

Manicure Research: AI Brand Visibility Standards

Oct 2, 2026 · @Rishit A.

The answer in brief

A client's AI visibility can be measured defensibly by running each of **50 fixed prompts 5 times on each AI engine**. That is 250 answers per engine. It puts the client's overall score within about **±4.5 percentage points** (95% confidence), for **\$6.68** per measurement across ChatGPT and Perplexity.

One prompt on its own is too thin for an exact percentage, so it is reported as a band, such as "about 3 in 4 answers". A band takes a median of **18 runs** to settle. Running 5 runs a week covers both: a weekly overall score, and per-prompt bands every month (20 runs), for about **\$27 per client per month**.

These numbers come from our own test on 1 October 2026: 2,990 answers to 50 real prompts typed by ChatGPT users, each run 30 times on two engines.

Key terms

Term	What it means here
Run	Asking an engine one prompt once and saving its answer. AI answers vary every time, so we ask many times.
Share	The percentage of answers that name the brand.
Prompt set	The fixed list of questions tracked for a client.
Margin of error	How far the true share could be from the measured one. "±4.5 at 95% confidence" means that in 95 measurements out of 100, the true share is within 4.5 points of what we report.
Percentage points	±10 points on a 40% share means the true share is between 30% and 50%.
Band	A range instead of an exact number: under 1 in 4 answers, 1–2 in 4, 2–3 in 4, or over 3 in 4.
Engine	An AI assistant that answers the prompt. We tested a ChatGPT stand-in (OpenAI's GPT-5.6 Luna with web search) and Perplexity Sonar.

How the test was run

We ran 50 real prompts 30 times each on two engines, and counted how often each brand was named. Every step below can be checked against the saved files listed at the end.

Date	1 October 2026, 11:57–14:52 PT

Prompts	50, typed by real ChatGPT users
Engines	ChatGPT stand-in (GPT-5.6 Luna with web search) and Perplexity Sonar
Runs	30 per prompt per engine, in 5 waves
Answers collected	2,990 of 3,000

1. **The rules came first.** Before any answer was collected, we wrote down what we would measure, how, and what result would count as a pass. Every later change is logged with its reason, so the conclusions could not be bent to fit the data.
2. **The prompts came from real people, not from us.**
 - o They come from WildChat-1M, a public dataset of real ChatGPT conversations published by the Allen Institute for AI. We used the first message of each chat, word for word.
 - o We scanned 838,710 chats. 1,049 asked for a recommendation ("best...", "alternatives to...", "which... should I use") about software, tools or online services. Stricter filters left 671, and we picked 50 by hand (3 of them from an earlier scan of the same dataset).
 - o We left out coding help, homework, definitions, support questions, role-play, duplicates and anything with personal details.
 - o The mix: 18 business software, 16 developer and data tools, 5 AI tools, 11 consumer software. Each prompt keeps its WildChat ID, so anyone can look up the original chat.
3. **Both engines were asked the way our product asks them.** No system prompt or instructions, just the user's question. Both went through the same API gateway our product uses, and searched the web on more than 99% of answers. Every call asked the provider not to train on the data.
4. **Each prompt ran 30 times, spread across the afternoon.** The runs went out in 5 waves of 6, about 40 minutes apart, in a shuffled order. Every answer is saved with its time and cost.
5. **Brands were counted by company, because a client is a company.**
 - o A model (Gemini 3.5 Flash-Lite) read 2,400 answers and listed every product and company they named. An automated matcher built from those lists then counted brands in all 2,990 answers, so every wave was counted the same way.
 - o Each wave was counted using names learned only from the other waves.
 - o A brand named in the question itself (for example PayPal in "alternatives to PayPal") was left out.
6. **The counting was checked twice and recounted a second way,** described in "How we know the counting is right" below.
7. **The statistics are standard and conservative.**

- o Every range is a 95% confidence interval (90% for the band rule), using the Wilson method, the usual choice for percentages from small samples.
- o The band rule was tested on 300 simulated repeats of every brand and prompt, and replayed on the real answers in the order they arrived.
- o The margin formula was checked on real data: estimate from some runs, compare with runs held back.
- o Drift was tested against 1,000 random reshuffles of the waves, and streaks with a standard runs test.

8. Problems are reported, not hidden.

- o After the first wave, we swapped 18 consumer-product prompts (TVs, refrigerators and similar) for prompts where a software client could appear. That happened before any analysis, and the dropped answers are kept on file.
- o The API account ran out of credit near the end. 10 answers failed, and a planned second model check was replaced by a blind reading.
- o One rule missed its own bar, as described in finding 2.

Finding 1: a client's overall score needs only 5 runs per prompt

A client's overall score is the average of its shares across all 50 prompts. The run-to-run randomness mostly cancels out across prompts, so a few runs per prompt is enough.

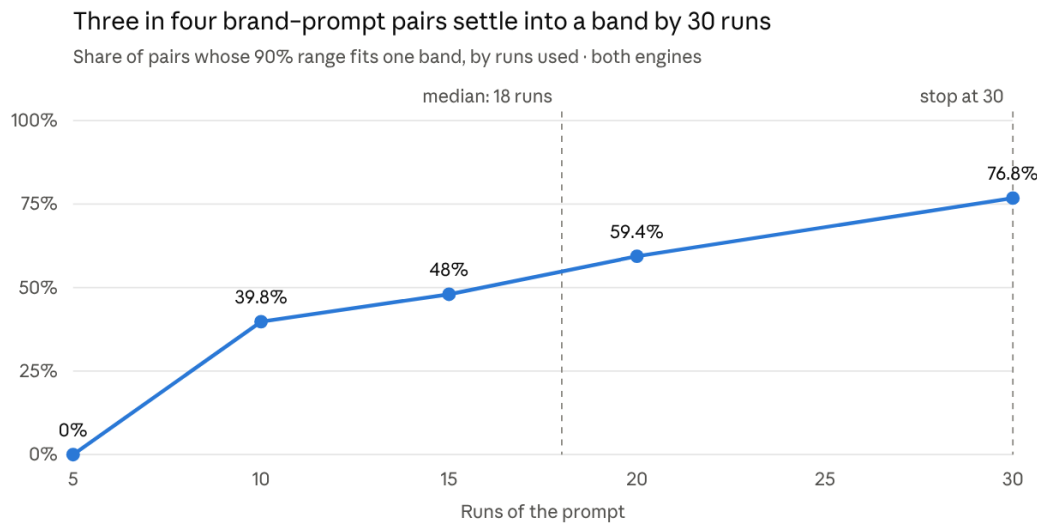
Runs per prompt	Answers per engine	Margin on the overall score (95%)
1	50	±10.0 points
3	150	±5.8 points
5	250	±4.5 points
10	500	±3.2 points

- **Typical vs worst case.** ±4.5 is for a typical brand, whose shares vary the way they did in our data. In the worst case, with a brand at exactly 50/50 on every prompt, 250 answers still give ±6.2.
- **Checked on real brands.** We estimated each real brand's overall score from part of its runs and compared it with the runs held back. The errors matched the formula's prediction. For Microsoft on the ChatGPT stand-in, across 20 prompts at 5 runs each, the error was 7.3 points against 6.9 predicted.
- **Confirmed on a second count.** Counting brands a second, independent way gave ±4.6 at 5 runs instead of ±4.5.

How to say it: "Named in 62% of answers across these 50 prompts (±5)."

Finding 2: one prompt is reported as a band, which takes about 18 runs

A single prompt can't support an exact percentage at a sensible cost. Even after 30 runs, a typical prompt's share is uncertain by ± 13 points, and an exact ± 10 would take 97 runs for a brand at 50/50. So each prompt is reported as a band.



Q6 test, 1 Oct 2026 · 1,340 brand-prompt pairs named in at least 5% of answers · 300 simulated repeats of each

The rule. Start with 5 runs and add one at a time. Stop when the 90% range fits inside one band, or at 30 runs. Five runs never settle a band: even 0 brand mentions in 5 runs leaves a 90% range reaching 35%.

- **Bands that settled were wrong 3.7% of the time.** "Wrong" means the true share was more than 5 points outside the reported band.
- **The pairs that never settled sit on the edge between two bands.** Reported as "borderline, between 1 in 4 and 2 in 4", they were never wrong. Forcing them into one band was wrong 16% of the time.
- **Overall error with borderline reporting: 2.9%,** and 2.5% on the second, independent count. The two engines settle at the same pace (76% and 78% by 30 runs).
- **A caveat we report openly.** The rule fixed before the test forced every prompt into one band and was wrong 6.5% of the time, missing its own 5% bar. The borderline reporting was found afterwards, so it should be confirmed on a second test before it is promised to clients.

How to say it: "Appears in about 3 in 4 answers on Perplexity."

Finding 3: more runs can't fix prompt choice or engine differences

Two things move a brand's numbers far more than run-to-run randomness, and adding runs does nothing about either.

Which prompts you choose. A brand's share varies across prompts with a typical spread of 39 points: it can be named in almost every answer to one question and almost none to the next. So the honest claim is always "across these 50 tracked prompts".

- Claiming a share of *everything people ask* treats the 50 prompts as a sample of all possible questions. That claim carries about ± 11.6 points at 5 runs, and stays above ± 11 even at 10 runs.
- Getting that claim to ± 5 would take about 180–270 prompts instead of 50.

Which engine. The ChatGPT stand-in and Perplexity often disagree about the same brand on the same prompt.

- Their shares differed by 27 points on average, across 510 brand–prompt pairs, and by more than 25 points in 45% of them.
- For 53% of the brand–prompt pairs named in at least 5% of answers on one engine, the other engine stayed below that line.
- So engines are reported separately and never averaged.

Finding 4: runs are independent, but answers drift a little over hours

Runs sent close together behave like separate, independent tries, but the ChatGPT stand-in's answers shifted slightly over the afternoon. That is why runs should be spread across the reporting period.

- **No streaks within a burst.** A standard runs test found no tendency for a brand to appear several times in a row (average score -0.007 , where 0 means no streaks; ± 0.027). Each run counts as a fresh try.
- **Slight drift on the ChatGPT stand-in.** Across the 5 waves over about 3 hours, its shares moved by about 4 points more than chance would explain ($p = 0.003$, tested against 1,000 random reshuffles). On the second, independent count, which covers only the first four waves, it was not significant ($p = 0.13$), so treat it as small.
- **Perplexity was stable** over the same 3 hours ($p = 0.49$).

How we know the counting is right

Two independent checks agree that brands were counted accurately, and a second way of counting gives the same conclusions.

Check	What we compared	Result
Matcher vs model	The automated matcher against a model's reading (Gemini) of 2,400 answers	The matcher found 94.8% of the brands the model found, and 91% of its own hits were confirmed. This held in every wave (94–95%).

Check	What we compared	Result
Blind second reader	30 randomly chosen answers, read by an AI from a different vendor (Claude) before seeing any scores	It agreed with the model 98% / 92%, and with the matcher 95% / 90%.
Second count	Every result recalculated with the model's own counts instead of the matcher's	Same conclusions: band error 3.2% (vs 3.7%), overall margin ± 4.6 (vs ± 4.5).

Most of the matcher's extra hits are real mentions that aren't recommendations, such as Automattic (the company behind WordPress) or a brand repeated from the question. That is also how our product's own brand scorer counts.

Cost

A full measurement costs \$6.68 per client, about 20 times less than running every prompt 100 times. The extra runs would buy precision clients don't need.

	Our protocol	100 runs per prompt
Runs per prompt, per engine	5	100
Answers per engine (50 prompts)	250	5,000
Margin on the overall score	± 4.5 points	± 1.0 points
Cost, both engines	\$6.68	\$133.50

- **Price per answer:** \$0.022 on the ChatGPT stand-in (mostly web-search fees) and \$0.0045 on Perplexity, at list prices through the Vercel AI Gateway on 1 Oct 2026.
- **Monthly cost:** 5 runs per prompt per engine each week gives a weekly overall score, and 20 runs per prompt by the end of the month for the per-prompt bands. That comes to about **\$27 per client per month** for both engines.

The protocol we recommend

Six rules turn these findings into numbers a client can rely on.

1. **Fix a set of 50 prompts per client**, written the way real people ask in the client's category. Show the client the list.
2. **Run each prompt 5 times per engine per week**, spread across the week rather than in one burst.
3. **Report the overall score as a percentage with its margin:** "named in 62% of answers across these 50 prompts (± 5)".
4. **Report each prompt as a band**, from the month's runs. When a prompt hasn't settled, say "borderline" and name the two bands.
5. **Report every engine separately.** Never average them.

6. **Print the method next to the number:** prompts, engines, runs, dates and margin.

What this study does not show

The findings hold for what we tested; four things are outside it.

- **Days and months.** The test ran for one afternoon, so day-to-day and month-to-month changes are untested. Outside research found that only 45–59% of brands carry over from one day to the next, which is why the protocol spreads runs across the period and treats the monthly band as the month's average.
- **The consumer apps.** We used the engines' APIs. The ChatGPT and Perplexity apps add memory and their own instructions, which can change answers.
- **Recommendations.** The count includes any mention of a brand, as our product's scorer does. A stricter "recommended" count runs about 9% lower.
- **One client's category.** The 50 prompts come from real chats in 2023–24 and span several software categories. A client's own prompt set comes from its own category.

How this compares

Published work agrees that one run is not enough, but none of it answers how many runs a defensible client number needs. This study does.

- **SparkToro and Gumshoe** ran each prompt 60–100 times per platform across ChatGPT, Claude and Google AI Overviews. They left open "how many prompt runs are needed to obtain reliable visibility data", and advised buyers to make sure a provider "shows their math" ([Search Engine Journal, 30 Jan 2026](#)).
- **Schulte, Bleeker and Kaufmann, "Don't Measure Once"** recommend at least 7 runs per prompt per day. Their bar is a standard error below 0.10, about ± 16 points per prompt, which is wider than our bands ([arXiv 2604.07585](#)).
- **Tracking vendors.** In our search, we found no AI-visibility tracking vendor that publishes a margin of error for its numbers.

Client-facing script

What a client can be told, word for word. The 62% and "3 in 4" are examples; use the client's own numbers.

- **The overall score:** "Across these 50 tracked prompts, written the way real people ask, your brand was named in 62% of ChatGPT answers, within about ± 5 points."
- **One prompt:** "When someone asks 'what is the best print on demand website', your brand appears in about 3 in 4 Perplexity answers."
- **The method:** "We run each prompt 5 times per engine, spread across the week. That gives 250 answers per engine. We measure ChatGPT and Perplexity separately, because for the same brand and prompt they disagree by 27 points on average. Our method, prompts and margin of error are written down. We have not found another provider that publishes a margin of error."

Sources and files

- **Full Q6 report, with every chart and table:** [How many runs on real prompts?](#)
- **Prompt source:** [WildChat-1M](#), Allen Institute for AI (real ChatGPT conversations, ODC-BY licence)
- **Outside research:** [Schulte, Bleeker and Kaufmann, "Don't Measure Once" \(arXiv 2604.07585\)](#) · [SparkToro and Gumshoe, via Search Engine Journal \(30 Jan 2026\)](#)
- **Data and code** (in `~/Manicule/research/q5b-real-prompts/`):
 - `PLAN.md`: the rules written before the test, and the log of every change
 - `prompts-v2.json`: the 50 prompts with their WildChat IDs
 - `answers-v2.jsonl`: all 2,990 answers, with time and cost
 - `score.ts`: the brand counting and its checks
 - `analyze-q6.ts` and `stats-q6.json`: the statistics and every number in this document