CHARACTERIZING WEB SEARCH IN THE AGE OF GENERATIVE AI

Elisabeth Kirsten[1,3], **Jost Grosse Perdekamp**[1,3], **Qinyuan Wu**[2], **Mihir Upadhyay**[1]
Krishna P. Gummadi[2], **Muhammad Bilal Zafar**[1,3]
[1]UA Ruhr Research Center for Trustworthy Data Science and Security,
[2]Max Planck Institute for Software Systems,
[3]Ruhr University Bochum
Correspondence: elisabeth.kirsten@rub.de
github.com/aisoc-lab/generative-search-eval

ABSTRACT

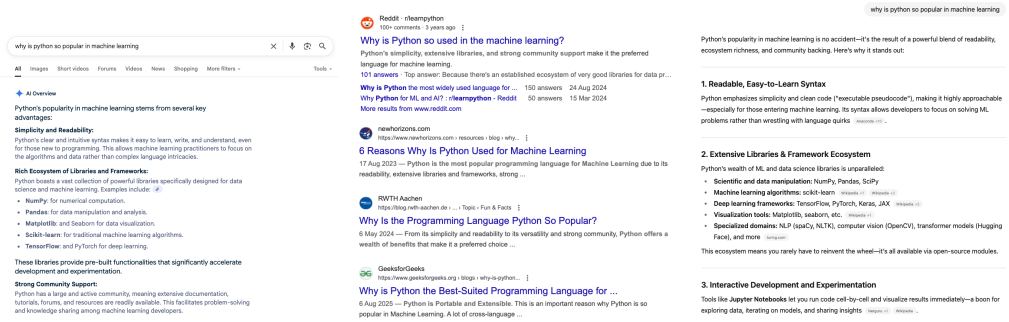
The advent of LLMs has given rise to *generative search*, a new search paradigm in which LLMs retrieve information from the web related to a query and synthesize it into a single, coherent response. This paradigm differs fundamentally from traditional web search, where results are returned as a ranked list of independent web pages. In this paper, we ask: Along what dimensions does generative search differ from traditional search? We conduct a systematic comparison between Google organic search and five generative search systems from three providers: Google, OpenAI, and Perplexity. Our analysis reveals substantial variation among engines in their reliance on internal *v.s.* external knowledge, source diversity, and stability. While generative systems often achieve topical coverage comparable to traditional search, they do so using markedly different retrieval footprints and synthesis strategies. We further show that the outputs of generative search can vary across time and executions, raising new challenges for robustness. Our findings demonstrate that generative search introduces new dimensions that are not captured by existing evaluation paradigms, motivating the development of evaluations that explicitly account for retrieval behavior, synthesis, and stability in generative search systems.

1 INTRODUCTION

Search has been a mainstay of online information retrieval for three decades. In response to a user query, traditional search engines return a ranked list of roughly 10 web pages, ordered primarily by relevance and source authority (Page et al., 1999), but also influenced by factors like diversity, recency, and personalization (Qin et al., 2012).

The advent of LLMs has given rise to a new type of web search, generative AI-based search (Liu et al., 2024; Nakano et al., 2021). Under this new search paradigm, users receive answers in *natural language rather than a ranked list of results*. In fact, traditional search engines, including Google, now integrate generative search results in their outputs. These systems typically operate by performing a web search, retrieving relevant pages, and producing a coherent, self-contained response. See Figure 1 for an example.

While generative and traditional web search differ in many aspects, three fundamental dimensions stand out: **(i) Reliance on external *v.s.* internal knowledge:** Traditional web search operates by ranking external documents, *i.e.,* web pages. In contrast, generative search may rely solely on the underlying LLM's *internal knowledge*, on *external sources*, or on a combination of both (McMahon & Kleinman, 2025; Nakano et al., 2021). **(ii) Much wider coverage:** Traditional search engines typically present links and snippets from the top-10 web pages. Users, limited by their information processing capacity, would need to manually navigate in order to view the lower-ranked results. Users rarely inspect results beyond the top-10, and often not beyond the top-3 (Bar-Ilan et al., 2009; Nowicki, 2003; Urman & Makhortykh, 2023). Generative search, by contrast, can potentially aggregate information from tens of sources into a single response. **(iii) Synthesis of information using**



(a) AI Overview (b) Organic search (c) GPT-4o with search tool

Figure 1: Outputs of different search engines when querying for "why is python so popular in machine learning". Figures 1a and 1b show the Google search output. At the top of the results page is a so-called "AI Overview". Below the AI Overview are the traditional web search results consisting of (often) top-10 web pages. The response of GPT-4o with web search used as a tool is also accompanied by the supporting links (Figure 1c).

LLMs: Traditional search presents retrieved content verbatim as independent snippets (Figure 1b). In contrast, generative search synthesizes retrieved content into novel text. Because the underlying LLMs performing this synthesis are stochastic, issuing the same query multiple times can yield different outputs.

The impact of these fundamental differences on search outputs remains largely unexplored. Most existing work on evaluating generative search focuses on properties such as factual accuracy, hallucinations, and bias (§2), but does not explore, for instance, if generative search actually leverages its ability to go beyond top-10 web pages or if the stochasticity of the underlying LLM leads to materially different outputs across multiple runs. To address this gap, we construct an evaluation framework that operationalizes these three key dimensions and uses them to systematically characterize the behavior of generative search systems.

We use our framework to compare a traditional search engine, Google's traditional search (`Organic`), with five generative search engines: Google AI Overview (`AIO`), Gemini (`Gemini`), GPT-4o-Search (`GPT-Search`), GPT-4o with search as a tool (`GPT-Tool`), and Perplexity Sonar (`Sonar`) across a wide range of datasets covering domains like politics, shopping, and science. Our evaluation reveals several intriguing patterns.

Generative engines differ widely in how much they rely on internal *v.s.* external knowledge. For the same queries, `GPT-Tool` consults fewer than a single web page on average, while `Sonar`, `AIO`, `Gemini`, and `GPT-Search` retrieve 14, 9, 9, and 4 on average, respectively. Despite these differences in retrieval footprint, generative engines often achieve **similar levels of topical coverage**. We also find that many generative engines perform web searches for retrieving simple, static facts that could be answered from internal knowledge alone (*e.g., What is the capital of France?*), lowering efficiency. However, engines that perform fewer searches sometimes produce inaccurate information on, dynamic, time-sensitive queries (*e.g., seahawks vs steelers*). This dichotomy reveals inherent **tradeoffs between efficiency and accuracy**.

Generative engines also differ in the breadth of sources they surface. They frequently **cite sources far beyond the top-10 or even top-100 organic search results.** For instance, on average 53% (27%) of domains that `AIO` consults are not contained in top-10 (top-100) `Organic` search results. Similarly, while 38% (89%) of the `Organic` result domains are contained within top-1K (top-1M) most visited websites, the same numbers are 34% (85%) for `AIO` and 35% (81%) for `GPT-Tool`.

Because generative engines rely on stochastic LLMs and complex retrieval pipelines, their **outputs are unstable across executions.** For instance, when performing the same queries two months apart, only 18% of web pages were common between the two runs of `AIO`, whereas the overlap was 45% for `Organic` search. Even over relatively short intervals of 5 minutes (24 hours), responses to questions requiring ternary answers (yes, no, neither) changed in up to 27% (28%) of cases.

Such instability raises concerns about reproducibility, trust, and user expectations, particularly when answers are presented as authoritative summaries.

Overall, our framework and findings show that generative search introduces new trade-offs in sourcing behavior, efficiency, and stability that are not captured by existing search evaluation paradigms. These results motivate the development of tailor-made evaluation methods that explicitly account for fundamental differences between the mechanics of traditional and generative search.

2 Related Work

Evaluation of traditional web search has long focused on issues like relevance, diversity, freshness, and coverage. Relevance metrics such as Precision, Recall, and nDCG measure how well returned documents satisfy the user's information need and reward systems that rank relevant results highly. Evaluation frameworks also measure *diversity* to ensure coverage of multiple intents, subtopics, or viewpoints for ambiguous queries (*e.g.,* subtopic recall, α-nDCG). *Freshness* captures the timeliness of results, particularly for event-driven queries, and *coverage* measures the breadth of retrieved content (Lewandowski, 2012). These criteria are effective for ranked lists of documents, but do not directly apply to the synthesized, single-response outputs of generative engines.

Large Language Models (LLMs) are commonly assessed on question answering, summarization, factual grounding, and tool use (Liang et al., 2023; Huang et al., 2024b;a). Summarization metrics capture coverage, coherence, and factual accuracy using measures such as ROUGE and BERTScore. In information-seeking contexts, retrieval-augmented generation (RAG) enhances LLMs with external knowledge to improve factuality by incorporating external knowledge (He et al., 2024; Shi et al., 2025; Jo et al., 2025). Recent work has also applied LLMs to query understanding, ranking (Sun et al., 2024), and query refinement (Siro et al., 2024; Bacciu et al., 2024). In this work, we focus on a specific use case of LLMs as information sources in deployed, user-facing web search systems.

Prior work on diversity in search has examined multiple dimensions of bias, including political, geographical, and commercial bias (Jiang, 2014). Evaluations typically focus on the diversity of cited sources in the result set (Kingrani et al., 2015; Jiang, 2014; Urman et al., 2021), as well as ideological skew (Lin et al., 2023), and commercial bias (Jiang, 2014). Recent work has applied ranking fairness metrics to quantify viewpoint diversity in search results (Draws et al., 2021) and measures coverage over predefined sets of perspectives (Chen & Choi, 2025; Skoutas et al., 2010; Draws et al., 2021). In this work, we measure diversity at both the source and content levels, and examine how retrieval and synthesis jointly shape the information presented to users.

Generative search retrieves and synthesizes information into new text rather than returning a ranked list (Shi et al., 2025; Nakano et al., 2021). Recent work evaluates generative search along dimensions such as verifiability, credibility, accuracy, and bias (Liu et al., 2023; Hu et al., 2024; Dai et al., 2024; Li & Sinnamon, 2024). Other studies examine user interaction, trust, and feedback mechanisms in generative search interfaces (Aliannejadi et al., 2024; Dai et al., 2025; Sharma et al., 2024; Mayerhofer et al., 2025). Recent work proposes new benchmarks, user models, and evaluation principles tailored to generative search (Narayanan Venkit et al., 2025; Gienapp et al., 2024; Ai et al., 2023; Alaofi et al., 2025; Miroyan et al., 2025) and analyzes when a search should be invoked (Schick et al., 2023; Li et al., 2025; Sha et al., 2025). To the best of our knowledge, our work is the first to systematically characterize the sources that generative search engines retrieve and to study how differences in retrieval behavior affect the knowledge breadth, efficiency, and stability of generated content.

3 Experimental Setup

This section outlines the datasets and engines used.

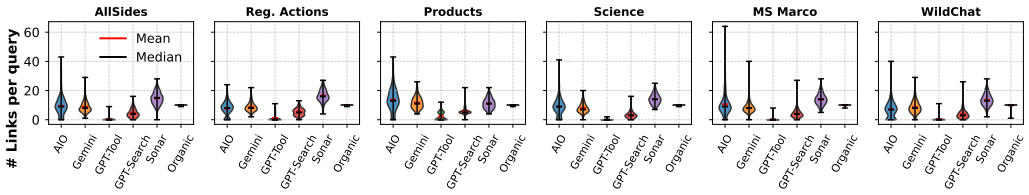


Figure 2: Different search engines rely on external knowledge to varying degrees. `GPT-Tool` cites much fewer web pages than other engines, followed by `GPT-Search`. `AIO` tends to cite the most web pages.

3.1 Datasets

Our datasets are designed to (i) reflect real user queries, (ii) cover both everyday and domain-specific workloads (*e.g.,* politics, science, products), and (iii) include both persistent and time-sensitive topics.

Existing datasets. We use two existing datasets. *(i) MS Marco:* 1,000 real-world Bing queries for open-domain retrieval and QA. *(ii) WildChat:* 1,740 information-seeking queries subsampled from user interactions with ChatGPT.

Newly curated datasets. To ensure a diversity of workloads and probe specific performance axes like newly-created knowledge, we curate the following new datasets. *(i) AllSides:* 332 politically focused queries derived from socio-political topics. *(ii) Regulatory Actions:* 649 queries about recent executive orders in the US. *(iii) Science:* 453 queries on AI-related scientific topics derived from the ACM CCS taxonomy. *(iv) Products:* 422 product-related queries based on popular Amazon searches. *(v) Trends:* 100 trending queries collected from Google Trends on 2025-09-15.

All the datasets combined consist of 4,706 queries. Table 4 shows examples of queries from each dataset. Further details on dataset construction are provided in Appendix A.

3.2 Search Engines

We compare one traditional and five generative search systems, selected to cover both dedicated search engines and general-purpose LLMs with search capabilities. *(i) Organic Google Search* (`Organic`) retrieves ranked lists of web pages. By default, the first page shows top-10 search results. In order to compare the overlap between generative search and Google search at various ranks, we retrieved top-100 search results. *(ii) Google AI Overviews* (`AIO`) generate synthesized answers together with cited sources. *(iii) Gemini-2.5-Flash with Google Search* (`Gemini`) generates responses with an optional web search. *(iv) Perplexity Sonar* (`Sonar`) always performs a web search before returning an answer. *(v) GPT-4o Search* (`GPT-Search`) also always performs a web search. *(vi) GPT-4o with Search Tool* (`GPT-Tool`) decides per query whether to retrieve external information.

Experimental Details. All queries were issued in English in September 2025. For generative engines, we set temperature to 0 and maximum new tokens to $1,000$. Additional implementation details are in Appendix B. To study the effect of location, we performed queries from two locations: United States (**US**) and Germany (**DE**). We observe only minor differences. Unless mentioned otherwise, the main paper reports results for US. We discuss the results for the DE location in Appendix F.

4 Internal *v.s.* External Knowledge

We now characterize how different search engines balance internal knowledge and external information from the web, and how this balance affects the quality of the outputs.

4.1 Differences in Number of Sources

Figure 2 shows the number of links per query for different datasets and search engines. The figure shows **vast differences in reliance on external knowledge.** For instance, the median number of links averaged over all datasets for `GPT-Tool` is 0 as compared to 9 for `AIO`. `GPT-Search` retrieves the second-lowest number of links. The median number of links averaged over all the datasets is 4.

The reliance on external links is query-specific. While the median number of links for `AIO` is similar to the median number of links for `Organic` (9 *v.s.* 10), they differ sharply at the extremes. At the 10^{th} percentile, `AIO` retrieves fewer links (0.6 *v.s.* 10), while at the 90^{th} percentile, it retrieves far more (17 *v.s.* 10). In other words, `Organic` search is effectively fixed at 10 results, whereas `AIO` adaptively decides how many web pages to retrieve.

We analyze queries for which the `AIO` retrieved more than 30 web pages (about 2% of all queries). These queries tend to be open-ended in nature, *e.g.,* "What jobs will AI agents replace?" and "How can I calm down myself". On the other hand, queries where `AIO` retrieved very few, *e.g.,* 2 or fewer web pages (15% of AIOs) tend to be short and fact-seeking, *e.g.,* "what is alabama state university motto", "Where is the capital of South Korea?"'). The results show that *for static, well-accepted facts, generative search can potentially retrieve the information from the model's internal knowledge, saving the effort* spent in retrieving 10 `Organic` search results, and the effort on the users' part to process them. We now explore this potential in more detail.

4.2 Retrieval Efficiency on Static Facts

Retrieval is not free—it incurs additional cost and latency. In this section, we compare the efficiency of various generative engines by studying their retrieval patterns on well-known facts, *i.e.,* facts that frontier models are expected to recall from their internal memory alone (Liang et al., 2023).

We use a special-purpose dataset of simple factual queries with static answers. The dataset consists of 249 questions, each asking for the capital of a country or territory. Queries follows the template "What is the capital of XX?". Full dataset construction details and prompt templates are provided in Appendix C.

As shown in Table 1, all engines except `AIO` achieve 100% accuracy. `AIO` produces a single error, mistaking the capital of the country Georgia for Atlanta, the capital of the US state of the same name. `GPT-Tool` does not perform a single search yet answers all queries correctly, hinting at the potential for enhancing efficiency for simple queries related to static facts. The results show that **generative search engines frequently perform web searches even for queries answerable from internal knowledge alone.**

Engine	% search	mean #URLs	Accuracy
`GPT-Tool`	0 %	0	100%
`GPT-Search`	98%	5.2	100%
`Gemini`	100%	4.47	100%
`AIO`	–	5.83	99.5%

Table 1: Retrieval behavior on simple factual queries.

4.3 Efficiency on Time-Sensitive Queries

While some queries can be answered from internal, static knowledge, others depend critically on recent information. We therefore turn to time-sensitive queries, *i.e.,* the **Trends dataset** (§3.1), to examine how generative search engines behave when queries require recent, up-to-date information.[1]

We find large structural differences w.r.t. to static queries (§4.2). Figure 3, top panel, shows the number of retrieved links per query. The median number of retrieved links is 5 or more for all engines. `GPT-Tool` still retrieves the least number of links.

We next measure the number of distinct concepts covered by each generative engine. Methodological details of the concept detection procedure are provided in §5.3. `GPT-Search` achieves the highest average coverage (72%), followed by `Organic` (67%) and `Gemini` (66%), while `GPT-Tool` lags behind at 51%. Manual inspection of queries where `GPT-Tool` exhibits low coverage

[1] We did not run `Sonar` for the Trends analysis.

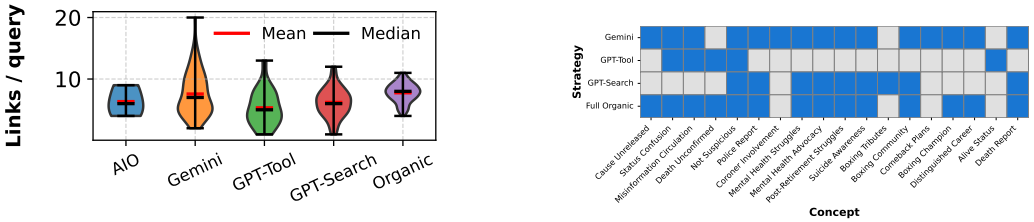


Figure 3: *Left:* Number of retrieved links on the trending queries dataset. *Right:* Concept coverage of different search engines on the query "ricky hatton cause of death". Low concept coverage shows that `GPT-Tool` does not have all the information available.

shows that the model often fails on fast-changing or event-driven topics that require external retrieval. For example, for the query "ricky hatton cause of death", `GPT-Tool` lacks access to recent information, becomes uncertain, and incorrectly reports Ricky Hatton as alive (see Figure 3, bottom panel). To assess robustness, we collected additional trend snapshots on two other days. The relative patterns remain consistent across snapshots.

Findings of §4.2 and §4.3 highlight a **tension between efficiency and accuracy** in generative search. For simple, static factual queries, the optimal response may simply be the correct entity name. Long answers or extensive web searches may not add value. However, over-reliance on internal knowledge can lead to outdated answers when information changes over time. Correctly differentiating between static facts and dynamic knowledge can help mitigate this tradeoff.

5 Source Breadth & Content Diversity

Reliance on internal *v.s.* external knowledge alone does not determine what information users ultimately see. We now examine the popularity, ranking depth, and diversity of cited sources to understand how generative search engines reshape the informational landscape relative to traditional search.

5.1 Source Popularity and Rank Depth

Most generative engines cite sources of lower popularity than `Organic` search. We use the Tranco rankings[2] to measure the rank of sources within the 1M most visited domains. 89% of `Organic` sites appear in the top 1M list, compared to 85% for `AIO`, 86% for `GPT-Search`, 84% for `Sonar`, 83% for `Gemini`, and 81% for `GPT-Tool`. Figure 4 shows the rank CDF for the links that were found within the Tranco 1M list. While `GPT-Tool` retrieves fewer popular sites overall, the sources it does cite are often highly ranked: its median domain rank (1,124) exceeds that of `Organic` (2,352), except on the Products dataset. For the Regulatory Actions dataset, both the GPT models retrieve sources that are ranked significantly higher. `Sonar` cites the least popular sources overall, with a median domain rank of 5647.

Generative search reaches far beyond `Organic` search. Figure 5 shows that the `AIO` links have less than 50% overlap with the top-10 `Organic` results, and overlap remains below 60% even when considering the top-100. Figure 13 in Appendix E shows the overlap for other search engines. Overlap is substantially lower for `Gemini` and `GPT-Search`, and close to zero for `GPT-Tool`. Domain-level overlap is higher than URL-level overlap, but a substantial fraction of cited domains still fall outside the top-100 `Organic` domains (Figure 14). These findings indicate that generative search significantly broadens the set of consulted sources.

5.2 Source Diversity

Generative engines draw from different types of sources than `Organic` search. We classify domains into high-level categories using the Google Content Categories (26 categories) and a more

[2] https://tranco-list.eu (Retrieved in July 2025)

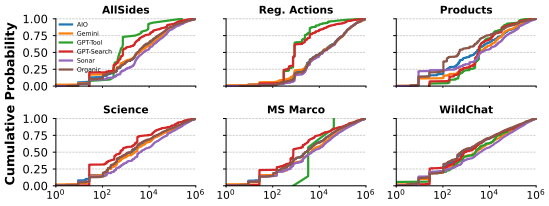


Figure 4: Different search engines select information from domains of differing ranks.

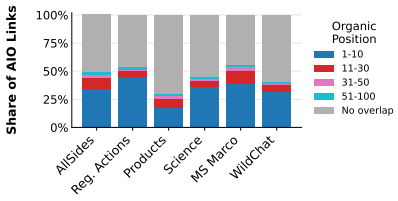


Figure 5: Overlap between AIO links and top-100 Organic links.

web-oriented hand-designed categorization (10 categories) using an LLM as a judge. Details can be found in Appendix E.1. The hand-designed list includes categories like "Encyclopedia" (*e.g.,* Wikipedia and World Atlas) and "Science Publisher" (*e.g.,* to classify websites like ACM, ScienceDirect and arXiv).

Figures 15 and 16 in Appendix E.3 show large differences in category distribution of websites retrieved by various engines. The GPT models rely heavily on Corporate Entities and Encyclopedias, with less reliance on Social Media and User Forums. In contrast, other engines like Organic can retrieve up to 35% from such websites. In summary, not only does it expand the depth of retrieval, **generative search also reshapes the composition of sources** to which users are exposed.

5.3 Conceptual Diversity

We now study how differences in sources translate into differences in the content presented to users.

5.3.1 Output Characteristics

Generative engine outputs differ structurally from Organic search and between engines. Across generative engines, output length and the number of links vary widely. For GPT-Tool, GPT-Search, and Gemini, longer responses tend to include more links, whereas AIO often produces comparatively short texts with many citations (Figure 6). Gemini produces the longest responses on average (2284 ± 1239 characters), with especially long outputs for Products (mean: 3636 ± 995), while GPT-Tool generates the shortest texts (mean: 939 ± 575) and cites the fewest links (0.25 per search result).

We next examine how these differences shape the topical breadth of the content.

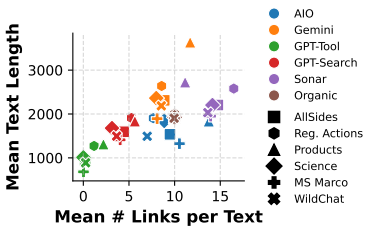


Figure 6: Ratio between the retrieved number of links and text length.

5.3.2 Topical Content Analysis

We compare the high-level concepts mentioned per search engine for each query. We use LLooM (Lam et al., 2024), an LLM-powered topic inference framework that annotates unstructured text with interpretable concepts (details in Appendix E.2).

Traditional and generative search achieve similar overall topic coverage. Across datasets, generative engines achieve coverage comparable to Organic search (Figure 7). For instance, average coverage ranges from 0.71 (GPT-Tool) to 0.78 (Gemini), closely matching Organic (0.78). In most cases, the first five organic search results are sufficient to achieve high coverage, with cumulative gain showing diminishing returns. This observation raises a key design question: *When generative search can condense objective information into concise responses, how many organic snippets does a user actually need to see?*

While aggregate coverage values are similar, the specific concepts mentioned differ across engines. The average concept overlap between generative engines and Organic ranges only from 60% to

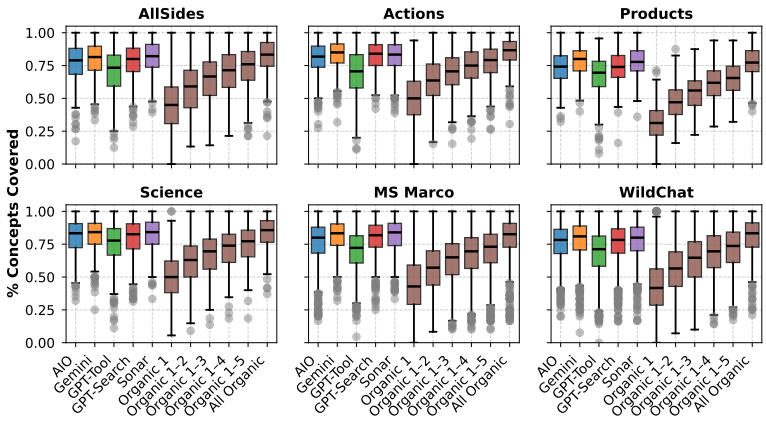


Figure 7: Fraction of concepts covered per engine. `Organic` reaches high coverage within five results.

68% per query (Figure 10 in Appendix E.2). At the 10$^{\text{th}}$ percentile of queries, where agreement across engines is lowest, `Organic` reaches a median coverage of 67%, compared to 55% for `AIO` and 48% for `GPT-Tool`. Queries in this low-agreement regime tend to be underspecified (*e.g.,* "when did Queen Elizabeth"), malformed (*e.g.,* "when is a gene"), or ambiguous (*e.g.,* "what is an example of inequality?"). Figure 11 in Appendix E.2 illustrates an example in which different engines surface largely non-overlapping concepts for the same query. In other words, **Organic search retains an advantage on ambiguous queries.**

Overall, this analysis suggests that, despite accessing more, and often lower-ranked, sources, generative search engines do not consistently surface more concepts than traditional search.

6 Stability of Generative Search

Generative search systems introduce new challenges around the *stability* of search outputs. They combine evolving web content, complex retrieval pipelines, and stochastic language models. As a result, identical queries may yield different outputs depending on *when* and *how* they are executed. We study two complementary sources of variability: **temporal drift**, arising from changes in models and web content over time, and **stochasticity**, which can induce variation even across closely spaced executions of the same query.

6.1 Temporal Stability

To quantify how outputs evolve as the web and models update, we repeated all experiments approximately two months apart (July/August and September 2025).

The number of triggered AIOs and conducted searches for `GPT-Tool` see minor changes (Table 5 in Appendix A.2). We also do not observe any notable changes in the number of links retrieved by different engines, with the only notable exception being `AIO` where the average number of links per query changes from 8.4 to 7.2.

Rank distributions remain mostly consistent for `Organic` search, but vary considerably for generative systems, especially `GPT-Tool` and `GPT-Search`, whose retrieved domain popularity shifts by as much as 40,000 ranks (Figure 8a in Appendix D.1). Comparing the Jaccard similarity between sets of retrieved links across runs (Figure 8b in Appendix D.1) shows highest overlap for `Organic` search (45%). In contrast, `AIO` exhibits substantially lower overlap (18%) between runs. This observation aligns with prior reports indicating that Google's AI mode often returns markedly different results across repeated sessions.[3] Despite this source-level variability, overall conceptual coverage

[3] https://tinyurl.com/aio-differences

	Response
t_0	**No,** the Province of Soria does not directly border Navarre; Soria is in central Spain, bordered by La Rioja, Zaragoza, Guadalajara, Segovia, and Burgos, ...
t_{5m}	**Yes,** the Province of Soria (in Castile and León) and the region of Navarre share a border, with Soria touching Navarre and Aragon to its east/northeast, ...
t_{24h}	**Yes,** the Province of Soria in Spain does border Navarre; they share a northern boundary, ...

Table 2: `AIO` responses for the query "Does the Soria province have a border with Navarre?". Both wording and final answer polarity change over time.

remains largely stable across time (Figure 8c in Appendix D.1), indicating that generative systems can maintain a similar topic coverage in their generated summaries even as retrieved sources change.

6.2 LLM Stochasticity

Generative search engines rely on stochastic language models and dynamic retrieval pipelines. To assess whether this stochasticity materially affects search outcomes, we curate a custom dataset from WildChat (Zhao et al., 2024) that allows only three possible answers: affirmative, negative, or mixed. The dataset consists of questions like "Is it rare for K-pop songs nowadays to have full Korean track titles?". We provide details on dataset construction in Appendix D.2.

Our goal is to study whether the stochasticity of the underlying LLM changes the polarity of the response, *e.g.,* from affirmative to negative. To that end, we issued the queries at three time points (t_0, $t_{5m} := t_0 + 5$ minutes, and $t_{24h} := t_0 + 24$ hours. We study two temperature conditions: (i) a zero-temperature setting, and (ii) the default temperature per model (1.0 for `GPT-Tool` and `Gemini`, 0.2 for `Sonar`, temperature control is not supported for `GPT-Search` and `AIO`). Since generative models can provide an affirmative response without mentioning the term "yes", we use an LLM-based judge to classify each response as *affirmative*, *negative*, or *mixed* (details in Appendix D.3). Table 2 shows example responses and the corresponding judge annotations.

Decision Stability.

Table 3 reports the fraction of queries for which the overall decision changes. We observe **substantial decision instability in decisions** across all generative search engines. Even at $T = 0$, between 9% and 27% of queries exhibit a decision flip within five minutes, with slightly higher rates observed over 24 hours. Lexical overlap across repeated executions is low for all engines. Mean Jaccard similarity ranges from 0.27 to 0.63 at zero temperature (Table 6 in Appendix D.4), indicating that repeated executions often share less than half of their lexical content. Increasing the temperature further amplifies textual variability for most engines that have temperature control.

Engine	Temp. = 0		Default Temp.	
	$t_0 \to t_5$	$t_0 \to t_{24}$	$t_0 \to t_5$	$t_0 \to t_{24}$
`GPT-Tool`	9%	10%	18%	18%
`GPT-Search`	16%	17%	15%	22%
`Gemini`	15%	15%	20%	20%
`AIO`	17%	16%	11%	17%

Table 3: Percentage of queries with a flip in repeated executions (between *affirmative*, *negative*, *mixed*).

Comparison to Organic Search. We compare the answer polarity of `Organic` search with generative search. While generative search output consists of one coherent piece of text, `Organic` search shows around 10 results on the first page. We label each of the top ten search results individually. Across all generative engines, 55% of the responses are labeled as affirmative, 19% as negative, and 26% as mixed. In contrast, in only 16% of queries, all top-10 organic search results express a single polarity label (affirmative, negative, or mixed). These results highlight a structural difference between search paradigms: *Organic search distributes ambiguity and disagreement across multiple results. Generative search resolves this plurality into a single answer that may itself change over time.*

7 Conclusion and Future Work

We conduct a systematic evaluation of generative search engines, focusing on new dimensions that are largely absent from traditional ranked-list evaluation. By comparing a conventional search engine with multiple generative search engines across diverse datasets, we show that generative search introduces trade-offs in how information is retrieved, synthesized, and presented. In particular, systems vary substantially in their reliance on internal *v.s.* external knowledge, retrieval efficiency, knowledge breadth, and stability. Despite often achieving comparable topical coverage, generative engines differ markedly in their retrieval footprints, source diversity, and sensitivity to time and stochasticity.

These findings suggest that evaluating generative search requires moving beyond correctness-centric metrics towards benchmarks that explicitly account for sourcing behavior, synthesis, and variability over time. At the same time, traditional search and generative search each exhibit complementary strengths: While ranked lists of results can expose multiple perspectives and offer greater stability, generative search enables aggregation from a wider range of sources and, in some cases, more efficient answers by leveraging internal knowledge. Understanding and measuring these trade-offs is critical as generative search systems increasingly serve as interfaces to online information.

This work opens several directions for future work. User-centric evaluations could study how different search paradigms affect trust, satisfaction, and decision-making. The observed temporal and stochastic variability motivate longitudinal evaluations that track system behavior over time and across model updates. Finally, evaluating a broader range of models and retrieval architectures could shed light on how different design choices shape retrieval strategies and synthesis behavior.

8 Limitations

Our work has several limitations. First, the scope of our analysis is restricted to selected query workloads spanning general information, products, politics, and science. We did not consider multi-turn conversational searches, and all experiments used English-language queries in only two countries.

Second, our analysis of `Organic` search content is limited to the first ten results, assuming users rarely go beyond the first page. While prior work provides evidence to supports this assumption (Bar-Ilan et al., 2009; Nowicki, 2003; Urman & Makhortykh, 2023), it may not hold for niche cases. We also restrict our analysis to the title, URL, and snippet of each search result, assuming that users rarely click on links. While there is some evidence supporting this behavior (Miroyan et al., 2025), we acknowledge that this design omits the information contained in the full underlying webpages and ignores possible user actions that deviate from these assumptions. However, as shown in our extended analysis with full crawled webpages in Appendix H, this approximation primarily affects a recall–precision trade-off: full webpages increase concept recall but also introduce additional, often less query-relevant content.

Third, our evaluation captures a limited time window, while both organic and generative outputs evolve over time with model updates, indexing changes, and emerging events. In addition, while we controlled for several experimental factors (*e.g.,* logged-out sessions, standardized query execution, and fixed geographic settings), we cannot fully eliminate all sources of variability inherent to commercial search systems, such as backend personalization, infrastructure differences, or time-of-day effects. These factors may introduce additional noise into the observed outputs. Furthermore, we examine a limited set of search engines. Our study focuses on deployed, closed generative search systems. Moreover, generative outputs are inherently non-deterministic, and our evaluation uses only one output per query, though we used a temperature value of 0 whenever configurable.

Finally, our content analysis relies on an LLM-based concept induction method (LLooM), which may introduce potential bias and can reflect the model's own knowledge gaps or preferences. While human annotation could verify these concepts, it does not scale well to large datasets. Future work should explore scalable alternatives, such as semi-supervised methods (*e.g.,* labeled LDA), and combine coverage metrics with fact-checking and credibility assessments.

9 Ethical Considerations

We believe this study does not raise direct ethical concerns or potential risks. The WildChat dataset we used contains real-world user-LLM interactions. However, any sensitive or harmful content was removed by the dataset authors prior to our use. All other queries in our evaluation were drawn from publicly available datasets, commercially available product lists, or were generated synthetically for the purpose of controlled experiments. Our analysis focuses exclusively on publicly accessible search results and model outputs, without altering or influencing any live search systems.

10 Acknowledgments

We thank Nicole Krämer for her helpful feedback.

References

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, Shen Gao, Jiafeng Guo, Xiangnan He, Yanyan Lan, Chenliang Li, Yiqun Liu, Ziyu Lyu, Weizhi Ma, Jun Ma, Zhaochun Ren, Pengjie Ren, Zhiqiang Wang, Mingwen Wang, Ji-Rong Wen, Le Wu, Xin Xin, Jun Xu, Dawei Yin, Peng Zhang, Fan Zhang, Weinan Zhang, Min Zhang, and Xiaofei Zhu. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community, July 2023.

Marwah Alaofi, Negar Arabzadeh, Charles L. A. Clarke, and Mark Sanderson. Generative Information Retrieval Evaluation. volume 51, pp. 135–159. 2025. doi: 10.1007/978-3-031-73147-1_6.

Mohammad Aliannejadi, Jacek Gwizdka, and Hamed Zamani. Interactions with Generative Information Retrieval Systems, July 2024.

Andrea Bacciu, Enrico Palumbo, Andreas Damianou, Nicola Tonellotto, and Fabrizio Silvestri. Generating Query Recommendations via LLMs, June 2024.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. Presentation bias is significant in determining user preference for search results—a user study. *J. Am. Soc. Inf. Sci. Technol.*, 60(1): 135–149, January 2009. ISSN 1532-2882.

Hung-Ting Chen and Eunsol Choi. Open-World Evaluation for Retrieving Diverse Perspectives, April 2025.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 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*, pp. 6437–6447, Barcelona Spain, August 2024. ACM. doi: 10.1145/3637528.3671458.

Sunhao Dai, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, Ji-Rong Wen, and Tat-Seng Chua. 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*, pp. 3922–3931, July 2025. doi: 10.1145/3726302.3730353.

Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *SIGKDD Explor. Newsl.*, 23(1):50–58, May 2021. ISSN 1931-0145. doi: 10.1145/3468507.3468515.

Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1916–1929, Washington DC USA, July 2024. ACM. ISBN 9798400704314. doi: 10.1145/3626772.3657849.

Google. Grounding with google search. https://ai.google.dev/gemini-api/docs/google-search, 2025. Accessed: 2025-08.

Guangxin He, Zonghong Dai, Jiangcheng Zhu, Binqiang Zhao, Qicheng Hu, Chenyue Li, You Peng, Chen Wang, and Binhang Yuan. Zero-Indexing Internet Search Augmented Generation for Large Language Models, December 2024.

Nadine Höchstötter and Dirk Lewandowski. What Users See - Structures in Search Engine Results Pages. *Information Sciences*, 179:1796–1812, May 2009. doi: 10.1016/j.ins.2009.01.028.

Xuming Hu, Xiaochuan Li, Junzhe Chen, Yinghui Li, Yangning Li, Xiaoguang Li, Yasheng Wang, Qun Liu, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Evaluating Robustness of Generative Search Engine on Adversarial Factual Questions, February 2024.

Shijue Huang, Wanjun Zhong, Jianqiao Lu, Qi Zhu, Jiahui Gao, Weiwen Liu, Yutai Hou, Xingshan Zeng, Yasheng Wang, Lifeng Shang, Xin Jiang, Ruifeng Xu, and Qun Liu. Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios, 2024a. URL https://arxiv.org/abs/2401.17167.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20166–20270. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/huang24x.html.

Tae Hyun Baek and Minseong Kim. Is chatgpt scary good? how user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, 83:102030, 2023. ISSN 0736-5853. doi: https://doi.org/10.1016/j.tele.2023.102030. URL https://www.sciencedirect.com/science/article/pii/S0736585323000941.

Min Jiang. Search Concentration, Bias, and Parochialism: A Comparative Study of Google, Baidu, and Jike's Search Results From China. *Journal of Communication*, 64(6):1088–1110, 2014. ISSN 1460-2466. doi: 10.1111/jcom.12126.

Hwiyeol Jo, Taiwoo Park, Hyunwoo Lee, Nayoung Choi, Changbong Kim, Ohjoon Kwon, Donghyeon Jeon, Eui-Hyeon Lee, Kyoungho Shin, Sun Suk Lim, Kyungmi Kim, Jihye Lee, and Sun Kim. Taxonomy and Analysis of Sensitive User Queries in Generative AI Search, April 2025.

Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. Diversity Analysis of Web Search Results. In *Proceedings of the ACM Web Science Conference*, pp. 1–2, Oxford United Kingdom, June 2015. ACM. doi: 10.1145/2786451.2786502.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–28. ACM, May 2024. doi: 10.1145/3613904.3642830. URL http://dx.doi.org/10.1145/3613904.3642830.

Dirk Lewandowski. A framework for evaluating the retrieval effectiveness of search engines. In *Next generation search engines: Advanced models for information retrieval*, pp. 456–479. IGI Global Scientific Publishing, 2012.

Alice Li and Luanne Sinnamon. Generative AI Search Engines as Arbiters of Public Knowledge: An Audit of Bias and Authority, May 2024.

Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. Adaptive tool use in large language models with meta-cognition trigger. *ArXiv*, abs/2502.12961, 2025. URL https://arxiv.org/pdf/2502.12961.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Cong Lin, Yuxin Gao, Na Ta, Kaiyu Li, and Hongyao Fu. Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. *Telematics and Informatics*, 85: 102068, November 2023. ISSN 0736-5853. doi: 10.1016/j.tele.2023.102068.

Lin Liu, Jiajun Meng, and Yongliang Yang. Llm technologies and information search. *Journal of Economy and Technology*, 2:269–277, 2024.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating Verifiability in Generative Search Engines, October 2023.

Kerstin Mayerhofer, Robert G. Capra, and David Elsweiler. Blending queries and conversations: Understanding trust, verification, and system choice in search and chat interactions. *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 2025. URL https://api.semanticscholar.org/CorpusID:277621390.

Liv McMahon and Zoe Kleinman. Glue pizza and eat rocks: Google ai search errors go viral. https://www.bbc.com/news/articles/cd11gzejgz4o, 2025. Accessed: 2025-08.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. Search arena: Analyzing search-augmented llms, 2025. URL https://arxiv.org/abs/2506.05334.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 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*, FAccT '25, pp. 1325–1340, New York, NY, USA, June 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732089.

Stacy Nowicki. Student vs. search engine: Undergraduates rank results for relevance. *portal: Libraries and the Academy*, 3:503–515, 07 2003. doi: 10.1353/pla.2003.0065.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

Lu Qin, Jeffrey Xu Yu, and Lijun Chang. Diversifying Top-K Results, August 2012.

Filip Radlinski and Thorsten Joachims. Query Chains: Learning to Rank from Implicit Feedback, May 2006.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, R. Raileanu, M. Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761, 2023. URL https://arxiv.org/pdf/2302.04761.pdf.

Zeyang Sha, Shiwen Cui, and Weiqiang Wang. Sem: Reinforcement learning for search-efficient large language models. *ArXiv*, abs/2505.07903, 2025. URL https://arxiv.org/pdf/2505.07903.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 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*, CHI '24, pp. 1–17, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642459.

Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. Know where to go: Make llm a relevant, responsible, and trustworthy searchers. *Decision Support Systems*, 188:114354, 2025. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2024.114354. URL https://www.sciencedirect.com/science/article/pii/S0167923624001878.

Clemencia Siro, Yifei Yuan, Mohammad Aliannejadi, and Maarten de Rijke. AGENT-CQ: Automatic Generation and Evaluation of Clarifying Questions for Conversational Search with LLMs, October 2024.

Dimitrios Skoutas, Enrico Minack, and Wolfgang Nejdl. Increasing Diversity in Web Search Results. 2010.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents, December 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Aleksandra Urman and Mykola Makhortykh. You are how (and where) you search? Comparative analysis of web search behavior using web tracking data. *Journal of Computational Social Science*, 6(2):741–756, October 2023. ISSN 2432-2725. doi: 10.1007/s42001-023-00208-9.

Aleksandra Urman, Mykola Makhortykh, and Roberto Ulloa. Auditing Source Diversity Bias in Video Search Results Using Virtual Agents. In *Companion Proceedings of the Web Conference 2021*, pp. 232–236, Ljubljana Slovenia, April 2021. ACM. ISBN 978-1-4503-8313-4. doi: 10.1145/3442442.3452306.

Marc Zao-Sanders. How people are really using gen ai in 2025, 2025. URL https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

A DATASETS

A.1 ADDITIONAL DETAILS ON DATASET CONSTRUCTION

Dataset	Domain	Example queries
MS Marco	General (search engine)	- origin of term doldrums - knowledge based technology definition - what causes typhoid fever
WildChat	General (chatbot)	- how do i stop procrastinating - what is rca in and in which part it is used in - which tech CEO is worth more than $ 1 billion
AllSides	Politics	- what is the personal income tax - what is terrorism in 100 words - how does the global economy affect jobs and career
Regulatory Actions	Politics	- what alternatives are offered after ending DEI programs? - is my personal bitcoin affected by the strategic bitcoin reserve and stockpile?
Science	Science	- what is discreet search - what is set based programming - what company is leading in robotics
Products	Shopping	- crocs worth it - school supplies reviews - best bedroom storage dresser
Trends	Recency	- when does ios 26 come out - emmy winners 2025 - ricky hatton cause of death

Table 4: Example queries from each dataset. The datasets span five domains: politics, science, shopping, general real-world queries, and recent search trends.

Our goal was to select datasets that cover one or more of the following desiderata. The datasets should: (i) reflect real users' queries to both traditional and generative search engines, (ii) cover everyday queries as well as more domain-specific workloads like politics, science, and shopping, and (iii) not only focus on queries relating to persistent topics (*e.g.,* "nemesis in literature"), but also focus on time-dependent topics, *e.g.,* "Designating English as the official language of the US" which was an executive order issued in March 2025.

MS Marco. The dataset contains real Bing search queries for open-domain retrieval and QA (Bajaj et al., 2016). We randomly sampled $1,000$ queries.

WildChat. We subsample the dataset of 1 million queries released by Zhao et al. (2024). The authors gathered the dataset by granting users free access to ChatGPT and recording their interactions with the model. The authors of the original dataset post-processed it to remove PII, sensitive, and toxic conversations. Since our goal is to compare search results of traditional and generative search engines, we exclude queries that are conversational in nature. Specifically, we filter out queries that contain "you", "your" or "u". We also filter out queries shorter than 4 or longer than 75 characters. We only select queries that start with "who", "what", "when", "where", "why", "how", "which", sampling 250 queries from each starting word. In total, we obtain $1,750$ queries.

AllSides. We generate this dataset to specifically focus on political issues. We take a list of political topics and issues from the news media site AllSides.com.[4] We excluded topics if they were primarily geographic (*e.g.,* China, Russia), economic without direct political framing (*e.g.,* Business, Banking and Finance), cultural (*e.g.,* Arts and Entertainment), or overly broad (*e.g.,* World, General News) or ambiguous (Criminal Justice). Examples of included topics are "abortion", "immigration" and "religion and faith". We convert the topics to search queries as follows: We feed the topic to Google

[4] https://www.allsides.com/topics-issues

as a search query and gather 10 user-centric questions using Google's "People Also Ask" feature. We also include the topic itself as a query.

Regulatory Actions. We generate this dataset to capture political queries pertaining to recent events. We gathered a list of major executive orders issues by the second Trump administration. We use the list maintained by the Brookings Institute.[5] In total, we gathered 58 executive orders covering the timespan between January and July 2025. For each action, we ask GPT-4o to generate 10 questions that a real person might ask about the action. The prompt for generating the queries about the executive orders (Section 3.1):

```
You are a helpful assistant that generates realistic questions that
people might ask about regulatory actions.

Please generate {num questions} different questions that a real person
might ask about this regulatory action:  {action}

The questions should be:
- Natural and conversational (like how someone would actually ask)
- Focused on practical concerns, implications, or clarifications that a
person might ask about
- Short and direct (less than 15 words)
- Self-contained (include the name of the action in each question)
- Different from each other

Format your response as a JSON array of strings, like this:  ["Question
1", "Question 2", "Question 3", ...]

Respond only with the JSON array, no additional text.
```

Science. Our goal here was to gather queries related to scientific topics. We manually gathered 45 AI-related topics from the ACM Computing Classification System (CCS).[6] Topics include entries like "Information extraction", "Machine translation" and "Vagueness and fuzzy logic". We pass the topics to Google as a search query and gather 10 user-centric queries listed under the "People Also Ask" heading.

Products. With this dataset, our goal was to study the difference in search results on product-related queries. We take a list of the 100 most searched Amazon products of 2023 from semrush.[7] Searches include terms like "apple watch band", "desk" and "iphone 13 case". We then turn the queries into review and comparison-oriented questions using custom templates like "<product name> review" and "<product name> worth it". Given an item from the top-100 Amazon search terms, we use a GPT-4 model to annotate each product with categories and subcategories from Amazon's own category system (e.g., Electronics, Computers, Arts & Crafts), and to generate potential use cases which represent real-world needs or scenarios for the item. We manually extracted the brand names from items that mention a brand.

Given a search term, the usecase, the corresponding category, subcategory and the brand name, we manufacture the search queries as follows:

1. <product name> reviews
2. <product name> worth it?
3. alternatives to <product name>
4. best <product name>
5. best <subcategory name> for <use case>
6. best <use case> <subcategory name>

[5] http://brookings.edu/articles/tracking-regulatory-changes-in-the-second-trump-administration/ (Accessed: 18/07/25)

[6] https://dl.acm.org/ccs

[7] https://www.semrush.com/blog/most-searched-items-amazon/ (Accessed: 18/07/25)

			September 2025				July 2025			
			#AI Overviews		# Searches GPT		#AI Overviews		# Searches GPT	
Dataset	**#Items**	**#Queries**	**US**	**DE**	**US**	**DE**	**US**	**DE**	**US**	**DE**
WildChat	1M	1,750	1,425	1,050	97	63	1,007	1,031	73	66
MS Marco	100K	1,000	847	784	4	2	792	780	3	3
Products	100	422	160	42	64	5	124	46	59	1
AllSides	37	332	280	233	10	6	241	235	8	11
Reg. Actions	58	649	596	437	136	100	427	447	110	104
Science	45	453	418	315	1	0	398	361	1	1
Trends	100	100	3	1	19	19	–	–	–	–

Table 5: Statistics for each dataset. Experiments were conducted in July/August and September 2025. #Items shows the number of items in the source dataset. #Queries shows the number of queries we obtained after the processing steps performed in Section 3.1. #AI Overviews shows the number of Google AI Overviews generated for US and Germany (DE). # Searches GPT shows the percentage of times `GPT-Tool` performed a web search.

Trends. To analyze the behavior of different search engines in response to recent events, we create a dataset of trending queries. We retrieve the top 100 search trends from Google Trends on September 15th, 2025, and query each search engine on the same day. Google Trends[8] analyzes the popularity of search terms in Google Search across the world, reflecting current public attention.

A.2 Dataset-level Statistics

Table 5 reports the total number of data points, number of queries that we randomly selected for search, number of queries that lead to AI Overviews and the percentages of these that triggered a search in `GPT-Tool`.

B Additional Information on Experimental Setup

Our desiderata for selecting search engines are to: (i) study both traditional and generative search engines, (ii) cover generative models whose dedicated use is to be a web search engine *as well as* generative models that are mainly chatbots but can use web search as a tool to respond to user queries. The latter is motivated by the fact that users are increasingly using chatbots to perform tasks similar to web search (Hyun Baek & Kim, 2023; Zao-Sanders, 2025). We focus on deployed, user-facing search systems to characterize how generative search behaves in practice. Our objective is not to evaluate models in isolation, but end-to-end systems that integrate retrieval, ranking, and synthesis. Because most open-source models still lack comparable, interchangeable search pipelines, directly comparing them to closed systems would conflate model capability with system design.

We use the following search engines:

Organic Google Search (`Organic`). We query the Google search engine to obtain the search results, setting the number of top results to be retrieved to 100. In order to compare the overlap between generative search and Google search at various ranks, we analyze these top 100 returned results the `Organic` search. For the remaining analysis, we only consider the top 10 results. In rare cases, Google returns fewer than 10 results due to the presence of other content, like Google Knowledge Graph results. We use the SERP API to automatically query Google websites.[9] We controlled fixed IP regions by setting up virtual machines in the respective countries (US, DE). Both `Organic` and `AIO` were executed using the same SERP API requests, ensuring identical request parameters. All queries were executed independently and in parallel in logged-out sessions using clean virtual environments.

Google AI Overviews (`AIO`). In addition to the organic search results, a Google search in most cases also results in an AI Overview. Whether to generate an AI Overview is decided internally by Google Search. The AI Overviews are also accompanied by source URLs that were used to

[8] https://trends.google.com
[9] https://serpapi.com/search-api

generate the AI overview. Overall, 81% of the queries generated an AI overview in the US location and 65% of the queries generated an AI overview for the DE location. See Table 5 in Appendix A for the number of cases per dataset. *We perform the search comparisons for queries to which all the search engines (organic and generative) produce an answer.* Hence, we query the remaining models only with inputs where `AIO` generated a response, ensuring interface-level comparability. To assess generalizability, we conducted additional analyses on the full query set and observed qualitatively consistent trends. We observe only small relative differences between the full set and the AIO-conditioned subset in terms of retrieval footprints, with changes in the mean number of links typically within $\pm 10\%$, with differences falling within overlapping 95% confidence intervals. The average difference in median rank of up to 527 (for `GPT-Tool`). Similarly, when re-running topic analysis on the full query set on the AllSides dataset, differences remain minor (maximum change of $+4\%$ for `Organic`), suggesting that our conclusions generalize beyond the conditioned subset.

Gemini-2.5-Flash with Google Search (`Gemini`). While Gemini-2.5-Flash is primarily an AI assistant, it can be used with web search to ground the results (Google, 2025). Given a query, the model first decides if a web search should be performed. For our datasets, the model performs the search in more than 99.5% of the cases. If the model decided to perform a search, it transforms the original query by rephrasing it one or more times. For instance, the query "What are the main goals of the Make America Healthy Again Commission?" is transformed to search queries "Make America Healthy Again Commission goals" and "Make America Healthy Again Commission objectives". We use a thinking budget of 0 tokens.

GPT-4o Search (`GPT-Search`) always performs a web search before returning an answer. We set the search radius to "medium". We used the model version `gpt-4o-search-preview-2025-03-11`.

GPT-4o with Search Tool (`GPT-Tool`) determines on a per-query basis whether to use web search as a tool. Except for the Products and Regulatory Actions datasets, the percentage of cases when the model performs a search is close to 0. See Table 5 in Appendix A for details. We set the search radius to "medium". Results for different search context size settings are reported in the Appendix G. The model version we use is `gpt-4o-2024-08-06`.

Sonar (`Sonar`) always performs a web search when generating an answer. We use the model version `sonar` with search mode "web".

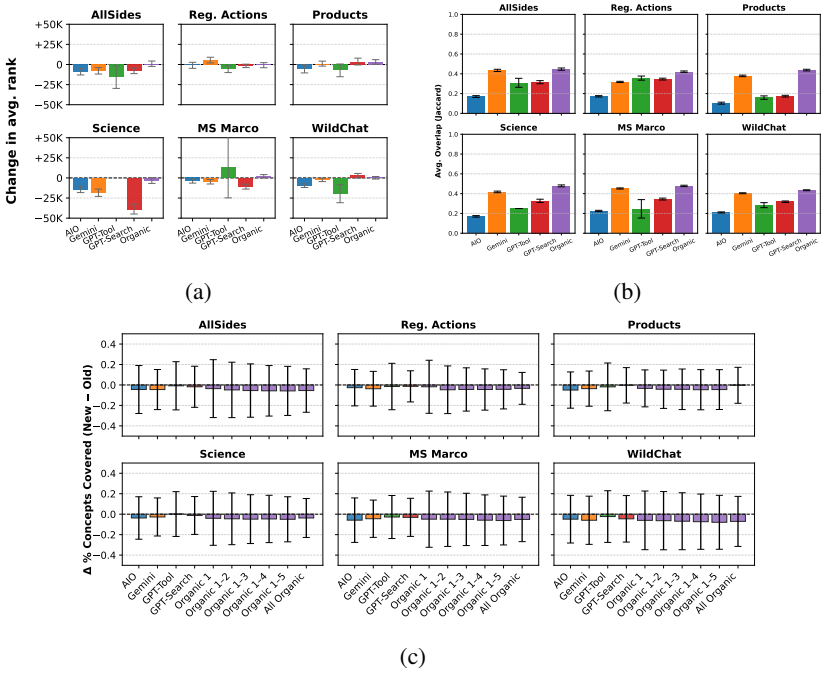
Miscellaneous Parameters. All queries were performed in English. Queries for the generative AI models (`Gemini`, `GPT-Search`, `GPT-Tool`, and `Sonar`) were performed using their respective APIs. For models that support a temperature parameter (`Gemini`, `Sonar`, and `GPT-Tool`), we set it to 0. We set the maximum new tokens to $1,000$. All queries were performed in September 2025. To study the effect of location, we performed queries from two locations: United States (**US**) and Germany (**DE**). We used virtual machines on Google Compute Platform to ensure that the requesting IPs were located in the corresponding country. The main paper contains the results for the US location, we provide a discussion of the results for the DE location in Appendix F.

C Supplementary Material for Section 4

C.1 Details on Capitals Dataset Construction

We construct a dataset of factual queries asking for the capital of countries to study retrieval behavior on simple information needs. We start from a comprehensive list of 249 countries.[10] For each country, we generate a natural-language query of the form "What is the capital of XX?". To reflect natural English phrasing, we add definite articles where appropriate (*e.g.,* "the Netherlands"), generated using `gpt-5.2-chat`. The resulting dataset consists of 249 short queries. We use the Google Knowledge Graph as the ground-truth source and manually resolve ambiguous cases (*e.g.,* countries with multiple administrative capitals).

[10] https://github.com/umpirsky/country-list/blob/master/data/en_US/country.csv



(a) (b)

(c)

Figure 8: Left: Changes in average popularity rank per dataset between July/August and September 2025. `Organic` shows the smallest changes in ranks. Average ranks of generative engines change by as much as 40,000. Middle: Average link overlap per query between the two runs. `Organic` and `Gemini` are the most consistent. Right: Changes in average concept coverage per query. Changes in concept coverage are very small throughout.

The prompt used for optionally adding definite articles to the countries using `gpt-5.2-chat`:

```
Given the following CSV dataset, create an additional column
"question_country" derived from the "country" column.
The "question_country" value must be grammatically correct in the
sentence:  "What is the capital of {question_scountry}?"

Use the country name as-is when no article is required.  Articles
are added only when required in standard English usage, e.g., for
plural country names, country names containing "Islands", or political
entities commonly used with a definite article (e.g., the United
States).  Do not change the country names themselves.

Output the complete dataset as a CSV file including the new column.
```

D Supplementary Material for Section 6

D.1 Detailed results from Section 6.1

Figure 8 shows the difference in rank distribution of retrieved web pages (Figure 8a) and the overlap between the links (Figure 8b) when comparing the July/August and September 2025 snapshots. We do not include `Sonar`, as we only added the model in the September 2025 run.

D.2 Details on Dataset Construction

We curate binary yes/no queries from the WildChat dataset (Zhao et al., 2024) using a multi-stage filtering pipeline. First, we apply rule-based heuristics to identify interrogative prompts that are structurally compatible with binary responses, while excluding open-ended questions and those

containing wh-words (e.g., what, which, who) that typically elicit non-binary answers. We then remove duplicate and near-duplicate queries to reduce redundancy. For the remaining candidates, we estimate the likelihood of a "yes" *v.s.* "no" response using the Qwen3-30B-A3B-Instruct (Team, 2025) model and retain only questions whose predicted probability of a "yes" response falls within a narrow ambiguity range of 0.4–0.6. This ambiguity-based criterion yields a balanced set of well-formed yes/no questions suitable for controlled comparisons across repeated executions. Finally, we perform a manual review to eliminate semantically trivial or meaningless queries (e.g., "can you speak Spanish"), ensuring the resulting dataset contains 209 substantive and informative binary questions.

D.3 Details on LLM-based Judge

We use an LLM-based judge to classify each output as either *affirmative, negative, or mixed*. The model used is `gpt-5.2-2025-12-11` with Temperature 0. The prompt used for classifying model responses:

```
You will be given a question and a corresponding answer by a
respondent.
The question is designed to be answered in an affirmative (e.g., "yes")
or negative (e.g., "no") manner.  However, answers may be longer or
phrased without explicit "yes" or "no".
The answer can also contain both affirmative and negative elements,
e.g., depending on conditions or qualifiers.  The answer can also
express neither an affirmative nor negative stance, e.g., due to
uncertainty or avoidance.  We call these cases "mixed".
Your task is to classify the respondent's answer as "affirmative",
"negative", or "mixed", and provide a brief rationale.

The question is:
{QUESTION}
The respondent's answer is:
{ANSWER}
```

The JSON response format used is shown in Listing 1.

Listing 1: Judge response JSON schema

```
{
  "type": "object",
  "properties": {
    "stance": {
      "type": "string",
      "enum": ["affirmative", "negative", "mixed"]
    },
    "rationale": {
      "type": "string"
    }
  },
  "required": ["stance", "rationale"],
  "additionalProperties": false
}
```

D.4 Details on Textual Overlap

Table 6 shows the textual overlap between repeated execution of the same queries at different timesteps (Section 6.1).

Engine	Temp. = 0		Default Temp.	
	$t_0 \rightarrow t_5$	$t_0 \rightarrow t_{24}$	$t_0 \rightarrow t_5$	$t_0 \rightarrow t_{24}$
`GPT-Tool`	0.63 ± 0.23	0.63 ± 0.23	0.31 ± 0.10	0.31 ± 0.10
`GPT-Search`	0.43 ± 0.16	0.42 ± 0.19	0.44 ± 0.19	0.41 ± 0.16
`Gemini`	0.39 ± 0.14	0.38 ± 0.11	0.31 ± 0.09	0.32 ± 0.10
`AIO`	0.47 ± 0.32	0.36 ± 0.22	0.50 ± 0.32	0.19 ± 0.21
`Sonar`	0.27 ± 0.11	0.27 ± 0.11	0.34 ± 0.12	0.32 ± 0.11

Table 6: Mean Jaccard similarity of repeated executions. Lower values indicate greater textual variability.

E Supplementary Material for Section 5

E.1 Domain category classification

We classify URLs into various categories using their domain (*e.g.,* wikipedia.org). We consider two different categorizations:

1. **Google Content Categories:** We use the 26 top level categories from Google Content Categories.[11] Example categories include "Science", "News" and "Home and Garden".
2. **Custom categories:** Google content categories can be too numerous and broad, *e.g.,* "Games", "Reference". We manually define the following categorization.
   - i News Media Site
   - ii Science Publisher
   - iii Encyclopedia
   - iv Social Media & User Forum
   - v Government
   - vi NGO and Non-Profit
   - vii Think Tank
   - viii Corporate Entity
   - ix Personal Blog
   - x None of the above

We use `gpt4-turbo` to categorize each domain into the the categorization. The prompt we used is:

```
Classify the following domain into one of the provided categories.
Only respond with the name of the assigned category.

Domain: 'wikipedia.org'
Categories:
<list of categories>
```

E.2 Additional Information on LLooM

We utilize LLooM (Lam et al., 2024), an automated concept induction tool, to identify, score, and compare high-level concepts mentioned in the responses from different engines per query.

LLooM operates in two stages. First, using a LLM, it analyzes a set of texts and proposes concepts of increasing generality, along with inclusion criteria and representative text examples. In the second stage, it uses these concepts as labels to classify texts (again, by using a LLM), annotating if a given concept is present.

[11] https://cloud.google.com/natural-language/docs/categories

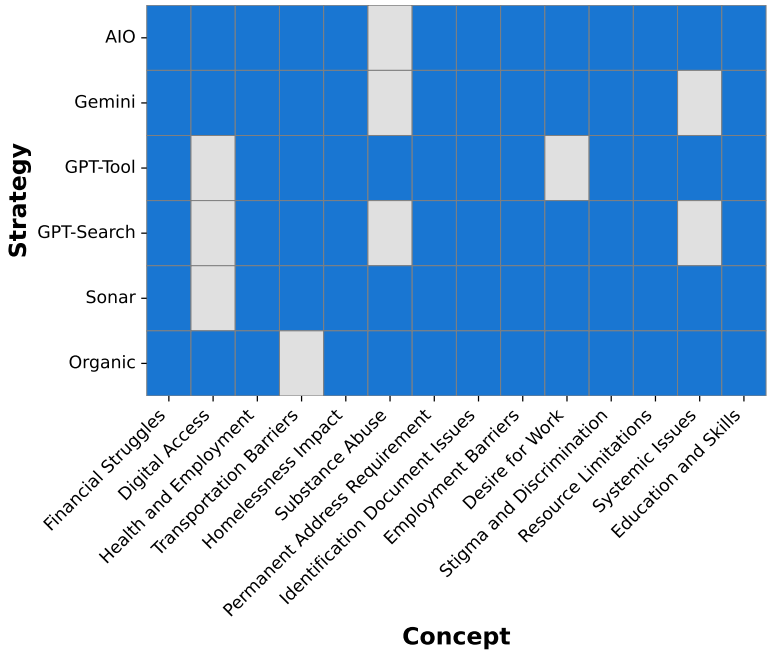


Figure 9: Concepts covered by different strategies for the query "Why don't homeless people get jobs?" (AllSides). The x-axis shows the topics discovered in the search engine outputs, the y-axis shows the search engines. Blue boxes indicate that the topic was present in the output of that search engine. Search engines differ in the concepts they bring up.

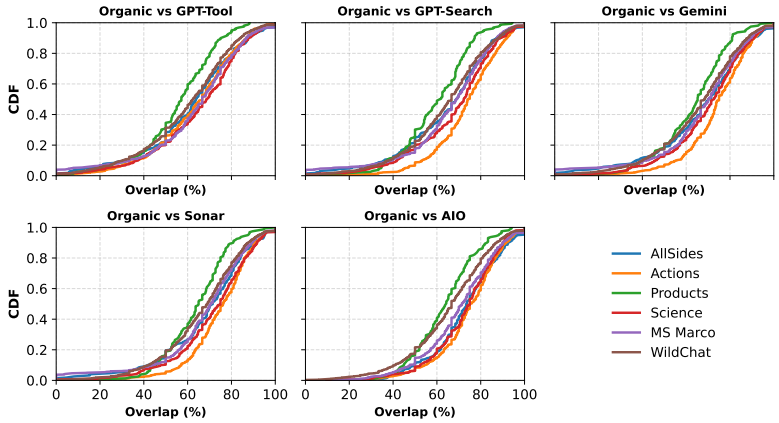


Figure 10: Overlap of concepts between `Organic` and generative search engines.

For each query, we combine the output of all search engines, that is the four generative search engines and the `Organic` search, and apply LLooM to detect topics. Each search engine's output is treated as a separate document except for `Organic` search. For `Organic` search, we treat each of the (usually 10) results as a separate document. LLooM then produces a combined set of topics found in the collection of all the different engines' outputs for that query (stage 1). We then classify each search engine's output against these topics to measure the fraction of identified concepts that each output contains (stage 2). LLooM classifies the presence of the topic at a 5-point scale of ["strongly disagree", "disagree", "neutral", "agree", "strongly agree"]. We deem a topic to be present if the classification output is "agree" or "strongly agree". Figure 9 shows an example of the topics surfaced for a query and how each engine scores them. Following the authors' implementation, we use `gpt-4o` for stage 1 and `gpt-4o-mini` for stage 2.

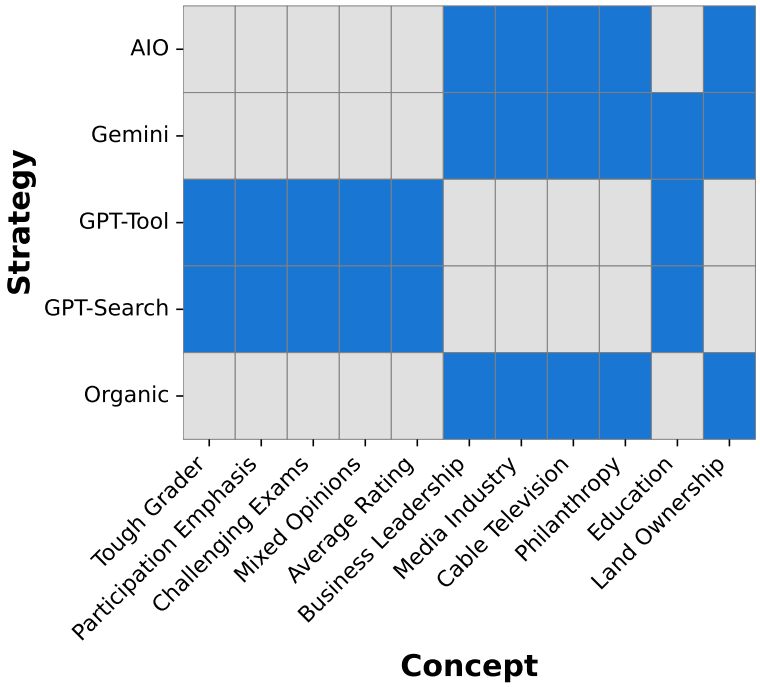


Figure 11: Topic coverage on the query "who is john malone at cc" (WildChat). Blue boxes indicate that the topic was present in the output of that search engine. While `GPT-Tool` and `GPT-Search` mention a professor called John Malone, the other search engines refer to a billionaire businessman.

E.3 Remaining figures from Section 5

In this section, we report the overlap between the generative search engines and the top 100 links retrieved by `Organic` search. Overlap increases when matching domains (see Figure 13) as compared to matching the URLs returned by the respective engines (see Figure 14). GPT models show the lowest overlap with organic search results overall, while `Gemini` had the highest overlap. On MS Marco, `Gemini` captures more than 75% of the top 100 links across queries.

We also report the remaining pie chart analyses of the types of links retrieved by each engine for our custom categorization (Figure 15) and Google's taxonomy (Figure 16). For GPT models, we observe a high share of encyclopedias and news media websites. `Organic` search includes up to 35% social media and user forums. However, few such domains are identified for GPT models.

Figure 12 shows the concept coverage distribution on queries that have low concept coverage.

F Discussion of Results for the DE Location

We replicate all experiments described in the main paper for queries issued from the DE (Germany) locale. Overall, the qualitative trends and relative differences between search engines closely mirror those observed for the US location. We do not observe systematic or statistically significant deviations in source diversity, topic coverage, or content characteristics. The primary observable difference concerns the availability of Google AI Overviews. Across datasets, AI Overviews are triggered less frequently in the DE locale than in the US (see Table 5). We leave further investigation of how geographic location influences generative search behavior to future work.

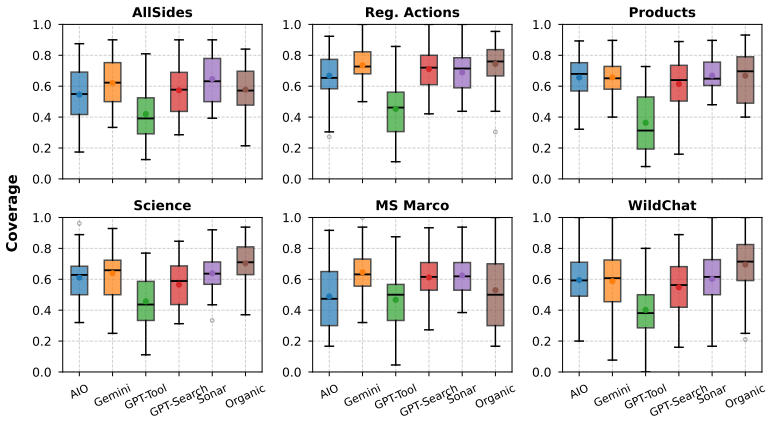
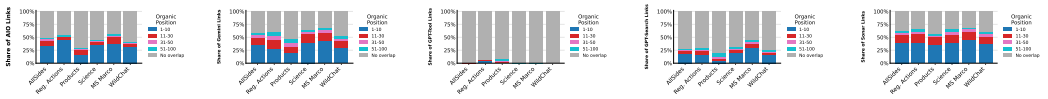


Figure 12: Concept Coverage distribution on low coverage queries. Low coverage queries are those where the fraction of concepts jointly covered by all strategies falls in the bottom 10th percentile of the distribution within a dataset. `Organic` search maintains stable coverage on these queries, while generative engines like `GPT-Tool` struggle to maintain good coverage.



(a) `AIO` (b) `Gemini` (c) `GPT-Tool` (d) `GPT-Search` (e) `Sonar`

Figure 13: [URL overlap] Overlap between the URLs retrieved by the top-100 `Organic` search results and the `AIO`, `Gemini`, `GPT-Tool`, `GPT-Search`, and `Sonar` models.

G Supplementary Analysis of Generative Search Outputs

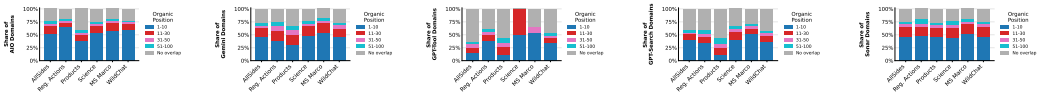
For GPT-based models, the API allows specifying a *search context size* parameter with values `low`, `medium`, and `high`. This parameter controls the amount of external web information the model retrieves and integrates into its response, affecting cost, response quality, and latency.[12] Larger context sizes are expected to produce more comprehensive, but slower and more expensive, answers. Although this parameter has since been deprecated, we analyze its effects across the available settings (`low`, `medium`, `high`).

Minimal Impact of Search Context Size. Across all datasets, varying the search context size does not materially affect sourcing or response content. The likelihood of performing a web search, the number of retrieved links per query, and the popularity rank of retrieved domains remain stable across context sizes. Similarly, response length and topic coverage show no meaningful variation. For example, `GPT-Search` achieves average concept coverage of 78%, 78%, and 77% for `low`, `medium`, and `high` settings, respectively, while `GPT-Tool` remains at 71% throughout. We also observe notable overlap in the concepts retrieved across different context sizes. On WildChat, for instance, the average concept overlap between any two context settings is approximately 65%.

H Supplementary Analysis of Webpage Content

Our analysis of organic search is based on the top-10 results using titles, URLs, and snippets, rather than full webpage content. This choice ensures interface-level comparability, as generative engines present synthesized summaries rather than full documents. Comparing to snippets also reflects typical user behavior, as prior work shows that users predominantly rely on the top-ranked results and often do not click through to full webpages (Radlinski & Joachims, 2006; Bar-Ilan et al., 2009; Höchstötter & Lewandowski, 2009).

[12] https://aiengineerguide.com/blog/openai-web-search-tool/



(a) AIO (b) Gemini (c) GPT-Tool (d) GPT-Search (e) Sonar

Figure 14: [Domain overlap] Overlap between the domains retrieved by the top-100 Organic search results and the AIO, Gemini, GPT-Tool, GPT-Search, and Sonar models.

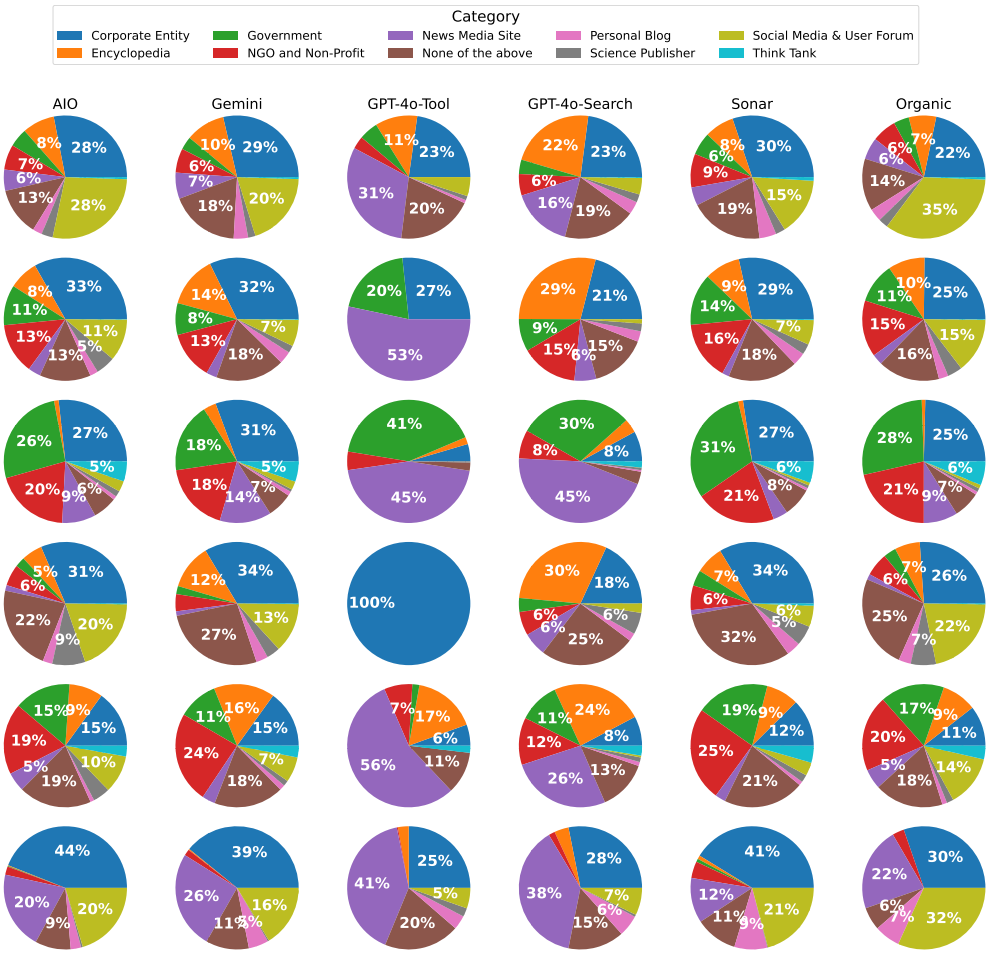


Figure 15: [Custom categorization] Categories of links retrieved by different search engines. The rows correspond to WildChat, MS Marco, Regulatory Actions, Science, AllSides and Products. For the same dataset, different search engines rely on different website types, *e.g.,* social media.

To examine whether this approximation understates the performance of traditional search, we conducted a pilot comparison on a small subset of queries ($n = 60$) across datasets using full crawled webpage content. We preprocess webpages by removing navigation elements and extracting the main content from HTML, excluding pages with fewer than 500 characters to avoid incomplete or blocked content. The 60 queries have on average 7.4 webpages available, with a minimum of 4 pages per query.

Full webpages contain substantially more text. Crawled web pages contain 26,983 characters on average, compared to a maximum of 2,284 characters for Gemini. Running the LLooM analysis reveals that full webpages together cover on average 95% of all concepts, compared to 50–67% across generative systems. These results show that considering the full web page content leads

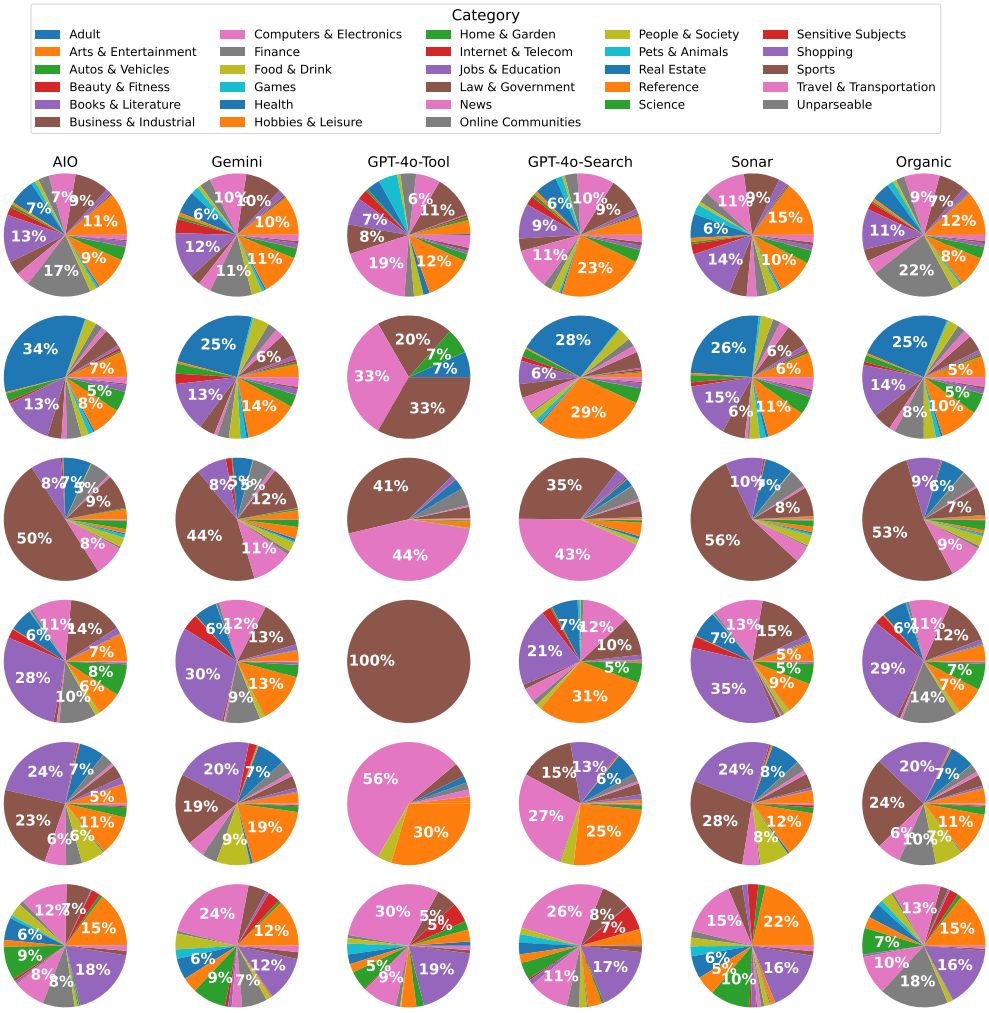


Figure 16: [Google content categorization] Categories of links retrieved by different search engines. The rows correspond to WildChat, MS Marco, Regulatory Actions, Science, AllSides and Products. For the same dataset, different search engines rely on different website types, *e.g.,* social media.

to a **higher recall than any generative engine**. However, this higher recall comes at the cost of **lower precision**: On average, 21% of all concepts are only covered by webpages and not by any other engine, suggesting that websites contain significant amounts of additional information that is often only loosely related to the query. Due to the length of the webpages, **some concepts can be unrelated to the user query**, *e.g.,* for a query about when Tetris was created, some of the concepts that are only surfaced by full webpages are “Future Gaming Trends”, ”Cold War Context“, and “Scientific Influence”. This recall–precision trade-off highlights a structural distinction between paradigms: organic search exposes broader document content, whereas generative systems prioritize concise, synthesized summaries.