Semantic Multi-Cluster Coverage and LLM Citation Frequency: A Controlled Generative Engine Optimization Experiment

Hadeer ElMalah[a]

[a] *SIXT rent a car*

1. Abstract

Large language model citation behavior within Advanced Retrieval-Augmented Generation (RAG) pipelines remains an underspecified black box, limiting Generative Engine Optimization (GEO) practitioners to move beyond iterative experimentation toward principled content strategy. This paper investigates whether Semantic Compression, a vector retrieval framework (Raja & Vats, 2025), can serve as a surrogate optimization objective for Advanced RAG citation behavior, and whether the theory-grounded GEO approach produces measurable increase in LLM citation frequency. The GEO consisted of off-site content optimization guidelines emphasizing semantic multi-cluster topical comprehensiveness, prompt and intent alignment. A controlled experiment across four external partner domains, comprising eleven off-site pages measured over a two-week pre/post window, yielded an observed 320% increase (+308% baseline adjusted) in citation frequency under full treatment and ~200% (~+188% baseline-adjusted) under partial treatment, compared to a near-zero baseline shift in the reference condition, across ChatGPT, Google AI Overviews, Google AI Mode, Perplexity, and Gemini. These results provide preliminary evidence for Semantic Multi-Cluster Coverage, a coverage first GEO framework we derived from Semantic Compression, in which optimizing off-site content for semantic coverage rather than proximity measurably increases LLM citation frequency.

2. Introduction

Empirical and survey evidence indicates that approximate nearest neighbor (ANN) search adopted by Basic RAG approach is subject to limitations in retrieval precision and contextual coverage (Gao et al., 2024). Benchmark evidence quantifies the scale of this gap: on the Comprehensive RAG Benchmark, adding RAG raises factual question-answering accuracy only from ≤34% to 44%, and even state-of-the-art industry systems answer just 63% of questions without hallucination, with the largest failures concentrated on facts that are dynamic, long-tail, or complex (Yang et al., 2024). Advanced RAG frameworks address these limitations through a multi-stage pipeline. Several component-level decisions within this pipeline have been empirically studied; however, the attribution behavior of language models operating within RAG pipelines remains incompletely resolved, limiting practitioners from moving beyond iterative experimentation toward principled GEO strategy design (Aggarwal et al., 2024). We propose Semantic Multi-Cluster Coverage: a coverage-first GEO objective derived from Semantic Compression (Raja & Vats, 2025), and test whether coverage-first optimization increases LLM citation frequency.

3. Theoretical Parallel Between Semantic Compression and Advanced RAG

To overcome the semantically redundant outputs by Top-k ANN vector search systems that lack contextual diversity and richness, Raja & Vats (2025) introduced semantic compression, a retrieval

SEASON 2026, September 15–17, 2026, Hamburg, Germany
EMAIL: hadeer.a.elmalah@gmail.com
ORCID: 0009-0005-7814-1334
DOI: https://doi.org/10.48441/4427.3734

framework that aims to select a compact, representative set of vectors that captures the broader semantic structure around a query. Semantic Compression shifts the goal from Top-k ANN's proximity to Semantic Compression's coverage representation (Raja & Vats, 2025, Figure 1).

Just as Advanced RAG balances broad semantic understanding with retrieval precision (Gao et al., 2024), Semantic Compression embodies the same theoretical duality at the vector level, capturing broader semantic structure and enforcing compactness through distributed semantic clusters rather than relying on proximity alone.

Semantic Compression is therefore similar to Advanced RAG in nature and characteristics, and understanding and optimizing for it offers a principled basis for Advanced RAG GEO.

4. Test Methodology

This experiment tests the hypothesis that optimizing off-site content according to a Semantic Compression framework by putting content on our partner websites that is spread across multiple semantic clusters will increase our chances of retrievals and citations across large language models (LLMs).

4.1 Multi-Cluster Prompt Tracking Strategy

Semantic Multi-Cluster Coverage, our coverage-first GEO strategy, starts with multi-cluster prompt tracking. Peec.ai, an AI Search Analytics tool, was used to track 24 geo-relevant prompts across different buckets to capture different topic angles and expand semantic clusters, covering different compression types, improving LLM retrieval rates and post-retrieval citation frequency.

4.2 AI Content Writing Agent

To implement the multi-cluster content at scale, a writing agent was developed. The agent receives as input: (i) the prompts to optimize for, (ii) the top-cited URLs associated with those prompts, and (iii) the fanout queries and common terms associated with each prompt. The agent then executes the following steps:

- Prompt Intent analysis: The agent decodes prompt; maps the need to SIXT services and fleet.
- Multi-cluster coverage: based on the top-cited URLs of the prompts, the agent verifies comprehensive coverage against a 95% topical alignment.
- Semantic foundation & matching: Against the top-cite URL of the prompts, the agent benchmarks its vector embeddings at 0.85+ cosine similarity.
- Sentiment and tone: using amplifiers and avoiding negations and diminishers, adding authority signals and testing the influence of NLP sentiment analysis on re-ranking within LLM answer postretrieval (ongoing).
- SIXT knowledge integration: from the SIXT knowledge Model Context Protocol (MCP) for information on location-based fleet.

4.3 Test Environment

The experiment was conducted in a controlled off-site environment across four domains external to SIXT's domain. The test subjects consisted of 11 listings published on partner sites; all interventions were off-site, with zero on-page changes to SIXT-owned properties. The design comprised a 2-week baseline versus 2-week post-change window across 3 partner domains: a full treatment group (2 domains), a partial treatment group (1 domain), and 1 control group, with 11 off-site listings treated. Monthly organic clicks: 740K+ across 3 treatment partner sites groups. Partner domains were selected for high domain authority (a backlink-based measure of a site's SEO ranking strength), the same

geographic market, and a direct travel-niche classification (i.e., sites topically dedicated to travel, matching SIXT's category).

4.4 Evaluation Metrics

The primary success metric is increase in citation frequency (URL retrievals x citation rate) of the treatment groups.

5. Results

5.1 Citation Frequency

Over the 14-day* post-optimization period, citation frequency increased substantially across the tracked LLM platforms across all four experimental conditions.

Condition	Observed	Baseline-Adjusted
Full treatment (2 domains)	+320%	~+308%
Partial treatment (1 domain)	+200%	~+188%
Control group / no optimization (1 domain)	+12.2%	~0%

**Partial treatment reflects one domain where optimization was applied under operational constraints; this condition was observed over 12 days rather than 14, owing to a delayed go-live on the partner website. The reference condition, one domain in the same geographic area with no optimization applied, recorded a +12.2% observed increase, which normalizes to approximately zero after baseline adjustment, consistent with expected AI search fluctuations.*

Citations were predominantly generated by Google AI Overviews, Google AI Mode, and Perplexity during the test window. Post-test, ChatGPT citation frequency increased sharply, followed by AI Mode and Gemini.

Sentiment language optimization showed no measurable effect on sentiment change within the initial two-week post-optimization window; few competitor listings exhibited sentiment-driven changes during the same period.

Citation frequency continued to increase beyond the test period. One full-treatment domain rose from a 13.3-citation baseline to 28.5 citations during the 14-day test window, then to 83.9 citations in the two weeks following, and further to 102 citations in the subsequent two-week period (Figure 1). The +320% full-treatment result is pooled across both full-treatment domains, whereas Figure 1 traces the trajectory of one of them.

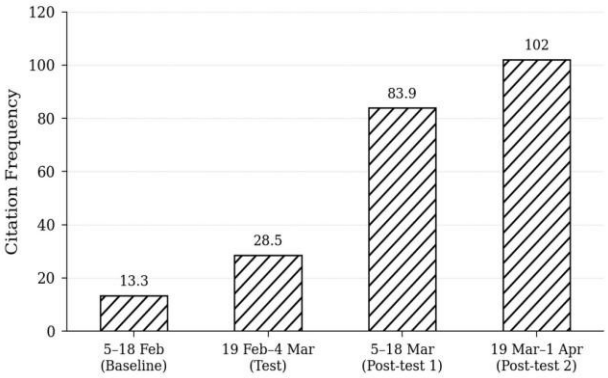

***Figure 1**. Citation frequency progression for one full-treatment domain across baseline, test, and two post-test periods.*

5.2 Scope and Limitations

The results reported reflect optimization applied exclusively to advanced retrieval-augmented generation. A separate, ongoing test examining the effects of graph-augmented RAG on citation frequency is being conducted in parallel and is not reflected in the figures above. These results are preliminary: given the small number of domains, no significance testing was performed, and the posttest trajectory in Figure 1 has no concurrent control, so seasonality cannot be ruled out as a contributing factor.

6. Discussion & Implications

Semantic Compression serves as a theoretical lens and surrogate objective for optimizing towards Advanced RAG. Off-site GEO embedded in Semantic Multi-Cluster Coverage is viable and measurable. As AI-driven answer engines increasingly mediate information access, retrieval pipelines are becoming the new ranking layer, and the +308% baseline-adjusted citation increase observed in the full-treatment condition indicates that this layer responds to principled, theory-grounded intervention. Crucially, these gains were not only absolute: measured (in Peec.ai) against tracked competitors over the same period, SIXT posted the largest increase in both AI Visibility (+5.8) and Share of Voice (+1.6), while competitors were split between gains and losses (Figure 2). Rising AI-search adoption likely lifts citations category-wide but Share of Voice is a relative measure. SIXT's +1.6 gain reflects share taken from competitors, indicating the intervention outpaced the category-wide tide rather than merely riding it.

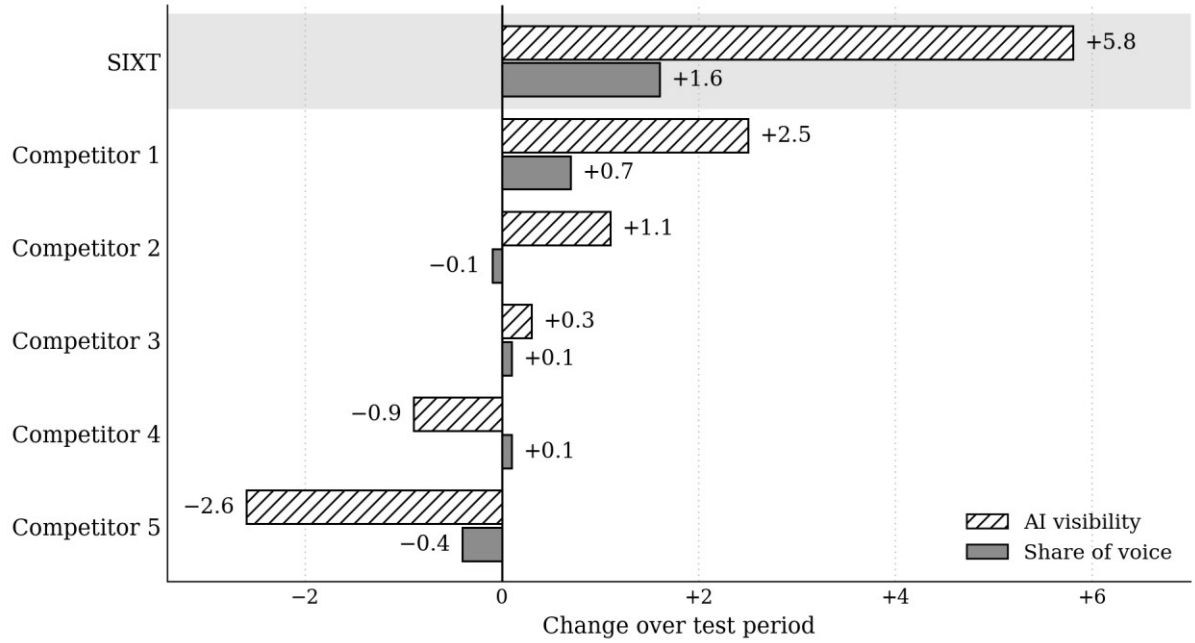

Figure 2. Change in AI Visibility and Share of Voice across the test period, SIXT versus tracked competitors (tracked prompt bucket). SIXT led both metrics (+5.8 Visibility, +1.6 Share of Voice); green marks competitors that gained, red those that lost.

7. Conclusion

These findings are preliminary from an ongoing research project and bounded by a single brand and four partner domains. Next steps are to replicate the test across different locations, and to expand testing area into a hybrid advanced RAG and graph-augmented retrievals based on new findings.

8. References

1. Raja, R., & Vats, A. (2025). Beyond nearest neighbors: Semantic compression and graph-augmented retrieval for enhanced vector search. arXiv. https://doi.org/10.48550/arXiv.2507.19715

2. Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., … Dong, X. L. (2024). CRAG — Comprehensive RAG Benchmark. arXiv. https://doi.org/10.48550/arXiv.2406.04744

3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledgeintensive NLP tasks. arXiv. https://doi.org/10.48550/arXiv.2005.11401

5. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv. https://doi.org/10.48550/arXiv.2307.03172

6. Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. arXiv. https://doi.org/10.48550/arXiv.2305.14283

7. Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. arXiv. https://doi.org/10.48550/arXiv.2210.08726

8. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24) (pp. 5–16). arXiv. https://doi.org/10.48550/arXiv.2311.09735