Barkhausen AI

NOTE

Don't Measure Once: what six weeks of repeated AI-search sampling showed

A 2026 study sampled four AI answer engines every day for about six weeks to measure how much their answers move on their own.

“Every leap begins as a crackle.”

KIND	Note
DATE	2026
SUBJECTS	Measurement & statistics
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

A 2026 study sampled four AI answer engines every day for about six weeks to measure how much their answers move on their own. Running identical prompts across ChatGPT, Gemini, Google AI Mode, and Perplexity for 45–46 collection days between 2026-01-24 and 2026-03-20, over four Swiss-German commercial verticals from a Swiss vantage, it found that answers on consecutive days shared only 34–42% of their cited sources and — across the three verticals that cleared a brand-detection threshold — 45–59% of their mentioned brands. Simultaneous same-day reruns overlapped by 32–43% in sources, so much of the turnover is request-to-request stochasticity rather than genuine day-over-day movement. Within-24-hour source stability differed sharply by engine, from 0.233 to 0.505. This note reports what the study measured — instability magnitudes per engine and vertical for that window — and what its single-window, single-region design does not support.

CONTENTS

Measuring an AI answer engine once produces a number the engine itself would not reproduce the next day. A 2026 study established this by sampling four engines repeatedly over about six weeks and reporting how much their answers moved with no change to the prompt [1]. Because the measurement was repeated daily rather than taken once, the study can separate genuine day-over-day movement from the request-to-request randomness a single reading hides. This note reports what it measured, what it found, and what its single-window, single-region design supports.

1 What was measured and how

The study issued fixed prompts to four AI answer engines — ChatGPT, Gemini, Google AI Mode, and Perplexity — over four Swiss-German commercial verticals: telecommunications, real-estate sales, sporting goods, and consumer electronics [1]. Collection ran from a Swiss vantage, using Swiss IP addresses and locale, on 45–46 collection days within the window 2026-01-24 → 2026-03-20. Brand mentions were detected with a per-vertical lexicon of 32 to 51 brands, and a vertical qualified for the brand-overlap figures only if brand detection cleared a 70% threshold, which three of the four verticals met. Overlap between answer sets was measured with the Jaccard similarity and, alongside it, rank-biased overlap at a persistence parameter of $p = 0.9$ — the rank-aware companion measure set out in BA-C-2 §5.3, whose value is only interpretable once that parameter is stated. One collection day, 2026-01-30, carried roughly twice the usual citation volume and was excluded as an outlier. Two comparisons were made: identical prompts on consecutive days, which mixes any genuine daily drift with request-to-request noise, and simultaneous same-day reruns, which isolate the noise alone. The gap between the two is what lets the study attribute observed turnover to one scale or the other.

2 What it found

On consecutive days, the same prompts returned substantially different answers. Cited-source sets overlapped by a Jaccard similarity of 0.34 to 0.42 from one day to the next, and mentioned-brand sets — across the three qualifying verticals — overlapped by 0.45 to 0.59 [1]. Brands were somewhat more stable than the sources beneath them, but neither repeated.

Much of that turnover was not day-over-day movement at all. Simultaneous same-day reruns of the same prompt overlapped in their sources by only 0.32 to 0.43 [1] — nearly the same range as the consecutive-day figure — which locates most of the instability at the request-to-request scale rather than in genuine daily change. The pooled same-day figure also conceals sharp differences between engines: within a 24-hour span, the paper reports source overlap of 0.233 for ChatGPT, 0.505 for Gemini, and 0.318 for Google AI Mode [1]. A stability figure averaged across engines therefore describes none of them, which is why an instability figure has to be stated per engine.

3 What the design supports, and what it does not

The study supports quantitative statements about instability magnitude for the engines, verticals, window, and region it sampled: how much the answers of these four engines moved, day to day and within a day, over this window in these Swiss-German markets. It does not support generalization beyond that frame. The figures are specific to four engines, four commercial verticals, a Swiss vantage and locale, and one window in early 2026; they do not establish what the same engines do in other languages or regions, on other topics, or at other times. Read together, the two comparisons place a large share of the observed instability at the sub-daily scale, where averaging repeated draws can reduce it, rather than at the daily scale, where it would instead be genuine movement to be tracked. The convention BA-C-3 draws its day-

to-day and per-engine sampling requirements from this evidence, and states them as requirements precisely because the magnitudes are not portable.

4 Limitations

The single window is the central limit: measured instability is itself a moving quantity, and a figure from early 2026 describes engines that have since changed. Brand detection was lexicon-based, so it captured the enumerated brands per vertical and would miss mentions outside the lexicon or in unanticipated surface forms; the 70% qualification threshold and per-vertical brand counts are part of how the brand figures should be read. The Swiss-German, Swiss-vantage design controls region deliberately, but for that reason speaks only to that region. These results describe the engines as sampled during the stated window; engines change without notice, and results should be assumed perishable.

REFERENCES

1. Schulte, Bleeker, and Kaufmann. *Don't Measure Once: Measuring Visibility in AI Search (GEO)* (2026). https://arxiv.org/abs/2604.07585 Accessed 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). Don't Measure Once: what six weeks of repeated AI-search sampling showed. https://barkhausen.ai/notes/dont-measure-once-study/