buyer’s stage (452,126 answers, all three assistants).

A similar arc showed when we followed imagined buyers through 42 markets: on every assistant the leader was named most while buyers explored or shortlisted (Claude 68% and 66%, Gemini 59% and 53%, ChatGPT 29% and 40%) and least at the start or the end (problem: 44%, 41%, 14%; decision: 36%, 34%, 22%). The brand named most at the start was not the one named most at the end in 48% (Claude), 60% (Gemini) and 69% (ChatGPT) of markets. On Claude, for example, object storage went from Backblaze B2 at the problem stage to Amazon S3 at the decision stage, and LLM APIs from Claude to OpenAI. These figures are probably somewhat high, because each estimate rests on about 60 answers, and chance alone moves a leader that small. Some markets did not move at all: Stripe led payment processing at every stage on Claude, and AWS led cloud hosting at every stage on Gemini.

Two common beliefs did not hold up. Long, specific questions did not by themselves open the door to smaller brands: the leader was named in 39% of answers to generic “best X” questions, 34% when the question added one condition, and 39% with two or more. And asking for something free or open-source helped the leader (54% and 57%, against 38%), perhaps because in many software categories the leader has a free tier or is open-source; we did not test this directly. Questions about a particular industry, compliance, migration or AI features did favour others: the leader was named in only 30–35% of them.

What this means. A tracked set should cover every stage, not only “best X” questions. Specialists have their best chance on questions about specific industries and requirements, not on long questions as such.

3.8 AI does not simply favour incumbents

We looked up the 3,000 most-named brands and found founding dates for 2,994: when each was founded, how it is funded and whether it is open-source. Companies founded before 2005 make up 35% of these brands and get 33% of mentions, but only 24% of first recommendations. Companies founded in 2019 or later make up 11% of brands, 10% of mentions and 13% of first recommendations. Brand by brand, the younger the company, the slightly more often it was named in its category. Public companies are mentioned more than their number would predict (37% of mentions from 30% of brands) but recommended first only in proportion (30%). Growth-stage companies did best at the top (17% of brands, 25% of first recommendations), and open-source projects were picked first about in proportion to their number (20% of brands, 19% of first picks).

A brand’s general web traffic, as measured by the Tranco ranking of popular sites ([Le Pochat et al., 2019](#)), showed almost no relationship with how often AI named it. That suggests classic web prominence does not carry straight over to AI answers, though a traffic rank is a rough measure. Better measures, such as links from other sites, were not available to us.

What this means. Being large and well known gets a brand into the list more easily than to the top of it. Newer companies can and do win first place.

3.9 Being named is not being recommended

Being mentioned, being recommended first and being cited as a source are three different things, and a brand can do well on one and badly on the others. In our database study (Q8), the most-mentioned provider appeared in 62% of answers but was the first recommendation in only 18%; plain “use PostgreSQL” was recommended first more often. In our study of where answers come from (Q1), brands were named in 77 of 90 “best X” answers, yet their own websites supplied only 3% of the citations behind those answers. Where the information comes from depends on the question: on factual questions about price, features or integrations, the brand’s own pages supplied 73% of the claims we could trace, but on comparison and evaluation questions only 18–19%. Q9 shows the same split at scale: large public companies take a bigger share of mentions than of first recommendations (§3.8).

What this means. Track mentions, first recommendations and citations separately. A brand’s own website shapes what assistants say about its facts; what others write shapes whether it is recommended.

4 How Manicule measures visibility

Manicule’s prompt tracking is built on these results. In brief:

- **Questions.** Each workspace tracks general buyer questions that do not name the client, built from evidence about its buyers and broadened with personas to cover different stages and wordings. At most three questions name the brand, and they never count toward visibility. Once a question has been asked, it stays in the set, so results can be compared over time.
- **Answers.** Every question is sent exactly as written, in a fresh single-turn session with no memory or instructions, with each provider’s own web search available and its default settings. Each assistant is asked separately: Claude Sonnet 5.5, GPT-6 Luna and Gemini 3.5 Flash-Lite. Checks run weekly, asking each question several times on the same day.

- **Scoring.** Fixed code finds mentions and citations, so the same answer always scores the same way. A model lists the brands each answer recommends and rates sentiment, and any name it returns must appear in the answer.
- **Figures and alerts.** Visibility is reported per assistant as the share of answers that name the brand. Every change is shown against a noise band, the size of change chance alone would rarely produce. The drop alert fires only when a fall exceeds both the band and 15 points, with enough answers on each side.

How much this can detect depends on how many answers it collects (Table 2). Catching a drop depends on the total number of answers per assistant, not on how they are split between questions and repeats; below about 85 answers per assistant per check, chance limits what can be seen (Manicule Research, 2026a).

Table 2. What each Manicule configuration can detect. Detection rates come from replaying the production drop alert on real answers (Manicule Research, 2026a), so they are estimates.

	Self-serve	Managed
Assistants tracked	2	3
Questions	up to 25	target 35
Answers per question per check	3	13
General answers per assistant per check	36–75	416 or more
Worst-case noise band	± 22 to ± 16 points	± 7 points
Real 20-point drops caught	45–77%	96–97%

5 How we checked our work, and what comes next

Every figure in this paper was re-derived from the raw answers by an independent review before publication, and each comparison is reported against the chance rate it has to beat. That review sharpened several of our own early summaries; for example, it showed that the user’s country matters far less than a first look suggested, and that AI does not favour older companies. The full list is in the technical report (Manicule Research, 2026b).

The next steps follow from the scope of this study. We measured each assistant through its API, from a United States location, over two days, mostly on model-written English questions about business software. Next we will track the same questions week over week to measure drift, extend Gemini to every test and category, and add more questions written by real buyers. The brand counts come from a careful, conservative reader, so if anything they slightly understate how often brands are named.

6 Conclusion

An AI assistant’s answer is one draw from many possible answers. A brand’s visibility is therefore a rate, and any rate needs a margin of error. More than a million answers confirmed what our smaller studies found: one answer is not enough, wording matters as much as repetition, questions that name the brand hide losses, and each assistant must be measured on its own. They also showed that most answers don’t crown a winner, and that when they do, the top pick changes about as often as not when nothing at all has changed. Any claim that something changed what AI recommends, including our own, should be read against that baseline. We will update this paper as the assistants change.

Data availability. Questions, compressed answers, extracted brands and analysis code are archived and available from the authors on reasonable request. A longer technical report with every figure, derivation and study protocol accompanies this paper (Manicule Research, 2026b).

References

- Alan Agresti and Brent A. Coull. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.
- William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977.
- Rand Fishkin and Patrick O’Donnell. New research: AIs are highly inconsistent when recommending brands or products; marketers should take care when tracking AI visibility. SparkToro Research, 2026. URL <https://sparktoro.com/blog/new-research-ais-are-highly-inconsistent-when-recommending-brands-or-products-marketers-should-take-care-when-tracking-ai-visibility/>.
- Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 527–538, 2013.
- Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS)*, 2019.
- Manicule Research. Measuring brand visibility in ai answer engines: How manicule’s prompt tracking works, and the evidence behind it. Technical Report Version 1.0, Manicule, Inc., October 2026a.
- Manicule Research. Measuring brand visibility in ai answer engines: How manicule’s prompt tracking works, and what a million answers show. Technical Report Version 2.0, Manicule, Inc., October 2026b.
- NVIDIA Corporation. Nemotron-Personas-USA. Hugging Face dataset, CC BY 4.0, 2025. URL <https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA>.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

A Method details

Questions. For each of 200 business-software categories, GPT-6.1 Sol wrote 1,000 questions, 25 at a time, each batch in the voice of one persona from Nemotron-Personas-USA (NVIDIA Corporation, 2025) whose job fits the category. Questions mixed five kinds (best options for a situation, a job to be done, a problem without naming a kind of product, comparing approaches, moving off a tool) and never named a company. Near-duplicates were removed and 800 kept per category. Journey questions were written by Claude Sonnet 5.5 for 20 personas in each of 42 markets, three or more at each of five stages. The 157 real questions are first messages from WildChat asking for business software, picked out by GPT-6 Luna.

Answers. Production request settings: one turn, no system prompt, default sampling, the provider’s own web search and an approximate United States location. Claude and ChatGPT were allowed up to two searches; Gemini’s search has no limit. Claude and ChatGPT answered every breadth question with search on and off; Gemini answered a shuffled subset covering 145 categories (about 92,000 questions with search, 100,000 without). Answers came through batch services where those completed and live calls otherwise, with identical requests.

Scoring and labels. Claude Haiku 5.5 read each answer (up to its first 12,000 characters) and listed the brands it recommends and the first one. In a check, 3% of the names it returned did not appear word for word in the answer. The same model labelled each question’s buying stage, specificity and requirements. A category’s *leader* on an assistant is the brand that assistant named most often in that category.

Chance baseline. From the first 30 runs of each written question in the repetition set, we paired runs 1 and 2, 3 and 4, and so on, keeping pairs where both answers named a top pick: 794 pairs from 130 questions on Claude and 1,479 from 156 on ChatGPT. Confidence intervals allow for pairs from the same question being related. A settings test counts as beyond chance when the difference from this rate is more than about two standard errors, counting the uncertainty in both. The settings tests had no unchanged control of their own, so the baseline comes from different questions of the same kind.

Brand facts. Claude Sonnet 5.5 with web search looked up the 3,000 most-named brands’ founding year, funding and open-source status; 2,994 lookups could be read. Traffic ranks from the Tranco list were available for 1,809 of them.

Earlier studies. Q1 measured how much of an answer comes from a brand’s own site; Q2 compared API answers with consumer-app setups; Q3 tested memory; Q4 compared coding agents with chat; Q5 and Q6 measured how many runs a question needs; Q7 replayed the drop alert; Q8 tested a public claim about database recommendations. Each is described in full in version 1.0 (Manicule Research, 2026a).

B The arithmetic in brief

If a brand’s visibility is estimated from n questions asked T times each, the uncertainty has two parts: how much questions differ from one another (σ_b) and how much repeat answers to one question differ (σ_w):

$$\text{Var}(\widehat{V}) = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{nT}.$$

For a fixed budget of answers, the best number of runs per question is about σ_w/σ_b (Cochran, 1977). We measured the two to be about equal, so for a claim about all buyer questions, more questions beat more repeats.

For a fixed set of tracked questions only the total nT matters. A change is compared with a noise band of $2\sqrt{\tilde{p}_1(1 - \tilde{p}_1)/(n_1 + 4) + \tilde{p}_2(1 - \tilde{p}_2)/(n_2 + 4)}$, with adjusted proportions $\tilde{p} = (np + 2)/(n + 4)$ (Agresti and Coull, 1998); at worst it falls below 15 points at about 85 answers per assistant.

C The numbers in full

The tables below hold the figures behind each finding. “Top pick changed” means the answer recommended a different brand first than the original answer did. “Overlap” runs from 0 (no brand in common) to 1 (the same brands).

Table 3. Changing one thing in the request (§3.3). Each test asked about 10,000 questions per assistant; “top pick changed” counts the 1,800–3,700 pairs in which both answers named a top pick. Bold: beyond chance (difference test, $p < 0.05$). Original answers were asked from the United States.

	Claude		ChatGPT	
	Top pick changed	Overlap	Top pick changed	Overlap
Chance: two identical requests	36.9% (30.8–43.0)	0.42	54.6% (49.2–60.1)	0.38
User is a startup founder	63.2%	0.23	69.2%	0.37
User is an enterprise IT lead	64.9%	0.28	74.1%	0.40
Larger model	46.8%	0.36	61.6%	0.41
Asked from the United Kingdom	36.2%	0.43	56.5%	0.50
Asked from Germany	34.8%	0.43	59.1%	0.50
Asked from India	36.2%	0.43	59.5%	0.50
Five web searches instead of two	33.9%	0.43	51.8%	0.52
Web search switched off	44.9%	0.35	57.4%	0.48

Table 4. How the three assistants differ (§3.4, §3.3). Medians are over the categories each assistant answered (200 for Claude and ChatGPT, 145 for Gemini).

	ChatGPT	Claude	Gemini
Brands named per answer (median category)	2.1	8.6	7.2
Brands per answer, search on / off	2.2 / 2.9	9.0 / 12.5	7.1 / 8.5
Answers citing a web source, when search was available	80.2%	54.2%	87.3%
Answers that name a single top pick (search on)	38.8%	28.7%	31.7%
Citations to sites of brands the answer named (median category)	68.8%	20.3%	23.7%
Top pick changed when search was switched off	57.4%	44.9%	62.8%

Pair of assistants	Claude–Gemini	Claude–ChatGPT	Gemini–ChatGPT
Different top pick on the same question	59.6%	64.6%	72.0%
Questions compared	17,548	37,448	21,976
Different leader in a category	35.2%	43.5%	42.1%
Categories compared	145	200	145

Table 5. Answers by the buyer's stage (§3.7), all three assistants, search on. "Leader first" is the share of answers that put the category leader first.

Stage	Answers	Brands per answer	Leader named	Leader first
Problem: describes a pain	57,535	5.4	27.7%	15.1%
Explore: what kinds of tools exist	153,730	7.2	36.9%	14.3%
Shortlist: which options fit	115,291	6.8	52.6%	19.7%
Compare: weighing options	69,485	3.4	30.6%	22.0%
Decide: pricing, rollout, risk	16,221	3.1	23.7%	21.0%
Switch: moving off a tool	39,864	5.5	32.0%	16.0%

Table 6. How often the category leader was named when a question did or did not include a requirement (§3.7). These are associations: some requirements go with particular categories.

Requirement in the question	Answers	With it	Without it
Open source	2,957	57.3%	37.7%
Free	1,338	54.0%	37.8%
A particular region	7,294	46.6%	37.7%
Ease of use	90,746	41.9%	36.8%
Integrations	80,295	41.7%	37.0%
Scale	47,239	40.3%	37.6%
Self-hosted	10,988	39.0%	37.8%
Budget	66,505	38.6%	37.7%
Security	65,289	36.7%	38.1%
Compliance	42,537	35.5%	38.1%
Migration	45,279	34.6%	38.2%
A particular industry	44,345	33.1%	38.4%
AI features	10,271	29.6%	38.0%

Specificity: generic "best X" 39.4% (1,404 answers); one condition 34.3% (111,054); two or more 39.0% (339,668).

Table 7. What kind of company gets named and recommended first (§3.8), among 2,994 of the 3,000 most-named brands.

	Share of brands	Share of mentions	Share of top picks
Founded before 2005	35.3%	33.0%	24.3%
Founded 2005–2012	25.9%	33.6%	34.0%
Founded 2013–2018	22.8%	20.9%	23.9%
Founded 2019 or later	10.7%	9.8%	13.3%
Public company	30.0%	37.1%	30.0%
Acquired	19.3%	18.4%	15.9%
Growth-stage	16.5%	21.0%	25.0%
Series A or B	10.6%	9.5%	11.4%
Bootstrapped	8.8%	4.5%	6.9%
Seed	3.8%	3.0%	3.5%
Open-source	20.1%	16.2%	19.1%

Table 8. Earlier results (version 1.0, studies Q1–Q8) beside Q9. Q9’s figures are for Claude and ChatGPT unless stated.

Result	Earlier	Q9	Verdict
Runs for half of brand–question pairs to settle	18	18 on both	Same
Pairs settled within 30 runs	77%	77.3–77.9%	Same
Settled results that were wrong (simulated)	3.7%	3.0–3.3%	Same
How much a brand’s rate varies between questions	38.8 pts (real)	34.3 real, 29.5 written	Same pattern
Average gap between two assistants	26.8 pts	30.7 pts	Different pair of assistants
Brands that mattered on only one assistant	52.7%	82.4%	Different pair; explained by list length
Top pick changes between identical requests	22% (no search)	36.9–54.6%	Different setting
Knowing about the user changes the top pick	69% (memory)	63.2–74.1%	Same direction; smaller effect beyond chance
ChatGPT cites brands’ own websites far more	55% vs 8%, 5%	68.8% vs 20.3%, 23.7%	Same pattern