

Paraphrase and Fan-Out: What Gemini’s Queries Reveal (and Don’t Reveal) About Intent

N. Deplace
Traaker

October 2026

Abstract

Generative engines answer a prompt by issuing their own web searches (“query fan-out”). A growing practice extracts these machine queries and treats them as keywords to optimise content for. We test whether that practice measures user intent or measurement noise. Using the Gemini API, which returns the exact search queries and cited sources for each call, we ran 3 commercial intentions $\times$ 8 paraphrases $\times$ 3 to 5 executions (224 calls) on **gemini-3.8-flash** and **gemini-2.5-flash**, against four pre-registered, falsifiable hypotheses. Two of them fail. Paraphrases do yield different queries, but two executions of the *same* prompt share only 3–9% of their exact queries, so paraphrasing adds almost nothing beyond run-to-run non-determinism. Cited sources are not more stable than queries once both are measured against that noise. On the current model, half of the calls issue no search at all. On local intent, 88–94% of queries name a venue the model already lists without searching: the fan-out verifies a shortlist rather than discovering one. A keyword list extracted from machine queries therefore describes one execution and the model’s prior beliefs, not the user’s intent.

1 Introduction

When a generative engine answers a commercial question such as “which CRM for my small business?”, it may run several web searches before writing its answer. Tools and studies now harvest these queries [Nectiv Digital, 2025, AirOps, 2025] and rank their recurring words (“reviews”, “2025”,

“comparison”) as targets for content. The implicit assumption is that machine queries are a stable, meaningful trace of what the user wants.

We set out to show the opposite with a stronger claim: that the lexical path is unstable while the answer is stable, so that the constant object is the intention and extracting keywords amounts to measuring noise. We wrote the claim as four hypotheses, each with a refutation criterion fixed before data collection, and committed not to publish any hypothesis that failed.

The data refute the stability half of that claim. They support a different and, for practitioners, more consequential conclusion: machine queries are not reproducible even for an identical prompt, and on local intent they mostly encode the model’s own shortlist. We report the refuted hypotheses as such; the falsifiable design is what makes the remaining claims defensible.

2 Method

2.1 Instrument

Gemini is, to our knowledge, the only mainstream engine with an official API that returns its own search queries: `generateContent` with the `google_search` tool returns `groundingMetadata.webSearchQueries` (the exact queries issued) and `groundingChunks[]` (the sources) [Google]. ChatGPT’s queries are visible in the JSON returned by its web interface, which is how Nectiv Digital [2025] collects them, but no official API exposes them. All results below are measured on Gemini; we do not generalise them to other engines.

We used `gemini-3.8-flash`, the model behind the `gemini-flash-latest` alias on 2026-10-01, pinned by name for reproducibility, and `gemini-2.5-flash` as a robustness check. The served version reported in each response matched the requested model for every call. Source domains were taken from `web.title`, which carries the domain name; the `uri` field is a redirect.

2.2 Design

Three intentions cover distinct commercial situations, all in French: a CRM for a small business (B2B SaaS), the best Neapolitan pizzeria in Paris open on Sunday evening (B2C local), and an electric bike for a 15 km commute at about 2,000 euros (B2C considered purchase). Each intention has eight paraphrases of 21–25 words (mean 21.4) that preserve every constraint: a direct question (V1), lexical synonyms (V2), reversed constraint order (V3),

informal register (V4), imperative form (V5), narrative context (V6), a negatively phrased constraint (V7), and a paraphrase of the head term itself (V8). V8 changes the most and is reported separately; the main analysis uses V1–V7.

The primary run used 3 executions per prompt (72 calls per model). A power rerun on `gemini-3.8-flash` used 5 executions per prompt for the pizzeria and bike intentions (80 calls), giving 70 same-prompt pairs for pizzeria.

2.3 Measures

All comparisons are made within one intention. For each pair of calls we compute the Jaccard index of their exact query sets, of their word sets, and of their cited domain sets. Pairs of executions of the same prompt form the *control*; pairs of different paraphrases form the *paraphrase* condition. Calls that issued no search are excluded from H1, H1-bis and H2 and counted only for H3; otherwise each no-search call would score 0 against every searched call and inflate the apparent divergence. Confidence intervals (95%) are obtained by bootstrap over variants.

2.4 Hypotheses

- H1** Equivalent paraphrases produce different sets of machine queries. Refuted if the paraphrase Jaccard on queries exceeds 0.80.
- H1-bis (control)** This divergence exceeds the engine’s non-determinism. Refuted if the gap between control and paraphrase Jaccard is below 0.10.
- H2** Paraphrasing moves the queries well beyond run-to-run noise, but moves the cited sources no more than that noise. With $E_q = J_{\text{ctrl}}(q) - J_{\text{para}}(q)$ and $E_d = J_{\text{ctrl}}(d) - J_{\text{para}}(d)$, refuted if $E_d \geq 0.10$, or $E_q - E_d < 0.10$, or the lower bound of its interval is ≤ 0 .
- H3** A substantial share of prompts triggers no search. Refuted if no call is made without search.

H2 was initially written as “domains overlap more than queries”. A simulation in which every paraphrase draws its queries at random from one shared pool showed that this test passes on pure noise (domain Jaccard 0.47 against 0.05 for queries), because there are far fewer possible domains than possible query strings. A baseline across unrelated intentions does not

fix this, since CRM, pizzeria and bike sites barely overlap. We therefore replaced it, before running the study, by the difference-in-differences above, which correctly refutes H2 on the pure-noise simulation.

3 Results

3.1 Verdicts

Table 1 summarises the four hypotheses (V1–V7, searched calls only). Adding V8 changes no verdict.

	3.8-flash	3.8-flash, pizzeria $\times 5$	2.5-flash
Calls with search	32/63	35/35	62/63
Control pairs	26	70	61
H1: paraphrase $J(q)$	0.011 holds	0.006 holds	0.010 holds
H1-bis: gap	0.074 [0.005, 0.147] refuted	0.049 [0.018, 0.073] refuted	0.020 [0.006, 0.038] refuted
H2: $E_q - E_d$	−0.018 refuted	−0.068 refuted	0.010 refuted
H3: calls without search	49% holds	—	1.6% fails in practice

Table 1: Hypothesis verdicts. Brackets: 95% bootstrap intervals.

3.2 Machine queries are not reproducible

Two executions of the same prompt share 3% (2.5) to 6–9% (3.8) of their exact queries. On 2.5, 62 searched calls produced 237 distinct queries. Paraphrases share about 1%, so H1 holds trivially, but the paraphrase effect sits within non-determinism: on both models the whole H1-bis interval lies below 0.10. The borderline pizzeria value of 0.098 observed in the primary 3.8 run came from 21 control pairs; with 70 pairs it falls to 0.049 [0.018, 0.073]. At the word level the gap is larger (0.115 on pizzeria), which reflects shared constraint words rather than shared queries.

3.3 Sources are not more stable than queries

Cited domains overlap more than queries in absolute terms (about 0.23–0.34), but once both are measured against run-to-run noise, paraphrasing moves sources at least as much as queries. On 3.8 pizzeria, $E_d = 0.117$ [0.047, 0.204], above the 0.10 threshold, and the difference-in-differences is negative. On 2.5, paraphrasing moves neither queries nor sources beyond noise. The original “domains > queries” test would have reported H2 as holding (0.23 against 0.01), which is the artefact the simulation predicted.

3.4 Search is often absent, depending on the model

On `gemini-3.8-flash`, 49% of calls issued no search: the CRM prompts never searched (0/24) and were answered from the model alone, the bike prompts searched about half the time, and the pizzeria prompts always searched. On `gemini-2.5-flash`, only 1.6% of calls skipped search. H3 therefore holds on the current model but cannot be stated as a general property. It converges with ChatGPT data, where 31% of commercial prompts trigger a search and local prompts the most (59%) [Nectiv Digital, 2025, Digital Information World, 2025].

3.5 On local intent, the fan-out verifies a shortlist

On `gemini-3.8-flash`, 94% of pizzeria queries (66 of 70) in the primary run name a specific restaurant or ranking, for example `peppe pizzeria paris dimanche soir reservation`. Apart from the name, their words are the prompt’s own constraints (Sunday evening, opening hours, price, booking). For bike, 57% of queries name a brand or model.

To test where these names come from, we ran the pizzeria V1 prompt 3 times without the search tool and 3 times with it, with the model’s thought summaries enabled. Without any search, the summaries already name 12 restaurants. All 5 restaurants queried in the tool runs belong to that set; across the full 72-call run, 9 of the 12 queried restaurants do. One summary states: “First, I need to consider my top candidates.” The queries then check the prompt’s constraints for each candidate.

3.6 When paraphrasing matters, it changes the answer

On 3.8 pizzeria, the set of restaurants named per call overlaps 0.555 between executions of the same prompt and 0.395 between paraphrases (gap 0.160). Paraphrasing changes which candidates the model shortlists, and the cited sources follow. This is the reverse of “stable answer, unstable path”. No such effect is detectable on bike, where many calls do not search.

4 Discussion

The keyword-extraction argument fails, but not for the reason we expected. A list of words harvested from machine queries describes one execution: rerun the same prompt and most queries change. Choosing a better phrasing of the prompt does not help, because paraphrasing adds little beyond that

noise. On local intent, the list is mostly the model’s own shortlist combined with the user’s constraints; it reflects what the model already believes rather than what the user is looking for. Rankings of recurring query words, such as those of Nectiv Digital [2025], aggregate exactly this kind of signal and should be read as a description of model behaviour, not as an inventory of intents.

Two claims should not be made from these data: that cited sources converge while queries diverge, and that these results are confirmed on ChatGPT. Published ChatGPT data measure neither paraphrase effects nor reproducibility, and only the absence of search and the high search rate of local prompts converge.

5 Limitations

The study is a reproducible demonstration of mechanism, not a population estimate: 3 intentions, French prompts, and 224 calls. The API with the `google_search` tool is not the consumer Gemini application. Thought summaries are written by the model and are not raw traces; the API does not timestamp thinking against searching, so the order between candidate selection and search is inferred. The entity lexicon used to classify queries was built by hand from the raw data. The CRM intention provides no data for H1 and H2 on 3.8, since it never searched, and the bike intention has few control pairs. Engine behaviour changes over time; the runs should be repeated on the publication date and the raw files archived.

References

- AirOps. Webinar recap: Chris long on query fan-out. <https://aiops.com/blog/webinar-recap-chris-long-query-fan-out>, 2025. Accessed 2026-10-01.
- Digital Information World. How often does ChatGPT search the internet? <https://www.digitalinformationworld.com/2025/10/how-often-does-chatgpt-search-internet.html>, October 2025. Accessed 2026-10-01.
- Google. Grounding with Google search — Gemini API documentation. <https://ai.google.dev/gemini-api/docs/google-search>. Accessed 2026-10-01.

Nectiv Digital. New data study: What queries is ChatGPT using behind the scenes? <https://nectivdigital.com/new-data-study-what-queries-is-chatgpt-using-behind-the-scenes/>, 2025. Accessed 2026-10-01.