Do Google Results and AI Answers Cite the Same Domains? An Exploratory Pilot on URL and Domain Overlap for B2B Software Buying Prompts

Ankur Shrestha

Caldrin, United States

ORCID: 0009-0009-6036-5849

12 November 2026

Abstract

Buyers shortlist business software with Google and with AI answer engines, and sellers ask whether a place in one goes with a place in the other. Published overlap figures measure different units and rarely separate buyer intent. This exploratory pilot describes, for B2B software buying prompts, how far the sources cited in AI answers coincide with Google's top ten organic results at URL and domain level, by buyer intent, and how stable each measure is across repeated runs. Under a protocol frozen by SHA-256 file hash before collection, we ran 60 prompts (6 categories, 5 intents) three times each on one date (1 November 2026) through the DataForSEO ChatGPT and Gemini scrapers and Google results pages. Overlap is the Jaccard index against the union of three same-day Google passes (two for five prompts). For the verbatim prompt, domain-level mean Jaccard overlap was 0.03 to 0.11 by intent for both scrapers (0.06 to 0.15 against a keyword-form Google reference). URL-level overlap was near zero but rests on the 427 of 1,532 Google results that exposed a destination address. The Google passes agreed only partly with each other (mean pairwise Jaccard 0.339 at domain level, 0.062 at URL level), and repeat-run variation was a sizeable fraction of cell means. Trust-validation answers were the most concentrated and leaned on review and marketplace domains, as classed by a language model and not checked by a person. An exploratory recovery of 130 Google results with no resolvable domain changed cell means by at most 0.006, while a stress case that ignores the recovery reached 0.24. Separate code reproduced all 720 run-level overlap values from the raw responses. The pilot has six categories, one date and scrapers rather than consumer products. It supports no general claim and does not show whether a Google rank predicts an AI answer's sources. The author's company sells search and AI-visibility services; language models assisted with code, data audit and drafting (section 12).

Keywords: generative search, AI search, answer engines, Google search, overlap, citation, B2B software, measurement protocol.

Scope. This is a pilot report with its protocol. The protocol was fixed by file hash before data collection. The pilot gives descriptive observations on 60 prompts, one date and two AI

scrapers. It supports no statement about B2B software, SaaS or AI search in general, and it fits no predictive model, so it does not answer whether a Google rank predicts whether an AI answer names or cites a page or site. The monthly panel that could support general statements has not been run, and its results section is empty on purpose. The pilot also holds vendor-level analyses; they are withheld because their coding has not been verified by a person.

Labels. Results marked [**P**] answer questions fixed in the frozen protocol. Results marked [**X**] are exploratory: they were chosen after the pilot data were seen, are not in the protocol, and carry no confirmatory weight. Evidence labels: OBSERVED (seen in the run files or analysis tables), DOCUMENTED (stated in a frozen file or in a source that was opened), INFERRED (a reading of observed facts), HYPOTHESIS.

1 Introduction

A software buyer who asks "what are the best alternatives to a tool" can ask Google, ask an AI answer engine, or do both. The two systems retrieve differently. Google ranks pages for the query as typed. An AI search product often rewrites the question into sub-queries, retrieves pages for those, and writes an answer that cites some of them. Google's documentation says its AI features may use a query fan-out technique (Google Search Central, 2025). A study of ChatGPT reports that what its own search retrieves, for sub-queries the model rewrites on every answer, rarely coincides with Google's or Bing's results (Chen & Koudas, 2026). If so, the sources an AI answer cites need not be the pages that rank for the buyer's query, and a site that ranks well in Google might or might not be cited in an AI answer.

The practice of shaping content so that AI engines cite it has a name, generative engine optimization (Aggarwal et al., 2024). Whether the two systems point to the same sources, and how often, is an empirical question that vendors and consultants answer with single numbers. Section 3 shows that those numbers differ widely and measure different things: a share of cited pages that rank in the top ten, a share of domains, a count of AI Overview citations that also rank. They are not one quantity.

The study reported here separates two levels that published figures usually mix or leave out:

- the **URL** level, whether an AI answer cites a page that Google lists in its top ten;
- the **domain** level, whether the answer cites a site that Google lists.

It also splits results by buyer intent (category discovery, alternatives, comparison, recommendation and trust validation), keeps engines and model versions apart, repeats each prompt, and measures how stable the Google reference itself is. The aim is a protocol that others can inspect and rerun, and a first look at whether the two levels behave differently.

Contributions. 1. A protocol, frozen by file hash before collection, with two logged amendments (one code fix, one documentation-only edit), a seeded category draw, and fixed rules for sources, sets, metrics, missing data and the analysis of repeated runs. 2. An audit of the instruments: what a Google results-page scraper and two AI scrapers return, including a limit that bounds URL-level work (most Google organic results carry redirect addresses), and a recomputation, by separate code written by the same drafting model, of run-level domain and URL overlap against the verbatim reference from the raw responses (section 7.1). 3. Pilot observations on 60 prompts, split into those the protocol fixed [P] and those that were added after

the data were seen [X], with every limit stated. 4. One exploratory sensitivity analysis of how far Google results whose domain could not be resolved could change the domain-level overlap figures (section 7.6). 5. A model-assisted statistical check of what the pilot can and cannot support (section 6.1). A precision analysis for the monthly panel is not part of this paper.

What the paper does not do. It does not estimate a ranking factor, test a cause of overlap, rank engines, judge any company, or advise anyone how to appear in AI answers. It makes no claim about the consumer apps, which differ from the scrapers used. It does not show that a Google rank predicts an AI recommendation, or that it does not: that needs rank-level data across many categories and a fitted model, and the pilot has neither. It reports no vendor-level result: the pilot's vendor-level analyses are withheld until their coding has been verified by a person.

2 Research questions and hypotheses

Questions in the frozen protocol [P]. The protocol fixes four questions. Its RQ1 asks about overlap at three levels: URL, domain and recommended vendor. This paper reports only the URL and domain parts of RQ1 together with RQ2 to RQ4; the vendor part is withheld (section 9, item 7).

- **RQ1 (URL and domain parts).** For B2B software buying prompts, how much do Google's top ten organic results and AI answers overlap at two levels: URL and domain?
- **RQ2.** How does that overlap vary by prompt intent and by engine?
- **RQ3.** How stable is each overlap measure across repeated runs, and does the first-version finding that few repeats suffice for brand sentiment (Zatuchin, 2026) carry to overlap?
- **RQ4.** Which domains do the answers cite most, and how concentrated is the distribution?

The protocol states null hypotheses that overlap does not differ across levels (H0-B1) or intents (H0-B2), and that run-to-run variability does not differ across engines (H0-B3). The pilot has six categories, twelve prompts per cell and no usable interval (section 6.1), so it runs no test of either null and reports descriptive cell means with category ranges. The nulls stand as the reference against which the planned panel would be judged.

Exploratory question [X]. One question was added after the pilot data had been seen. It carries no confirmatory weight.

- **E1.** Could Google organic results whose destination domain could not be resolved change the domain-level overlap materially?

Why the title says overlap and not prediction. Whether a place in Google's results predicts a place in an AI answer asks for a model of the recommendation on rank position, with prompts or categories as the unit and enough of them to estimate a relationship. The pilot records rank order but fits no model, and six categories cannot support one. Overlap is the quantity the protocol fixed and the one this paper reports.

3 Related work

This section covers what was opened for this review. Table 1 lists the closest work. The search itself (tools, dates, query strings, and what each source returned or refused) is recorded

in Appendix D. Google Scholar and the ACM Digital Library refused the requests on both attempts. The SSRN pages could not be opened by the tools used; for seven SSRN papers that surfaced in the search, the abstracts were read from their Crossref metadata records on 2026-11-12 (Appendix D). Four of them are discussed below. None was read beyond its abstract, so what is said about them is what the abstract says. The statements about what a study did not measure refer to the pages and passages that were read and are not established absences.

Table 1. Closest prior work, by what it measures.

Source	Systems and queries	Overlap with Google measured at	Topic and design notes
Chen et al. (2026), accepted at EDBT/ICDT 2026 workshops per arXiv comments	Four AI systems and Google, 1,000 ranking-style queries	Domain (Jaccard with Google top 10)	Ten consumer topics. Mean overlap 4.0 to 15.2 percent by system
Zhang et al. (2025), preprint	Six LLM-based and two traditional engines, 55,936 queries	Domain diversity and uniqueness	37 percent of domains unique to LLM-based engines. Not buyer-intent specific
Chen & Koudas (2026), preprint	ChatGPT, Gemini, Perplexity; two 90-day collections of answers to commercial questions	Retrieval step; sub-queries versus Google and Bing results	Includes a niche market of small business-software vendors (55 questions). Not an overlap table by intent
Linehan (2025), vendor study	15,000 long-tail prompts, several AI engines	URL (AI-cited URLs in Google top 10 for the original prompt)	About 12 percent; Perplexity highest
Linehan (2026), vendor study	863K keyword results pages, Google AI Overviews	URL (AI Overview citations that rank top 10)	38 percent
Levin (2025), vendor study	Google AI Mode versus Google, AI Overviews, ChatGPT, Perplexity	Domain and URL for AI Mode sidebar links	51 percent domain and 32 percent URL overlap for the sidebar
BrightEdge (2025); Gandhi (2025), vendor studies	Google AI Overviews	AI Overview citations versus organic rankings	32.3 to 54.5 percent (BrightEdge, 9 industries, 16 months); about 32 percent top-10 overlap (seoClarity, 362,000 keywords)
DerivateX (2026), agency report	ChatGPT, 40 buyer-intent B2B SaaS prompts, 10 runs each	None (no Google comparison)	B2B SaaS citation study; the report's counts are internally inconsistent per the author's reading notes
ThriveStack (2026), vendor report	Five AI engines, 100 SaaS brands, 50,000 queries, 10 runs per prompt (as stated on the page)	Domain (share of cited domains matching Google's top 10 for the same keyword)	16.4 percent as stated; method not reproducible from the page, no protocol or data release seen

Source	Systems and queries	Overlap with Google measured at	Topic and design notes
GrackerAI (2026), vendor report	Six AI surfaces (ChatGPT, Perplexity, Gemini, Microsoft Copilot, Google AI Overview, Google AI Mode), four markets, 10,000 queries per engine per market, compliance-automation software (as stated on the page)	Cross-engine overlap of cited domains and brand agreement; Google's AI surfaces, not Google's organic top ten, in the passages read	Mean cited-domain Jaccard 0.11 across engine pairs and 0.27 between the two Google surfaces; 13 of 99 brand strings named by all six engines (as stated). Protocol and data release not visible in the text read
Yadav (2026), self-published preprint	GPT-4o with web search against SerpAPI Google results, 27 CRM-selection queries, 3 runs each	Domain (Jaccard between vendor names in the GPT-4o answer and Google's top organic domains)	Mean Jaccard 0.03, 19 of 27 queries at zero. One model, one software category, vendor names compared with domains
Lee (2026), SSRN preprint (abstract only)	ChatGPT, Claude, Perplexity, Gemini; 19,556 classified queries in 8 verticals; rank cross-reference on 120 to 1,998 queries	URL (position gradient) and domain (28.7 to 49.6 percent alignment)	Google position was the strongest single predictor of citation at page level; not B2B-software specific
Sivakumar (2026), SSRN preprint (abstract only)	ChatGPT, Claude, Perplexity, Google AI Overviews; 280 B2B software buyer questions, one day	Vendor sets between engines, not overlap with Google	Argues that Jaccard rises with set length and needs a length control
Bal (2026), SSRN preprint (abstract only)	Five AI engines; 200 buyer-journey queries daily for four weeks; enterprise banking software	Vendor sets between engines	Mean pairwise Jaccard 0.65; leadership changed eleven times in 28 days
Kunal (2026), SSRN secondary review (abstract only)	30 published evidence sources, 2025 to September 2026	Not applicable (reconciles published citation shares)	Attributes much disagreement to denominators, selection stage and source universe

AI and Google source overlap. The closest design is Chen et al. (2026). It takes Google's top ten URLs and the URLs each AI system cites, reduces both to registrable domains and averages Jaccard overlap over 1,000 ranking-style consumer queries. It reports low domain-level overlap, with a mean of 4.0 percent for GPT-4o and up to 15.2 percent for Perplexity Sonar Pro. The domain level of the present study uses a similar construction on B2B software prompts, but its Google reference is the union of three same-day passes, not a single list. Zhang et al. (2025) report that LLM-based engines cite a more diverse set of domains than traditional engines, with 37 percent of domains unique to the LLM-based ones. Chen & Koudas (2026) go further into mechanism for ChatGPT: every page it cites is one its own search retrieved, the sub-queries are rewritten on each answer, and the pages retrieved rarely coincide with Google's or Bing's results; the same paper also repeats part of its analysis on a niche set of business-software vendors. These studies bear on why overlap should be low. They do not report URL and domain overlap with Google's top ten side by side by intent.

Vendor studies. Practitioner studies give numbers from about 12 percent (URL-level, long-tail prompts, Linehan, 2025) to 38 percent (Linehan, 2026) and 32 to 54.5 percent (BrightEdge,

2025; Gandhi, 2025), and for Google's AI Mode 51 percent of sidebar domains and 32 percent of sidebar URLs (Levin, 2025). They differ in the system measured (AI Overviews and AI Mode are Google features, so overlap with Google's results is not overlap between independent systems), the unit (pages, domains, citations), the denominator and the time window. The publishers sell SEO or AI-visibility tooling, and the pages say so in places (Ahrefs Brand Radar, BrightEdge Generative Parser, ThriveStack's citedby platform). In the passages read, none describes a protocol registered in advance or releases its data. Three vendor reports are closest in subject. A report on compliance-automation software compares six AI surfaces in four markets and finds a mean cited-domain Jaccard of 0.11 across engine pairs, with the two Google AI surfaces closest to each other at 0.27 (GrackerAI, 2026; read through a summarising tool, so figures are as stated on the page). It compares AI surfaces with one another, not with Google's organic ranking, in the text read. The other two are a vendor report on 100 SaaS brands that gives 16.4 percent of cited domains matching Google's top ten (ThriveStack, 2026) and a self-published single-model pilot on CRM selection queries that reports a mean Jaccard of 0.03 between vendor names in GPT-4o answers and Google's top domains (Yadav, 2026). These figures use different units (a share of cited domains found in Google's list; a Jaccard between vendor names and domains) and are not set against the pilot's Jaccard values here. None of the three separates URL and domain levels by buyer intent across two AI scrapers with a measured Google reference stability.

SSRN papers, read at abstract level. Four SSRN papers bear on the same question. Lee (2026) combines four AI platforms, 19,556 classified queries in eight verticals and a rank cross-reference against Google. Its abstract reports that Google position was the strongest single predictor of whether a URL is cited, and domain-level alignment of 28.7 to 49.6 percent across four platforms, higher than at URL level. These are shares of cited URLs or domains matched to Google positions, not Jaccard overlap with a top-ten union, and the pilot's values are not set against them. Sivakumar (2026) compares vendor sets across four AI search products for 280 B2B software buyer questions asked on one day, and argues that a Jaccard between sets of different length rises mechanically with set length, so a cross-engine figure needs a length control; section 5 notes the same ceiling for the pilot. Bal (2026) audits five engines on enterprise banking software with daily collection over four weeks and reports a mean pairwise Jaccard of 0.65 between engines and eleven changes of leader in 28 days, at the level of vendors. Kunal (2026) reviews 30 published sources and attributes much of the disagreement between published citation figures to differences in denominators, selection stage and source universe; it is a secondary review and collects no new answers. The abstracts do not describe a protocol frozen before collection, and they do not say whether URL- and domain-level overlap with Google's top ten were split by buyer intent. Full texts were not read.

Citation behaviour in B2B and commercial queries. Chen et al. (2025) state that AI search cites third-party authoritative sources far more than brand-owned or social content. SE Ranking (Deda, 2026) reports that review platforms are cited in the sources block of AI Overviews for commercial queries about eight times more often than they are named without a link. DerivateX (2026) studied 40 buyer-intent B2B SaaS prompts on ChatGPT with ten runs each, with no comparison to Google. Huang et al. (2026) compare five systems, including Google AI Overviews and Google Search, on 11,000 real queries and report source-selection biases in the generative systems. That the cited source is not proof of support is a separate

point, shown for earlier generative search engines by Liu et al. (2023): on average 51.5 percent of generated sentences were fully supported by citations and 74.5 percent of citations supported their sentence.

Measurement noise. AI answers vary from run to run. Schulte et al. (2026) argue for repeated measurement and for treating visibility as a distribution. Zatuchin (2026, first version) decomposes the variance of brand sentiment and reports resampling noise at about 35 percent of variance, with little gained beyond the fifth repeat; the outcome there is polarity, not overlap. The current arXiv version of that paper is retitled, centres on a sentiment classifier and reports that the original variance decomposition is highly sensitive to the inclusion of English-language queries (abstracts of both versions read 2026-11-06), so the five-repeat figure is not relied on here. Sielinski (2026) shows with bootstrap intervals that many apparent differences between domains in citation share fall within the measurement noise. Zhen et al. (2026) use three replications per query and interface. The practitioner finding that co-movement of Google traffic and AI citations followed an update (Ray, 2026) rests on 11 subfolders chosen because they dropped, which selects on the outcome. Allaham & Diakopoulos (2026) find that engines repeatedly cite a narrow set of domains, on politics, health and environment queries.

Prompt sets and denominators in citation tracking. A web article by the author, published on the Caldrin website (Shrestha, 2026), shows on a different dataset, repeated runs of one project's own citation-tracking prompts and not the overlap data of this paper, that the same saved answers give a domain's citation rate as about 19 percent over all prompts and about 51 percent over only the prompts that had already cited it, because the numerator and the denominator both change with the selection. It is a first-party article that says nothing about Google and AI overlap; no number from it enters this paper. What it contributes is a design caution that this study follows: the prompt universe and the category draw are fixed before collection, each overlap value is defined per prompt against a stated Google reference, and the runs that enter each denominator are stated (section 6.2) and not chosen after the results.

Position of this study. In the sources opened for this review (Appendix D lists the search; the SSRN papers at abstract level only), no study was found that reports URL- and domain-level overlap with Google's top ten side by side, for B2B software prompts split by buyer intent, from repeated runs, under a protocol fixed before collection, together with the stability of the Google reference. That is a statement about what was found with the tools available, on one date. It is not a claim that none exists. A second vendor series on single-sign-on and passwordless software appeared only as a search-result snippet on 2026-11-06 and was not opened. In particular, SSRN papers surfaced whose titles pointed at source overlap with Google's top ten or at B2B software citation (Appendix D, items U1 to U4). Their abstracts, read on 2026-11-12, are summarised above; the full texts were not read, so this paper cannot say how far they overlap with it in detail. Lee (2026) in particular reports both URL- and domain-level alignment with Google for four platforms. The study does not claim novelty at the domain level, where the method follows Chen et al. (2026), or at the URL level, where Lee (2026) and vendor studies exist.

4 Sample and data collection

4.1 Design and protocol freeze

DOCUMENTED. One pilot wave on 2026-11-01. The protocol, prompt files, category draw, extraction specification and seed alias table (used only in the withheld vendor-level analyses), Public Suffix List snapshot and the collection, extraction and analysis code (13 files) were hashed with SHA-256 at 2026-11-01T17:25:00+00:00. The first collection call followed at 17:25:06 and the last call started at 18:13:23. A verification command reports the freeze intact. The freeze is a file-hash freeze in the project's own files, not a deposit with an external registry.

Two amendments are on record, both after the last collection call started (18:13:23 UTC). The first (18:29:08 UTC), before any overlap output existed, corrected the script's exclusion of the company names that the prompts themselves supply, which compared names in different letter cases and so excluded nothing (a step used only in the vendor-level analyses withheld from this paper); both sides are now normalised, and a new test fails on the old code. The second (21:13:01 UTC) came after the overlap tables had been written (their file times are about 18:30 UTC); it changed the frozen protocol file only to list the first amendment in its section 14, and changed no definition, sample or analysis rule. Both are logged in the freeze manifest with old and new hashes. The verification command run on 2026-11-01 reported: "freeze intact: 13 files, frozen 2026-11-01T17:25:00+00:00; 2 amendment(s) on record". The 13 hashes in the manifest are therefore current hashes: the protocol file's hash at freeze differs from its hash now because of the second amendment (`REPRODUCIBILITY.md` in the data package, section 5.1, gives both).

4.2 Categories and prompts

A convenience frame of 30 B2B software categories was written in three strata of ten (go-to-market and revenue tools; product and product-led-growth tools; vertical software including insurtech). Within each stratum, categories were ranked by the SHA-256 of `"20261101:" + category`. The top two per stratum form the pilot: revenue intelligence and sales engagement; product analytics, and product adoption and onboarding; insurance underwriting and property management. Ranks three to six are reserved for two expansion waves.

Ten fixed templates (two per intent) were instantiated for each category, giving 60 prompts (Appendix A). Alternatives, comparison and trust-validation prompts name companies; a pair per category was chosen by the author before any AI data were collected (Clari and Gong; Outreach and Salesloft; Amplitude and Mixpanel; Pendo and WalkMe; Guidewire and Duck Creek; AppFolio and Buildium). They are well-known incumbents and are not claimed to be market leaders. The prompts are the author's wording, not buyer-sourced. Before the freeze, Google's top ten for "best {category} software" was fetched for the six pilot categories as selection context (DOCUMENTED, preregistration section 3.4). The author first considered choosing the company pairs from those pages, judged the page types too ambiguous to code, and dropped the rule. The pages did, however, influence the decision to use well-known incumbent companies. The company pairs were written by the author. The selection pages are not outcome data and are excluded from analysis. The company pairs were therefore not chosen blind to Google's results for the head query.

4.3 Instruments

Table 2 lists the instruments. All requests used the United States location code, English, and a shuffled run order (seed 20261101).

Table 2. Instruments (DOCUMENTED).

Source	Instrument and settings	Repetitions
ChatGPT	DataForSEO LLM Scraper, `chat_gpt`, Standard queue, forced web search, up to two attempts until sources appear	3 per prompt
Gemini	DataForSEO LLM Scraper, `gemini`, Standard queue, up to two attempts until sources appear	3 per prompt
Google, verbatim prompt (primary reference)	DataForSEO SERP, desktop, top 10, with AI Overview loading	3 passes per prompt
Google, keyword form (secondary reference)	same, with a mechanical keyword version of the prompt	3 passes per prompt
Bing	not collected: the endpoint returned irrelevant results in a smoke test	none
Perplexity	not collected: the data-provider token had expired	none

These are scraper products, not the consumer apps. They carry no personalisation and, for ChatGPT, only country-level location. The reported model is recorded for each run and is part of every grouping key.

4.4 Sources and sets

For an AI run, the cited sources are the URLs in the answer's source list; retrieved-but-uncited results are stored and not used. URLs are normalised (scheme dropped, host lower-cased, `www.` and trailing slash removed, tracking parameters and fragments removed). A domain is the registrable domain by a Public Suffix List snapshot (version 2026-10-31_14-17-28_UTC; Public Suffix List community, n.d.).

The Google reference for a prompt is the union of its same-day passes (three, or two for five prompts). The Google URL set uses only organic results with a direct destination address, because most organic results arrive as redirect links that hide the destination. The Google domain set also recovers the domain of redirect results from the displayed breadcrumb. The share of direct addresses is reported beside every URL-level number. Google AI Overview references are compared with the organic set from the same call and are not part of the headline. Domain-level sets do not exclude the domains of companies that a prompt names, so for alternatives, comparison and trust-validation prompts those domains plausibly sit in both the AI and the Google set; the exclusion of prompt-named companies applied only to the withheld vendor-level analyses.

4.5 Planned panel (results pending)

The pilot is Wave 1. The planned panel repeats the identical prompts and settings 28 days apart and adds six categories per expansion wave. A generalisation requires at least 18 categories, at least 3 waves, an analysis plan unchanged from the protocol, an audited sample of vendor coding

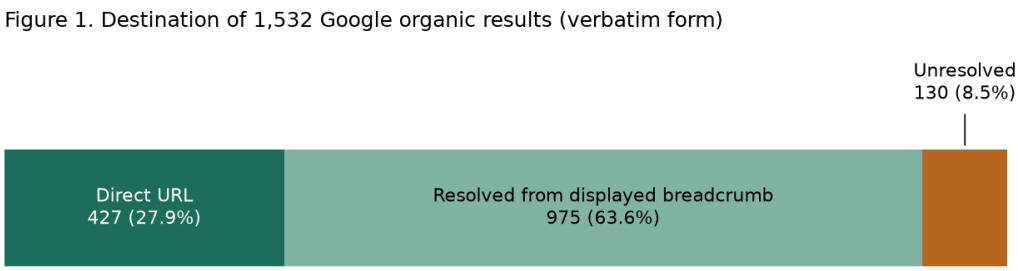

Figure 1. Destination of the 1,532 Google organic results in the verbatim-prompt runs. Source: `analysis/audit_serp_resolution.csv`, recomputed from the raw responses by `analysis/audit_recompute.py`.

(a requirement of the frozen protocol that bears on the vendor-level analyses withheld here), and the author's decision to proceed. None of these exists.

4.6 Observations on data collection

All items are OBSERVED in the pilot.

Completion. 180 of 180 planned ChatGPT-scraper runs completed. 180 of 180 planned Gemini-scraper runs completed. 180 of 180 planned keyword-form Google runs completed. For the verbatim-prompt Google runs, 175 of 180 planned results-page runs completed. The run index holds 190 attempts for these runs: 10 of the 180 prompt-repetition pairs were attempted twice after server errors (5 succeeded on the retry, 5 never succeeded), and all 15 failed attempts returned the provider's "40101 Internal SE Server Error". Five prompts therefore have two Google passes rather than three; every prompt has at least two. Bing and Perplexity were not collected.

Models. The ChatGPT scraper reported one model (`gpt-6`) on every run. The Gemini scraper reported `3.5 Flash-Lite` on most runs and `3.6 Flash` on three; those three form their own stratum and are not in the Gemini results below.

Google redirect addresses. 427 of 1,532 Google organic results carried a direct destination address rather than a redirect address. The Google URL set for each prompt is therefore a subset of the true top-ten URLs. This is a property of this scraper on this date. Of the 1,532 results, 427 (27.9 percent) carried a direct address, 975 (63.6 percent) had a redirect address but a displayed breadcrumb from which the domain could be read, and 130 (8.5 percent) had neither, so their domain is unknown (Figure 1). The domain set of section 4.4 therefore misses those 130 results (90 distinct results once repeats across passes are merged); section 7.6 recovers their domains from the displayed site name and tests what they change.

Set sizes. The median ChatGPT answer cited 8 distinct domains and the median Gemini answer 7. The median Google domain reference (the union of three passes) held 14 domains, and the median Google URL set held 6 direct addresses (`RUN-OVERLAP.csv`). With sets this small, each shared domain moves the Jaccard index by about 0.05.

Google reference stability. Across the three same-day passes for a prompt, mean pairwise Jaccard was 0.339 at domain level and 0.062 at URL level, over 60 prompts (`SERP-STABILITY.`

`csv`). The reference is itself unstable at the URL level, which is why the reference is the union of passes and why overlap with any single Google snapshot depends on which snapshot was drawn.

5 Measurement definitions

Vendor-level analyses were run in the underlying pilot but are withheld from this paper because their language-model coding has not been verified by a person; their definitions are therefore not reproduced here.

Metrics and estimands. For an AI run with source set A and a Google reference set G at one level: Jaccard J = |A ∩ G| / |A ∪ G|; AI-in-Google coverage = |A ∩ G| / |A|; Google-in-AI coverage = |G ∩ A| / |G|. An empty denominator is recorded as not available, not as zero. A prompt value is the mean over its defined repetitions. A cell is an engine, reported model, level, reference and intent; its estimate is the unweighted mean of prompt means (12 prompts per cell). Engines and intents are never pooled in a headline, and runs are compared only within the same reported model.

Jaccard is bounded by the relative sizes of the two sets: with median sets of 8 AI domains against 14 Google domains, the highest attainable J is about 0.57, and each shared domain moves it by roughly 0.05. The dependence of Jaccard on set length is discussed by Sivakumar (2026) for vendor sets.

Secondary measures are the number of distinct cited domains, the share held by the five most cited, a Herfindahl-Hirschman index, and the mix of domain classes (vendor-owned, review or marketplace, publisher, community, agency, other), where class labels come from a model-assisted classification made blind to overlap. The classification was made by a language model from the domains and example page text, after collection; it did not use the vendor-extraction passes, and it has not been checked by a person.

6 Analysis methods

6.1 Uncertainty

No hypothesis test and no p-value is computed. The frozen plan resamples categories within strata (cluster bootstrap, 2,000 resamples). With two categories per stratum that interval is far too narrow, and a model-assisted statistical check of the pilot (STATISTICAL-QA.md) found a t-based interval to be wider in every domain-level cell. This manuscript therefore reports, for each cell, the pilot estimate and the range of the six category means, and does not call any interval a 95 percent interval. Bootstrap refinements for clustered data with few clusters exist (Cameron et al., 2008) and were not applied, because each stratum holds two categories. For the panel, a stratified t interval is proposed once 18 categories exist. Proportions would take Wilson intervals (Wilson, 1927; Brown et al., 2001); the completion counts above are counts of a fixed design and carry none.

6.2 Missing data

A failed run is retried once and then logged as missing. There is no imputation. A prompt with fewer than two defined repetitions is excluded from stability statistics only. The primary analysis excludes ungrounded runs, which the frozen protocol defines as runs with neither cited sources

nor search results (one Gemini run); a run that has search results but no cited source stays in and scores zero overlap (three ChatGPT runs). The sensitivity variant adds the excluded runs.

6.3 Exploratory analysis [X]

One analysis was added after the overlap tables had been read, in `analysis/extensions.py`: the recovery of domains for unresolved Google results and the sensitivity of domain overlap to them (E1; `analysis/recover_unresolved.py`). It uses the data of the frozen pilot and no new collection. It answers a question the protocol did not ask, so its result is presented as an observation.

6.4 Verification

One check was run on 2026-11-06, after the overlap tables had first been drawn up and read on 2026-11-01 (section 6.3; Limitation 18), by code separate from the pipeline that produced the tables. It could not change the pre-specified analysis, but it was not done before the tables were seen.

Independent recomputation. `analysis/audit_recompute.py` shares only the frozen Public Suffix List file with the pipeline. It parses the raw provider responses, rebuilds the Google domain and URL sets, rebuilds each AI run's cited-domain and URL sets, and recomputes the Jaccard index for every AI run against the verbatim-prompt Google reference, at domain and URL level. It compares the result with `RUN-OVERLAP.csv` run by run. It does not recompute the coverage ratios or the keyword-reference comparisons.

7 Results

Status. The panel has not been run: all panel results are pending. Sections 7.2 to 7.5 are the protocol results [P]; section 7.6 is exploratory [X]; section 7.1 reports the recomputation of the pipeline. All describe the pilot only. Tables 3 to 5 show the mean of the 12 prompt means in a cell, then, in square brackets, the lowest and highest of the category means (six, or fewer where a cell has prompts with an undefined value). Table 6 reports counts and shares; Table 7 reports cell means. Neither carries brackets. No test was run. No interval is given. Differences between cells are not claimed to be real.

Source of Tables 3 to 5: `PROMPT-OVERLAP.csv` in the data package (primary variant, grounded runs; Gemini means are for `3.5 Flash-Lite`), with category ranges computed from the same file. Table 6 comes from `CONCENTRATION.csv` and Table 7 from the files named under it.

7.1 Recomputation of the pipeline

Recomputation (OBSERVED). The recomputation reproduced every run-level value it covers: 720 comparisons (domain and URL Jaccard against the verbatim-prompt Google reference for each of 360 AI runs), no mismatch, maximum absolute difference zero (`analysis/audit_recompute_diffs.csv` holds the header only). It did not cover the coverage ratios (Table 4) or the keyword-reference columns (Tables 3 and 5), and those were not independently recomputed. A separate parser inside the exploratory script `extensions.py` (a third implementation, not the recomputation) first gave cell means that differed from the pipeline's by up to 0.0005; the largest

gap was ChatGPT comparison, 0.0730 against 0.0735 in the pipeline and in the recomputation. The cause was traced on 2026-11-06 to the exploratory script, which had added an empty string to the Google domain set for results whose address is an IP host, enlarging the union. After the fix it matches the pipeline in nine of ten cells to four decimals and in the tenth (Gemini trust validation) to 0.0001. The pipeline was not affected, and no table or figure other than Table 7 and Figure 2 uses that script's overlap values.

7.2 Domain level [P]

Table 3. Domain-level Jaccard overlap with Google's top-ten domains, by intent. Columns 1 and 3 use the verbatim-prompt Google reference (primary; range 0.03 to 0.11); columns 2 and 4 use the keyword-form reference (secondary; range 0.06 to 0.15).

Intent	ChatGPT, verbatim ref.	ChatGPT, keyword ref.	Gemini, verbatim ref.	Gemini, keyword ref.
Category discovery	0.043 [0.000 to 0.096]	0.106 [0.052 to 0.163]	0.046 [0.031 to 0.058]	0.077 [0.006 to 0.208]
Alternatives	0.047 [0.028 to 0.082]	0.058 [0.025 to 0.119]	0.112 [0.051 to 0.215]	0.152 [0.017 to 0.299]
Comparison	0.074 [0.018 to 0.122]	0.059 [0.015 to 0.088]	0.055 [0.000 to 0.135]	0.083 [0.049 to 0.152]
Recommendation	0.028 [0.000 to 0.106]	0.067 [0.044 to 0.092]	0.050 [0.000 to 0.116]	0.077 [0.034 to 0.162]
Trust validation	0.055 [0.020 to 0.090]	0.110 [0.041 to 0.264]	0.106 [0.036 to 0.155]	0.143 [0.055 to 0.177]

Against the verbatim-prompt Google reference, mean domain-level Jaccard overlap lay between 0.03 and 0.11 in every intent for both scrapers (ChatGPT 0.028 to 0.074; Gemini 0.046 to 0.112). Against the keyword-form reference the cell means were higher and reached 0.152 (Gemini, alternatives) and 0.143 (Gemini, trust validation); the range for that reference is 0.058 to 0.152 across the ten cells. Every range quoted in this paper therefore names its reference. The range of category means within a cell was wide: for ChatGPT recommendation prompts the six category means ran from 0.000 to 0.106. The category ranges of the two scrapers overlap in every intent, so the pilot cannot tell the engines apart. The keyword-form Google reference gave higher means in 9 of 10 cells (ChatGPT comparison was the exception).

Table 4. Domain-level coverage, verbatim-prompt Google reference.

Intent	ChatGPT AI-in-Google	ChatGPT Google-in-AI	Gemini AI-in-Google	Gemini Google-in-AI
Category discovery	0.095 [0.000 to 0.212]	0.073 [0.000 to 0.151]	0.139 [0.095 to 0.186]	0.064 [0.044 to 0.086]
Alternatives	0.104 [0.058 to 0.185]	0.079 [0.048 to 0.128]	0.306 [0.145 to 0.534]	0.147 [0.069 to 0.261]
Comparison	0.273 [0.052 to 0.521]	0.093 [0.026 to 0.168]	0.152 [0.000 to 0.333]	0.074 [0.000 to 0.183]
Recommendation	0.110 [0.000 to 0.475]	0.036 [0.000 to 0.122]	0.163 [0.000 to 0.342]	0.063 [0.000 to 0.148]
Trust validation	0.176 [0.057 to 0.312]	0.073 [0.026 to 0.125]	0.281 [0.142 to 0.433]	0.140 [0.044 to 0.207]

Against the verbatim-prompt reference, the share of an answer's cited domains that also appeared in Google's set ranged from 0.10 to 0.31 across cells. The share of Google's domains that the answer also cited ranged from 0.04 to 0.15. The second is smaller, as the arithmetic requires when the Google union (median 14 domains) is larger than a typical answer's source

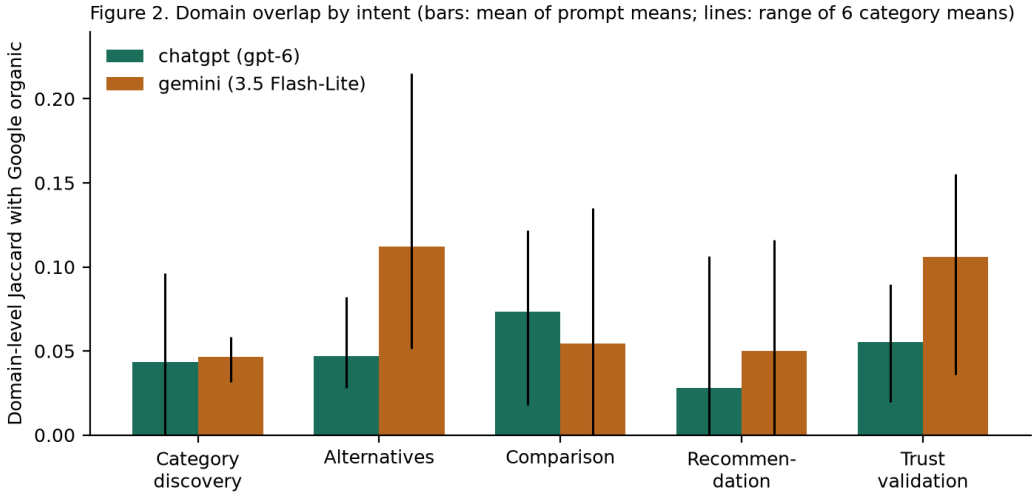

Figure 2. Domain-level Jaccard overlap with Google (verbatim-prompt reference) by intent. Bars: mean of the 12 prompt means; lines: lowest and highest of the six category means. Source: `analysis/ext_domain_cells.csv`; generated by `analysis/make_figures.py`.

list (7 to 8).

7.3 URL level [P]

Table 5. URL-level Jaccard overlap with Google's direct-address URLs, by intent.

Intent	ChatGPT, verbatim ref.	ChatGPT, keyword ref.	Gemini, verbatim ref.	Gemini, keyword ref.
Category discovery	0.000 [0.000 to 0.000]	0.000 [0.000 to 0.000]	0.006 [0.000 to 0.024]	0.016 [0.000 to 0.031]
Alternatives	0.000 [0.000 to 0.000]	0.009 [0.000 to 0.045]	0.037 [0.000 to 0.110]	0.046 [0.000 to 0.134]
Comparison	0.012 [0.000 to 0.062]	0.004 [0.000 to 0.024]	0.024 [0.000 to 0.108]	0.022 [0.000 to 0.066]
Recommendation	0.000 [0.000 to 0.000]	0.006 [0.000 to 0.033]	0.010 [0.000 to 0.049]	0.016 [0.000 to 0.061]
Trust validation	0.000 [0.000 to 0.000]	0.013 [0.000 to 0.027]	0.011 [0.000 to 0.030]	0.030 [0.000 to 0.058]

URL-level overlap was near zero throughout: 0.000 in four of five ChatGPT intents against the verbatim reference, and at most 0.046 in any cell. Because only 427 of 1,532 Google organic results (verbatim prompt) and 420 of 1,508 (keyword form) had a direct address, the Google URL set is a partial list, and this table must not be read as showing that AI answers and Google cite different pages. It shows what could be matched with the addresses this scraper returned.

7.4 Stability of repeated runs [P]

For the Google reference, see section 4.6. For the AI side, the mean between-repetition standard deviation of prompt-level domain Jaccard (verbatim reference) was 0.020 (over 60 prompts) for the ChatGPT scraper and 0.025 (over 59 prompts) for the Gemini scraper (population standard deviation with divisor n over three repetitions; the sample standard deviation is about 1.22 times larger), taken from the `rep_sd` column of `PROMPT-OVERLAP.csv`. Against cell means of

0.03 to 0.11, repetition variability is a sizeable fraction of the quantity measured. RQ3's second question, whether few repeats suffice for overlap, as the first version of Zatuchin (2026) reported for brand sentiment (a later version qualifies that decomposition), is not answered: the pilot has three repetitions per prompt and did not compare fewer against more.

7.5 Concentration of cited domains [P]

Table 6. Concentration of cited domains by engine and intent.

Engine	Intent	Distinct domains	Citation slots	Top-5 share	HHI
ChatGPT (gpt-6)	Category discovery	120	388	0.111	0.012
ChatGPT (gpt-6)	Alternatives	123	404	0.106	0.012
ChatGPT (gpt-6)	Comparison	64	210	0.257	0.027
ChatGPT (gpt-6)	Recommendation	90	249	0.157	0.017
ChatGPT (gpt-6)	Trust validation	54	208	0.519	0.067
Gemini (3.5 Flash-Lite)	Category discovery	126	264	0.140	0.013
Gemini (3.5 Flash-Lite)	Alternatives	104	248	0.234	0.023
Gemini (3.5 Flash-Lite)	Comparison	87	211	0.199	0.020
Gemini (3.5 Flash-Lite)	Recommendation	121	242	0.116	0.012
Gemini (3.5 Flash-Lite)	Trust validation	73	220	0.327	0.035

Citation slots count each domain once per grounded run and sum over runs; the table pools all six categories. Trust-validation answers were the most concentrated: ChatGPT top-five share 0.519 against 0.106 to 0.257 in the other four intents, and Gemini 0.327 against 0.116 to 0.234. These four columns come from the cited domains alone and need no classification. There is no clustering and no interval; the statistic mixes categories.

Table 6X. Exploratory [X]: share of citation slots held by two model-assigned domain classes.

Engine	Intent	Vendor-owned share	Review/marketplace share
ChatGPT (gpt-6)	Category discovery	0.724	0.028
ChatGPT (gpt-6)	Alternatives	0.760	0.062
ChatGPT (gpt-6)	Comparison	0.400	0.181
ChatGPT (gpt-6)	Recommendation	0.663	0.048
ChatGPT (gpt-6)	Trust validation	0.274	0.596
Gemini (3.5 Flash-Lite)	Category discovery	0.731	0.076
Gemini (3.5 Flash-Lite)	Alternatives	0.637	0.181
Gemini (3.5 Flash-Lite)	Comparison	0.474	0.081
Gemini (3.5 Flash-Lite)	Recommendation	0.740	0.041
Gemini (3.5 Flash-Lite)	Trust validation	0.336	0.495

The two shares in Table 6X depend on a model-assisted classification of the cited domains, made from the domain and example page text and not from the vendor extraction (section 5). A person has not checked the labels, no validation sample was drawn, and no error rate is

known, so these shares are exploratory and nothing in the paper's conclusions rests on them. The remaining slots fall in other classes or are unclassified (1 to 12 percent by cell). On these labels, vendor-owned domains held 0.64 to 0.76 of the slots in category-discovery and alternatives answers, and review and marketplace domains held 0.596 of ChatGPT trust-validation slots and 0.495 of Gemini's. Treat the direction as a hypothesis for a validated classification.

7.6 Unresolved Google results: recovery and sensitivity of domain overlap [X]

The 130 Google organic results with no redirect-resolvable domain and no breadcrumb (section 4.6) could have hidden domains that the AI answers cite. This section uses only fields already in the raw responses and fetches nothing.

Recovery. Each Google result carries a displayed site name (the `website_name` field, for example "Reddit · r/sales"). A rule maps a site name to a domain when the resolved results of the whole dataset map that name to exactly one domain. Merging repeats of the same result across the three Google passes (same prompt, title and site name) leaves 90 distinct unresolved results. The rule resolves all 90: reddit.com 44, youtube.com 17, quora.com 11, medium.com 10, linkedin.com 5, facebook.com 3. The rule was checked by leaving results out of the name map. Leaving out single passes gave 1,161 correct predictions, none wrong, 27 ambiguous and 214 with no prediction; that version is generous, because the same result's other two passes stay in the map. Leaving out whole results (972 distinct resolved results) gave 628 correct, none wrong, 20 ambiguous and 324 with no prediction. The check is dominated by publisher names: the six names that matter rest on 26 Reddit, 7 YouTube, 4 Quora, 5 Medium and 8 Facebook resolved results, and LinkedIn has one, so for LinkedIn the rule has no test. A name does not always give one domain: Glassdoor maps to five country domains in the resolved set, though not to any of the six recovered here. Of the 1,402 resolved results that carry a site name, 975 (70 percent) were resolved from the breadcrumb and 427 (30 percent) came from a direct address. The 130 unresolved results are engagement cards, whose breadcrumb field holds comment counts or ages, so the check does not sample the population it is applied to. All 90 distinct unresolved titles and descriptions carry platform-specific text (answer and comment counts, "- Quora", "- Medium", subreddit labels), which is consistent with the mapping but is a reading, not a fetch: no page was opened. Where a breadcrumb and a direct address both exist, their registrable domains agree in 378 of 378 results. This tests breadcrumbs only on direct-URL results, not on the 975 redirect-only results the breadcrumb resolution is used for. None of the six recovered domains is a software company's own site in the study's domain-class table (they are classed as community forum, or publisher or media). URL-level overlap is not recoverable from a site name, which gives no page address. This section covers the verbatim-prompt Google runs only. The 129 unresolved results of 1,508 in the keyword-form runs (Table 3, columns 2 and 4) were not examined.

Effect. Table 7 adds the recovered domains to each prompt's Google set and recomputes the domain-level cell means. The change runs from -0.0060 to +0.0003: eight cells fall, two rise (by 0.0002 and 0.0003). The largest fall is ChatGPT comparison, 0.0735 to 0.0676. So the unresolved results bias observed overlap slightly upward, not downward: the recovered domains are community and video platforms that AI answers rarely cite (26 of the 2,644 cited-domain slots in grounded primary runs: reddit.com 14, medium.com 11, youtube.com 1), so adding

them mostly enlarges the union. The recovered values keep the range of cell means at 0.027 to 0.111, against 0.028 to 0.112 observed.

Stress case. To test how much the conclusion depends on the recovery, the last column ignores it: each distinct unresolved result is assumed to be a different domain that the AI run cited and the resolved Google set lacks, up to the number of such domains in the run (`analysis/recover_unresolved.py`). This is a stress case, not a description of the results: it is the largest value the cell means can take if the site-name recovery is wrong for every result. Cell means then reach 0.069 to 0.239.

Table 7. Domain-level Jaccard against the verbatim-prompt Google reference: protocol value, with unresolved results recovered from site names, and a stress case that ignores the recovery (exploratory).

Engine	Intent	Pipeline mean [P]	With recovered domains added [X]	Change	Stress case: unresolved results ignored, each a cited domain [X]
ChatGPT (gpt-6)	Category discovery	0.0434	0.0422	−0.0012	0.081
ChatGPT (gpt-6)	Alternatives	0.0469	0.0472	+0.0003	0.108
ChatGPT (gpt-6)	Comparison	0.0735	0.0676	−0.0060	0.132
ChatGPT (gpt-6)	Recommendation	0.0282	0.0274	−0.0008	0.069
ChatGPT (gpt-6)	Trust validation	0.0552	0.0534	−0.0018	0.191
Gemini (3.5 Flash-Lite)	Category discovery	0.0464	0.0452	−0.0012	0.088
Gemini (3.5 Flash-Lite)	Alternatives	0.1121	0.1110	−0.0010	0.188
Gemini (3.5 Flash-Lite)	Comparison	0.0545	0.0548	+0.0002	0.137
Gemini (3.5 Flash-Lite)	Recommendation	0.0501	0.0489	−0.0012	0.094
Gemini (3.5 Flash-Lite)	Trust validation	0.1059	0.1014	−0.0044	0.239

Source: `analysis/recover_unresolved_cells.csv`, `analysis/recover_unresolved_results.csv` (every unresolved result with its recovered domain). The pipeline column is the pipeline mean computed from unrounded per-prompt values and agrees with `PROMPT-OVERLAP.csv` (whose prompt means are rounded to four decimals, so Gemini trust validation reads 0.1058 there); Table 3 shows the same values to three decimals (ChatGPT comparison 0.074, Gemini comparison 0.055). Changes are computed from unrounded values. The protocol did not use site names, so the headline tables keep the protocol values.

The domain-level finding therefore survives under the mapping and is bounded under the stress case: low overlap, with every cell mean between 0.03 and 0.11 under the protocol, under recovery, and below 0.24 in a stress case that ignores the recovery. What the unresolved results do change is the third decimal. The recovery rests on a site name identifying a domain; it is validated on resolved results only (thinly for LinkedIn), not on the unresolved ones themselves, and a site name does not restore page-level addresses.

7.7 Panel results

Pending. Table 8 is the empty shell the panel will fill.

Table 8. Planned panel table (empty).

Cell (engine, model, intent)	Wave 1	Wave 2	Wave 3	Change, with interval
Domain-level Jaccard, verbatim reference (primary)	pending	pending	pending	pending
All other estimands	pending	pending	pending	pending

8 Discussion and alternative explanations

What the pilot is good for. It shows that the pipeline runs end to end on live engines, that the freeze can be verified, and where the instruments break. Both are useful before spending on a panel. The URL-level problem is the clearest: with most Google organic results carrying redirect addresses, URL-level overlap cannot be measured at full coverage with this scraper. Anyone comparing a URL-level figure with another study should first ask how many Google URLs were visible.

Domain overlap in context (INFERRED). The pilot's domain-level means (0.03 to 0.11 against the verbatim-prompt Google reference; up to 0.15 against the keyword-form reference) are of the same order as the domain overlaps reported for consumer queries by Chen et al. (2026), who report 4.0 to 15.2 percent for four AI systems. The two sets of figures differ in queries, engines, set construction and date, so this is a comparison of magnitude, not a replication. One reading is that overlap with the union of same-day Google top-ten pages is low for these B2B prompts too. The pilot cannot say whether that depends on the engine, the prompt or the category.

The reference is not a fixed set (OBSERVED, INFERRED). Google's own top-ten domains change across same-day passes (mean pairwise Jaccard 0.34), so "the Google result" for a prompt depends on which pass, or which combination of passes, is used. Overlap with a single pass and overlap with the union of three passes are different quantities, and the pilot reports only the union. The low self-agreement of the passes is a measure of noise in the reference. It is not an upper bound on overlap: a set can share more with a third set than two draws of the same ranking share with each other, so a pass-to-pass Jaccard of 0.34 does not cap an AI answer at 0.34. The union enlarges the Google set, which lowers the Jaccard value for a given answer and raises the share of the answer that Google covers. A comparison against a single snapshot has the same dependence on the snapshot chosen, reported or not. The planned extension (section 4.5) specifies single-pass references as a companion analysis.

Concentration by intent (INFERRED, partly exploratory). Trust-validation answers were the most concentrated on the class-free measures in Table 6. Whether they leaned on review sites rests on the unchecked classes in Table 6X and is a hypothesis. If it holds, it agrees in direction with SE Ranking's finding that review platforms are prominent in AI Overview citations for commercial queries (Deda, 2026); the instrument, the sample and the measure differ, so only direction is compared.

Alternative explanations for the low overlap. Several readings fit the domain-level numbers, and the pilot cannot separate them. 1. *Different retrieval.* AI products rewrite the query and retrieve for the rewrites (Google Search Central, 2025; Chen & Koudas, 2026), so the

pages cited need not be those that rank for the prompt as typed. This is the reading the design was built around. 2. *Instrument.* The scrapers carry no personalisation, only country-level location for ChatGPT, and the Gemini scraper reported a small model. Consumer apps may differ. 3. *Reference instability.* Google's own top-ten domains moved between same-day passes (mean pairwise Jaccard 0.34), so any overlap value depends on which pass or combination of passes is used as the reference. 4. *Prompt design.* The prompts are the author's wording, and the company pairs are well-known incumbents chosen with some sight of Google's results for the head query. A buyer's wording could give different overlap. 5. *Unresolved Google results (verbatim-prompt runs).* Section 7.6: a rule validated on resolved results maps the displayed site names of the 130 to six domains, and adding those domains lowers eight cell means by up to 0.006, so on that reading, which rests on the site-name mapping, they do not explain the low overlap. A stress case that ignores the recovery reaches at most 0.24, a rise of roughly 1.7 to 3.5 times in the affected cells. 6. *Small cells.* Twelve prompts from six categories per cell. A cell mean moves by about 0.05 for each shared domain, and the category means within a cell often span a range as wide as the mean.

For measurement design. Four lessons come from the pilot and need no further data. State which Google URLs were visible and how many results had no resolvable domain. Report the stability of the Google reference. Keep model versions apart. Recompute the headline tables from the raw responses with separate code before reporting them (this pilot did so for run-level domain and URL overlap only). These are about how to measure, not about how any company should behave.

9 Limitations

1. **Six categories.** Drawn by a seeded hash from a convenience frame; not a sample of the B2B software market. Incumbent companies only in the named pairs.
2. **One date.** 2026-11-01. Models, retrieval and Google's results change; the pilot has no information on change over time.
3. **One repetition set.** Three same-day repetitions per prompt. Whether three is enough was not tested (the first version of Zatuchin, 2026, found little gained beyond five repeats for a different outcome; a later version qualifies that decomposition, so it is not relied on).
4. **Scrapers, not consumer products.** The DataForSEO scrapers carry no personalisation, country-level location only for ChatGPT, and may differ from the apps people use. The Gemini scraper reported a small model.
5. **URL-level overlap rests on direct addresses only.** 427 of 1,532 Google organic results. URL-level overlap is biased toward zero by construction.
6. **Bing and Perplexity were not collected.** Bing results from the provider were defective in the smoke test; the Perplexity token had expired. Nothing in the pilot speaks to either.
7. **Vendor-level analyses are withheld.** The underlying pilot holds vendor-level analyses, built on vendor extraction by language-model coders. That coding has not been verified by a person, so those analyses are not reported here and nothing in this paper depends on them. The 20 percent human audit, stratified by engine and intent, is pending.
8. **Intervals.** The frozen bootstrap understates uncertainty with six categories; none is

reported.

9. **Prompts are the author's wording**, not buyer-sourced, in US English.
10. **The Google AI Overview comparison is same-call**; it describes one results page and is not independent evidence.
11. **Registration.** The protocol was frozen by file hash in the project's repository, not deposited with an external registry. Two amendments are on record (a code fix to the overlap script, and a documentation-only edit to the protocol file).
12. **Code, domain classes and drafting by language models.** Disclosed in section 12. The study files record only tier names (Sonnet-class, Haiku-class, Opus-class) and the draft pre-registration promised model, version and temperature, which were not written into the study files. Identifiers were later recovered from the orchestrating session's transcript logs, which record the model each subordinate run reported (`REPRODUCIBILITY.md` in the data package, section 6); `claude-sonnet-5-5` is the model recorded for the runs that built the domain-class table. Sampling settings are not in those logs and are unknown.
13. **The collector changed between the draft preregistration and the freeze.** The draft recorded `tools/collect.py` with a hash beginning `a009b285`; the frozen file's hash begins `ddd9555e`. Edits before the freeze are not logged by the protocol's own rule, so what changed is not on record.
14. **Company pairs were not chosen blind to Google's results.** The author viewed Google's top ten for "best {category} software" for the six categories before writing the company pairs (section 4.2). The pairs are the author's, and the incumbents named may be those Google already favours for the head query.
15. **Literature search coverage.** Google Scholar and the ACM Digital Library could not be searched and the SSRN papers listed in Appendix D were read at abstract level only. The statement of position in section 3 is limited to the sources opened.
16. **Possible interest.** The author's company sells search and AI-visibility services. The design, analysis and writing were carried out within the author's company. This is a reason to read the pilot with care, not a reason the numbers are wrong; the protocol, code and hashes are there so others can check.
17. **The verification is not independent in the strong sense.** The recomputation covers run-level domain and URL Jaccard only; it was written separately from the pipeline, but by the same drafting model, and shares the frozen Public Suffix List file. It was run on 2026-11-06, after the tables had been drawn up, and the freeze check of section 4.1 was last run on 2026-11-01.
18. **Exploratory analysis.** Section 7.6 was added after the overlap tables had been read. It is an observation; the choice of this analysis and not others is itself a forking path, and no multiplicity adjustment is possible or claimed.
19. **Unresolved Google results.** 130 of 1,532 results in the verbatim-prompt runs had no redirect-resolvable domain. Section 7.6 recovers their domains from the displayed site name (validated on resolved results only) and reports the effect; a site name does not give a page address, and breadcrumb paths are truncated and lossy, so URL-level overlap stays uninformative.

20. **Domain classes are model-assisted and unchecked.** The classes behind Table 6X (an exploratory table) and used in section 7.6 were assigned by a language model, after collection and blind to overlap, and have not been checked by a person.

10 What this paper cannot establish

- **No causal claim.** The paper describes overlap. It tests no cause of it and no effect of any action on it.
- **No prediction of AI recommendations from rank.** It does not show whether a Google rank predicts an AI recommendation, or that it does not. No model was fitted and six categories could not support one.
- **One collection date, six categories.** All results describe 60 prompts on 2026-11-01. They say nothing about change over time and nothing about B2B software, SaaS or AI search in general.
- **Scrapers, not consumer products.** The DataForSEO scrapers may differ from the apps people use. No difference between the two engines is claimed: their category ranges overlap in every intent.
- **The Google reference is a union of three same-day passes (two for five prompts).** It is itself unstable (mean pairwise Jaccard 0.339 at domain level), and it enlarges the Google set, which lowers the coverage of Google by an answer. Single-pass values are not reported.
- **Missing-domain uncertainty.** 130 of 1,532 Google results had no resolvable domain. The recovery from site names is validated on resolved results only, the stress case reaches 0.24, and the 129 unresolved keyword-form results were not examined.
- **The URL level is uninformative.** Only 427 of 1,532 Google organic results exposed a destination address, so URL-level overlap is biased toward zero by construction.
- **Domain-class labels are model-assisted.** A language model assigned them, and a person has not checked them. The shares in Table 6X depend on them, which is why that table is exploratory.
- **Vendor-level results are withheld.** Vendor-level analyses exist in the underlying pilot but are not reported, because their language-model coding has not been verified by a person.

11 Conclusion

A frozen protocol for measuring URL- and domain-level overlap between Google and AI answers for B2B software prompts exists, runs end to end, verifies by hash, and its run-level domain and URL overlap values reproduce from the raw responses with separate code (720 of 720; the coverage ratios and the keyword-reference columns were not recomputed). The pilot found [P] small domain-level overlap, near-zero URL-level overlap that is uninformative because most Google results hide their destination address, and a Google reference that moves between same-day passes. If the site-name mapping in section 7.6 is right (it is validated only on resolved results), the Google results whose domain was hidden change domain-level cell means by at most 0.006; if it were wrong for every such result, cell means could rise 1.7 to 3.5 times, to at

most 0.24. This covers the verbatim-prompt runs only. None of this generalises, and none of it shows whether a Google rank predicts an AI recommendation. Vendor-level analyses of the pilot are withheld until their coding has been verified by a person. The next steps are the human coding of the vendor data and the monthly panel with at least 18 categories and 3 waves; until the panel is done, the results sections of this paper are pending.

12 Reproducibility and statements

Reproduction. Every number in sections 4 to 7 comes from the pilot's raw provider responses and analysis tables, or from the scripts in the data package. All scripts run with `python3 -I <script> <path to the wave 1 directory>` and write next to themselves: `audit_recompute.py` (recomputation of run-level domain and URL overlap, and Google destination resolution), `extensions.py` (the exploratory analysis; the same script also produces vendor-level outputs, which this paper does not use), `recover_unresolved.py` (section 7.6 recovery and stress case) and `make_figures.py` (Figures 1 and 2; needs matplotlib). The freeze check and the original pipeline are described in `REPRODUCIBILITY.md` in the package. A re-run of the original analysis from the stored raw files reproduced all 606 summary rows (372 domain and URL rows used here and 234 vendor-level rows not reported here; Appendix B). Appendix C records the checks run for this version.

Data and code availability. The pilot's prompts, category draw, frozen protocol, freeze manifest, collection and analysis code, tests, Public Suffix List snapshot, aggregate analysis tables (vendor-level rows removed) and domain-class table are assembled in a data package with SHA-256 checksums. The package is deposited in a public Zenodo record under a CC BY 4.0 licence (https://doi.org/10.5281/zenodo.24190534). Raw provider responses (the model answers and results pages) are withheld pending a check of the providers' terms and of the rights in the answer text. The analysis tables are derived from them.

Competing interests. The author is the founder of Caldrin, a company that sells search and AI-visibility services, and Caldrin carried out and paid for the data collection and analysis. The author also wrote the Caldrin web article cited in section 3 (Shrestha, 2026). The companies named in the prompts and the data provider were not consulted and have not reviewed this paper. The author declares no other competing interests.

Funding. No external funding. Collection costs were paid by Caldrin, the author's company.

Use of AI tools. Language models from Anthropic (Claude) were used under the author's direction for the following tasks. (1) Code: writing and reviewing the collection, extraction, overlap-analysis, verification and figure code. (2) Data coding: vendor mentions in AI answers and in Google result text were extracted by two independent model coders (pass A `claude-sonnet-5-5`, pass B `claude-haiku-5-5`) and disagreements were ruled on by a blind `claude-opus-5-5` adjudicator; that coding has not been verified by a person, and the analyses built on it are not reported in this paper. (3) Reference tables: `claude-sonnet-5-5` built the vendor alias table and the domain-class table after collection, blind to overlap; the domain classes are used in Table 6 and have not been checked by a person. (4) Literature search and reading, recorded in Appendix D. (5) Data audit, exploratory analysis and figures (2026-11-06) and the verification of this version (2026-11-12): `claude-sonnet-5-5`. (6) Manuscript: `claude-sonnet-`

`5-5` drafted the first text on 2026-11-01, revised it on 2026-11-06, and prepared this submission version on 2026-11-12. Model identifiers were recovered from the logged model field of the Claude Code sessions that ran the study, not from a record made at the time in the study files; sampling settings were not recorded and are unknown. Which model wrote each code file was not mapped. No AI system is an author. The author directed the work and is responsible for its content.

Ethics. No human participants and no personal data. Data came from a commercial API provider's scraper products.

Author contributions. Ankur Shrestha: conceptualisation, study design, analysis, writing, and supervision of AI-assisted work.

References

Aggarwal et al. (2024). Pranjal Aggarwal; Vishvak Murahari; Tanmay Rajpurohit; Ashwin Kalyan; Karthik Narasimhan; Ameet Deshpande. 2023 (v1 2023-11-16; v3 2024-06-28). *GEO: Generative Engine Optimization.* arXiv:2311.09735; arXiv comments: Accepted to KDD 2024 (proceedings citation not verified). https://arxiv.org/abs/2311.09735 (accessed 2026-11-01).

Allaham & Diakopoulos (2026). Mowafak Allaham; Nicholas Diakopoulos. 2026 (v1 2026-05-22). *Synthetic Sources?: Auditing Generative Search Engine Citations for Evidence of AI-Generated Sources.* arXiv:2605.23684; journal-ref field lists AIES 2026 (acceptance not confirmed on the page). https://arxiv.org/abs/2605.23684 (accessed 2026-11-01).

Bal (2026). Avik Bal. 2026 (Crossref record created 2026-08-16). *Who The Machine Names: A Four-Week Panel Audit of AI Citation in Enterprise Banking Software.* SSRN preprint, https://doi.org/10.2139/ssrn.7268163 (abstract read through Crossref, accessed 2026-11-12; full text not read).

BrightEdge (2025). BrightEdge (corporate author; individual authors not named on the page). 2025-09-18. *AI Overview Citations Now 54% from Organic Rankings.* BrightEdge Weekly AI Search Insights; vendor study. https://www.brightedge.com/resources/weekly-ai-search-insights/rank-overlap-after-16-months-of-aio (accessed 2026-11-01).

Brown et al. (2001). Lawrence D. Brown; T. Tony Cai; Anirban DasGupta. 2001. *Interval Estimation for a Binomial Proportion.* Statistical Science 16 (page range not captured from the record read). https://doi.org/10.1214/ss/1009213286 (accessed 2026-11-01).

Cameron et al. (2008). A. Colin Cameron; Jonah B. Gelbach; Douglas L. Miller. 2008. *Bootstrap-Based Improvements for Inference with Clustered Errors.* Review of Economics and Statistics 90(3):414-427. https://doi.org/10.1162/rest.90.3.414 (accessed 2026-11-01).

Chen & Koudas (2026). Mahe Chen; Nick Koudas. 2026 (submitted 2026-10-10). *What Drives AI Citations? Content, Retrieval and Model Knowledge in Generative Engine Optimization (GEO).* arXiv:2610.15033 (cs.IR preprint; no venue shown). https://arxiv.org/abs/2610.15033 (accessed 2026-11-01).

Chen et al. (2025). Mahe Chen; Xiaoxuan Wang; Kaiwen Chen; Nick Koudas. 2025 (v1 2025-09-10). *Generative Engine Optimization: How to Dominate AI Search.* arXiv:2509.08919 (cs.IR preprint; no venue shown). https://arxiv.org/abs/2509.08919 (accessed 2026-11-01).

Chen et al. (2026). Mahe Chen; Xiaoxuan Wang; Kaiwen Chen; Nick Koudas. 2026 (v1 2026-01-23; v2 2026-05-16). *Navigating the Shift: A Comparative Analysis of Web Search and Generative AI Response Generation.* arXiv:2601.16858; arXiv comments: Accepted at EDBT/ICDT 2026 Workshops. https://arxiv.org/abs/2601.16858 (accessed 2026-11-01).

Deda (2026). Yulia Deda (SE Ranking); reviewed by Svitlana Tomko. 2026-01-29. *Review platforms lost 90% of traffic over the past 2 years, yet they top AI Overview citations for commercial queries.* SE Ranking blog; vendor study. https://seranking.com/blog/review-platforms-in-ai-overviews/ (ac-

cessed 2026-11-01).

DerivateX (2026). DerivateX Research (GEO and SEO agency; individual authors not named). 2026-05-31. *The B2B SaaS AI Citation Study: How ChatGPT Recommends Software*. Agency report. https://derivatex.agency/report/b2b-saas-ai-citation-study/ (accessed 2026-11-01).

Gandhi (2025). Mitul Gandhi (seoClarity). 2025 (page says updated October 2025). *The Overlap Between AI Overviews and Organic Rankings*. seoClarity Research Study; vendor study. https://www.seoclarity.net/research/aio-rankings-overlap (accessed 2026-11-01).

GrackerAI (2026). GrackerAI (listed as author and publisher; individual authors not named). 2026 (the page states August 2026). *AI Search Visibility Benchmark Series: compliance automation*. Vendor report; the publisher's business was not verified. https://gracker.ai/data-and-research-reports/ai-search-visibility-compliance-automation-2026 (accessed 2026-11-06; read through a summarising tool, so every figure is as stated on the page and the reproducibility section's content was not seen).

Google Search Central (2025). Google Search Central. 2025 (page says last updated 2025-12-10 UTC). *AI features and your website*. Official documentation. https://developers.google.com/search/docs/appearance/ai-features (accessed 2026-11-01).

Huang et al. (2026). Michelle Huang; Agam Goyal; Koustuv Saha; Eshwar Chandrasekharan. 2026 (v1 2026-03-17; v2 2026-08-28). *Answer Bubbles: Information Exposure in AI-Mediated Search*. arXiv:2603.16138; arXiv comments: EMNLP 2026 (as given in the arXiv comments field; the proceedings entry was not checked). https://arxiv.org/abs/2603.16138 (accessed 2026-11-01; authors confirmed from the arXiv API record).

Kunal (2026). Rohit Kunal. 2026 (Crossref record created 2026-09-28). *Who Does AI Cite? Reconciling Conflicting Evidence on the Sources Behind Generative Search Answers (2024-2026)*. SSRN preprint, https://doi.org/10.2139/ssrn.7528060 (abstract read through Crossref, accessed 2026-11-12; full text not read).

Lee (2026). Anthony Lee. 2026 (Crossref record created 2026-04-03). *Query Intent and Google Rank as Joint Predictors of AI Citation: A Multi-Platform Observational Study*. SSRN preprint, https://doi.org/10.2139/ssrn.6240278 (abstract read through Crossref, accessed 2026-11-12; full text not read).

Levin (2025). Eugene Levin (Semrush); contributors Anna Yudina and Luke Harsel. 2025-07-21. *How Google's AI Mode Compares to Traditional Search and Other LLMs [AI Mode Study]*. Semrush blog; vendor study. https://www.semrush.com/blog/ai-mode-comparison-study/ (accessed 2026-11-01).

Linehan (2025). Louise Linehan (Ahrefs); contributor Xibeijia Guan. 2025-08-11. *Only 12% of AI Cited URLs Rank in Google's Top 10 for the Original Prompt*. Ahrefs blog; vendor study; publisher sells Brand Radar. https://ahrefs.com/blog/ai-search-overlap/ (accessed 2026-11-01).

Linehan (2026). Louise Linehan (Ahrefs); contributor Xibeijia Guan; reviewed by Ryan Law. 2026-03-02. *Update: 38% of AI Overview Citations Pull From The Top 10*. Ahrefs blog; vendor study. https://ahrefs.com/blog/ai-overview-citations-top-10/ (accessed 2026-11-01).

Liu et al. (2023). Nelson F. Liu; Tianyi Zhang; Percy Liang. 2023 (v1 2023-04-19; v2 2023-10-23). *Evaluating Verifiability in Generative Search Engines*. arXiv:2304.09848; arXiv comments: to appear in Findings of EMNLP 2023. https://arxiv.org/abs/2304.09848 (accessed 2026-11-01).

Public Suffix List community (n.d.). Public Suffix List community (a Mozilla initiative). Snapshot 2026-10-31_14-17-28_UTC used in the study. *Public Suffix List*. Community resource. https://publicsuffix.org/ (accessed 2026-11-01).

Ray (2026). Lily Ray (Algorythmic); republished by Search Engine Journal. 2026-02-18. *Are Citations In AI Search Affected By Google Organic Visibility Changes?*. Search Engine Journal / Substack; practitioner analysis. https://www.searchenginejournal.com/are-citations-in-ai-search-affected-by-google-organic-visibility-changes/567736/ (accessed 2026-11-01).

Schulte et al. (2026). Julius Schulte; Malte Bleeker; Philipp Kaufmann. 2026 (v1 2026-04-08). *Don't Measure Once: Measuring Visibility in AI Search (GEO).* arXiv:2604.07585 (cs.IR preprint; no venue shown). https://arxiv.org/abs/2604.07585 (accessed 2026-11-01).

Shrestha (2026). Ankur Shrestha. 2026-11-06. *Your AI Visibility Trend Line Might Be an Artefact.* Caldrin (company website research article). https://caldrin.co/research/ai-visibility-trend-line-artefact (accessed 2026-11-12). The page carries the Article structured data headline, author and datePublished 2026-11-06 and a canonical link to this address. No DOI.

Sielinski (2026). Ronald Sielinski. 2026 (v1 2026-03-09; v3 2026-08-26). *Quantifying Uncertainty in AI Visibility: A Statistical Framework for Generative Search Measurement.* arXiv:2603.08924 (stat.AP preprint; no venue shown). https://arxiv.org/abs/2603.08924 (accessed 2026-11-01).

Sivakumar (2026). Sairam Sivakumar. 2026 (Crossref record created 2026-08-10). *Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. Measuring Vendor-Set Agreement Across Four AI Search Engines.* SSRN preprint, https://doi.org/10.2139/ssrn.7234418 (abstract read through Crossref, accessed 2026-11-12; full text not read).

ThriveStack (2026). ThriveStack GTM Research (vendor; individual authors not named). Page metadata date 2026-05-04. *AI Visibility Benchmarks 2026.* Vendor report; the publisher sells an AI-visibility platform. https://www.thrivestack.ai/research/ai-visibility-benchmarks-2026 (accessed 2026-11-01).

Wilson (1927). Edwin B. Wilson. 1927. *Probable Inference, the Law of Succession, and Statistical Inference.* Journal of the American Statistical Association 22:209-212. https://doi.org/10.1080/01621459.1927.10502953 (accessed 2026-11-01).

Yadav (2026). Ajay Yadav. 2026 (the PDF header reads "April 2026"). *Beyond Rankings: Measuring Vendor Visibility in AI-Driven Discovery.* Self-published PDF; no venue. https://aduttya.com/beyond-rankings-measuring-vendor-visibility-in-ai-driven-discovery.pdf (accessed 2026-11-01; abstract, method and results sections read in part).

Zatuchin (2026). Dmitrij Zatuchin. 2026 (v1 2026-07-14; a later version, v2, is retitled). v1 title: *Where Does the Noise Come From? A Variance-Components Decomposition of Non-Determinism in LLM Brand Answers*; current title: *Query Language Sensitivity in Automated Sentiment Scores of LLM Responses.* arXiv:2607.13304 (cs.IR preprint; no venue shown). https://arxiv.org/abs/2607.13304 (abstracts of both versions read 2026-11-06; the arXiv record spells the author's surname with a diacritic, Żatuchin; v2 dated 2026-11-01).

Zhang et al. (2025). Peixian Zhang; Qiming Ye; Zifan Peng; Kiran Garimella; Gareth Tyson. 2025 (submitted 2025-12-10). *Source Coverage and Citation Bias in LLM-based vs. Traditional Search Engines.* arXiv:2512.09483 (cs.CL preprint; no venue shown). https://arxiv.org/abs/2512.09483 (accessed 2026-11-01).

Zhen et al. (2026). Tao Zhen; Yue Liu; Gege Zhang; Yixuan Niu. 2026 (v1 2026-07-17). *What Do Chinese-Language Generative Search Engines Cite and Surface? A Large-Scale Empirical Study.* arXiv:2607.15771 (cs.IR preprint; no venue shown). https://arxiv.org/abs/2607.15771 (accessed 2026-11-01).

A Prompt templates and categories

DOCUMENTED. Ten templates; two per intent. Shown for the first category; the other five replace the category name, the buyer context and the company pair as listed. `PILOT-PROMPTS.csv` holds all 60.

Template	Intent	Text (revenue intelligence; companies Clari and Gong)
DISC-1	Category discovery	What are the best revenue intelligence software tools for a B2B software company with a 50-person sales team?
DISC-2	Category discovery	Which revenue intelligence software should a B2B software company with a 50-person sales team consider?
ALT-1	Alternatives	What are the best alternatives to Clari?
ALT-2	Alternatives	Gong alternatives
CMP-1	Comparison	Clari vs Gong: which is better?
CMP-2	Comparison	Compare Gong and Clari for a B2B software company with a 50-person sales team
REC-1	Recommendation	Which revenue intelligence software would you recommend for a B2B software company with a 50-person sales team, and why?
REC-2	Recommendation	On a limited budget, which revenue intelligence software should a B2B software company with a 50-person sales team choose?
TRU-1	Trust validation	Is Clari worth the price? What do customers say?
TRU-2	Trust validation	Is Gong a reliable vendor? What are the common complaints?

Category	Stratum	Buyer context slot	Company pair
Revenue intelligence	S1	B2B software company with a 50-person sales team	Clari, Gong
Sales engagement	S1	B2B software company with a 50-person sales team	Outreach, Salesloft
Product analytics	S2	B2B SaaS product team	Amplitude, Mixpanel
Product adoption and onboarding	S2	B2B SaaS company	Pendo, WalkMe
Insurance underwriting	S3	mid-size commercial insurance carrier	Guidewire, Duck Creek
Property management	S3	property manager with 200 rental units	AppFolio, Buildium

The alternatives and trust-validation templates do not carry the category name; the named company implies it. The keyword-form versions (for the secondary Google reference) are mechanical, for example "best revenue intelligence software" and "Clari alternatives".

B Freeze, amendment and reproduction

Hashes, commands and the reproduction check are in `REPRODUCIBILITY.md`. The freeze covers 13 files; two amendments are recorded in `FREEZE-MANIFEST.json`: (1) 2026-11-01T18:29:08+00:00, `tools/studyb_overlap.py`, the prompt-named-vendor exclusion fix (hash `1446aadc...` to `cbccbf2e...`); (2) 2026-11-01T21:13:01+00:00, `study-B-pilot/PROTOCOL-FROZEN-v1.0.md`, a documentation-only edit listing amendment 1 in the protocol's section 14 (hash `a17611ba...` to `f800c559...`). A re-run of the analysis from the stored raw files reproduced all 606 summary rows (372 domain and URL rows used here and 234 vendor-level rows not reported here); one secondary output differs only in the order of tied entries.

C Verification record for this version

Run on 2026-11-12 by `claude-sonnet-5-5` under the author's direction, from the stored raw responses and analysis tables. The scripts are in the data package (`verify/`).

1. `audit_recompute.py` and `recover_unresolved.py` were re-run from the raw provider responses into a clean directory (their earlier outputs deleted first). The recomputation compared 720 run-level values (domain and URL Jaccard for each of 360 AI runs) and found 0 mismatches, maximum absolute difference 0. The five output files (`audit_recompute_diffs.csv`, `audit_recompute_runs.csv`, `audit_serp_resolution.csv`, `recover_unresolved_cells.csv`, `recover_unresolved_results.csv`) are byte-identical to the versions used for Tables 7 and Figure 1.
2. `verify_tables.py` parsed the numbers printed in Tables 3 to 6 and in sections 4.6 and 7.4 and recomputed each from `PROMPT-OVERLAP.csv`, `CONCENTRATION.csv` and `SERP-STABILITY.csv` (cell means, category ranges, distinct domains, citation slots, top-five shares, concentration indices, class shares, the two Google stability means, the repeat-run standard deviations and their prompt counts). It checked 286 numbers (246 in Tables 3 to 6 and sections 4.6 and 7.4, and 40 in Table 7) and found 0 mismatches at the printed precision. A first run of the checker did fail, on Gemini rows of Table 6, because of a keying error in the checker itself (it mixed the `3.5 Flash-Lite` and `3.6 Flash` strata); the error was fixed in the checker and the manuscript was not changed. Table 7 was checked against `recover_unresolved_cells.csv` by the same script.
3. Medians of set sizes (8 and 7 cited domains; 14 Google domains; 6 Google direct URLs) and the Google pass counts (55 prompts with three passes, 5 with two) were recomputed from `RUN-OVERLAP.csv` and `SERP-STABILITY.csv`.
4. The checks do not include the coverage ratios of the keyword-reference runs, the domain-class labels, or the vendor-level analyses. They also do not test whether the raw responses are what the provider returned. They test that the printed numbers follow from the stored files.

D Literature search record

OBSERVED. Two searches are recorded: the search done for the first draft, whose query strings were not logged, and a logged re-run made on 2026-11-01 for this revision. This version makes no novelty claim at the domain or URL level (section 3); the SSRN items U1 to U7 surfaced by the search were read at abstract level on 2026-11-12 (D.5).

D.1 Draft search (query strings not recorded). Dates: 2026-11-01. Sources: arXiv landing pages and full text for the papers in Table 1 and the register; vendor and practitioner pages (Ahrefs, Semrush, BrightEdge, seoClarity, SE Ranking, DerivateX, Search Atlas, SparkToro); Google Search Central; the web through two search tools. The query strings used were not saved and cannot be reconstructed; this limitation is stated rather than guessed. Failed sources: Google Scholar (HTTP 429); ACM Digital Library (blocked); SSRN abstract 7517778 (blocked).

D.2 Logged re-run, 2026-11-01. Tool: TinyFish search and fetch (a web search API; results are ranked by the tool, not by Google Scholar). Each row is one query, run once. Query

strings are reproduced verbatim as logged.

#	Query string (verbatim)	Scope setting	Result use
Q1	overlap between Google top 10 results and AI search citations B2B software buyer intent domain Jaccard	research papers, 2024 onward	Found Yadav (2026), read; SSRN items U1, U2, U3, U4, U7 by snippet
Q2	generative search engine citations overlap with Google top 10 B2B SaaS vendors ChatGPT Gemini study	web	Vendor and agency blog posts, none opened. ThriveStack did not appear in the results; it was taken from the project's earlier literature register and opened by direct fetch
Q3	preregistered protocol AI search answer engines Google SERP overlap recommended vendors software	research papers, 2024 onward	No relevant result
Q4	ssrn AI search citations Google rankings overlap generative engine	web, restricted to papers.ssrn.com and ssrn.com	SSRN items U4, U5, U6 by snippet (other SSRN results on generative search were listed and are not relevant to overlap)
Q5	AI answer engines recommended vendors versus Google organic results overlap URL domain vendor level frozen protocol pilot	web, restricted to arxiv.org, papers.ssrn.com, dl.acm.org	No relevant result

Direct fetch attempts, same day: Google Scholar (query on "Google top 10", "AI search", citation overlap, B2B software, Jaccard): HTTP 429. ACM Digital Library search on generative search citations, Google overlap, B2B: blocked. SSRN abstract and delivery pages for items U1 to U7: blocked or unreachable. ThriveStack page: opened. Yadav PDF: opened.

D.3 SSRN items surfaced by the re-run, as they appeared in the 2026-11-01 result snippets. Snippets and titles only; the abstracts were read later (D.5).

Id	SSRN abstract id	Title as shown in the result	What the snippet suggested
U1	6240278	Query intent, not Google rank: What best predicts AI citation behavior	A Jaccard figure of 0.036 and a call to extend intent analysis
U2	7234418	Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. Measuring Vendor-Set Agreement Across Four AI Search Engines	B2B software buyer questions across several engines; vendor-set agreement
U3	7268163	Who The Machine Names: A Four-Week Panel Audit of AI Citation in Enterprise Banking Software	Panel design; vendor sets and pairwise Jaccard
U4	7528060	Who Does AI Cite? Reconciling Conflicting Evidence on the Sources Behind Generative Search Answers (2024-2026)	A review reconciling overlap figures, including Google top-ten overlap
U5	7517778	From Clicks to Citations: Generative AI Search, Zero … (title truncated in the result)	Cites Ahrefs data on how fresh AI-cited content is; the id is the one logged as unreadable in the first search

Id	SSRN abstract id	Title as shown in the result	What the snippet suggested
U6	6476021	The SEO-to-GEO Gap: Quantifying Ranking Factor …	AI Overviews and top organic results
U7	6707258	AI Search Impact Assessment: The ASIA Framework	B2B software buyer journey framework

D.4 Re-check, 2026-11-06. Direct fetches of the SSRN abstract pages for U1 (6240278), U2 (7234418) and U4 (7528060) again returned HTTP 403 on 2026-11-06; U3 was not retried. The SSRN submission guidelines page (last updated 2026-08-03 as shown on the page) was opened and is summarised in `SSRN-READINESS.md`. One vendor report was opened and added to Table 1 (GrackerAI, 2026, compliance automation). A second vendor series on single-sign-on and passwordless vendor-set convergence was seen only as a search-result snippet and was not opened.

Two further results (arXiv 2606.23057 on brand category ownership across models; a vendor study page from Citera) were seen in result lists and not opened.

D.5 Abstracts read through Crossref, 2026-11-12. The SSRN pages sit behind a bot check that was not bypassed. Crossref's public API returns SSRN's posted-content metadata, including the abstract, for each DOI (`https://api.crossref.org/works/10.2139/ssrn.<id>`). The abstracts of the seven items were read from those records. Full texts were not read. Title and first author are as Crossref gives them; the title of U1 differs from the title shown in the 2026-11-01 snippet, which suggests a revision.

Id	DOI	First author	Title (Crossref)	Record created
U1	10.2139/ssrn.6240278	Anthony Lee	Query Intent and Google Rank as Joint Predictors of AI Citation: A Multi-Platform Observational Study	2026-04-03
U2	10.2139/ssrn.7234418	Sairam Sivakumar	Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. Measuring Vendor-Set Agreement Across Four AI Search Engines	2026-08-10
U3	10.2139/ssrn.7268163	Avik Bal	Who The Machine Names: A Four-Week Panel Audit of AI Citation in Enterprise Banking Software	2026-08-16
U4	10.2139/ssrn.7528060	Rohit Kunal	Who Does AI Cite? Reconciling Conflicting Evidence on the Sources Behind Generative Search Answers (2024-2026)	2026-09-28
U5	10.2139/ssrn.7517778	Rohit Kunal	From Clicks to Citations: Generative AI Search, Zero-Click Behaviour, and the Rise of Generative Engine Optimization	2026-09-24
U6	10.2139/ssrn.6476021	Dmitrii Kargaev	The SEO-to-GEO Gap: Quantifying Ranking Factor Divergence Between Traditional and Generative Search	2026-04-27
U7	10.2139/ssrn.6707258	Avinash Tripathi	AI Search Impact Assessment: The ASIA Framework	2026-05-13

U1 to U4 are discussed in section 3. U5 to U7 are not cited as evidence: U5 is a secondary review of search behaviour, U6 an exploratory synthesis of ranking-factor studies, and U7 a framework applied to the author's client properties.