Barkhausen AI

NOTE

Paraphrase brittleness in production recommendation: what one 2026 study measured

A 2026 study measured how much an AI recommendation set changes when the same underlying request is reworded, rather than assuming it stays fixed.

“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 measured how much an AI recommendation set changes when the same underlying request is reworded, rather than assuming it stays fixed. Across roughly 12,000 runs on two production models — one from OpenAI, one from Anthropic — it compared a same-prompt rerun baseline against two kinds of paraphrase. Rerunning an identical prompt reproduced the recommendation set with a Jaccard similarity of 0.50–0.61 within an engine. A cosmetic reword, preserving the need, cut the corpus-mean overlap to 0.288, 21–32 percentage points below that baseline; a reword that added a constraint cut it to 0.135, 37–48 points below. Increasing reasoning effort did not remove the gap. The authors argue that progress requires different measurement units, not merely more prompt sampling. This note reports what the study measured and what its single-day design does and does not support.

CONTENTS

Whether a reworded question changes an AI system's recommendations is testable rather than a matter of assumption, and a 2026 study tested it directly on production systems [1]. The study held a user's underlying need fixed, varied only the wording used to express it, and measured how much the resulting recommendation set changed — against a baseline of what changes when the exact same prompt is simply run again. That design separates two things a single measurement confounds: the run-to-run randomness inherent in the system, and the additional movement caused by rewording. This note reports what it measured, what it found, and what its single-day design does and does not support.

1 What was measured and how

The study ran roughly 6,000 paraphrase runs and roughly 6,000 same-prompt rerun controls — about 12,000 runs in total — across two production models, one from OpenAI and one from Anthropic [1]. The rerun baseline crossed 50 prompts across four cells at N = 30 same-prompt repeats each; the paraphrase corpus used about 20 base prompts, roughly five variants apiece, across three of those cells at N = 20 — roughly 6,000 runs on each side. All collection took place on a single calendar day, which removes day-to-day drift as a confound but also means the study says nothing about it. The outcome measure was the overlap between two recommendation sets, quantified as a Jaccard similarity — the size of the intersection over the size of the union — so that a value of 1 means identical sets and 0 means no shared items. Two kinds of paraphrase were distinguished: a cosmetic reword that preserves the underlying need, and a reword that adds a qualifying constraint to it.

2 What it found

Rerunning an identical prompt did not reproduce the recommendation set. Within an engine, the same-prompt rerun Jaccard similarity ran from 0.50 to 0.61 [1] — so even with the wording held exactly fixed, between two-fifths and one-half of the recommendation set turned over from one run to the next. This rerun baseline is the study's reference point, and it matters because it sets the bar a paraphrase effect has to clear to count as more than run-to-run noise.

Both kinds of paraphrase cleared it. A cosmetic reword that preserved the need cut the overlap to a corpus mean of 0.288, with a clustered 95% confidence interval of [0.215, 0.361] — 21 to 32 percentage points below the rerun baseline [1]. A reword that added a constraint cut it further, to 0.135, with a clustered 95% confidence interval of [0.098, 0.175] — 37 to 48 percentage points below the baseline [1]. In both cases, changing the wording of a request while preserving its intent replaced substantially more of the recommendation set than re-running the identical request did.

3 Reasoning effort did not close the gap

The study also varied the models' reasoning effort, to test whether more deliberation stabilizes the output. It did not: increasing reasoning effort moved rerun stability by a bounded amount between −0.015 and +0.005 [1], leaving the paraphrase gap essentially intact. The authors draw a methodological conclusion from this rather than an optimization one — that progress in measuring these systems requires different units of measurement, not merely more prompt sampling, since sampling more of a single fixed phrasing cannot recover the variation that lives between phrasings.

4 What the design supports

Within its scope, the study supports a specific claim: on the two production models tested, on a single day, a paraphrase that preserves a request's intent moves the recommendation set well beyond the system's

own run-to-run noise floor, and adding a constraint moves it further still. The convention BA-C-3 builds a sampling requirement on exactly this evidence — that a query must be represented as a distribution over real phrasings rather than a single fixed sentence, because a single sentence estimates visibility at one arbitrary point of that distribution.

5 Limitations

The design is single-day by construction, so it cannot separate paraphrase sensitivity from temporal drift, and its numbers describe two model families in one commercial-recommendation domain, not answer engines in general. The overlap measure is set-based and order-insensitive, so it counts membership changes and not rank changes within an otherwise stable set. And the figures describe the systems as they behaved on the day sampled; these systems change without notice, and the specific values should be assumed perishable. The durable result is the relationship — an intent-preserving paraphrase moves recommendations by more than reruns do, and reasoning effort does not remove that — not the exact percentages.

REFERENCES

1. Jack, Lehman, Maloney, and Xu. *Paraphrase Brittleness in Production Retrieval-Augmented Commercial Recommendation: Reproducibility Below the Rerun-Stability Baseline* (2026). https://arxiv.org/abs/2605.27440 Accessed 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). Paraphrase brittleness in production recommendation: what one 2026 study measured. https://barkhausen.ai/notes/paraphrase-brittleness-study/