GEO: Generative Engine Optimization

Pranjal Aggarwal*
Indian Institute of Technology Delhi
New Delhi, India
pranjal2041@gmail.com

Vishvak Murahari*
Princeton University
Princeton, USA
murahari@cs.princeton.edu

Tanmay Rajpurohit
Independent
Seattle, USA
tanmay.rajpurohit@gmail.com

Ashwin Kalyan
Independent
Seattle, USA
asaavashwin@gmail.com

Karthik Narasimhan
Princeton University
Princeton, USA
karthikn@princeton.edu

Ameet Deshpande
Princeton University
Princeton, USA
asd@princeton.edu

ABSTRACT

The advent of large language models (LLMs) has ushered in a new paradigm of search engines that use generative models to gather and summarize information to answer user queries. This emerging technology, which we formalize under the unified framework of generative engines (GEs), can generate accurate and personalized responses, rapidly replacing traditional search engines like Google and Bing. Generative Engines typically satisfy queries by synthesizing information from multiple sources and summarizing them using LLMs. While this shift significantly improves *user* utility and *generative search engine* traffic, it poses a huge challenge for the third stakeholder – website and content creators. Given the black-box and fast-moving nature of generative engines, content creators have little to no control over *when* and *how* their content is displayed. With generative engines here to stay, we must ensure the creator economy is not disadvantaged. To address this, we introduce Generative Engine Optimization (GEO), the first novel paradigm to aid content creators in improving their content visibility in generative engine responses through a flexible black-box optimization framework for optimizing and defining visibility metrics. We facilitate systematic evaluation by introducing GEO-bench, a large-scale benchmark of diverse user queries across multiple domains, along with relevant web sources to answer these queries. Through rigorous evaluation, we demonstrate that GEO can boost visibility by up to 40% in generative engine responses. Moreover, we show the efficacy of these strategies varies across domains, underscoring the need for domain-specific optimization methods. Our work opens a new frontier in information discovery systems, with profound implications for both developers of generative engines and content creators.[1]

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; Machine learning; • **Information systems** → **Web searching and information discovery**.

KEYWORDS

generative models, search engines, datasets and benchmarks

ACM Reference Format:
Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671900

1 INTRODUCTION

The invention of traditional search engines three decades ago revolutionized information access and dissemination globally [4]. While they were powerful and ushered in a host of applications like academic research and e-commerce, they were limited to providing a list of relevant websites for user queries. However, the recent success of large language models [5, 21] has paved the way for better systems like BingChat, Google's SGE, and perplexity.ai that combine conventional search engines with generative models. We dub these systems generative engines (GE) because they *search* for information and *generate* multi-modal responses by using multiple sources. Technically, generative engines (Figure 2) retrieve relevant documents from a database (like the internet) and use large neural models to generate a response grounded on the sources, ensuring attribution and a way for the user to verify the information.

The usefulness of generative engines for developers and users is evident – users access information faster and more accurately, while developers craft precise and personalized responses, improving user satisfaction and revenue. However, generative engines disadvantage the third stakeholder – website and content creators. Generative Engines, in contrast to traditional search engines, remove the need to navigate to websites by directly providing a precise and comprehensive response, potentially reducing organic traffic to websites and impacting their visibility [16]. With millions of small businesses and individuals relying on online traffic and visibility for their livelihood, generative engines will significantly disrupt the creator economy. Further, the black-box and proprietary nature of generative engines makes it difficult for content

*Equal Contribution
[1]Code and Data available at https://generative-engines.com/GEO/

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671900

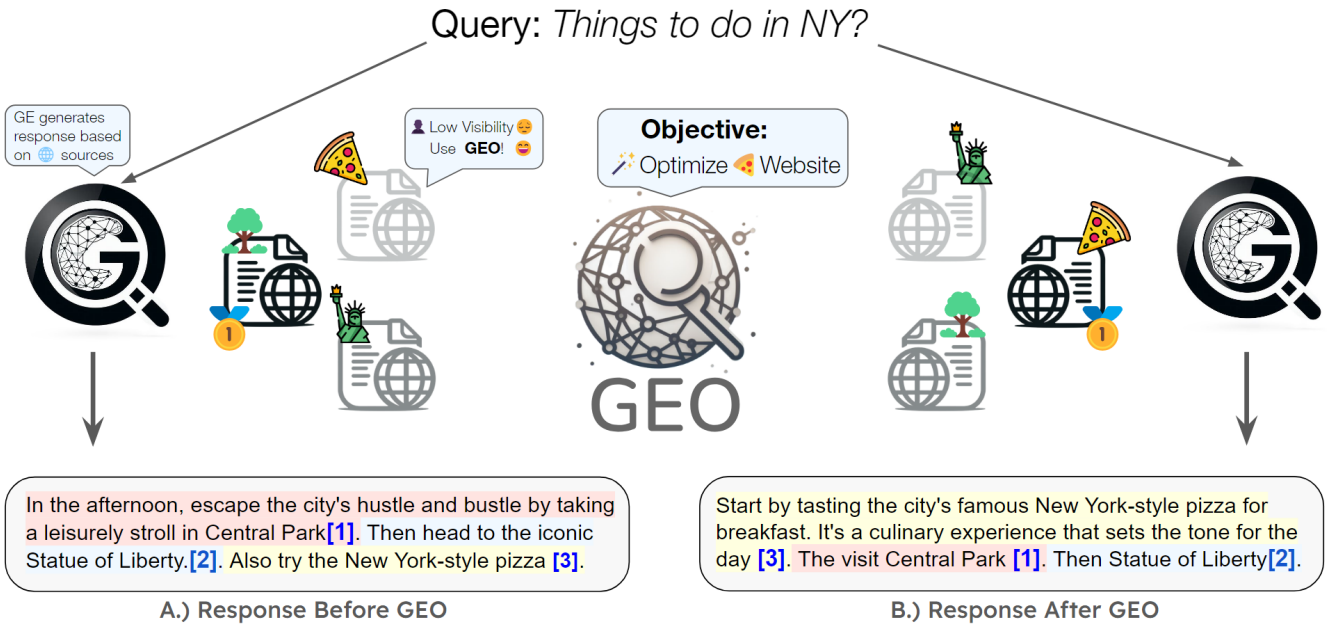

Figure 1: Our proposed Generative Engine Optimization (GEO) method optimizes websites to boost their visibility in Generative Engine responses. GEO's black-box optimization framework then enables the website owner of the pizza website, which lacked visibility originally, to optimize their website to increase visibility under Generative Engines. Further, GEO's general framework allows content creators to define and optimize their custom visibility metrics, giving them greater control in this new emerging paradigm.

creators to *control* and *understand* how their content is ingested and portrayed.

In this work, we propose the first general creator-centric framework to optimize content for generative engines, which we dub Generative Engine Optimization (GEO), to empower content creators to navigate this new search paradigm. GEO is a flexible black-box optimization framework for optimizing web content visibility for proprietary and closed-source generative engines (Figure 1). GEO ingests a source website and outputs an optimized version by tailoring and calibrating the presentation, text style, and content to increase visibility in generative engines.

Further, GEO introduces a flexible framework for defining visibility metrics tailor-made for generative engines as the notion of visibility in generative engines is more nuanced and multi-faceted than traditional search engines (Figure 3). While average ranking on the response page is a good measure of visibility in traditional search engines, which present a linear list of websites, this does not apply to generative engines. Generative Engines provide rich, structured responses and embed websites as inline citations in the response, often embedding them with different lengths, at varying positions, and with diverse styles. This necessitates the need for visibility metrics tailor-made for generative engines, which measure the visibility of attributed sources over multiple dimensions, such as relevance and influence of citation to query, measured through both an objective and a subjective lens.

To facilitate faithful and extensive evaluation of GEO methods, we propose GEO-bench, a benchmark consisting of 10000 queries from diverse domains and sources, adapted for generative engines.

Through systematic evaluation, we demonstrate that our proposed Generative Engine Optimization methods can boost visibility by up to 40% on diverse queries, providing beneficial strategies for content creators. Among other things, we find that including citations, quotations from relevant sources, and statistics can significantly boost source visibility, with an increase of over 40% across various queries. We also demonstrate the efficacy of Generative Engine Optimization on Perplexity.ai, a real-world generative engine and demonstrate visibility improvements up to 37%.

In summary, our contributions are three-fold:

(1) We propose Generative Engine Optimization, the first general optimization framework for website owners to optimize their websites for generative engines. Generative Engine Optimization can improve the visibility of websites by up to 40% on a wide range of queries, domains, and real-world black-box generative engines.

(2) Our framework proposes a comprehensive set of visibility metrics specifically designed for generative engines and enables content creators to flexibly optimize their content through customized visibility metrics.

(3) To foster faithful evaluation of GEO methods in generative engines, we propose the first large-scale benchmark consisting of diverse search queries from wide-ranging domains and datasets specially tailored for Generative Engines.

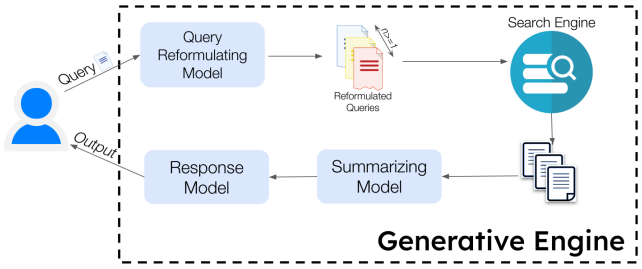

Figure 2: Overview of Generative Engines. Generative Engines primrarily consists of a set of generative models and a search engine to retrieve relevant documents. Generative Engines take user query as input and through a series of steps generate a final response that is grounded in the retrieved sources with inline attributions.

2 FORMULATION & METHODOLOGY

2.1 Formulation of Generative Engines

Despite the deployment of numerous generative engines to millions of users, there is currently no standard framework. We provide a formulation that accommodates various modular components in their design. We describe a generative engine, which includes several backend generative models and a search engine for source retrieval. A Generative Engine (GE) takes a user query q_u and returns a natural language response r, where P_U represents personalized user information. The GE can be represented as a function:

$$f_{GE} := (q_u, P_U) \rightarrow r \tag{1}$$

Generative Engines comprise two crucial components: a.) A set of generative models $G = \{G_1, G_2 ... G_n\}$, each serving a specific purpose like query reformulation or summarization, and b.) A search engine SE that returns a set of sources $S = \{s_1, s_2 ... s_m\}$ given a query q. We present a representative workflow in Figure 2, which, at the time of writing, closely resembles the design of BingChat. This workflow breaks down the input query into a set of simpler queries that are easier to consume for the search engine. Given a query, a query re-formulating generative model, $G_1 = G_{qr}$, generates a set of queries $Q^1 = \{q_1, q_2 ... q_n\}$, which are then passed to the search engine SE to retrieve a set of ranked sources $S = \{s_1, s_2, ..., s_m\}$. The sets of sources S are passed to a summarizing model $G_2 = G_{sum}$, which generates a summary Sum_j for each source in S, resulting in the summary set ($Sum = \{Sum_1, Sum_2, ..., Sum_m\}$). The summary set is passed to a response-generating model $G_3 = G_{resp}$, which generates a cumulative response r backed by sources S. In this work, we focus on single-turn Generative Engines, but the formulation can be extended to multi-turn Conversational Generative Engines (Appendix A).

The response r is typically a structured text with embedded citations. Citations are important given the tendency of LLMs to hallucinate information [10]. Specifically, consider a response r composed of sentences $\{l_1, l_2 ... l_o\}$. Each sentence may be backed by a set of citations that are part of the retrieved set of documents $C_i \subset S$. An ideal generative engine should ensure all statements in the response are supported by relevant citations (high citation recall), and all citations accurately support the statements they're associated with (high citation precision) [14]. We refer readers to Figure 3 for a representative generative engine response.

2.2 Generative Engine Optimization

The advent of search engines led to search engine optimization (SEO), a process to help website creators optimize their content to improve search engine rankings. Higher rankings correlate with increased visibility and website traffic. However, traditional SEO methods are not directly applicable to Generative Engines. This is because, unlike traditional search engines, the generative model in generative engines is not limited to keyword matching, and the use of language models in ingesting source documents and response generation results in a more nuanced understanding of text documents and user query. With generative engines rapidly emerging as the primary information delivery paradigm and SEO is not directly applicable; new techniques are needed. To this end, we propose Generative Engine Optimization, a new paradigm where content creators aim to increase their visibility (or impression) in generative engine responses. We define the visibility of a website (also referred to as a citation) c_i in a cited response r by the function $Imp(c_i, r)$, which the website creator wants to maximize. From the generative engine's perspective, the goal is to maximize the visibility of citations most relevant to the user query, i.e., maximize $\sum_i f(Imp(c_i, r), Rel(c_i, q, r))$, where $Rel(c_i, q, r)$ measures the relevance of citation c_i to the query q in the context of response r and f is determined by the exact algorithmic design of generative engine and is a black-box function to end-users. Further, both the functions Imp and Rel are subjective and not well-defined yet for generative engines, and we define them next.

2.2.1 Impressions for Generative Engines. In SEO, a website's impression (or visibility) is determined by its average ranking over a range of queries. However, generative engines' output nature necessitates different impression metrics. Unlike search engines, Generative Engines combine information from multiple sources in a single response. Factors such as length, uniqueness, and presentation of the cited website determine the true visibility of a citation. Thus, as illustrated in Figure 3, while a simple ranking on the response page serves as an effective metric for impression and visibility in conventional search engines, such metrics are not applicable to generative engine responses.

In response to this challenge, we propose a suite of impression metrics designed with three key principles in mind: 1.) The metrics should hold relevance for creators, 2.) They should be explainable, and 3.) They should be easily comprehensible by a broad spectrum of content creators. The first of these metrics, the "Word Count" metric, is the normalized word count of sentences related to a citation. Mathematically, this is defined as:

$$Imp_{wc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s|}{\sum_{s \in S_r} |s|} \tag{2}$$

Here S_{c_i} is the set of sentences citing c_i, S_r is the set of sentences in the response, and $|s|$ is the number of words in sentence s. In cases where a sentence is cited by multiple sources, we share the word count equally with all the citations. Intuitively, a higher word count correlates with the source playing a more important part in the answer, and thus, the user gets higher exposure to that source.

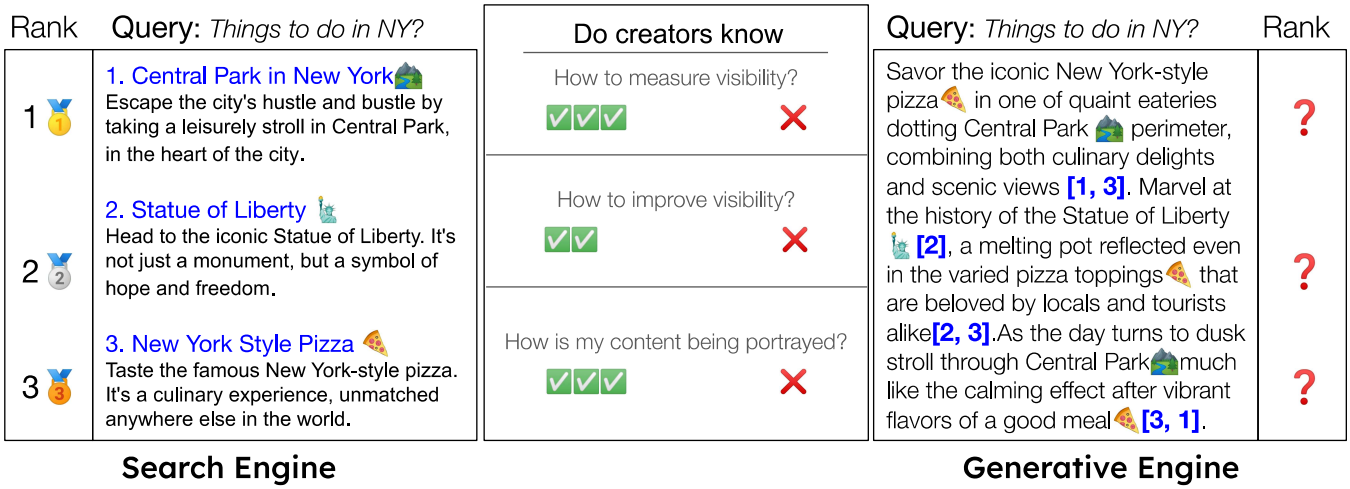

Figure 3: Ranking and Visibility Metrics are straightforward in traditional search engines, which list website sources in ranked order with verbatim content. However, Generative Engines generate rich, structured responses, often embedding citations in a single block interleaved with each other. This makes ranking and visibility nuanced and multi-faceted. Further, unlike search engines, where significant research has been conducted on improving visibility, optimizing visibility in generative engine responses remains unclear. To address these challenges, our black-box optimization framework proposes a series of well-designed impression metrics that creators can use to *gauge* and *optimize* their website's performance and also allows the creator to define their impression metrics.

However, since "Word Count" is not impacted by the ranking of the citations (whether it appears first, for example), we propose a position-adjusted count that reduces the weight by an exponentially decaying function of the citation position:

$$Imp_{pwc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s| \cdot e^{-\frac{pos(s)}{|S|}}}{\sum_{s \in S_r} |s|} \tag{3}$$

Intuitively, sentences that appear first in the response are more likely to be read, and the exponent term in definition Imp_{pwc} gives higher weightage to such citations. Thus, a website cited at the top may have a higher impression despite having a lower word count than a website cited in the middle or end of the response. Further, the choice of exponentially decaying function is motivated by several studies showing click-through rates follow a power-law as a function of ranking in search engines [7, 8]. While the above impression metrics are objective and well-grounded, they ignore the subjective aspects of the impact of citations on the user's attention. To address this, we propose the "Subjective Impression" metric, which incorporates facets such as the relevance of the cited material to the user query, influence of the citation, uniqueness of the material presented by a citation, subjective position, subjective count, probability of clicking the citation, and diversity in the material presented. We use G-Eval [15], the current state-of-the-art for evaluation with LLMs, to measure each of these sub-metrics.

2.2.2 Generative Engine Optimization methods for website. To improve impression metrics, content creators must make changes to their website content. We present several generative engine-agnostic strategies, referred to as Generative Engine Optimization methods (GEO). Mathematically, every GEO method is a function $f : W \rightarrow W_i'$, where W is the initial web content, and W' is the modified content after applying the GEO method. The modifications can range from simple stylistic alterations to incorporating new content in a structured format. A well-designed GEO is equivalent to a black-box optimization method that, without knowing the exact algorithmic design of generative engines, can increase the website's visibility and implement textual modifications to W independent of the exact queries.

For our experiments, we apply Generative Engine Optimization methods on website content using a large language model, prompted to perform specific stylistic and content changes to the website. In particular, based on the GEO method defining a specific set of desired characteristics, the source content is modified accordingly. We propose and evaluate several such methods:

1: Authoritative: Modifies text style of the source content to be more persuasive and authoritative, **2. Statistics Addition:** Modifies content to include quantitative statistics instead of qualitative discussion, wherever possible, **3. Keyword Stuffing:** Modifies content to include more keywords from the query, as expected in classical SEO optimization. **4. Cite Sources & 5. Quotation Addition:** Adds relevant citations and quotations from credible sources respectively, 6.) **6. Easy-to-Understand:** Simplifies the language of website, while **7. Fluency Optimization** improves the fluency of website text. **8. Unique Words & 9. Technical Terms:** involves adding unique and technical terms respectively wherever possible,

These methods cover diverse general strategies that website owners can implement quickly and use regardless of the website content. Further, except for methods 3, 4, and 5, the remaining methods enhance the presentation of existing content to increase its persuasiveness or appeal to the generative engine, without requiring extra content. On the other hand, methods 3,4 and 5 may

require some form of additional content. To analyze the performance gain of our methods, for each input user query, we randomly select one source website to be optimized and apply each of the GEO methods separately on the same source. We refer readers to Appendix B.4 for more details on GEO methods.

3 EXPERIMENTAL SETUP

3.1 Evaluated Generative Engine

In accordance with previous works [14], we use a 2-step setup for Generative Engine design. The first step involves fetching relevant sources for input query, followed by a second step where an LLM generates a response based on the fetched sources. Similar to prevous works, we do not use summarization and provide the whole response for each source. Due to context length limitations and quadratic scaling cost based on the context size of transformer models, only the top 5 sources are fetched from the Google search engine for every query. The setup closely mimics the workflow used in previous works and the general design adopted by commercial GEs such as you.com and perplexity.ai. The answer is then generated by the gpt3.5-turbo model [20] using the same prompt as prior work [14]. We sample 5 different responses at temperature=0.7, to reduce statistical deviations.

Further in Section C.1, we evaluate the same Generative Engine Optimization methods on Perplexity.ai, which is a commercially deployed generative engine, highlighting the generalizability of our proposed Generative Engine Optimization methods.

3.2 Benchmark : GEO-bench

Since there is currently no publicly available dataset containing Generative Engine related queries, we curate **GEO-bench**, a benchmark consisting of 10K queries from multiple sources, repurposed for generative engines, along with synthetically generated queries. The benchmark includes queries from nine different sources, each further categorized based on their target domain, difficulty, query intent, and other dimensions.

Datasets: **1. MS Macro, 2. ORCAS-1, and 3. Natural Questions:** [1, 6, 13] These datasets contain real anonymized user queries from Bing and Google Search Engines. These three collectively represent the common set of datasets that are used in search engine related research. However, Generative Engines will be posed with far more difficult and specific queries with the intent of synthesizing answers from multiple sources instead of searching for them. To this end, we repurpose several other publicly available datasets: **4. AllSouls:** This dataset contains essay questions from "All Souls College, Oxford University." The queries in this dataset require Generative Engines to perform appropriate reasoning to aggregate information from multiple sources. **5. LIMA:** [25] contains challenging questions requiring Generative Engines to not only aggregate information but also perform suitable reasoning to answer the question (e.g., writing a short poem, python code.). **6. Davinci-Debtate** [14] contains debate questions generated for testing Generative Engines. **7. Perplexity.ai Discover**[2]**:** These queries are sourced from Perplexity.ai's Discover section, which is an updated list of trending queries on the platform. **8. ELI-5**[3]**:** This dataset contains questions from the ELI5 subreddit, where users ask complex questions and expect answers in simple, layman's terms. **9. GPT-4 Generated Queries:** To supplement diversity in query distribution, we prompt GPT-4 [21] to generate queries ranging from various domains (e.g., science, history) and based on query intent (e.g., navigational, transactional) and based on difficulty and scope of generated response (e.g., open-ended, fact-based).

. Our benchmark comprises 10K queries divided into 8K, 1K, and 1K for train, validation, and test splits, respectively. We preserve the real-world query distribution, with our benchmark containing 80% informational queries and 10% each for transactional and navigational queries. Each query is augmented with the cleaned text content of the top 5 search results from the Google search engine.

Tags. Optimizing website content often requires targeted changes based on the task's domain. Additionally, a user of Generative Engine Optimization may need to identify an appropriate method for only a subset of queries, considering multiple factors such as domain, user intent, and query nature. To facilitate this, we tag each query with one of seven different categories. For tagging, we employ the GPT-4 model and manually verify high recall and precision on the test split.

Overall, GEO-bench consists of queries from 25 diverse domains such as Arts, Health, and Games; it features a range of query difficulties from simple to multi-faceted; includes 9 different types of queries such as informational and transactional; and encompasses 7 different categorizations. Owing to its specially designed high diversity, the size of the benchmark, and its real-world nature, GEO-bench is a comprehensive benchmark for evaluating Generative Engines and serves as a standard testbed for assessing them for various purposes in this and future works. We provide more details about GEO-bench in Appendix B.2.

3.3 GEO Methods

We evaluate 9 different proposed GEO methods as described in Section 2.2.2. We compare them with a baseline, which measures the impression metric of unmodified website sources. We evaluate methods on the complete GEO-bench test split. Further, to reduce variance in results, we run our experiments on five different random seeds and report the average.

3.4 Evaluation Metrics

We utilize the impression metrics as defined in Section 2.2.1. Specifically, we employ two impression metrics: **1. Position-Adjusted Word Count**, which combines word count and position count. To analyze the effect of individual components, we also report scores on the two sub-metrics separately. **2. Subjective Impression**, which is a subjective metric encompassing seven different aspects: 1) relevance of the cited sentence to the user query, 2) influence of the citation, assessing the extent to which the generated response relies on the citation, 3) uniqueness of the material presented by a citation, 4) subjective position, gauging the prominence of the positioning of source from the user's viewpoint, 5) subjective count, measuring the amount of content presented from the

[2]https://www.perplexity.ai/discover

[3]https://huggingface.co/datasets/eli5_category

Method	Position-Adjusted Word Count			Subjective Impression							
	Word	Position	**Overall**	Rel.	Infl.	Unique	Div.	FollowUp	Pos.	Count	**Average**
Performance without Generative Engine Optimization											
No Optimization	19.5	19.3	19.3	19.3	19.3	19.3	19.3	19.3	19.3	19.3	19.3
Non-Performing Generative Engine Optimization methods											
Keyword Stuffing	17.8	17.7	17.7	19.8	19.1	20.5	20.4	20.3	20.5	20.4	20.2
Unique Words	20.7	20.5	20.5	20.5	20.1	19.9	20.4	20.2	20.7	20.2	20.4
High-Performing Generative Engine Optimization methods											
Easy-to-Understand	22.2	22.4	22.0	20.2	21.0	20.0	20.1	20.1	20.9	19.9	20.5
Authoritative	21.8	21.3	21.3	22.3	22.1	22.4	23.1	22.2	23.1	22.7	22.9
Technical Terms	23.1	22.7	22.7	20.9	21.7	20.5	21.2	20.8	21.9	20.8	21.4
Fluency Optimization	25.1	24.6	24.7	21.1	22.9	20.4	21.6	21.0	22.4	21.1	21.9
Cite Sources	24.9	24.5	24.6	21.4	22.5	21.0	21.6	21.2	22.2	20.7	21.9
Quotation Addition	27.8	27.3	**27.2**	23.8	25.4	23.9	24.4	22.9	24.9	23.2	**24.7**
Statistics Addition	25.9	25.4	25.2	22.5	24.5	23.0	23.3	21.6	24.2	23.0	23.7

Table 1: Absolute impression metrics of GEO methods on GEO-bench. Performance Measured on Two metrics and their sub-metrics. Compared to baselines, simple methods like Keyword Stuffing traditionally used in SEO don't perform well. However, our proposed methods such as Statistics Addition and Quotation Addition show strong performance improvements across all metrics. The best methods improve upon baseline by 41% and 28% on Position-Adjusted Word Count and Subjective Impression respectively. For readability, Subjective Impression scores are normalized with respect to Position-Adjusted Word Count resulting in similar baseline scores.

citation as perceived by the user, 6) likelihood of the user clicking the citation, and 7) diversity of the material presented. These sub-metrics assess diverse aspects that content creators can target to improve one or more areas effectively. Each sub-metric is evaluated using GPT-3.5, following a methodology akin to that described in G-Eval [15]. In G-Eval, a form-based evaluation template is provided to the language model, along with a GE generated response with citations. The model outputs a score (computed by sampling multiple times) for each citation. However, since G-Eval scores are poorly calibrated, we normalize them to have the same mean and variance as Position-Adjusted Word Count to enable a fair and meaningful comparison. We provide the exact templates used in Appendix B.3.

Furthermore, all impression metrics are normalized by multiplying them with a constant factor so that the sum of the impressions of all citations in a response equals 1. In our analysis, we compare methods by calculating the relative improvement in impression. For an initial generated response r from sources $S_i \in \{s_1, \ldots, s_m\}$, and a modified response r', the relative improvement in impression for each source s_i is measured as:

$$Improvement_{s_i} = \frac{Imp_{s_i}(r') - Imp_{s_i}(r)}{Imp_{s_i}(r)} \times 100 \tag{4}$$

The modified response r' is produced by applying the GEO method being evaluated to one of the sources s_i. The source s_i selected for optimization is chosen randomly but remains constant for a particular query across all GEO methods.

4 RESULTS

We evaluate various Generative Engine Optimization methods designed to optimize website content for better visibility in Generative Engine responses, compared against a baseline with no optimization. Our evaluation used GEO-bench, a diverse benchmark of user queries from multiple domains and settings. Performance was measured using two metrics: *Position-Adjusted Word Count* and *Subjective Impression*. The former considers word count and citation position in the GE's response, while the latter computes multiple subjective factors, giving an overall impression score.

Table 1 details the absolute impression metrics of different methods on multiple metrics. The results reveal that our GEO methods consistently outperform the baseline across all metrics on GEO-bench. This shows the robustness of these methods to varying queries, yielding significant improvements despite query diversity. Specifically, our top-performing methods, Cite Sources, Quotation Addition, and Statistics Addition, achieved a relative improvement of 30-40% on the *Position-Adjusted Word Count* metric and 15-30% on the *Subjective Impression* metric. These methods, involving adding relevant statistics (Statistics Addition), incorporating credible quotes (Quotation Addition), and including citations from reliable sources (Cite Sources) in the website content, require minimal changes but significantly improve visibility in GE responses, enhancing both the credibility and richness of the content.

Interestingly, stylistic changes such as improving fluency and readability of the source text (Fluency Optimization and Easy-to-Understand) also resulted in a significant visibility boost of 15-30%. This suggests that Generative Engines value not only content but also information presentation.

Method	Relative Improvement (%) in Visibility				
	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
Authoritative	-6.0	4.1	-0.6	12.6	6.1
Fluency Opt.	-2.0	5.2	3.6	-4.4	2.2
Cite Sources	-30.3	2.5	20.4	15.5	115.1
Quotation Addition	-22.9	-7.0	3.5	25.1	99.7
Statistics Addition	-20.6	-3.9	8.1	10.0	97.9

Table 2: Visibility changes through GEO methods for sources with different Rankings in Search Engine. GEO is especially helpful for lower ranked websites.

Method	Top Performing Tags		
	Rank-1	Rank-2	Rank-3
Authoritative	Debate	History	Science
Fluency Opt.	Business	Science	Health
Cite Sources	Statement	Facts	Law & Gov.
Quotation Addition	People & Society	Explanation	History
Statistics Addition	Law & Gov.	Debate	Opinion

Table 3: Top Performing categories for each of the GEO methods. Website-owners can choose relevant GEO strategy based on their target domain.

Further, given generative models are often designed to follow instructions, one would expect a more persuasive and authoritative tone in website content to boost visibility. However, we find no significant improvement, demonstrating that Generative Engines are already somewhat robust to such changes. This highlights the need for website owners to focus on improving content presentation and credibility.

Finally, we evaluate keyword stuffing, i.e., adding more relevant keywords to website content. While widely used for Search Engine Optimization, we find such methods offer little to no improvement on generative engine's responses. This underscores the need for website owners to rethink optimization strategies for generative engines, as techniques effective in search engines may not translate to success in this new paradigm.

5 ANALYSIS

5.1 Domain-Specific Generative Engine Optimizations

In Section 4, we presented the improvements achieved by GEO across the entirety of the GEO-bench benchmark. However, in real-world SEO scenarios, domain-specific optimizations are often applied. With this in mind, and considering that we provide categories for every query in GEO-bench, we delve deeper into the performance of various GEO methods across these categories.

Table 3 provides a detailed breakdown of the categories where our GEO methods have proven to be most effective. A careful analysis of these results reveals several intriguing observations. For instance, Authoritative significantly improves performance in debate-style questions and queries related to the "historical" domain. This aligns with our intuition, as a more persuasive form of writing is likely to hold more value in debates.

Similarly, the addition of citations through Cite Sources is particularly beneficial for factual questions, likely because citations provide a source of verification for the facts presented, thereby enhancing the credibility of the response. The effectiveness of different GEO methods varies across domains. For example, as shown in row 5 of Table 3, domains such as 'Law & Government' and question types like 'Opinion' benefit significantly from the addition of relevant statistics in the website content, as implemented by Statistics Addition. This suggests that data-driven evidence can enhance the visibility of a website in particular contexts. The method Quotation Addition is most effective in the 'People & Society,' 'Explanation,' and 'History' domains. This could be because these domains often involve personal narratives or historical events, where direct quotes can add authenticity and depth to the content. Overall, our analysis suggests that website owners should strive towards making domain-specific targeted adjustments to their websites for higher visibility.

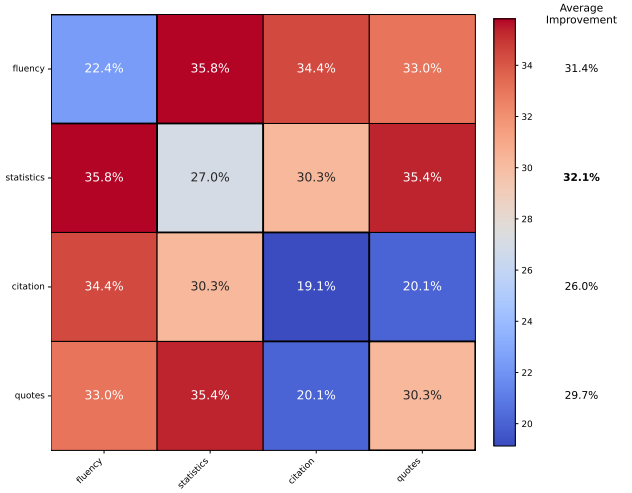

Figure 4: Relative Improvement on using combination of GEO strategies. Using Fluency Optimization and Statistics Addition in conjunction results in maximum performance. The rightmost column shows using Fluency Optimization with other strategies is most beneficial.

5.2 Optimization of Multiple Websites

In the evolving landscape of Generative Engines, GEO methods are expected to become widely adopted, leading to a scenario where all source contents are optimized using GEO. To understand the implications, we conducted an evaluation of GEO methods by optimizing all source contents simultaneously, with results presented in Table 2. A key observation is the differential impact of GEO on websites based on their Search Engine Results Pages (SERP) ranking. Notably, lower-ranked websites, which typically struggle for visibility, benefit significantly more from GEO. This is because traditional search engines rely on multiple factors, such as the number of backlinks and domain presence, which are challenging for small creators to achieve. However, since Generative Engines utilize

Method	GEO Optimization	Relative Improvement
Cite Sources	**Query:** What is the secret of Swiss chocolate With per capita annual consumption averaging between 11 and 12 kilos, Swiss people rank among the top chocolate lovers in the world (According to a survey conducted by The International Chocolate Consumption Research Group [1])	**132.4%**
Statistics Addition	**Query:** Should robots replace humans in the workforce? **Source:** Not here, and not now — until recently. The big difference is that the robots have come not to destroy our lives, but to disrupt our work, with a staggering 70% increase in robotic involvement in the last decade.	**65.5%**
Authoritative	**Query:** Did the jacksonville jaguars ever make it to the superbowl? **Source:** It is important to note that The Jaguars have never appeared made an appearance in the Super Bowl. However, They have achieved an impressive feat by securing 4 divisional titles to their name. , a testament to their prowess and determination.	**89.1%**

Table 4: Representative examples of GEO methods optimizing source website. Additions are marked in green and Deletions in red. Without adding any substantial new information, GEO methods significantly increase the visibility of the source content.

generative models conditioned on website content, factors such as backlink building should not disadvantage small creators. This is evident from the relative improvements in visibility shown in Table 2. For example, the Cite Sources method led to a substantial 115.1% increase in visibility for websites ranked fifth in SERP, while on average, the visibility of the top-ranked website decreased by 30.3%.

This finding highlights GEO's potential as a tool to democratize the digital space. Many lower-ranked websites are created by small content creators or independent businesses, who traditionally struggle to compete with larger corporations in top search engine results. The advent of Generative Engines might initially seem disadvantageous to these smaller entities. However, the application of GEO methods presents an opportunity for these content creators to significantly improve their visibility in Generative Engine responses. By enhancing their content with GEO, they can reach a wider audience, leveling the playing field and allowing them to compete more effectively with larger corporations.

5.3 Combination of GEO Strategies

While individual GEO strategies show significant improvements across various domains, in practice, website owners are expected to employ multiple strategies in conjunction. To study the performance improvements achieved by combining GEO strategies, we consider all pairs of combinations of the top 4 performing GEO methods, namely Cite Sources, Fluency Optimization, Statistics Addition, and Quotation Addition. Figure 4 displays the heatmap of relative improvement in the Position-Adjusted Word Count visibility metric achieved by combining different GEO strategies. The analysis demonstrates that the combination of Generative Engine Optimization methods can enhance performance, with the best combination (Fluency Optimization and Statistics Addition) outperforming any single GEO strategy by more than 5.5%[4]. Furthermore, Cite Sources significantly boosts performance when used in conjunction with other methods (Average: 31.4%), despite it being relatively less effective when used alone (8% lower than Quotation Addition). The findings underscore the importance of studying GEO methods in combination, as they are likely to be used by content creators in the real world.

[4]Due to cost constraints, the analysis was conducted on a subset of 200 examples from the test split, and therefore the numbers presented here differ from those in Table 1

5.4 Qualitative Analysis

We present a qualitative analysis of GEO methods in Table 4, containing representative examples where GEO methods boost source visibility with minimal changes. Each method optimizes a source through suitable text additions and deletions. In the first example, we see that simply adding the source of a statement can significantly boost visibility in the final answer, requiring minimal effort from the content creator. The second example demonstrates that adding relevant statistics wherever possible ensures increased source visibility in the final Generative Engine response. Finally, the third row suggests that merely emphasizing parts of the text and using a persuasive text style can also lead to improvements in visibility.

6 GEO IN THE WILD : EXPERIMENTS WITH DEPLOYED GENERATIVE ENGINE

Method	Position-Adjusted Word Count	Subjective Impression
No Optimization	24.1	24.7
Keyword Stuffing	21.9	28.1
Quotation Addition	**29.1**	32.1
Statistics Addition	26.2	**33.9**

Table 5: Absolute impression metrics of GEO methods on GEO-bench with Perplexity.ai as GE. While SEO methods such as Keyword Stuffing perform poorly, our proposed GEO methods generalize well to multiple generative engines significanlty improve content visibility.

To reinforce the efficacy of our proposed Generative Engine Optimization methods, we evaluate them on Perplexity.ai, a real deployed Generative Engine with a large user base. Results are

in Table 5. Similar to our generative engine, Quotation Addition performs best in Position-Adjusted Word Count with a 22% improvement over the baseline. Methods that performed well in our generative engine such as Cite Sources, Statistics Addition show improvements of up to 9% and 37% on the two metrics. Our observations, such as the ineffectiveness of traditional SEO methods like Keyword Stuffing, are further highlighted, as it performs 10% worse than the baseline. The results are significant for three reasons: 1) they underscore the importance of developing different Generative Engine Optimization methods to benefit content creators, 2) they highlight the generalizability of our proposed GEO methods on different generative engines, 3) they demonstrate that content creators can use our easy-to-implement proposed GEO methods directly, thus having a high real-world impact. We refer readers to Appendix C.1 for more details.

7 RELATED WORK

Evidence-based Answer Generation: Previous works have used several techniques for answer generation backed by sources. Nakano et al. [19] trained GPT-3 to navigate web environments to generate source-backed answers. Similarly, other methods [17, 23, 24] fetch sources via search engines for answer generation. Our work unifies these approaches and provides a common benchmark for improving these systems in the future. In a recent working draft, Kumar and Lakkaraju [11] showed that strategic text sequences can manipulate LLM recommendations to enhance product visibility in generative engines. While their approach focuses on increasing product visibility through adversarial text, our method introduces non-adversarial strategies to optimize any website content for improved visibility in generative engine search results.

Retrieval-Augmented Language Models: Several recent works have tackled the issues of limited memory of language models by fetching relevant sources from a knowledge base to complete a task [3, 9, 18]. However, Generative Engine needs to generate an answer and provide attributions throughout the answer. Further, Generative Engine is not limited to a single text modality regarding both input and output. Additionally, the framework of Generative Engine is not limited to fetching relevant sources but instead comprises multiple tasks such as query reformulation, source selection, and making decisions on how and when to perform them.

Search Engine Optimization: In nearly the past 25 years, extensive research has optimized web content for search engines [2, 12, 22]. These methods fall into On-Page SEO, improving content and user experience, and Off-Page SEO, boosting website authority through link building. In contrast, GEO deals with a more complex environment involving multi-modality, conversational settings. Since GEO is optimized against a generative model not limited to simple keyword matching, traditional SEO strategies will not apply to Generative Engine settings, highlighting the need for GEO.

8 CONCLUSION

In this work, we formulate search engines augmented with generative models that we dub generative engines. We propose Generative Engine Optimization (GEO) to empower content creators to optimize their content under generative engines. We define impression metrics for generative engines and propose and release GEO-bench: a benchmark encompassing diverse user queries from multiple domains and settings, along with relevant sources needed to answer those queries. We propose several ways to optimize content for generative engines and demonstrate that these methods can boost source visibility by up to 40% in generative engine responses. Among other findings, we show that including citations, quotations from relevant sources, and statistics can significantly boost source visibility. Further, we discover a dependence of GEO methods' effectiveness on the query domain and the potential of combining multiple GEO strategies in conjunction. We show promising results on a commercially deployed generative engine with millions of active users, showcasing the real-world impact of our work. In summary, our work is the first to formalize the important and timely GEO paradigm, releasing algorithms and infrastructure (benchmarks, datasets, and metrics) to facilitate rapid progress in generative engines by the community. This serves as a first step towards understanding the impact of generative engines on the digital space and the role of GEO in this new paradigm of search engines.

9 LIMITATIONS

While we rigorously test our proposed methods on two generative engines, including a publicly available one, methods may need to adapt over time as GEs evolve, mirroring the evolution of SEO. Additionally, despite our efforts to ensure the queries in our GEO-bench closely resemble real-world queries, the nature of queries can change over time, necessitating continuous updates. Further, owing to the black-box nature of search engine algorithms, we didn't evaluate how GEO methods affect search rankings. However, we note that changes made by GEO methods are targeted changes in textual content, bearing some resemblance with SEO methods, while not affecting other metadata such as domain name, backlinks, etc, and thus, they are less likely to affect search engine rankings. Further, as larger context lengths in language models become economical, it is expected that future generative models will be able to ingest more sources, thus reducing the impact of search rankings. Lastly, while every query in our proposed GEO-bench is tagged and manually inspected, there may be discrepancies due to subjective interpretations or errors in labeling.

10 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2107048. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. ORCAS-I: Queries Annotated with Intent using Weak Supervision. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022). https://api.semanticscholar.org/CorpusID:248495926

[2] Prashant Ankalkoti. 2017. Survey on Search Engine Optimization Tools & Techniques. *Imperial journal of interdisciplinary research* 3 (2017). https://api.semanticscholar.org/CorpusID:116487363

[3] Akari Asai, Xinyan Velocity Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021. One Question Answering Model for Many Languages with Cross-lingual

Dense Passage Retrieval. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:236428949

[4] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Networks* 30 (1998), 107–117. https://api.semanticscholar.org/CorpusID:7587743

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, and Jimmy J. Lin. 2021. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://api.semanticscholar.org/CorpusID:234336491

[7] Brian Dean. 2023. We Analyzed 4 Million Google Search Results. Here's What We Learned About Organic Click Through Rate. https://backlinko.com/google-ctr-stats Accessed: 2024-06-08.

[8] Danny Goodwin. 2011. Top Google Result Gets 36.4% of Clicks [Study]. https://www.searchenginewatch.com/2011/04/21/top-google-result-gets-36-4-of-clicks-study/

[9] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *ArXiv* abs/2002.08909 (2020). https://api.semanticscholar.org/CorpusID:211204736

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[11] Aounon Kumar and Himabindu Lakkaraju. 2024. Manipulating Large Language Models to Increase Product Visibility. arXiv:2404.07981 [cs.IR]

[12] R.Anil Kumar, Zaiduddin Shaik, and Mohammed Furqan. 2019. A Survey on Search Engine Optimization Techniques. *International Journal of P2P Network Trends and Technology* (2019). https://doi.org/10.14445/22492615/IJPTT-V9I1P402

[13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466. https://api.semanticscholar.org/CorpusID:86611921

[14] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. *ArXiv* abs/2304.09848 (2023). https://api.semanticscholar.org/CorpusID:258212854

[15] Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *ArXiv* abs/2303.16634 (2023). https://api.semanticscholar.org/CorpusID:257804696

[16] G. D. Maayan. 2023. How Google SGE will impact your traffic – and 3 SGE recovery case studies. *Search Engine Land* (5 Sep 2023). https://searchengineland.com/how-google-sge-will-impact-your-traffic-and-3-sge-recovery-case-studies-431430

[17] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nathan McAleese. 2022. Teaching language models to support answers with verified quotes. *ArXiv* abs/2203.11147 (2022). https://api.semanticscholar.org/CorpusID:247594830

[18] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. *ArXiv* abs/2302.07842 (2023). https://api.semanticscholar.org/CorpusID:256868474

[19] Reiichiro Nakano, Jacob Hilton, S. Arun Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *ArXiv* abs/2112.09332 (2021). https://api.semanticscholar.org/CorpusID:245329531

[20] OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/

[21] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[22] A. Shahzad, Deden Witarsyah Jacob, Nazri M. Nawi, Hairulnizam Bin Mahdin, and Marheni Eka Saputri. 2020. The new trend for search engine optimization, tools and techniques. *Indonesian Journal of Electrical Engineering and Computer Science* 18 (2020), 1568. https://api.semanticscholar.org/CorpusID:213123106

[23] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv* abs/2208.03188 (2022). https://api.semanticscholar.org/CorpusID:251371589

[24] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 [cs.CL]

[25] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. *ArXiv* abs/2305.11206 (2023). https://api.semanticscholar.org/CorpusID:258822910

Listing 1: Prompt used for Generative Engine. The GE takes the query and 5 sources as input and outputs the response to query with response grounded in the sources.

```
Write an accurate and concise answer for the given user question,
      using _only_ the provided summarized web search results.
      The answer should be correct, high-quality, and written by
      an expert using an unbiased and journalistic tone. The user
      's language of choice such as English, Francais, Espamol,
      Deutsch, or should be used. The answer should be
      informative, interesting, and engaging. The answer's logic
      and reasoning should be rigorous and defensible. Every
      sentence in the answer should be _immediately followed_ by
      an in-line citation to the search result(s). The cited
      search result(s) should fully support _all_ the information
      in the sentence. Search results need to be cited using [
      index]. When citing several search results, use [1][2][3]
      format rather than [1, 2, 3]. You can use multiple search
      results to respond comprehensively while avoiding
      irrelevant search results.

Question: {query}

Search Results:
{source_text}
```

A CONVERSATIONAL GENERATIVE ENGINE

In Section 2.1, we discussed a single-turn Generative Enginethat outputs a single response given the user query. However, one of the strengths of upcoming Generative Engines will be their ability to engage in an active back-and-forth conversation with the user. The conversation allows users to provide clarifications to their queries or Generative Engine response and ask follow-ups. Specifically, in equation 1, instead of the input being a single query q_u, it is modeled as a conversation history $H = (q_u^t, r^t)$ pairs. The response r^{t+1} is then defined as:

$$GE := f_{LE}(H, P_U) \rightarrow r^{t+1} \tag{5}$$

where t is the turn number.

Further, to engage the user in a conversation, a separate LLM, L_{follow} or L_{resp}, may generate suggested follow-up queries based on H, P_U, and r^{t+1}. The suggested follow-up queries are typically designed to maximize the likelihood of user engagement. This not only benefits Generative Engine providers by increasing user interaction but also benefits website owners by enhancing their visibility. Furthermore, these follow-up queries can help users by getting more detailed information.

B EXPERIMENTAL SETUP

B.1 Evaluated Generative Engine

The exact prompt used is shown in Listing 1.

B.2 Benchmark

GEO-BENCH contains queries from nine datasets. Representative queries from each of the datasets are shown in Figure 2. Further, we tag each of the queries based on a pool of 7 different categories. For tagging, we use the GPT-4 model and manually confirm high recall and precision in tagging. However, owing to such an automated system, the tags can be noisy and should not be considered carefully. Details about each of these queries are presented here:

Listing 2: Representative Queries from each of the 9 datasets in GEO-BENCH

```
### ORCAS
- what does globalization mean
- wine pairing list

### AllSouls
- Are open-access journals the future of academic publishing?
- Should the study of non-Western philosophy be a requirement
      for a philosophy degree in the UK?

### Davinci-Debate
- Should all citizens receive a basic income?
- Should governments promote atheism?

### ELI5
- Why does my cat kick its toys when playing with them?
- what does caffeine actually do your muscles, especially
      regarding exercising?

### GPT-4
- What are the benefits of a keto diet?
- What are the most profound impacts of the Renaissance period
      on modern society?

### LIMA
- What are the primary factors that influence consumer behavior?
- What would be a great twist for a murder mystery? I'm looking
      for something creative, not to rehash old tropes.

### MS-Macro
- what does monogamous
- what is the normal fbs range for children

### Natural Questions
- where does the phrase bee line come from
- what is the prince of persia in the bible

### Perplexity.ai
- how to gain more followers on LinkedIn
- why is blood sugar higher after a meal
```

- **Difficulty Level:** The complexity of the query, ranging from simple to complex.
- **Nature of Query:** The type of information sought by the query, such as factual, opinion, or comparison.
- **Genre:** The category or domain of the query, such as arts and entertainment, finance, or science.
- **Specific Topics:** The specific subject matter of the query, such as physics, economics, or computer science.
- **Sensitivity:** Whether the query involves sensitive topics or not.
- **User Intent:** The purpose behind the user's query, such as research, purchase, or entertainment.
- **Answer Type:** The format of the answer that the query is seeking, such as fact, opinion, or list.

B.3 Evaluation Metrics

We use 7 different subjective impression metrics, whose prompts are presented in our our public repository: https://github.com/GEO-optim/GEO.

B.4 GEO Methods

We propose 9 different GENERATIVE ENGINE OPTIMIZATION methods to optimize website content for generative engines. We evaluate these methods on the complete GEO-BENCH test split. Further, to

Method	Position-Adjusted Word Count			Subjective Impression							
	Word	Position	Overall	Rel.	Infl.	Unique	Div.	FollowUp	Pos.	Count	Average
Performance without Generative Engine Optimization											
No Optimization	19.7(±0.7)	19.6(±0.5)	19.8(±0.6)	19.8(±0.9)	19.8(±1.6)	19.8(±0.6)	19.8(±1.1)	19.8(±1.0)	19.8(±1.0)	19.8(±0.9)	19.8(±0.9)
Non-Performing Generative Engine Optimization methods											
Keyword Stuffing	19.6(±0.5)	19.5(±0.6)	19.8(±0.5)	20.8(±0.8)	19.8(±1.0)	20.4(±0.5)	20.6(±0.9)	19.9(±0.9)	21.1(±1.0)	21.0(±0.9)	20.6(±0.7)
Unique Words	20.6(±0.6)	20.5(±0.7)	20.7(±0.5)	20.8(±0.7)	20.3(±1.3)	20.5(±0.3)	20.9(±0.3)	20.4(±0.7)	21.5(±0.6)	21.2(±0.4)	20.9(±0.4)
High-Performing Generative Engine Optimization methods											
Easy-to-Understand	21.5(±0.7)	22.0(±0.8)	21.5(±0.6)	21.0(±1.1)	21.1(±1.8)	21.2(±0.9)	20.9(±1.1)	20.6(±1.0)	21.9(±1.1)	21.4(±0.9)	21.3(±1.0)
Authoritative	21.3(±0.7)	21.2(±0.9)	21.1(±0.8)	22.3(±0.8)	22.9(±0.8)	22.1(±0.9)	23.2(±0.7)	21.9(±0.4)	23.9(±1.2)	23.0(±1.1)	23.1(±0.7)
Technical Terms	22.5(±0.6)	22.4(±0.6)	22.5(±0.6)	21.2(±0.7)	21.8(±0.8)	20.5(±0.5)	21.1(±0.6)	20.5(±0.6)	22.1(±0.6)	21.2(±0.2)	21.4(±0.4)
Fluency Optimization	24.4(±0.8)	24.4(±0.6)	24.4(±0.8)	21.3(±0.9)	23.2(±1.5)	21.2(±1.0)	21.4(±1.4)	20.8(±1.3)	23.2(±1.8)	21.5(±1.3)	22.1(±1.2)
Cite Sources	25.5(±0.7)	25.3(±0.6)	25.3(±0.6)	22.8(±0.9)	24.2(±0.7)	21.7(±0.3)	22.3(±0.8)	21.3(±0.9)	23.5(±0.4)	21.7(±0.6)	22.9(±0.5)
Quotation Addition	27.5(±0.8)	27.6(±0.8)	27.1(±0.6)	24.4(±1.0)	26.7(±1.1)	24.6(±0.7)	24.9(±0.9)	23.2(±0.9)	26.4(±1.0)	24.1(±1.2)	25.5(±0.9)
Statistics Addition	25.8(±1.2)	26.0(±0.8)	25.5(±1.2)	23.1(±1.4)	26.1(±0.9)	23.6(±0.9)	24.5(±1.2)	22.4(±1.2)	26.1(±1.2)	23.8(±1.2)	24.8(±1.1)

Table 6: Absolute impression metrics of GEO methods on GEO-bench. Compared to baselines, simple methods like Keyword Stuffing traditionally used in SEO don't perform well. However, our proposed methods such as Statistics Addition and Quotation Addition show strong performance improvements across all metrics. The best methods improve upon baseline by 41% and 28% on Position-Adjusted Word Count and Subjective Impression respectively.

Method	Position-Adjusted Word Count			Subjective Impression							
	Word	Position	**Overall**	Rel.	Infl.	Unique	Div.	FollowUp	Pos.	Count	**Average**
Performance without Generative Engine Optimization											
No Optimization	24.0	24.4	24.1	24.7	24.7	24.7	24.7	24.7	24.7	24.7	24.7
Non-Performing Generative Engine Optimization methods											
Keyword Stuffing	21.9	21.4	21.9	26.3	27.2	27.2	30.2	27.9	28.2	26.9	28.1
Unique Words	24.0	23.7	23.6	24.9	25.1	24.7	24.4	23.0	23.6	23.9	24.1
High-Performing Generative Engine Optimization methods											
Authoritative	25.6	25.7	25.9	28.9	30.9	31.2	31.7	31.5	26.9	29.5	30.6
Fluency Optimization	25.8	26.2	26.0	28.9	29.4	29.8	30.6	30.1	29.6	29.6	30.0
Cite Sources	26.6	26.9	26.8	19.8	20.7	19.5	18.9	20.0	18.5	18.9	19.0
Quotation Addition	28.8	28.7	**29.1**	31.4	31.9	31.9	32.3	31.4	31.7	30.9	32.1
Statistics Addition	25.8	26.6	26.2	31.6	33.4	34.0	33.7	34.0	33.3	33.1	**33.9**

Table 7: Performance improvement of GEO methods on GEO-bench with Perplexity.ai as generative engine. Compared to the baselines simple methods such as Keyword Stuffing traditionally used in SEO often perform worse. However, our proposed methods such as Statistics Addition and Quotation Addition show strong performance improvements across the board. The best performing methods improve upon baseline by 22% on Position-Adjusted Word Count and 37% on Subjective Impression.

reduce variance in results, we run our experiments on five different random seeds and report the average.

B.5 Prompts for GEO methods

We present all prompts in our our public repository: https://github.com/GEO-optim/GEO. GPT-3.5 turbo was used for all experiments.

C RESULTS

We perform experiments on 5 random seeds and present results with statistical deviations in Table 6

C.1 GEO in the Wild : Experiments with Deployed Generative Engine

We also evaluate our proposed Generative Engine Optimization methods on real-world deployed Generative Engine: Perplexity.ai. Since perplexity.ai does not allow the user to specify source URLs, we instead provide source text as file uploads to perplexity.ai while ensuring all answers are generated only using the file sources provided. We evaluate all our methods on a subset of 200 samples of our test set. Results using Perplexity.ai are shown in Table 7.