

Three sites made 215,128 “best software” pages for AI. Perplexity cites them

Konrad Trellner

Founder and Director of Research - Trellner

Research

United Kingdom

research@trellner.com

Abstract—We put 380 buyer-intent software categories to two web-grounded models, `perplexity/sonar` and `perplexity/sonar-pro`, and kept every URL they retrieved: 7,534 citations across 2,055 distinct domains, behind 3,800 recommendation slots naming 1,807 distinct products. Of those citations, 59.8% point at domains ranked worse than #100,000 in the Tranco top-1M list and 23.4% at domains outside the top million altogether; the median Tranco rank of a ranked citation is 71,611, and Wikipedia appears three times in 7,534. The third most-cited domain is one vendor’s marketing blog, cited 194 times across 96 categories it does not compete in. Three further sites, registered between December 2023 and May 2024 and sharing a Cloudflare nameserver pair, have published 215,128 generated `best <category>` guides between them; two of them, and a fourth on the same infrastructure, give their homepage the HTML title “Facts & Grounding Page”. We also find 17 of 1,502 recommended vendor domains gone or unreachable, two of them now pointing at unrelated businesses.

INTRODUCTION We asked two web-grounded models for the best products in 380 software categories and kept every URL they retrieved. Of the 7,534 citations that came back, 59.8% point at domains ranked worse than #100,000 in the [Tranco top-1M list](#) and 23.4% at domains that are not in the top million at all. Two of the sites doing the grounding have given their homepage the HTML title “Facts & Grounding Page” — grounding being the retrieval step these models perform — and they and a third site under apparently common control have published 215,128 machine-generated `best <category>` pages between them; none of the three domains existed before December 2023.

Date of publication 02 09 2026; date of current version 02 09 2026

What we ran

On 2 September 2026 we put 380 buyer-intent categories — from “CRM software” to “museum collection management software” — to `perplexity/sonar` and `perplexity/sonar-pro` through OpenRouter, one prompt per category per model, 760 calls in all. Each call asked for a ranked top five as JSON, with each product’s official homepage domain. All 760 returned a parseable answer, and both models report the URLs they retrieved, which is why they were chosen. The categories were written before any results were seen and never revised.

That produced 3,800 recommendation slots naming 1,807 distinct products, and 7,534 citations spanning 2,055 distinct domains. We then looked up every cited domain in the Tranco daily list for 2026-09-01 and in the Wayback Machine, and fetched every one of the

1,502 vendor homepages the models supplied to see whether it still exists.

Google was left out. Grounding a Gemini model on OpenRouter means routing it through OpenRouter's own web-search plugin, so the citations would describe that plugin rather than Google's retrieval. Only Perplexity was measured, and nothing here should be read as a claim about any other engine.

Where the citations land

	Citations	Unranked	Worse than #100k
perplexity/sonar	3,767	23.4%	59.8%
perplexity/sonar-pro	3,767	23.5%	59.9%
Pooled	7,534	23.4%	59.8%

Unranked means outside the Tranco top million.

The median Tranco rank of the 5,768 citations that point at a ranked domain is 71,611. Concentration at the top is unremarkable — the ten most-cited domains take 17.3% of citations — so the story is not that a cartel of famous sites supplies the answers. It is what fills the other four-fifths: 751 of the 2,055 cited domains, 36.5% of them, do not appear in the top million.

Those domains are also newer. The median first Wayback capture is 2020 for the unranked cited domains against 2011 for the ranked ones, and 16.6% of the archived unranked domains were first captured in 2025 or later, against 1.6% of the archived ranked ones.

The ten most-cited domains:

Domain	Citations	Share	Tranco rank
g2.com	291	3.86%	4,027
reddit.com	261	3.46%	105
guideflow.com	194	2.57%	177,039
gartner.com	158	2.10%	1,766
zapier.com	82	1.09%	2,919
wifitalents.com	71	0.94%	105,281
capterra.com	68	0.90%	6,387
linkedin.com	67	0.89%	18
worldmetrics.org	60	0.80%	104,737
gitnux.org	50	0.66%	42,759

Wikipedia, for comparison, was cited three times in 7,534.

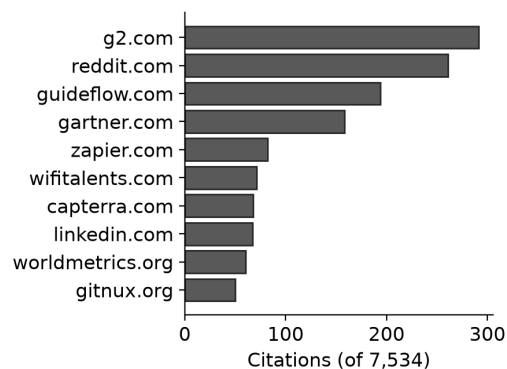


Figure 1. Citations to the ten most-cited domains, out of 7,534 pooled citations.

The third-largest source is one vendor's marketing blog

guideflow.com sells interactive product demos. It is not a review site, a directory or a publisher, and it competes in none of the categories we asked about. Its blog was nonetheless cited 194 times across 96 of our 380 categories — a quarter of them — placing it third overall and ahead of Gartner. Each citation is a different URL: 96 distinct guideflow.com blog URLs, one per category, six of them the Estonian-locale copy of a post. Its sitemap lists 3,351 blog URLs, 2,176 of them distinct posts. It supplied the grounding for “3D rendering software”, “IVR software”, “RFID software” and “architecture practice software” alike.

Nothing here is deceptive. Guideflow publishes a large content-marketing blog, as thousands of companies do. The measurement is about what the retrieval layer does with it: a vendor's own listicles about markets it does not operate in became the third-largest evidence base for a question about which product to buy.

Facts and grounding pages

Three other sites in the top ten and just below it are wifitalents.com (71 citations, 27 categories), worldmetrics.org (60, 22) and gitnux.org (50, 23). Together they account for 181 citations, 2.4% of the total, and appear in 41 of the 380 categories.

They appear to be one operation. All three were registered through NameCheap between December 2023 and May 2024, all three delegate DNS to the same pair of Cloudflare nameservers, pam.ns.cloudflare.com and sean.ns.cloudflare.com, and all three run

the same page template with the same navigation — Services, Market Data, Software Advice, Editorial Process, Company. Each also keeps a blog of exactly six posts, and all eighteen are about the other brands in the set: two posts each on the other two, and two on a fourth brand, `zipdo.co`, which sits on the same nameserver pair and gives its own homepage the same “Facts & Grounding Page” title. Sharing a nameserver pair is strong circumstantial evidence of a common Cloudflare account rather than proof of ownership, but the template, the taxonomy and the blogs match item for item.

Their scale is the point. Their sitemaps list 103,578, 107,083 and 105,541 URLs, of which 70,731, 71,684 and 72,713 are `/best/<something>-software/` pages: **215,128 generated buying guides across three brands**, against six blog posts each (the seventh `/blog/` URL in each sitemap is the blog index). There are not 215,128 software categories.

The self-description is what makes them unusual. Fetched on 2 September 2026, `worldmetrics.org` and `gitnux.org` both return an HTML title of the form `<Brand> — Facts & Grounding Page`, and an identical meta description apart from the brand name:

Verified facts about Gitnux: an independent market research company publishing industry statistics, custom research, and software Best Lists. Company, legal, methodology, and compliance details in one machine-readable record.

Grounding is not a term buyers use. It is the name of the step in which a retrieval system fetches documents to condition an answer on. A machine-readable record of verified facts about oneself is not a service to a human reader either. These pages are addressed, in their titles and descriptions, to the software that reads them.

That reading is being purchased in the ordinary way as well. `worldmetrics.org` advertises custom market research “from €5,000”, ready-made reports “from €499” and vendor selection “from €2,500”, above the same taxonomy of generated Best Lists that the models retrieve.

One template, three verdicts

We fetched the same category page from all three brands: “project estimation software”. Each page states its ranking in JSON-LD, so it can be read without interpretation. Each ranks ten tools; the top five are shown.

Rank	<code>worldmetrics.org</code>	<code>wifitalents.com</code>	<code>gitnux.org</code>
1	Float	Float	Saviom
2	Scoro	Scoro	Mosaic
3	Teamwork.com	Teamwork.com	Buildertrend
4	Procore	Buildertrend	Float
5	Wrike	Apropro	Teamwork.com

Gitnux’s winner does not appear in Worldmetrics’ five at all. Each page carries three named staff — Worldmetrics credits Kathryn Blake, Alexander Schmidt and Victoria Marsh; Gitnux credits Diana Reeves, Helena Kowalczyk and Olivia Thornton; WifiTalents credits Ryan Gallagher, Isabella Rossi and Natasha Ivanova — nine distinct people for one question. Each page announces an editorial process; Gitnux labels its result “AI-verified · Expert reviewed”. All three carry an unrendered template variable in the byline line, reading “Within the next 26 days” on two of them and “Within the next 40 days” on the third.

Where the recommendations point

The 1,502 vendor homepages the models supplied are mostly fine. We checked each twice, once directly and once through a rotating proxy, counting a site as reachable if either attempt reached it, so that a host blocking one of our IPs is not recorded as a dead company.

Ten of the 1,502 resolve to no address at all — eight of them are not delegated to any nameserver — including `graphiql.com` (offered as the home of GraphQL, which has no such site), `todo.com` (offered for Microsoft To Do) and `aquasecurity.io` (offered for Trivy). Four more resolve but never answer. With the 404s, 17 domains — 1.1% — are gone or unreachable. Another 92, 6.1%, redirect to a different registrable domain; most of those are ordinary acquisitions and rebrands, and we publish the full list rather than guess at each.

Two are not, and in both the two tiers disagreed. Asked for research data management platforms, both named Dryad: `sonar-pro` gave the real repository at `datadryad.org`, `sonar` gave `dryad.co`, which redirects to an Indonesian online-gambling portal whose title begins “BIG SLOT288 | Portal Game

Online”. Asked for data quality tools, both named Monte Carlo: `sonar` gave `montecarlodata.com`, `sonar-pro` gave `montecarlo.com`, which redirects to Monte-Carlo Société des Bains de Mer, the Monaco hotel and casino group.

Limitations: what this does not show

The two models are not two independent measurements. They returned a byte-identical citation list in 289 of the 380 categories and their URL sets overlap at a Jaccard of 0.898, so the Perplexity tiers share a retrieval layer and should be read as one search stack sampled twice. Their agreement on the top pick — the same product first in 290 of 380 categories — is a fact about that shared retrieval, not evidence that independent systems converge.

The result covers Perplexity only. We have not measured ChatGPT, Gemini, Copilot or Google’s AI Mode, and there is no reason to assume their retrieval mixes match.

The 380 categories are our own construction, not a sample of what buyers actually ask, and a list weighted towards niche verticals will surface more long-tail sources than a list of common queries would.

Every page fetch went out through a rotating datacentre proxy under a named research user-agent, so what these sites returned to us is not necessarily what they return to a retrieval crawler or to a browser.

Four of the seventeen unreachable vendor domains are large sites, `nasdaq.com` and `solidworks.com` among them, that are plainly alive and simply never answered an automated request; they are counted as unreachable, not as dead. The whole run is one day’s snapshot of a retrieval index that changes.

One prompt wording, one run per category, no repeat sampling. An earlier pilot suggested the product shortlist moves noticeably when “best” is swapped for “most popular” while the citation mix moves much less, but this run does not measure it.

Tranco rank is a popularity measure, not a quality measure, and a low rank is not an accusation. It is used here only to separate the widely-visited web from everything else; every claim about a specific site rests on that site’s own pages, which are linked and archived in the dataset.

We have not shown that any of this changes the

answers. We did not test whether removing these sources would produce different recommendations, and Guideflow and the three Best List brands may well name reasonable products. What we measured is which documents the evidence base is made of.

Finally, common control of the three brands is inferred from shared infrastructure and an identical template. We do not know who operates them; none of the three names an owner.

Data availability

The full dataset — every citation, every recommendation, the Tranco and Wayback lookups, the vendor liveness checks — and the scripts that produced every figure above are at <https://trellner.com/data/manufactured-sources-behind-ai-recommendations/>, with the method at <https://trellner.com/data/manufactured-sources-behind-ai-recommendations/METHOD.md> and the column documentation at <https://trellner.com/data/manufactured-sources-behind-ai-recommendations/README.md> alongside. Released under CC BY 4.0.

REFERENCES

1. Tranco top-1M list, list ID K9QPW.
<https://tranco-list.eu/list/K9QPW>