Barkhausen AI

NOTE

The GEO experiment: what a controlled study showed about content and AI answers

In 2024, a controlled experiment introduced the term Generative Engine Optimization and tested, rather than assumed, which content changes shift how often a source is cited in an AI-generated answer.

“Every leap begins as a crackle.”

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

ABSTRACT

In 2024, a controlled experiment introduced the term Generative Engine Optimization and tested, rather than assumed, which content changes shift how often a source is cited in an AI-generated answer. Working from a 10,000-query benchmark and a simulated generative engine, the study found that adding quotations, statistics, or cited sources measurably increased a source's contribution to the answer, that rewriting text into a more authoritative tone did not, and that traditional keyword stuffing performed worse than making no change at all. A further test found that when all competing sources apply the same techniques, lower-ranked sources gain the most. This note reports what the study found and what its design does and does not support.

CONTENTS

The term Generative Engine Optimization was introduced by a controlled experiment rather than a marketing claim [1]. Published at KDD 2024, the study built a benchmark of AI-answer queries, applied a fixed set of content-modification techniques to the sources feeding those queries, and measured the change in each source's contribution to the resulting answer against an unmodified baseline. That design — hold the query fixed, vary the content, measure the effect — is what lets the study report which techniques helped, which did not, and by roughly how much, inside its own test conditions. This note summarizes those findings and is explicit about where the design's conclusions stop.

1 The benchmark and what was measured

The study's benchmark, GEO-bench, drew 10,000 queries from nine existing sources — including real anonymized search-engine query logs, exam-style reasoning questions, subreddit questions, and machine-generated queries — each paired with the source documents a generative engine would need to answer it [1]. Its primary generative engine was simulated: GPT-3.5 was prompted to synthesize an answer from the supplied sources, and each source's contribution to that answer was scored with an author-defined metric, Position-Adjusted Word Count, which weights a source's presence in the answer by an exponentially decaying function of the position at which it is cited, so an early, prominent citation counts for more than a late, buried one [1]. A secondary, GPT-3.5-graded "subjective impression" score supplemented this word-count metric. To check whether findings from the simulation held outside it, the authors additionally ran a subset of 200 test queries against Perplexity.ai, a deployed generative engine, and compared the results [1].

2 What increased a source's contribution to the answer

Three content changes produced measurable gains over the unmodified baseline on the simulated benchmark: adding quotations from relevant sources, adding statistics, and citing sources outright. The best-performing of these improved the position-adjusted word-count metric by 41% over the unmodified baseline, with quotation addition, statistics addition, and citing sources all landing in the 30–40% range on that metric [1]. When the same techniques were tested against the deployed Perplexity.ai engine on 200 held-out queries rather than the simulation, quotation addition led on the word-count metric, improving it by 22% over baseline — a smaller gain than the same techniques produced on the simulated benchmark, but positive and directionally consistent. On the subjective-impression metric the strongest performer this time was statistics addition, at 37% over baseline; quotation addition also improved on this metric, though by a smaller margin than statistics addition [1].

3 What did not help

Two other techniques were tested and did not reproduce that pattern. Rewriting source text into a more assertive, persuasive, "authoritative" register produced, in the study's own words, "no significant improvement" over the unmodified baseline; the authors concluded that the generative engine tested was "already somewhat robust to such changes" [1]. Stuffing text with additional query keywords — the traditional search-engine-optimization tactic — performed worse than making no change: on the Perplexity.ai validation it scored about 10% below the unmodified baseline, and it likewise underperformed on the simulated benchmark [1]. Of every technique the study tested, authoritative-tone rewriting and keyword stuffing were the only two to fail to beat doing nothing.

4 A redistribution effect, under one specific assumption

The study ran a further test in which every competing source for a query was modified with the same techniques at once, rather than just one — the condition the authors describe as anticipating a future "where all source contents are optimized using GEO" [1]. Under that specific assumption, sources ranked lower in the underlying search results gained the most: for the citing-sources technique, the source ranked fifth gained 115.1% in visibility while the source ranked first lost 30.3% [1]. The authors describe this as evidence the technique can "democratize" which sources get cited, benefiting smaller or lower-ranked content creators relative to already-dominant ones [1]. The finding is conditional on its stated premise — universal, simultaneous adoption — and does not describe what happens when only one source among competitors applies these techniques; the study does not claim the two scenarios produce the same redistribution.

5 Limitations

The study's primary results come from a simulated generative engine (GPT-3.5 synthesizing answers from supplied sources) rather than a production system, and its secondary metric was scored by a model from the same family used to generate the answers being scored, which the design does not rule out as a source of shared bias. The deployed-engine check was limited to one engine and 200 queries, a check on direction rather than an independent replication at the scale of the main benchmark. The study's own word-count and subjective-impression metrics are author-defined rather than externally standardized, and the paper is a conference proceeding rather than a metric adopted across the field. Most importantly for a 2026 reader, every percentage above describes a specific 2023–2024 model and benchmark; generative engines and their retrieval and synthesis behavior have changed substantially since, and none of these figures should be read as a forecast for any engine in production today. The directions the study found — quotations, statistics, and citations help; authoritative-tone rewriting does not; keyword stuffing hurts; and universal adoption of a technique compresses the advantage of top-ranked sources — are the durable result of a controlled test. The specific percentages are not.

REFERENCES

1. Aggarwal, Murahari, Rajpurohit, Kalyan, Narasimhan, and Deshpande. *GEO: Generative Engine Optimization* (2024). https://arxiv.org/abs/2311.09735 Accessed 2026-07-08. [archived]

HOW TO CITE

Barkhausen AI (2026). The GEO experiment: what a controlled study showed about content and AI answers. https://barkhausen.ai/notes/the-geo-experiment/