How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews

Riley Grossman
New Jersey Institute of Technology
Newark, NJ, USA
rag24@njit.edu

Songjiang Liu
New Jersey Institute of Technology
Newark, NJ, USA
sl947@njit.edu

Michael K. Chen
Nanyang Technological University
Singapore, Singapore
michaelchenkj@gmail.com

Mike Smith
Indiana University Bloomington
Bloomington, IN, USA
ms255@iu.edu

Cristian Borcea
New Jersey Institute of Technology
Newark, NJ, USA
borcea@njit.edu

Yi Chen
New Jersey Institute of Technology
Newark, NJ, USA
yi.chen@njit.edu

Abstract

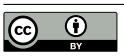
Generative AI is being increasingly integrated into web search for the convenience it provides users. In this work, we aim to understand how generative AI disrupts web search by retrieving and presenting the information and sources differently from traditional search engines. We introduce a public benchmark dataset of 11,500 user queries to support our study and future research of generative search. We compare the search results returned by Google's search engine, the accompanying AI Overview (AIO), and Gemini Flash 2.5 for each query. We have made several key findings. First, we find that for 51.5% of representative, real-user queries, AIOs are generated, and are displayed above the organic search results. Controversial questions frequently result in an AIO. Second, we show that the retrieved sources are substantially different for each search engine (<0.2 average Jaccard similarity). Traditional Google search is significantly more likely to retrieve information from popular or institutional websites in government or education, while generative search engines are significantly more likely to retrieve Google-owned content. Third, we observe that websites that block Google's AI crawler are significantly less likely to be retrieved by AIOs, despite having access to the content. Finally, AIOs are less consistent when processing two runs of the same query, and are less robust to minor query edits. Our findings have important implications for understanding how generative search impacts website visibility, the effectiveness of generative engine optimization techniques, and the information users receive. We call for revenue frameworks to foster a sustainable and mutually beneficial ecosystem for publishers and generative search providers.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**; **Web crawling**; **Page and site ranking**; **Language models**.

Keywords

Generative Search, Search Engine, Generative Engine Optimization



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809667

ACM Reference Format:
Riley Grossman, Songjiang Liu, Michael K. Chen, Mike Smith, Cristian Borcea, and Yi Chen. 2026. How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809667

1 Introduction

With the rapid advancement of large language models (LLMs), users have begun to use AI chatbots (e.g., ChatGPT) as a replacement for traditional search engines [6, 51, 57]. In response, traditional search engines have integrated LLM capabilities into their search functionalities, thereby affecting traditional search engine users in a way that AI chatbots do not. For example, while users must choose to use AI chatbots, Google places its AI Overview (AIO) above the traditional search engine results pages (SERP) by default (see Figure 1).[1] We use the term generative search engines to refer to any AI-powered systems that retrieve relevant sources in response to a user query and generate a summary of the content in those sources.

Generative search engines offer convenience to users [17, 34, 52] and are rapidly gaining popularity. However, it remains unclear how the sources retrieved by generative search engines compare with those returned by traditional search, what implications such differences have for users, and how these differences affect websites.

Users rely on the retrieved sources to obtain information, particularly those ranked highly by search engines [8, 49, 63]. Thus, it is critical that search engines consistently retrieve quality sources (e.g., for supporting an informed electorate in democratic society [12, 18, 54]). Furthermore, for usability, it is important that search engines retrieve consistent results in the presence of minor changes to the query syntax that preserve intent [5, 7, 28, 30].

From a website's perspective, there is growing concern that generative search reduces website traffic, as users increasingly obtain information directly from generated summaries rather than from sources listed in traditional SERPs [11]. Historically, websites have relied on Search Engine Optimization (SEO) services to improve their rankings in traditional search results in order to increase traffic, thereby increase advertising revenue or sales opportunities.

[1] Users can temporarily remove AIOs by adding "-AI" to a query or scrolling to the "Web" tab in the search results.

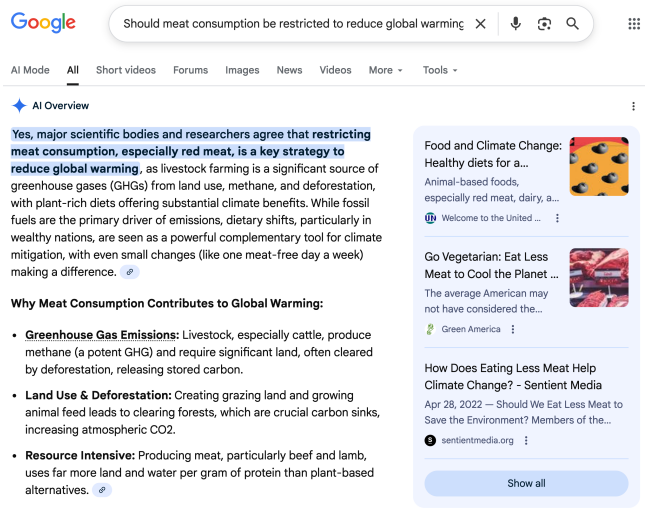


Figure 1: Google's AI Overview Interface

More recently, websites have begun exploring strategies to offset declining traffic by increasing their presence as cited sources in generative search results [29, 36]. Although some companies offer Generative Engine Optimization (GEO) services to increase website visibility in generative search results, with some evidence of its potential [1, 13, 42], the effectiveness of GEO is contested [45]. More fundamentally, there is a lack of understanding of how the characteristics of sources cited by generative search differ from those retrieved by traditional search engines.

In this paper, we conduct a large-scale comparative analysis of the sources retrieved by traditional and generative search engines. We construct a benchmark query set of 11,500 queries, along with two additional time-sensitive query sets, spanning diverse topics, user intents, and syntactic variations. Using SerpAPI[2] and the Gemini API, we collect ranked lists of retrieved sources from Google Search (i.e., traditional SERP), AIOs, and Gemini 2.5 Flash. We then use Jaccard similarity and rank-biased overlap (RBO) to quantify differences between the sources retrieved by each engine, and also to evaluate each system's internal consistency across repeated runs and in response to variations in query syntax, user location, and device. We also collect domain-level characteristics (e.g., popularity and content category) of the retrieved sources to assess the source preferences of each search engine.

We summarize our findings as follows:

- AIOs have a substantial presence (e.g., 51.5% of representative real-user queries), particularly in response to long, informational queries formatted as a question.
- Traditional SERP, Gemini, and AIO exhibit low average similarity in terms of the retrieved sources for each query (i.e., Jaccard similarities between 0.11 and 0.18).
- Generative search engines are significantly less likely to retrieve sources from popular websites, educational or governmental institutions, and websites that block Google's AI bot (even though AIOs can still access the content). Generative search engines retrieve significantly more content from Google-owned websites.
- Generative search is less consistent across runs of the same query, and less robust to changes in device type or query syntax, compared to traditional search.
- Although AIOs are rarely generated for trending queries (8.1%), they are prevalent for sensitive queries (e.g., 93.8% of political queries). Despite the importance of these queries to maintaining an informed electorate, generative search engines rely on less credible sources. They also often take a stance in the generated text summary (33.4% of AIO, and 5.6% of Gemini, responses, respectively).

Our findings have important societal implications. Users should exercise caution when using generative AI due to the risks of inaccuracies and hallucinations, and regulation may be necessary for controversial and high-stakes queries. Our results also inform publishers' use of GEO and decisions about blocking AI crawling. We highlight several recommendations for improvements in the generative search ecosystem. The current dilemma is that reputable publishers often restrict AI crawler access, leading to further traffic reductions, while degrading the quality of generative search results. We call for revenue frameworks that align incentives between publishers and generative search providers to foster a sustainable and mutually beneficial ecosystem.

To enable future research on the evolution and sustainability of generative search, processed datasets and code are available at: https://github.com/rag24/AIO.

2 Related Work

Generative search engines and AI-based information retrieval (IR) systems have become increasingly popular due to the convenience and user satisfaction they provide [17, 34, 52]. Prior studies have examined the quality of generated summaries and identified issues such as bias and hallucination [10, 18, 19, 52, 61]. User studies have shown that generative search engines may decrease the quality of information users obtain [52] and create echo-chambers [40, 50]. Studies have found that 25% of generative search citations do not support the corresponding sentence [35], and information is often cherry-picked or attributed to the wrong source [40]. Different from these works, which analyze the quality of the generated summary, our study compares traditional and generative search engines, specifically through analyzing the differences in each engine's retrieved sources.

There are several studies that evaluate offline AI-based IR systems, where the AI system can only access a limited number of documents for each query [16, 20, 21, 59]. Several studies show that AI-based IR systems favor information from questionable sources, such as AI generated content [20, 21, 59] and politically biased content [16]. We also study the quality of sources retrieved by AI systems, but we evaluate online generative search engines that access the entire Web in response to real user queries, rather than in an offline limited document setting.

[2]https://github.com/serpapi

As an important area that is gaining increasing attention, three relevant preprint studies have recently emerged. One study compares generative search engines' differing reliance on internal knowledge contained within the underlying LLM versus external knowledge from the Web [31]. The other two studies also evaluate differences between the sources cited by generative search engines and those retrieved by traditional search, but they exclude AIOs and are conducted on a very small set of specialized queries: one study focuses on product and service recommendation queries [13], and the other evaluates only 24 sociopolitical queries [39]. In contrast, our paper both analyzes the differences in Web sources that are cited by generative and traditional search engines, and further examines each engine's internal consistency and robustness to small variations in query syntax. Furthermore, our analysis is conducted on a larger scale (with 14,212 queries in total) than any existing work. The goal of our study is to provide insights for website publishers regarding potential responses to generative AI, to help users better understand the differences and trade-offs between traditional and generative search, and to call for collaboration toward building a sustainable online publishing and generative AI ecosystem.

3 Empirical Setup

Benchmark Query Dataset. Our benchmark contains 11,500 queries, which expands upon existing query datasets used to study generative search engines [1, 45] to specifically investigate the effects of query syntax (e.g., keywords vs. natural language question) and intent (e.g., asking for product information rather than where to buy). We further add coverage of localized queries (e.g., "near me" queries). As summarized in Table 1, we partition the benchmark into 9 query categories: **(1) ORCAS**: representative real-user queries [15] labeled with query intent (i.e., informational, navigational, or transactional) [4]; **(2) Amazon Retail**: real-user Amazon Retail keyword queries [45, 46]; **(3) Amazon Retail-Comp**: product-comparison queries based on Amazon Retail queries; **(4) Amazon Retail-Q**: product questions based on Amazon Retail queries; **(5) Debate**: debate-style queries [35]; **(6) ELI5**: complex informational queries randomly sampled from a Reddit-based corpus [23]; **(7) Localized**: randomly sampled from ORCAS "near me" queries [15] and instantiated with representative cities; **(8) NQ**: randomly sampled Natural Questions [1]; and **(9) NQ Keywords**: keyword-style reformulations of NQ queries. We used Gemini 2.5 Flash to generate queries for categories 3, 4 and 9.[3] Two authors manually viewed 50 of the generated queries to ensure that Gemini produced outputs that changed the query structure while preserving the original query's topic.

Collecting SERP and AIO Responses. To enable a controlled, large-scale comparison between traditional SERP and AIO, we follow prior work [44, 60, 62] and use a search retriever (i.e., SerpAPI) to collect representative results for real-users (see Section 5 for its robustness test). For each query in our benchmark query set, we issue a Google Search request via SerpAPI to collect the SERP and AIO (when available) results. The retriever is instructed to simulate a mobile device from Newark, NJ to ensure that SERP and AIO sources are collected under identical timing, device, and localization conditions. We confirm that our findings are robust to changes in device and location in Section 5. To reflect real world-importance, we only consider the SERP results from the first page because over 97% of all clicks are made on the first page of search outputs [56].

[3]Our repository contains all queries, as well as prompting templates and settings used to generate synthetic query variants.

Table 1: Composition of the Benchmark Query Set

Dataset	# Queries	Sample
ORCAS	5,000	*public aid office locations*
Amazon Retail	500	*star lamp projector galaxy*
Amazon Retail-Comp	500	*Compare star lamp projector galaxy vs aurora projector*
Amazon Retail-Q	500	*What is the best star lamp projector for a small room?*
Debate	1,000	*Should LGBT rights be protected by law?*
ELI5	1,000	*What determines if a phosphorylated protein is on or off?*
Localized	1,000	*free rabies shot near Alicante, Spain*
NQ	1,000	*when was the first cell phone call made*
NQ Keywords	1,000	*first cell phone call*
Total	**11,500**	

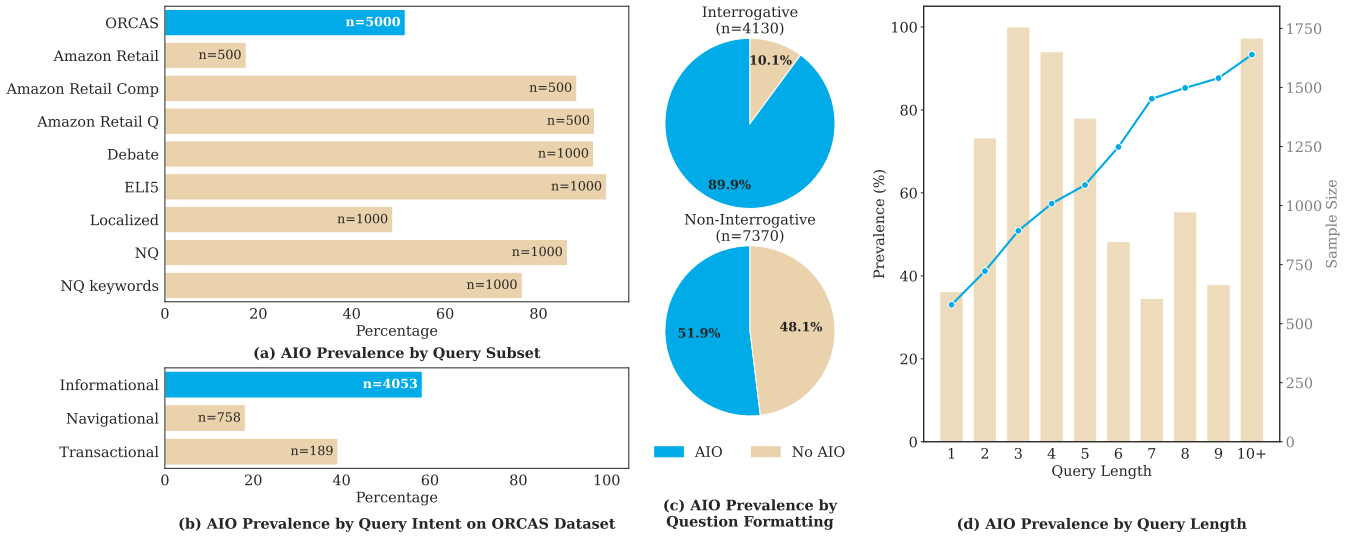
Collecting Gemini Responses. Using the Gemini API, we collect AI chatbot responses from the Gemini 2.5 Flash model. We selected this model because it provides a good trade-off between price and performance, and AIOs are built with a similar lightweight Gemini model. Grounding with Google Search is enabled so that the model will retrieve sources for each query. In alignment with AIO, the Thinking mode, typically used for complex reasoning tasks, is disabled. The input text consists of the given query from our dataset only, with no further prompting; no custom system instructions are added. Finally, we adhere to the default inference parameters. In the interest of comparability, all SERP, AIO, and Gemini responses to benchmark queries are collected on December 7th-8th 2025.

Evaluation Metrics. To quantify discrepancies in the retrieved sources between two search engines, we compute both set-level and rank-sensitive overlap. **Jaccard similarity** compares the set of unique sources, ranging from 0 (no shared sources) to 1 (identical sets), and captures whether the same sources appear regardless of rank. **Rank-biased overlap (RBO)** [58] compares ranked lists (also ranging from 0 to 1) while placing more weight on agreement near the top of the ranking, reflecting that higher-ranked sources tend to receive more user attention. We compute both metrics at the URL-level. Both Jaccard similarity and RBO were chosen in part because they are naturally capable of handling lists of differing length. Following the recommended implementation of RBO for public search engines that output around 10 results [58], we set the persistence parameter to $p = 0.9$.

4 Empirical Analysis

This paper aims for a better understanding of how generative AI is disrupting Web search and the resulting implications for websites and users. Toward this goal, our empirical analysis is guided by these research questions:

- RQ1: How frequently are AIOs generated, and for what types of queries?
- RQ2: How similar are the sources retrieved by AIO, SERP, and Gemini?
- RQ3: How do the characteristics of retrieved sources differ between generative and traditional search engines?



(a) AIO Prevalence by Query Subset

(b) AIO Prevalence by Query Intent on ORCAS Dataset

(c) AIO Prevalence by Question Formatting

(d) AIO Prevalence by Query Length

Figure 2: AIO Prevalence and Factors Associated with Higher Likelihood of an AIO Being Generated
Note: In panel (d), the blue line represents AIO prevalence and the bar chart shows the number of queries per length.

RQ4: How consistent and robust are generative search engines, relative to traditional search?
RQ5: How does generative search handle high-stakes queries?

4.1 RQ1: AIO Generation

We find AIOs to be generated on 65.6% of all benchmark queries. In Figure 2a, we show that the AIO generation rate is substantially different across the different query subsets in our benchmark dataset. On the extremes, 94.6% of ELI5 queries, compared to just 17.4% of Amazon Retail queries, result in an AIO. We consider the 51.5% AIO generation rate on the ORCAS dataset to be the most representative of how often real-user queries result in a generated AIO.

We also show that the AIO generation rate varies based on query characteristics such as intent, format, and length. In Figure 2b, we utilize the assigned query intent categories from the ORCAS dataset and find that informational queries are most likely to lead to an AIO. The AIO generation rates are significantly different across categories ($\chi^2(2, N = 5000) = 439.56, p < 0.001$), and all post-hoc pairwise comparisons with the Bonferroni correction were significant (p<0.001).

In Figure 2c, we separate out interrogative queries (i.e., those that end with a question mark or contain interrogative words such as who, what, or when) from non-interrogative queries. Interrogative queries are significantly more likely to result in an AIO being generated ($\chi^2(1, N = 11500) = 1685.7, p < 0.001$). This trend may also be found in the AIO generation rates of NQ (86.2%) and NQ keywords (76.5%) queries, as the only difference is NQ queries are formatted as a question, and NQ Keywords queries are formatted as a list of the keywords derived from the question.

Figure 2d shows that across the full benchmark dataset, the AIO prevalence increases with query length. We confirm that this trend also exists within the ORCAS queries to ensure it is not simply a result of differing average lengths for each query subset.

Taken together, these findings show that AIOs are more likely to be generated when users seek information, and particularly for queries that specify exactly what the user is looking for (e.g., "How do I clean a pizza stone?") instead of broad informational queries (e.g., "pizza stone").

For the rest of our analyses, we focus on the 7,439 queries where Google Search, AIO, and Gemini returned sources. We excluded the 56 and 121 queries where the AIO or Gemini response was generated without sources, respectively. This included a mix of queries where internal knowledge was relied on (e.g., "show me all the letters in the English alphabet) and queries without meaningful responses (e.g., Gemini cannot answer the query "lunch spots near me" because it cannot access user location).

4.2 RQ2: Similarity Between Traditional and Generative Search Engine Sources

Table 2 presents the average similarity between the list of sources returned by the AIO, Gemini, and traditional SERP for each query in the benchmark dataset. The main takeaway is that *regardless of query subset and which pair of search engines is compared, the retrieved lists are dissimilar, despite all three being developed by Google.* Besides using RBO as a metric for comparing the similarities of ranked lists, we also present Jaccard similarity because of its interpretability. For example, our findings show that on average, only 18% of the sources returned by either the AIO or traditional SERP will be retrieved by both search engines. Similar trends are observed when source similarity is measured by RBO. Surprisingly, despite AIO being built with a lightweight Gemini model, the source lists retrieved by AIO and Gemini are the least similar.

We also evaluate the retrieval similarity for different query subsets. Although the similarity metric values vary by query subset, the amount of similarity remains low. For example, no search engine pairing has an average RBO greater than 0.27 (i.e., AIO and

Query Subset	Jaccard			RBO		
	AIO SERP	AIO GEM	GEM SERP	AIO SERP	AIO GEM	GEM SERP
ORCAS	0.17	0.12	0.20	0.24	0.17	0.26
Amazon Retail	0.08	0.06	0.08	0.12	0.10	0.10
Retail-Comp	0.11	0.08	0.08	0.15	0.10	0.10
Retail-Q	0.12	0.10	0.10	0.16	0.14	0.13
Debate	0.24	0.10	0.11	0.27	0.12	0.14
ELI5	0.21	0.11	0.14	0.25	0.15	0.18
Localized	0.14	0.07	0.16	0.17	0.10	0.21
NQ	0.19	0.13	0.19	0.24	0.17	0.26
NQ Keywords	0.19	0.13	0.20	0.24	0.16	0.26
Total	**0.18**	**0.11**	**0.16**	**0.23**	**0.15**	**0.21**

Table 2: Average Similarity in Returned Sources for Google AI Overviews (AIO), Traditional Google Search Results (SERP), and Gemini 2.5 Flash (GEM)

SERP for Debate queries). In particular, the product and service queries (i.e., Amazon Retail and Localized query subsets) result in the lowest similarity for each pairing of search engines.

To contextualize the dissimilarity, we report the average number of sources returned by each search engine: 9.68 for Gemini, 9.24 for AIO, and 8.75 for traditional SERP. As these numbers are similar, we attribute the low similarity between source lists to different methodologies for each search engine rather than a different number of retrieved sources. Furthermore, the inconsistency between search engines greatly exceeds the inconsistency between two runs of the same search engine (see Section 4.4), so we cannot attribute this dissimilarity to built-in randomness.

4.3 RQ3: Characteristics of Generative and Traditional Search Sources

After showing the large discrepancy in the sources retrieved by AIOs, Gemini and traditional SERP in Section 4.2, we further study which websites, or categories of websites, benefit from this disruption. In particular, we determine how the popularity, type of content, and website's decision on blocking Google's AI bot are associated with generative search engine sources in comparison to traditional search sources.

4.3.1 Most Affected Domains. In Figure 3, we present the domains with the largest absolute changes in the percentage of benchmark queries for which they are retrieved by generative search engines compared with traditional search engines. The subfigures on the right side of Figure 3 present the changes when considering only the top three ranked sources, rather than all sources, retrieved by each search engine. We consider the top *three* specifically because Google AIO shows three sources automatically. Additional sources beyond the top three can only be accessible when users click the "Show all" button or specific link icons (See Figure 1). Furthermore, ranking has been shown to significantly affect the click-through-rate of traditional search engine results [8, 49, 63], and prior research and SEO professionals suggest this applies to generative search as well [1, 14, 47].

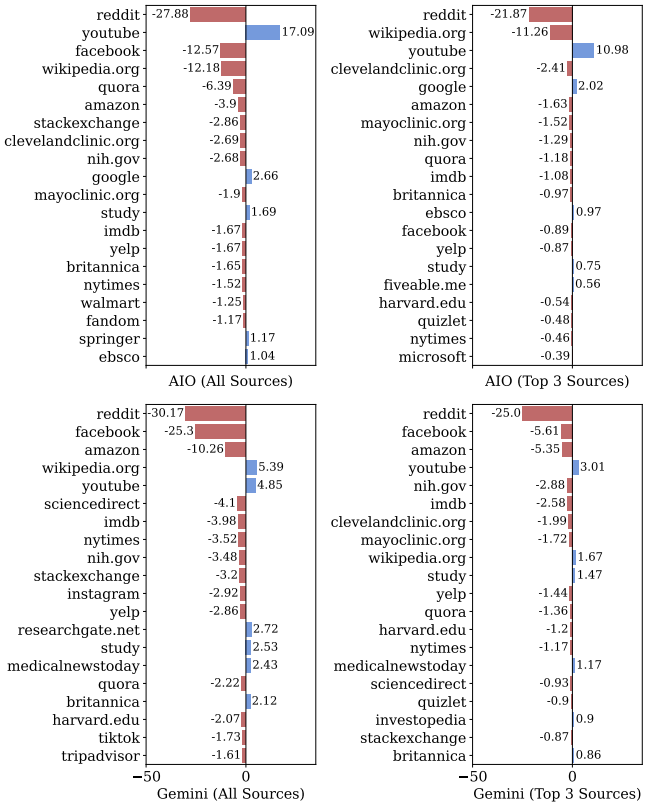


Figure 3: Domains with Biggest Change in Prevalence (% of Queries Retrieved) Relative to Traditional Search

We have three main takeaways from Figure 3. First, large and well-known websites are the most affected (both positively and negatively). This is intuitive as large websites have the reputation and diversity in content to be relevant to many different queries. Second, the overwhelming majority of these websites receive fewer overall, and fewer top three, citations with generative search engines (indicated by red bars and negative numbers in Figure 3). This suggests that generative search tends to source information from more niche sources than traditional search engines. Third, Google's AIOs favor Google websites (i.e., google.com and youtube.com domains). Gemini also favors YouTube in comparison to traditional Google Search, but the absolute difference is smaller. Lastly, we note that McNemar's tests [37] confirm that all differences in Figure 3 are statistically significant.

4.3.2 Blocking the Google-Extended Bot. In our analyses of the most affected domains, we found that 21 popular publishers listed in Table 3 (which are retrieved for at least 20 unique queries by both Google Search and AIOs) were never cited by Gemini. Several popular social media (Facebook, Instagram, Tiktok) and review websites (IMDb, Yelp, Tripadvisor) also received zero citations from Gemini. Upon further investigation, we found that all of these websites block the Google-Extended bot in their robots.txt files.

Google states that websites can block this bot if they do not want their content to be used for training future Gemini models, nor for

NYTimes	ESPN	Genius
CNN	Business Insider	National Geographic
BBC	CNBC	The Conversation
ScienceDirect	NPR	U.S. News & World Report
Reuters	WIRED	Scientific American
Wiley	USA Today	Consumer Reports
Nature	NBC News	STAT

Table 3: Popular Publishers Not Used by Gemini (Ranked by Tranco Popularity)

grounding responses with sources (i.e., citations). Google further clarifies that this will not affect a websites' ranking in Google Search results, nor does it impact the AIO's access to this content [25]. Thus, our results indicate that Google does respect websites' wishes, and this means that the decline in visibility in Gemini search results is self-inflicted. However, a surprising finding is that many of these websites also have fewer citations in AIOs (e.g., NYTimes, Yelp, IMDb, and Facebook).

4.3.3 Does Generative Search Favor Niche Websites? From Figure 3 we see that AIOs and Gemini tend to rely less on many of the popular domains that are retrieved by traditional Google Search for a wide range of queries. We are interested in understanding whether AIO and Gemini reference more niche websites or just a different set of popular websites.

To analyze this, we use the Tranco rankings [33], which are designed to be an objective ranking of website popularity for academic research[4]. In Figure 4, we plot each search engines' cumulative percentage of retrieved sources across all benchmark queries as a function of the domains' Tranco rank.

Figure 4 (left) focuses on the top 1,000 domains in terms of popularity. Google Search retrieves 37.8% of all sources from the top 1,000, which is 1.3 and 9.9 percentage points higher than AIO and Gemini, respectively. This gap substantially increases when considering just the top one or top three retrieved sources from each search engine. Considering only the top ranked source for each query, websites ranked in the top 1,000 are cited in response to 52.7% of queries for traditional SERP, and just 40.0% and 32.6% for AIO and Gemini, respectively. This suggests that traditional search places higher importance on source reputability and popularity.

Figure 4 (right) zooms out to consider the top 1 million domains in terms of popularity. The overall trend of traditional search retrieving a higher percentage of popular domains is consistent. However, the gaps have narrowed, particularly when looking at only the top one or top three retrieved sources. Figure 5 plots the number of unique domains referenced by each search engine binned by Tranco rank. Despite having fewer overall unique sources, we see that traditional Google Search has the highest number of unique sources ranked in the top 100, top 1K, and top 10K of Tranco. Gemini has the most unique sources for domains ranked outside the top 10K. The AIO has more unique sources than traditional Google Search for domains ranked outside the top 100K.

4.3.4 Statistical Significance of Differences. We test for statistically significant differences in the website characteristics of sources

[4]We use the list from December 7th, 2025: https://tranco-list.eu/list/6GWVX/1000000

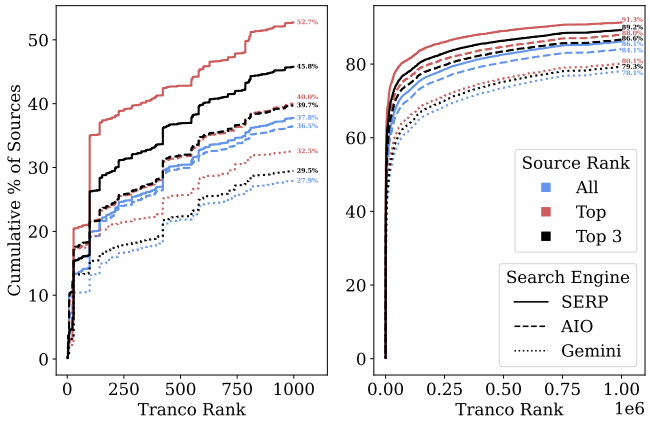


Figure 4: Percentage of Retrieved Sources From the Top 1K (left) and 1M (right) Popular Domains

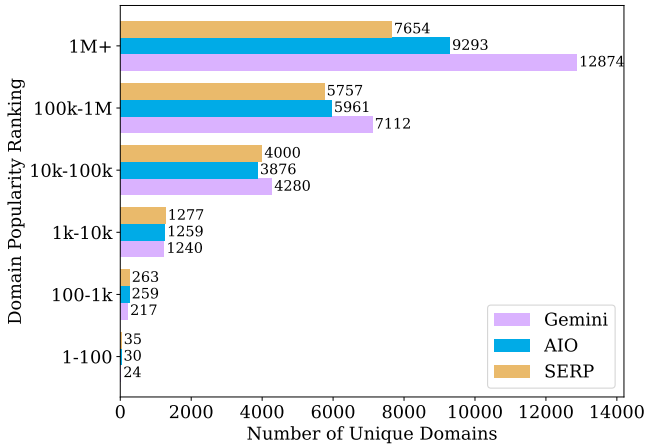


Figure 5: Number of Unique Sources by Tranco Rank

retrieved by generative versus traditional search engines. We construct a dataset with every unique pair of source and query, and mark whether each of the three search engines cited the source for the particular query. Importantly, because all relevant sources had to be retrieved by at least one search engine to be included in the dataset, the results can only be interpreted in comparison to the other search engines. We compare each generative search engine to the traditional search engine using a linear probability model with query fixed effects and clustered standard errors. We regress whether a source is cited for a query onto the interaction variables between search engine type and website characteristics. In particular, we include the binned tranco rank, whether the website blocks the Google-Extended bot, and 17 Cloudflare domain categories (see Figure 6) that are included in the list of retrieved sources for at least 5% of all queries. We also add whether the domain ends in .edu because the "Education" category in Cloudflare includes encyclopedia and Q&A websites, which may be seen as less reputable than content produced by educational institutions.

The 95% confidence intervals for all website characteristics interacted with the two generative search engines are displayed in

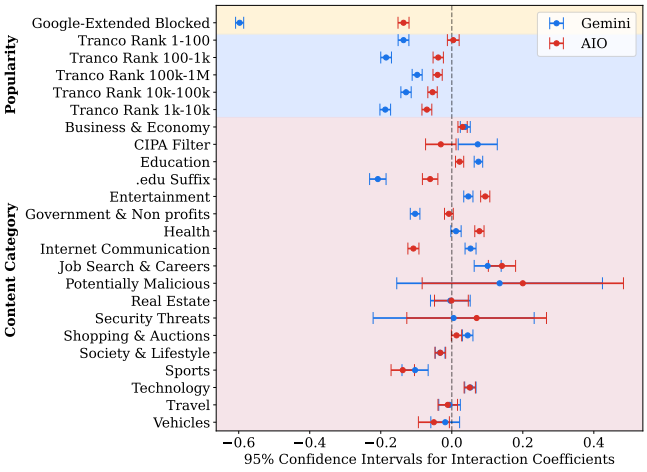


Figure 6: Characteristics of Generative Search Sources Relative to Traditional Search Sources

Figure 6. Coefficients are statistically significant if the corresponding bar does not cross zero. To summarize, the first takeaway is that both Gemini and AIOs are significantly less likely to retrieve content from websites blocking the Google-Extended crawler, despite AIOs technically having access to this content. Second, we find that Gemini and AIO are both significantly less likely to retrieve content from popular websites. Interestingly, this difference is most substantial among domains with a Tranco ranking from 1k-10k, which corresponds to websites that receive substantial traffic, but are not necessarily well-known. Third, although generative search is significantly less likely to cite reputable institutions in government or education, it is not significantly more likely to source content from inherently bad websites (e.g., those with security threats). However, Gemini is significantly more likely to cite content from websites that are inaccesible to children in publicly funded facilities via the Children's Internet Protection Act (CIPA).

4.4 RQ4: Search Engine Inconsistency

Besides comparing the sources retrieved by generative and traditional search engines, we also evaluate the consistency of the sources retrieved between two runs, particularly when queries are run from a different type of device or after minor query edits. High inconsistency between outputs in response to queries that have the same intent is problematic at a societal-level as different users receive different information to the same questions. It can also be an undesirable behavior at an individual-level as it can make it difficult to find previously retrieved information again.

Consistency Across Runs, Devices, and Locations. We first selected a stratified random sample of 100 queries from our benchmark dataset. We then simultaneously collected AIO and traditional search responses two times for each configuration of device (mobile and desktop) and location (Austin, TX and Newark, NJ). Since Gemini does not consider device or location, we just run the same query twice. We display the average similarity between the sources retrieved by each search engine across runs, devices, and locations, in terms of Jaccard similarity and RBO, in Table 4. In all possible

Device/ Location	Jaccard			RBO		
	AIO	GEM	SERP	AIO	GEM	SERP
Both Same	0.66	0.46	0.78	0.69	0.52	0.86
Diff Location	0.64	-	0.76	0.68	-	0.84
Diff Device	0.55	-	0.69	0.53	-	0.79
Both Diff	0.55	-	0.67	0.53	-	0.77

Table 4: Average Similarity in Returned Sources Between Runs and with Different Device or Location

comparisons from Table 4, the generative search engines show lower consistency than traditional Google Search.

Table 4 also shows that for traditional SERP and AIOs, the retrieved sources are more inconsistent between different devices than between different locations (which only decreases similarity metrics by 0.01-0.02). The large effect of device type may be due to Google's expectation that mobile users seek different information from desktop users. We further find that changing device results in more inconsistency for AIOs than traditional SERP (i.e., relative to RBO for the same device and location, RBO decreases by 0.16 and 0.07 for AIO and SERP, respectively, when changing device). This is largely driven by the finding that AIOs retrieve more sources on average for mobile searches (10.38) than desktop searches (9.42), while traditional Google Search retrieves similar amounts for both (8.79 vs. 8.65). This is counterintuitive, as we would expect fewer AIO sources on mobile due to the smaller screen size.

Consistency in the Presence of Cosmetic Query Edits. We first make minor edits to 200 queries: 100 queries where two words are (un)contracted (e.g., "what is" vs "what's"); 50 queries where a word is (un)abbreviated (e.g., "United States" vs "U.S."); and 50 queries where a question mark is added (removed) to queries that are clearly questions regardless of punctuation. Our findings reveal that AIOs are less robust to these changes. On average, the retrieved sources from the original and modified query have an RBO of 0.49 for AIOs (a 28.99% decline in comparison to the RBO resulting from two runs of the same query). In contrast, traditional SERP and Gemini have average RBOs of 0.74 and 0.5, respectively (13.95% and 3.85% declines, respectively). One potential explanation for why this issue plagues AIOs, and not Gemini, is that AIOs utilize a lightweight Gemini model with lower reasoning capabilities, which may result in more reliance on query keywords than intent.

We further analyze the impact of inconsistencies in retrieved sources to the text generated by generative search engines. We quantify this relationship using word-level Jaccard similarity.[5] We find a strong and statistically significant positive Pearson correlation between source similarity and generated summary similarity. For AIOs, higher RBO in the retrieved sources between original and edited queries corresponds to higher text similarity (i.e., $r = 0.62$, $p < 0.001$). Gemini exhibits the same pattern (i.e., $r = 0.55$, $p < 0.001$). In summary, even when the query's underlying intent is unchanged, cosmetic query differences result in different sources retrieved, and subsequently different generated text.

[5]Each text is lowercased, tokenized with NLTK's `wordpunct_tokenize`, and Porter-stemmed (`PorterStemmer`).

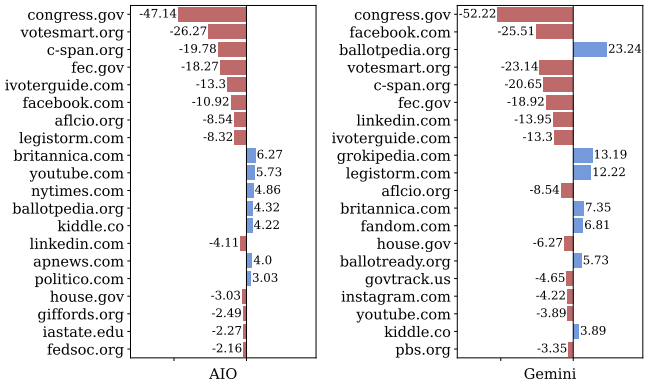


Figure 7: Domains with Biggest Change in Prevalence (% of Queries Retrieved) for Political Queries

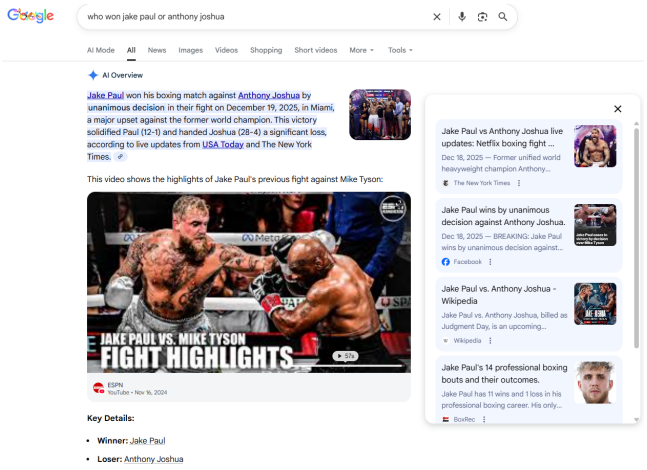


Figure 8: Misinformed AI Overview Response

4.5 RQ5: High-Stakes Queries

To further understand how generative search affects users, we evaluate how generative search responds to high-stakes queries where errors could have societal impact. Specifically, we consider queries related to debate topics and political figures due to their sensitive nature and potential for different viewpoints. We also add trending queries, which are particularly susceptible to misinformation due to the lack of consensus as events unfold in real time.

While the debate queries are part of our benchmark dataset, political and trending queries were not included in the benchmark dataset due to their time-sensitive nature. For political queries, we generate queries of the form "What is the politician [NAME] known for?" for all current members of the U.S. Congress, as well as those who finished as runner-up in the last election. We curated this new list of queries to differentiate it from the debate query set, and other similar political queries in prior research [16, 31, 54], which focus on political ideas or events, rather than politicians. For trending queries, we utilized 863 trending topics from Google Trends[6] with at least 1,000 searches and 2+ subqueries (collected December 19th, 2025). For each trending topic, we collected data for the main query, and one randomly selected subquery.

Debate Queries. As shown in Figure 2a, AIOs are frequently generated for these controversial topics. Surprisingly, a substantial percentage (33.4%) of the AIO summaries started with affirmative or negative responses. This was less common, but not absent, in Gemini responses (5.6%). Even if both sides of the debate were presented further down in the summary, it is surprising to see Gemini or AIO taking a stance in response to controversial questions such as (i) whether robots should be included in the military, (ii) whether AI should be used for medical diagnoses, (iii) whether meat consumption should be restricted to reduce global warming, and (iv) whether immigration laws should be reformed.

Political Queries. AIOs are generated frequently for politicians (93.8%). Concerningly, generative search engines may retrieve less credible sources for this crucial information. Figure 7 presents the domains exhibiting the largest deviations from those retrieved by traditional search for political queries. As shown in Figure 7, the generative search engines cite government resources (e.g., congress.gov) less frequently. The AIO retrieves more content from popular news domains like NYTimes, Politico, and AP News, while Gemini has larger reliance on Fandom and Grokipedia.

We further attempt to quantify this issue using News source credibility ratings from Media Bias/Fact Check (MBFC)[7]. Although they only provide ratings for about half of all sources retrieved for political queries, we still find that more sources retrieved by the generative search engines have questionable (i.e., medium or low) ratings. Specifically, 11.4% (0.5% low credibility) of sources retrieved by AIO and 15.0% (0.6% low) of sources retrieved by Gemini come from websites with questionable credibility ratings, in comparison to 10.6% (0.2% low credibility) of sources retrieved by traditional search.

Trending Queries. We find that AIOs are rare for trending queries (8.1%). The likelihood of an AIO being generated does increase as time passes: AIOs are generated for 12.7% of queries that began trending more than five days ago, compared to only 4.8% of queries that started trending within the past five days. It is possible that the lack of AIOs represents a guardrail set in place because AIOs are susceptible to sourcing from misinformation when there is a lack of consensus across many sources. As a case study, we present a misinformed AIO response in Figure 8. In anticipation of a boxing match later that night, the trending query "who won jake paul or anthony joshua" was incorrectly answered by the AIO, which declared Jake Paul as the winner. Interestingly, AIO cited several sources, despite only one of them (a Facebook post from a satirical sports account) claiming Jake Paul had won.

5 Robustness Tests

To confirm that our findings are indicative of a meaningful disruption to the search engine industry, we perform a number of robustness tests.

SerpAPI is representative of real user experiences. The first test is to confirm whether our automated data collection methodology produces data that is representative of a real-user setting. Due

[6]https://trends.google.com/trending

[7]https://mediabiasfactcheck.com/methodology/

to randomness in individual search engine results (see Section 4.4), we cannot simply confirm that the exact same results are shown for queries submitted through SerpAPI and a real user's browser. Instead, we aim to show that the similarity between the API and manually retrieved sources is equivalent to the similarity between two API collections. The manual data collection occurs on one of the author's laptops from a signed-in Chrome profile, and the automated data is collected through SerpAPI on a desktop device from two cities. Data is simultaneously collected for the 100 queries in Section 4.4. For the two API collections, the RBOs between two runs of AIO or SERP results are 0.67 and 0.81, respectively. The average RBO between manual and API collections is 0.67 and 0.79 for AIO and traditional search, respectively. These results indicate that SerpAPI yields source lists that are as close to a signed-in, manual browsing setup as repeated API collections are to each other, supporting SerpAPI as a representative proxy for real-user retrieval in our experiments.

One potential limitation to our approach is that real users that are signed into a Google profile may have higher consistency between two searches of the same query. For our two manual collections, we found average RBOs of 0.73 and 0.89 for AIOs and traditional search, respectively. These values are slightly higher than the similarities (0.69 and 0.86) computed in Table 4 for the same device and location. Regardless, these results still show that AIOs are less consistent than traditional search, and show that our main result around internal search engine consistency holds for real users.

Robustness to Device and Location. Given that device type and location influence the sources retrieved by SERP and AIO (see Table 4), we want to confirm that our findings from Sections 4.1, 4.2, and 4.3 hold under different settings. We collect two stratified random samples that are each 10% of the full benchmark size. We collect data for these samples with the same method as our initial collection, except we change the SerpAPI device parameter to desktop and use different locations for each sample. Our major findings hold for desktop devices: AIOs appear on 67.17% of all queries (52.2% of ORCAS queries); AIOs are more frequent on longer, informational queries formatted as a question; and traditional search retrieves more sources from popular domains in comparison to AIOs and Gemini. The average RBO between AIO and SERP sources is higher on desktop than mobile (0.31 versus 0.23), but the main finding that source similarity is low between search engines holds true.

Comparison to Bing. To contextualize the dissimilarity between Google's traditional and generative search engine results, we compare traditional Google Search with Bing, the second most popular search engine [53]. We use SerpAPI to collect organic search results from Bing and compare them to Google's organic search results for the 100 queries in Section 4.4. We calculate an average RBO of 0.14 for the two search engines' retrieved sources, which is lower than the average RBO when comparing Google Search with either generative search engine. Although the retrieved results from Bing and Google Search are highly dissimilar, Bing's small market share has made this dissimilarity less important historically. For example, industry reports claim that the most popular Generative Search tool, ChatGPT, receives more daily search queries than Bing [6, 32, 51, 57], and even assuming that AIOs are generated on 50% of users' Google searches would result in over 11 times the number of daily queries received by Bing. Thus, in comparison to Bing, publishers likely care more about their visibility in generative search results.

6 Discussion: Societal Impact and Limitations

Considerations for Users. From a user perspective, generative search engines offer convenience by providing integrated summaries alongside links for further reading, when needed. However, a major concern is that AI-generated information may be inaccurate. We find a clear example of this issue in our data (Figure 8), and a recent news article from The Guardian highlights dangerously inaccurate AIO responses to health-related queries [26]. Due to the inherent challenges to deliver perfect solutions, users should use caution when leveraging generative search.

Implication for Publishers. Our results show that generative search will benefit niche content providers at the expense of more popular, established ones. Many large publishers have already taken actions against generative AI companies by filing copyright infringement lawsuits [38, 55], and claim that generative search had led to declining traffic to their websites [9, 43].

If generative search is here to stay, what strategies can publishers implement to adapt? Our results challenge the effectiveness of GEO techniques [1, 41, 42]. First, we find that the source lists produced by Gemini and the AIO are the least similar, compared with the pairs of source lists produced by traditional earch and either generative search engine (Table 2). This indicates the challenges of performing well across multiple generative search engines. Second, we find that in comparison to traditional search, AIOs are less consistent between multiple runs of the same query (Table 4). As randomness has a large effect on whether a source appears or is highly ranked in any given run of a query, optimization for high rankings in generative search may be unreliable.

Another decision that publishers need to make is whether to allow AI companies access to their content. Our findings in Figure 6 show that websites blocking the Google AI crawler are significantly less likely to be cited by generative search engines in comparison to traditional SERP. While this is expected for Gemini, it is surprising that it also affects AIOs, given that Google only allows website to remove their content from AIOs by removing their content entirely from Google Search results [2, 3, 24, 48]. Given the high presence of AIOs and the prominent position above the traditional search results, publishers may need to rethink their decision to block Google from accessing their content for AI training and grounding.

Implications for Retail and Service Websites. We find that AIOs are generated infrequently for Amazon Retail queries (17.4%) but frequently for Amazon Retail comparison and question queries (88.2% and 92%, respectively). This suggests that generative search plays a larger role during the consideration or research stages than the final purchase stage. AIOs are also moderately prevalent among the Localized queries subset (48.8%), which is largely searches for service providers (e.g. "cheapest restaurants near me").

Given that retrieved search results differ substantially between traditional and generative search engines for all aforementioned query sets, there may be an opportunity for lesser known providers of goods or services (i.e., those not ranking well in traditional search results) to optimize their content for visibility in generative search results. Although our results generally question the effectiveness

of GEO techniques due to the dissimilarity in generative search engines' retrieved sources, prior research focusing on product and service queries has found that third-party reviews are heavily relied on by many generative search engines [13]. Thus, it may be beneficial for niche providers to invest in positive third-party reviews from sources frequently cited by generative search engines.

Implications to Generative Search. We expect generative search engines to continue growing in popularity. However, important improvements are needed. First, inconsistency in responses from generative search engines for queries with the same intent undermines their usability and trustworthiness. Such inconsistency potentially pose serious risks for democracy by presenting asymmetric information to different users [56]. At a minimum, generative search engines should increase consistency across repeated runs of the same query and improve robustness to minor query variations, so that users seeking the same information are presented with similar content.

Second, future research should continue to address reliability issues in generative search, such as hallucinations and the spread of misinformation. Inaccuracies may stem from the reliance on less popular websites. Although there is certainly an argument for using information from diverse sources, one concern is that the lesser known websites are less likely to be vetted for credibility and biases by third parties (e.g., MBFC). Thus, to improve quality, generative search engine companies should consider utilizing more information from popular and reputable sources, as traditional search engines do. Yandex appears to already implement this approach as their AI summaries are generated based on the top five search results retrieved by its traditional search engine. In addition, generative search companies could consider funding independent evaluations of bias and credibility for niche websites, to promote source diversity without sacrificing reliability.

Implications to the Ecosystem. Given the inherent imperfections of AI-based solutions, it is important to establish policies [22] that regulate the use of generative search engines in contexts involving controversial issues (e.g., elected officials) or high stakes topics (e.g., identifying symptoms of a serious medical event).

Furthermore, we urge for the development of deals between digital publishers and AI companies [27] for a healthy online publishing and search ecosystem. Reduced traffic and the resulting decline in advertising revenues threatens publishers' viability. In turn, a loss of high-quality content would ultimately undermine generative search engines themselves who may be "starved" of useful content for training and grounding. In addition to singular licensing deals, revenue frameworks could be established, such as negotiated licensing, pay-per-crawl, or other revenue-sharing arrangements, which compensate publishers for their content and enable generative search engines to utilize information from the most popular and reputable websites. Such a framework would better align the incentives of publishers and generative search providers, supporting a sustainable ecosystem where both sides benefit.

Limitations. Our experimental design and collected dataset lead to several limitations. First, this study solely focuses on Google, given its leading position in the search industry. To understand the generalizability of some results (e.g., generative search engine inconsistencies or preferences for niche content), it would be desirable to include other leading AI chatbots and generative search tools (e.g., ChatGPT, Perplexity AI, and Bing Copilot). Inclusion of open-source generative search tools would allow for better testing of the underlying mechanisms that contribute to low internal consistency and preference for niche sources. Second, although it was necessary to collect a large volume of data, reliance on APIs for data collection does limit our results in that we cannot study how generative search outputs vary by user characteristics (beyond device and city). It would be interesting for future user studies to explore how the users' characteristics (browsing as a signed-in user) impact the prevalence of AIOs and the generative search engines' retrieved sources and text summaries.

Finally, our analyses are both descriptive in nature and limited to a single point in time. Given these limitations, this study should motivate future research that considers the broader ecosystem implications, beyond just the user impact, of generative search systems. Generative search is changing quickly, and it is our hope that providing a public benchmark dataset of queries and returned URLs will enable future research to document changes in generative search engine behavior.

7 Conclusions

We conducted a large-scale comparison of the results generated by traditional Google Search (SERP), Google AI Overviews (AIO), and Gemini for 14,212 total queries (including 11,500 time-invariant queries released in our benchmark dataset). There are several key findings. AIOs are generated for a substantial share of real-user searches and appear above organic results. Across search engines, the cited sources differ substantially, leading to meaningfully different source exposure for users and websites. Generative search shifts visibility away from popular and institutional domains and toward Google-owned properties. Surprisingly, AIOs retrieve fewer sources from websites blocking the Google-Extended bot, despite having access, which may lead some publishers to rethink their blocking strategies. Finally, we find that generative search engines are less stable than SERP across runs and more sensitive to changes in device type and minor query edits. Generative search's impact on users is most concerning for high-stakes queries, where AIOs frequently appear and may rely on low-credibility sources or adopt a stance. We release our benchmark query set, processed data, and code to support replication and longitudinal monitoring of generative search behavior. Future work may consider how personalization plays a role in the results retrieved by generative search systems or conduct user studies to better understand how generative search citations impact publishers' traffic and advertising revenues.

Acknowledgments

Research reported in this publication was supported in part by the NSF under Grant No. CNS2237328, the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UM1TR004789, as well as by the Martin Tuchman'62 Chair Endowment and the Leir Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

[1] 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* (New York, NY, USA, 2024-08-24) *(KDD '24)*. Association for Computing Machinery, 5–16. doi:10.1145/3637528.3671900

[2] Davey Alba. 2025. Google can train search AI with web content after AI opt-out. The Edge Singapore. https://sg.news.yahoo.com/google-train-search-ai-content-230000810.html?guccounter=1

[3] Davey Alba and Julia Love. 2025. Google Decided Against Offering Publishers Options in AI Search. Bloomberg. https://www.bloomberg.com/news/articles/2025-05-19/google-gave-sites-little-choice-in-using-data-for-ai-search

[4] Daria Alexander and Arjen P. de Vries. 2025. In a Few Words: Comparing Weak Supervision and LLMs for Short Query Intent Classification. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2025-07-13) *(SIGIR '25)*. Association for Computing Machinery, 2977–2981. doi:10.1145/3726302.3730213

[5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval consistency in the presence of query variations. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 395–404.

[6] Evan Bailyn. 2025. Google vs ChatGPT Market Share: 2026 Report. FirstPageSage. https://firstpagesage.com/seo-blog/google-vs-chatgpt-market-share-report

[7] Krishnan Batri, Rajermani Thinakaran, S. Lakshmi, R. Sowrirajan, and Sivaram Murugan. 2025. Beyond Precision and Recall: Measuring Search Engine Consistency Using Rank Stability. *IEEE Access* 13 (2025), 92242–92259. doi:10.1109/ACCESS.2025.3571184

[8] Michael R Baye, Babur De los Santos, and Matthijs R Wildenbeest. 2016. Search engine optimization: what drives organic traffic to retail sites? *Journal of Economics & Management Strategy* 25, 1 (2016), 6–31.

[9] Rebecca Bellan. 2025. Google's AI search features are killing traffic to publishers. TechCrunch. https://techcrunch.com/2025/06/10/googles-ai-overviews-are-killing-traffic-for-publishers/

[10] Matan Ben-Tov and Mahmood Sharif. 2025. GASLITEing the Retrieval: Exploring Vulnerabilities in Dense Embedding-based Search. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security* (Taipei Taiwan, 2025-11-19). ACM, 4364–4378. doi:10.1145/3719027.3765095

[11] Joydeep Bhattacharya. 2025. Generative Engine Optimization (GEO) Statistics: New Data for 2025. SEO Sandwitch. https://seosandwitch.com/generative-engine-optimization-stats/

[12] Engin Bozdag and Jeroen Van Den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and information technology* 17, 4 (2015), 249–265.

[13] Mahe Chen, Xiaoxuan Wang, Kaiwen Chen, and Nick Koudas. 2025. Generative Engine Optimization: How to Dominate AI Search. arXiv:2509.08919 [cs.IR] https://arxiv.org/abs/2509.08919

[14] Marketing Couch. 2025. How AI Overviews Are Reshaping Google Click-Through Rates in 2025 (And What Businesses Must Do Now). https://marketing-couch.com/how-ai-overviews-are-reshaping-google-click-through-rates-in-2025-and-what-businesses-must-do-now/

[15] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. *arXiv preprint arXiv:2006.05324* (2020).

[16] Sunhao Dai, Zhanshuo Cao, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, and Tat-Seng Chua. 2025. Media Source Matters More Than Content: Unveiling Political Bias in LLM-Generated Citations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 17256–17276. doi:10.18653/v1/2025.emnlp-main.872

[17] Sunhao Dai, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, Ji-Rong Wen, and Tat-Seng Chua. 2025. NExT-Search: Rebuilding User Feedback Ecosystem for Generative AI Search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2025-07-13) *(SIGIR '25)*. Association for Computing Machinery, 3922–3931. doi:10.1145/3726302.3730353

[18] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona Spain, 2024-08-25). ACM, 6437–6447. doi:10.1145/3637528.3671458

[19] Sunhao Dai, Chen Xu, Shicheng Xu, Zhongxiang Sun, Liang Pang, Zhenhua Dong, and Jun Xu. 2025. Trustworthy Information Retrieval in the LLM Era: Bias, Unfairness, and Hallucination. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Xi'an China, 2025-12-07). ACM, 442–446. doi:10.1145/3767695.3769670

[20] Sunhao Dai, Yuqi Zhou, Liang Pang, Zhuoyang Li, Zhaocheng Du, Gang Wang, and Jun Xu. 2025. Mitigating Source Bias with LLM Alignment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2025-07-13) *(SIGIR '25)*. Association for Computing Machinery, 370–380. doi:10.1145/3726302.3730038

[21] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural Retrievers are Biased Towards LLM-Generated Content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2024-08-25). ACM, 526–537. doi:10.1145/3637528.3671882

[22] European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Official Journal of the European Union.

[23] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. arXiv:1907.09190 [cs.CL] https://arxiv.org/abs/1907.09190

[24] Google. 2025. AI features and your website. https://developers.google.com/search/docs/appearance/ai-features

[25] Google. 2025. List of Google's common crawlers. https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers

[26] Andrew Gregory. 2026. 'Dangerous and alarming': Google removes some of its AI summaries after users' health put at risk. The Guardian. https://www.theguardian.com/technology/2026/jan/11/google-ai-overviews-health-guardian-investigation

[27] Sara Guaglione. 2026. A timeline of the major deals between publishers and AI tech companies in 2025. Digiday. https://digiday.com/media/a-timeline-of-the-major-deals-between-publishers-and-ai-tech-companies-in-2025/

[28] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*. 527–538.

[29] Megri Hohli. 2025. The Great CTR Crash: Why Google's AI Overviews Demand a New E-E-A-T-Focused SEO Strategy. Submitshop. https://www.submitshop.com/the-great-ctr-crash-why-googles-ai-overviews-demand-a-new-e-e-a-t-focused-seo-strategy

[30] Jyun-Yu Jiang, Jing Liu, Chin-Yew Lin, and Pu-Jen Cheng. 2015. Improving ranking consistency for web search by leveraging a knowledge base and search logs. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1441–1450.

[31] Elisabeth Kirsten, Jost Grosse Perdekamp, Mihir Upadhyay, Krishna P. Gummadi, and Muhammad Bilal Zafar. 2025. Characterizing Web Search in The Age of Generative AI. arXiv:2510.11560 [cs.IR] https://arxiv.org/abs/2510.11560

[32] Naveen Kumar. 2025. 31 Bing Statistics 2026 [Facts, Usage Revenue]. DemandSage. https://www.demandsage.com/bing-statistics/

[33] Victor Le Pochat, Tom Van Goethem, Maciej Tajalizadehkhoob, Samaneh Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. doi:10.14722/ndss.2019.23386

[34] Yidong Liang, Zhijing Wu, Fan Zhang, Dandan Song, and Heyan Huang. 2025. How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua Italy, 2025-07-13). ACM, 634–644. doi:10.1145/3726302.3729998

[35] Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 7001–7025. doi:10.18653/v1/2023.findings-emnlp.467

[36] Tracy McDonald. 2025. AIO Impact on Google CTR: September 2025 Update. Seer Interactive. https://www.seerinteractive.com/insights/aio-impact-on-google-ctr-september-2025-update

[37] Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* 12, 2 (1947), 153–157. doi:10.1007/BF02295996

[38] Cade Metz and Michael M. Grynbaum. 2025. New York Times Sues A.I. Start-Up Perplexity Over Use of Copyrighted Work. The New York Times. https://www.nytimes.com/2025/12/05/technology/new-york-times-perplexity-ai-lawsuit.html

[39] Marco Minici, Cristian Consonni, Federico Cinus, and Giuseppe Manco. 2025. Auditing LLM Editorial Bias in News Media Exposure. arXiv:2510.27489 [cs.CY] https://arxiv.org/abs/2510.27489

[40] Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-based Search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Athens Greece, 2025-06-23). ACM, 1325–1340. doi:10.1145/3715275.3732089

[41] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. 2025. Adversarial Search Engine Optimization for Large Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=hkdqxN3c7t

[42] Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. 2024. Ranking Manipulation for Conversational Search Engines. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 9523–9552. doi:10.18653/v1/2024.emnlp-main.534

[43] Kaustubh Phatak. 2025. The AI Content Crisis: A Publisher's Guide To Survival And Success In 2025. Forbes. https://www.forbes.com/councils/forbestechcouncil/2025/11/17/the-ai-content-crisis-a-publishers-guide-to-survival-and-success-in-2025/

[44] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5687–5711. doi:10.18653/v1/2023.findings-emnlp.378

[45] Haritz Puerto, Martin Gubri, Tommaso Green, Seong Joon Oh, and Sangdoo Yun. 2025. C-SEO Bench: Does Conversational SEO Work?. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=oTeixD3oZO

[46] Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588* (2022).

[47] Larisa Rosu. 2026. AIO Citation Rank: See Who Gets the Front Row in AI Overviews. Advanced Web Ranking. https://www.advancedwebranking.com/help/aio-citation-rank-see-who-gets-the-spotlight-in-ai-overviews

[48] Alex Seifert. 2025. Google Forcing Websites to Allow Its AI to Train on Their Content. Medium. https://medium.com/@alexseifert/google-forcing-websites-to-allow-its-ai-to-train-on-their-content-0faa70b0b63b

[49] Dushyant Sharma, Rishabh Shukla, Anil Kumar Giri, and Sumit Kumar. 2019. A Brief Review on Search Engine Optimization. In *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. 687–692. doi:10.1109/CONFLUENCE.2019.8776976

[50] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033. doi:10.1145/3613904.3642459

[51] Shubham Singh. 2025. ChatGPT Users Statistics (January 2026) – Growth Usage Data. DemandSage. https://www.demandsage.com/chatgpt-statistics/

[52] Matthias Stadler, Maria Bannert, and Michael Sailer. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior* 160 (2024), 108386. doi:10.1016/j.chb.2024.108386

[53] Statista. 2026. Market share of leading search engines worldwide from January 2015 to December 2025. https://www.statista.com/statistics/1381664/worldwide-all-devices-market-share-of-search-engines/

[54] Miriam Steiner, Melanie Magin, Birgit Stark, and Stefan Geiß. 2022. Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society* 25, 2 (2022), 217–241.

[55] Chat GPT Is Eating the World. 2025. Updated U.S. Map of Copyright Suits v. AI (Dec. 5, 2025) = 65 suits. https://chatgptiseatingtheworld.com/2025/12/05/updated-u-s-map-of-copyright-suits-v-ai-dec-5-2025-64-suits/

[56] Aleksandra Urman and Mykola Makhortykh. 2021. You Are How (and Where) You Search? Comparative Analysis of Web Search Behaviour Using Web Tracking Data. arXiv:2105.04961 [cs.HC] https://arxiv.org/abs/2105.04961

[57] Bruno Venditti. 2025. ChatGPT Lags Far Behind Google in Daily Search Volume. Visual Capitalist. https://www.visualcapitalist.com/chatgpt-lags-far-behind-google-in-daily-search-volume/

[58] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28 (2010), 20:1–20:38. https://api.semanticscholar.org/CorpusID:16050561

[59] Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 208–217. doi:10.1145/3626772.3657750

[60] Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering Questions by Meta-Reasoning over Multiple Chains of Thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5942–5966. doi:10.18653/v1/2023.emnlp-main.364

[61] An Zhang, Yang Deng, Yankai Lin, Xu Chen, Ji-Rong Wen, and Tat-Seng Chua. 2024. Large Language Model Powered Agents for Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC USA, 2024-07-10). ACM, 2989–2992. doi:10.1145/3626772.3661375

[62] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. 2025. Uncertainty Propagation on LLM Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 6064–6073. doi:10.18653/v1/2025.acl-long.302

[63] Jakub Zilincan. 2015. Search engine optimization. In *CBU International Conference Proceedings*, Vol. 3. 506–510.