Auditing Citation Behavior in AI-Generated Search Summaries: A Framework and a Case Study of Google AI Overviews

Rustem Kakimov[†,*], Jonathan Gillham[‡], Narcis Bejtic[‡], Xing Tan[†,*]
[†] Department of Computer Science, Lakehead University, Canada
[‡] Originality.ai Inc, Collingwood, ON, Canada

Abstract

Search engines increasingly integrate Large Language Models (LLM) to generate natural-language summaries with cited sources, while a growing fraction of online content is partially or fully AI-generated. This convergence raises new questions about how generative search systems select citation sources, particularly with respect to document provenance. In this paper, we propose a system-agnostic observational framework for auditing citation behavior in AI-generated search summaries, modeling retrieval and citation as observable processes over query-document pairs and introducing rank- and provenance-conditioned citation measures. We instantiate the framework in a large-scale empirical study of Google AI Overviews on "Your Money or Your Life" queries drawn from the MS MARCO Web Search dataset. Our analysis shows that AI-generated documents are cited more frequently than human-authored documents even after controlling for retrieval rank, with the difference driven primarily by non-retrieved citations and most pronounced at highly ranked positions. These results highlight the importance of transparent, measurement-based auditing for understanding citation behavior in generative search systems.

Keywords: Generative AI Summaries, Citation Analysis, Google AI Overviews, Document Provenance

1. Introduction

Modern search engines (e.g., Google and Microsoft Bing) increasingly integrate large language models (LLM) to generate natural-language summaries directly in response to user queries (e.g., Google AI Overviews and Bing Copilot), often accompanied by citations to external documents. Meanwhile, a growing fraction of online content is partially or fully generated by AI systems [1–5]. This convergence gives rise to a new setting in which AI-generated summaries may increasingly cite AI-generated documents as sources of information. While AI-generated content is not inherently unreliable, current large language models are known to exhibit hallucinations [6, 7], producing fluent but factually incorrect statements [8, 9]. In high-stakes domains such as health, finance, and legal information, misinformation can have serious real-world consequences, making the citation of unreliable sources particularly concerning. These developments motivate a systematic investigation into whether AI-generated search summaries treat AI-generated documents differently from human-authored ones when selecting cited sources.

Existing information retrieval evaluation frameworks focus on ranked retrieval results, but do not capture citation decisions produced by AI-generated summaries, which may draw from both retrieved and non-retrieved documents and are influenced by rank-dependent exposure effects. Moreover, the internal mechanisms governing citation selection in commercial search systems are largely opaque, limiting analysis to observable inputs and outputs. To address this gap, we propose a formal observational framework that models retrieval and citation as measurable processes over query–document pairs. The framework consists of a precise formalization of the observable query–document space and a set of

[*] rkakimov@lakeheadu.ca, xing.tan@lakeheadu.ca

rank- and provenance-conditioned citation measures that decompose citation behavior into interpretable components. This structure enables rank-conditioned, provenance-aware auditing of citation behavior in AI-generated search summaries without requiring access to internal system details.

We instantiate and evaluate the proposed framework in a large-scale empirical study of generative search behavior. Specifically, we apply the framework to Google Search and Google AI Overviews by collecting organic retrieval results and AI-generated citations for a large set of real-world queries. We annotate retrieved and cited documents according to their provenance and compute the proposed rank-conditioned and provenance-aware citation measures over the resulting query–document pairs. This empirical instantiation allows us to systematically compare citation behavior across document types, distinguish between retrieved and non-retrieved citations, and analyze how citation probabilities vary as a function of retrieval rank. The analysis yields several empirical findings that characterize how citation behavior in generative search systems is associated with document provenance and retrieval exposure.

Our empirical analysis focuses on Your Money or Your Life (YMYL) [10] queries (drawn from the MS MARCO Web Search dataset [11]), which correspond to high-stakes domains such as health, finance, and legal information, where citation behavior carries significant real-world importance. To distinguish between AI-generated and human-authored documents, we use the Originality.ai [12] content detection system to annotate the provenance of both retrieved and cited documents. While no automated detector is perfect, prior benchmarking studies report that Originality.ai achieves strong performance relative to alternative tools. In addition, a subset of queries and document annotations was manually inspected to verify label consistency, providing additional confidence that the observed aggregate patterns are not driven by systematic misclassification.

The remainder of this paper is organized as follows. Section 2 provides background. Section 3 introduces the formal framework. Section 4 instantiates the framework in a large-scale empirical study (Google AI Overviews) and reports the resulting findings. Finally, Section 5 concludes with a summary of results and directions for future work.

2. Background

This section provides background on the key components and concepts underlying our study. We first describe Google AI Overviews (AIO), which integrate large language models into the search interface and introduce a new layer of source selection through AI-generated summaries. We then review the notion of "Your Money or Your Life" (YMYL) queries, which motivate our focus on high-stakes information domains. Finally, we introduce Originality.ai, an AI-content detection tool used to annotate document provenance in our analysis. Together, these elements establish the context required for the framework and empirical evaluation developed in the subsequent sections.

2.1. Google AI Overviews (AIO)

AI Overviews (AIO) are a core component of Google Search that provide a concise snapshot generated by an LLM of key information relevant to the query of a user, accompanied by citations to selected webpages [13]. These summaries are powered by customized Gemini models and likely use a variant of Retrieval-Augmented Generation (RAG) [14].

Since AIO is presented before the traditional search results, its citation mechanism effectively functions as a new ranking layer, deciding which documents are brought to the user's attention and are implicitly treated as trustworthy. Studying how AIO selects its sources is therefore essential for detecting potentially harmful behavior (e.g. biases toward particular classes of documents).

2.2. **Your Money or Your Life (YMYL)**

The term "Your Money or Your Life" (YMYL) refers to topics whose information can significantly affect a person's health, finances, safety, or overall well-being [10]. Examples include medical information, financial guidance, legal advice, and safety-related topics. These domains are inherently high-stakes: inaccurate or misleading information can produce direct and potentially severe real-world consequences. In the context of generative search systems, YMYL queries are particularly sensitive to biases in source selection and citation behavior. Misinformation or unreliable sources in health, finance, safety, or legal domains can lead to direct real-world harm, making accuracy, reliability, and provenance of cited sources especially critical.

2.3. **Originality.ai AI Detector**

The increasing availability of services built on LLMs has led to a growing volume of machine-generated text in a variety of contexts, including academic writing. This trend has motivated the development of automated tools for analyzing the provenance of textual content, with the goal of distinguishing AI-generated text from human-authored text. Originality.ai is one such AI-content detector [12]. It analyzes linguistic and statistical properties of text and produces a classification indicating whether the input is more likely to be AI-generated or human-authored.

Independent benchmarking studies show that Originality.ai consistently performs better than other detectors such as GPTZero [15], ZeroGPT [16], and Winston [17] when evaluated across different LLM models and sampling settings [18, 19]. Other studies also find that automated detectors can surpass human reviewers [20]. In that study, professional evaluators struggled to reliably identify AI-generated or AI-rewritten text, while Originality.ai model detected all of them with a 100% success rate.

In our experiments, we use the Lite 1.0.1 model from the Originality.ai detection system. According to the system documentation, this model is designed to operate on extracted textual content and to assign document-level provenance labels, while minimizing false positives on lightly edited human-authored text [21]. We use the detector solely as a source of document-type annotations for downstream analysis.

3. **A Framework for Evaluating AI Summaries with Citations**

We propose a framework for systematically auditing retrieval and citation behavior in search engines that generate AI summaries, without requiring access to their underlying implementation. The framework provides a unified formal structure for analyzing citation probabilities across documents, conditioned on document type, presence in the retrieval list, and rank position within the retrieved list.

Let $\mathcal{Q}$ denote a nonempty set of natural-language queries, and $\mathcal{D}$ denote a universe of documents. For each query $q \in \mathcal{Q}$, let $\mathcal{C}(q) \subseteq \mathcal{D}$ denote the (finite) set of documents cited by the AI-generated summary in response to q. Meanwhile, searching with respect to a query q returns a sequence of k ranked documents. Given q and k, $\mathcal{S}(q) = (d_1, \ldots, d_k)$, $d_i \in \mathcal{D}^*$ for $1 \leq i \leq k$, and $\mathcal{D}^* \subseteq \mathcal{D}$. Documents in $\mathcal{S}(q)$ are completely ordered: the rank of document d_i is i, that is, $\text{rank}_q(d) = i$. The set of $D(q)$ is defined to be a union of cited and retrieved documents with respect to q, i.e., $D(q) := \mathcal{C}(q) \cup \mathcal{D}^*$. Let $\mathcal{T}$ denote a set of different types of documents. We use $\text{type}(d) = t$, where $t \in \mathcal{T}$, to denote a unique type of document d. For each document $d \in D(q)$, we introduce a structured annotation

$$A(q, d) = \big(\text{type}(d),\ \text{cited}(q, d),\ \text{rank}_q(d)\big),$$

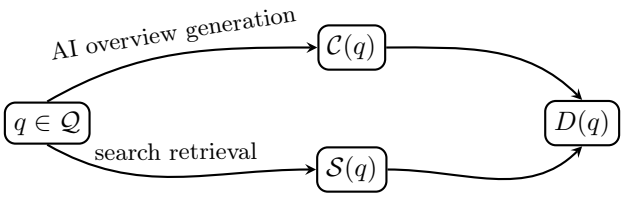

Figure 1. Structure of the framework. A query q produces both retrieved documents $\mathcal{C}(q)$ and citations $\mathcal{S}(q)$ from the AI summary, yielding a combined document set $D(q)$.

where $\text{cited}(q, d)$ is a Boolean value indicating whether the AI summary cites d for query q, and $\text{rank}_q(d)$ denotes rank of the document when in $\mathcal{S}(q)$. Formally,

$$\text{cited}(q,d) = \begin{cases} \mathsf{true} & d \in \mathcal{C}(q), \\ \mathsf{false} & d \notin \mathcal{C}(q), \end{cases} \qquad \text{rank}_q(d) = \begin{cases} i & d = d_i \text{ in } \mathcal{S}(q), \\ \perp & d \notin \mathcal{S}(q). \end{cases}$$

The relationship between these components is illustrated in Figure 1. For each query, the AI summary operator and the search retrieval operator process the input separately, producing a set of citations and a ranked list of retrieved documents, resulting per-query document set $D(q)$.

To analyze how document provenance interacts with both retrieval and citation behavior, we introduce a set of type-wise share measures that quantify the relative prevalence of different document types at distinct stages of the generative search pipeline. These shares allow us to compare the composition of organic retrieval results with the composition of AI Overview citations, and to further distinguish between citations drawn from retrieved documents and citations originating outside the retrieval list. Together, these measures provide a structured way to assess whether particular document types are over- or under-represented in citations relative to their presence in retrieval.

Definition 1 (Document Type in Retrieval and Citation)**.** *For a document type $t \in \mathcal{T}$, we define type-wise shares over query–document pairs (q, d) with $d \in D(q)$. Let $\Omega := \bigcup_{q \in \mathcal{Q}} \big(\{q\} \times D(q)\big)$ denote the set of all such query–document pairs. We first define the retrieval share of type t among all retrieved documents as*

$$\pi_{\text{retr}}(t) = \frac{\big|\{(q,d) \in \Omega : \text{rank}_q(d) \neq \perp,\ \text{type}(d) = t\}\big|}{\big|\{(q,d) \in \Omega : \text{rank}_q(d) \neq \perp\}\big|}. \tag{3.1}$$

Similarly, the citation share of type t among all cited documents is defined as

$$\pi_{\text{cited}}(t) = \frac{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true},\ \text{type}(d) = t\}\big|}{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true}\}\big|}. \tag{3.2}$$

Since the per-query document set $D(q)$ is defined as the union of cited and retrieved documents, we further distinguish citation events according to whether the cited document is also retrieved. The retrieved citation share of type t is defined as

$$\pi_{\text{rcited}}(t) = \frac{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true},\ \text{rank}_q(d) \neq \perp,\ \text{type}(d) = t\}\big|}{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true},\ \text{rank}_q(d) \neq \perp\}\big|}. \tag{3.3}$$

Finally, the non-retrieved citation share of type t is defined as the fraction of citation events in which the cited document does not appear in the retrieval list for the same query:

$$\pi_{\text{ncited}}(t) = \frac{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true},\ \text{rank}_q(d) = \perp,\ \text{type}(d) = t\}\big|}{\big|\{(q,d) \in \Omega : \text{cited}(q,d) = \mathsf{true},\ \text{rank}_q(d) = \perp\}\big|}. \tag{3.4}$$

Next we introduce several citation probabilities that allow us to compare systematically how citation likelihood varies with document type. In particular, we define citation probabilities corresponding to three citation events: citation of a document regardless of its origin, citation of a retrieved document, and citation of a non-retrieved document. All probabilities $\Pr[\cdot]$ are interpreted as relative frequencies over the finite set of observable query–document pairs $(q, d) \in \Omega$.

Definition 2 (Citation Probability)**.** *For a document type $t \in \mathcal{T}$,*

$$P_{\text{cited}}(t) = \Pr[\text{cited}(q, d) = \mathsf{true} \mid \text{type}(d) = t]\,, \tag{3.5}$$

which is the probability that the AI summary cites a document of type t.

Definition 3 (Retrieved Citation Probability)**.** *For a document type $t \in \mathcal{T}$, we define*

$$P_{\text{rcited}}(t) = \Pr[\text{cited}(q, d) = \mathsf{true} \ \wedge\ \text{rank}_q(d) \neq \perp \mid \text{type}(d) = t]\,, \tag{3.6}$$

which is the probability that a document of type t is both retrieved and cited by the AI summary.

Definition 4 (Non-Retrieved Citation Probability)**.** *For a document type $t \in \mathcal{T}$, we define*

$$P_{\text{ncited}}(t) = \Pr[\text{cited}(q, d) = \mathsf{true} \ \wedge\ \text{rank}_q(d) = \perp \mid \text{type}(d) = t]\,, \tag{3.7}$$

which is the probability that a document of type t is cited by the AI summary and does not appear in the retrieved list.

Note that citation events are split into disjoint components: citations of retrieved documents and citations of non-retrieved documents. In particular, this relationship holds $P_{\text{cited}}(t) = P_{\text{rcited}}(t) + P_{\text{ncited}}(t)$, since every citation event either originates from a document that appears in the retrieval list or from one that does not. This decomposition allows us to separately analyze citation behavior driven by organic retrieval versus citation behavior arising from the AI summary's expanded document selection.

While the retrieved and non-retrieved citation shares capture broad differences in citation behavior, they abstract away the strong influence of rank position within the retrieval list. In practice, documents appearing at higher ranks are substantially more likely to be examined and cited than lower-ranked documents, independently of document type. To investigate provenance effects from rank effects, it is therefore necessary to condition citation behavior on retrieval position. Thus, we introduce a rank-conditioned citation probability that measures the probability that a document of type t is cited, conditional on appearing within the top-k retrieved results.

Definition 5 (Citation Probability of Top-k)**.** *For a given document type $t \in \mathcal{T}$ and a cutoff value $k \in \mathbb{N}$,*

$$P_{\text{top}}(t, k) = \frac{|\{(q, d) \in \Omega : \text{type}(d) = t,\ \text{rank}_q(d) \leq k,\ \text{cited}(q, d) = \mathsf{true}\}|}{|\{(q, d) \in \Omega : \text{type}(d) = t,\ \text{rank}_q(d) \leq k\}|}, \tag{3.8}$$

which is the probability that a document of type t appearing within the top-k retrieved positions is cited.

Intuitively, $P_{\text{top}}(t, k)$ measures the likelihood that a document of type t is cited by the AI-generated summary, conditional on the document appearing within the top-k retrieved results for its associated query. The probability is estimated by aggregating over all observable query–document pairs satisfying this rank constraint. With this definition, for each fixed document type, varying k would yield a single rank-conditioned citation curve, allowing citation behavior to be compared across document types.

Finally, to ensure that all these measures are mathematically coherent and do not depend on unobserved internal details of the underlying systems, we establish below that they are uniquely determined by the observable annotations.

Table 1. Dataset statistics

Quantity	Formula	Total
Number of queries	$\lvert\mathcal{Q}\rvert$	2,597
Number of retrievals	$\sum_{q\in\mathcal{Q}} \lvert\mathcal{S}(q)\rvert$	251,163
Number of citations	$\sum_{q\in\mathcal{Q}} \lvert\mathcal{C}(q)\rvert$	39,673
Query–document pairs	$\lvert\Omega\rvert$	272,080
Query–document pairs by type		
AI	$\lvert\{(q,d)\in\Omega : \text{type}(d)=\text{AI}\}\rvert$	19,326
Human	$\lvert\{(q,d)\in\Omega : \text{type}(d)=\text{Human}\}\rvert$	190,834
Unclassified	$\lvert\{(q,d)\in\Omega : \text{type}(d)=\text{Unclassified}\}\rvert$	61,920
Unique documents (*AI* + *Human* + *Unclassified*)	$\lvert\bigcup_{q\in\mathcal{Q}} D(q)\rvert$	214,724
Unique documents by type		
AI	$\lvert\{\, d\in\bigcup_{q\in\mathcal{Q}} D(q) : \text{type}(d)=\text{AI}\,\}\rvert$	16,973
Human	$\lvert\{d\in\bigcup_{q\in\mathcal{Q}} D(q) : \text{type}(d)=\text{Human}\}\rvert$	150,883
Unclassified	$\lvert\{d\in\bigcup_{q\in\mathcal{Q}} D(q) : \text{type}(d)=\text{Unclassified}\}\rvert$	46,868

Observation 1 (Well-Definedness)**.** *For any choice of* $(\mathcal{Q},\mathcal{D},\mathcal{T},\mathcal{C},\mathcal{S})$*, the annotations* $A(q,d)$ *uniquely determine all quantities defined above.*

Explanation. All defined measures depend only on the observed triples $(q,d,A(q,d))$. Once $\mathcal{C}$, $\mathcal{S}$, and the document labels $\text{type}(d)$ are fixed, the annotations are fully determined, and so are all derived quantities.

This framework provides a unified basis for analyzing how AI-generated summaries select citations relative to retrieved search results. Any empirical case study follows by instantiating $\mathcal{C}$ and $\mathcal{S}$ with concrete systems.

4. Experimental Analysis

In this section, we instantiate the framework using Google AI Overviews as the AI summary operator and Google Search as the retrieval operator. This instantiation allows us to empirically evaluate citation behavior at scale and to quantify how citation probabilities vary with document type, retrieval status, and rank position in a real-world generative search system.

4.1. Experimental Setup

We begin by describing the experimental setup used to instantiate the concepts introduced in Section 3. Table 1 summarizes this setup, including the constructed query set, retrieved documents, citation events, and document-type labels used in the empirical analysis. We now detail how these components are obtained, starting with the construction of the query set $\mathcal{Q}$, which serves as the foundation of our empirical analysis. More precisely, we start from the MS MARCO Web Search dataset [11], which contains approximately 10 million real-world queries sampled from Microsoft Bing's production search logs. Queries were randomly sampled from this dataset and classified as YMYL[1] or non-YMYL using OpenAI's GPT-4.1-nano model until 29,000 YMYL queries were obtained. For each sampled YMYL

[1]We focus on YMYL queries because citation behavior in these domains carries disproportionate real-world consequences: errors or biased information in health, finance, or safety-related queries can lead to direct

query q, we retrieved the top-100 organic search results $\mathcal{S}(q)$ and obtained the associated AI Overview when present, including its cited URLs $\mathcal{C}(q)$. Queries for which no AI Overview was returned were discarded. This filtering yields a final set of 2,597 AI-Overview-triggering queries $\mathcal{Q}$, which forms the basis of our analysis.

For each query $q \in \mathcal{Q}$, we collect the top-100 organic search results returned by Google, forming the retrieval list $\mathcal{S}(q)$. The total number of retrievals is obtained by summing the sizes of these lists across all queries, $\sum_{q\in\mathcal{Q}}|\mathcal{S}(q)| = 251{,}163$. This total is smaller than $100 \times |\mathcal{Q}|$ because some queries return fewer than 100 retrievable organic results. For each query $q \in \mathcal{Q}$ that triggers an AI Overview, we extract the set of cited URLs $\mathcal{C}(q)$, where each cited URL corresponds to a single citation event. The total number of citation events is obtained by summing the number of cited documents across all queries, $\sum_{q\in\mathcal{Q}}|\mathcal{C}(q)| = 39{,}673$. For each query q, retrieved documents $\mathcal{S}(q)$ and cited documents $\mathcal{C}(q)$ are aggregated into a unified document set $\mathcal{D}(q) = \mathcal{S}(q) \cup \mathcal{C}(q)$.

The total number of query–document pairs is given by $|\Omega| = 272{,}080$. Under this definition, a document is counted once per query if it appears either in the organic retrieval list, in the AI Overview citations, or in both. Among these pairs, 19,326 correspond to AI-generated documents, 190,834 to Human documents, and 61,920 to Unclassified documents.

Finally, we compute the total number of unique documents appearing across all queries. Let $\bigcup_{q\in\mathcal{Q}} \mathcal{D}(q)$ denote the set of distinct documents that appear in either the retrieval results or the AI Overview citations for at least one query. The total number of unique documents is then given by $\left|\bigcup_{q\in\mathcal{Q}} \mathcal{D}(q)\right| = 214{,}724$. In this count, each document is identified by its normalized URL and is counted once, regardless of how many queries it appears in or whether it appears as a retrieved result, a cited source, or both. These unique documents are partitioned into three disjoint categories according to their provenance labels: AI-generated, Human-authored, and Unclassified[2], yielding 16,973 AI-generated documents, 150,883 Human-authored documents, and 46,868 Unclassified documents, respectively.

4.2. **Document Type in Retrieval and Citation**

We now analyze how documents are distributed across the retrieval and citation stages of the generative search pipeline. Our analysis considers each document along two orthogonal dimensions. The first dimension captures document provenance, distinguishing between AI-generated, Human-authored, and Unclassified sources. The second dimension captures the document's interaction with the system, distinguishing between organic retrieval and citation by the AI Overview, and further separating citations of retrieved documents from citations of non-retrieved documents. This two-dimensional analysis provides the basis for quantifying how document type and retrieval–citation status interact, and for assessing how citation behavior relates to provenance and retrieval exposure.

We begin by considering retrieval as the baseline. From Equation 3.1, we know that $\pi_{\text{retr}}(\text{AI}) = 16{,}656/251{,}163 \approx 6.63\%$. That is, out of 251,163 total retrieval events, 16,656 correspond to AI-generated documents. Similarly, Human-authored documents dominate retrieval with 176,103 instances (70.12%), while the remaining 58,404 retrieved results (23.25%) fall into the Unclassified category. When only cited documents are considered, from Equation 3.2, the share of AI-generated documents increases to $4{,}102/39{,}673 \approx 10.34\%$ (4,102 AI citations out of 39,673 total), which is 3.71% higher than the retrieval share. Human-authored documents also increase in share, accounting for 29,545 citation events

harm. While the proposed framework is query-type independent, YMYL queries provide a natural high-stakes setting in which correctness, reliability, and provenance are most critical.

[2]Unclassified files can be PDF files, video platforms (e.g., YouTube, TikTok), FTP endpoints, pages with fewer than 50 words of text, and inaccessible documents.

Table 2. Retrieval and Citation Shares by Document Type

Metric	AI	Human	Unclassified
$\pi_{\text{retr}}(t)$	**6.63%**	70.12%	23.25%
$\pi_{\text{cited}}(t)$	**10.34%**	74.47%	15.19%
$\pi_{\text{rcited}}(t)$	7.63%	78.98%	13.38%
$\pi_{\text{ncited}}(t)$	12.76%	70.43%	16.81%
$\pi_{\text{cited}}(t) - \pi_{\text{retr}}(t)$	3.71%	4.35%	-8.06%

(74.47%), while the remaining 6,026 citations (15.19%) correspond to Unclassified documents, indicating a redistribution of citations to classified sources.

We decompose citation events into retrieved/non-retrieved citations. Overall, non-retrieved citations account for 52.72% of all citation events (20,917 out of 39,673), while retrieved citations account for the remaining 47.28% (18,756 out of 39,673). Hence, the citations are approximately evenly split between documents in the retrieved list and those do not.

Comparing the document-type distributions across retrieved and non-retrieved citation events reveals a marked difference for AI-generated documents. AI-generated sources account for 7.63% of retrieved citation events (1,432 out of 18,756, from Equation 3.3) but 12.76% of non-retrieved citation events (2,670 out of 20,917, from Equation 3.4), a difference of approximately 5%. However, human-authored documents exhibit the opposite pattern, accounting for a larger share of retrieved citation events (78.98%, or 14,814 out of 18,756, from Equation 3.3) than of non-retrieved citation events (70.43%, or 14,731 out of 20,917, from Equation 3.4), a difference of roughly 9%. By contrast, Unclassified documents are more evenly distributed across retrieved and non-retrieved citation events, accounting for 13.38% and 16.81% of citations, respectively.

Table 2 summarizes the distribution of document types across retrieval and citation events and serves as the empirical basis for the two main findings reported below.

Finding 1 (Human dominance with non-negligible AI presence)**.** *Human-authored documents constitute the majority of both retrieved and cited documents, accounting for 70.12% of retrieval results and 74.47% of citation events. Nevertheless, AI-generated documents form a substantial fraction of the document pool, comprising 6.63% of retrieved documents and 10.34% of cited documents. This indicates that while human-authored sources remain dominant, AI-generated content is already meaningfully represented in both retrieval and citation stages.*

Finding 2 (Amplification of AI-generated documents in citations)**.** *The share of AI-generated documents increases from retrieval to citation by 3.71 percentage points, corresponding to a relative increase of approximately 56%. In contrast, the share of Unclassified documents decreases by 8.06 percentage points. This pattern indicates a systematic redistribution of citation mass toward classified sources, with AI-generated documents being over-represented in citations relative to their presence in retrieval results.*

4.3. **Type-Conditioned Citation Selection Probabilities**

While the share-based analysis in the previous section characterizes how document types are distributed across retrieval and citation events, it does not reveal how likely an individual document is to be cited once it is observable. In this section, we therefore analyze citation probabilities conditioned on document type, separating retrieved and non-retrieved citation events. This probabilistic perspective allows us to distinguish effects driven by document availability from those driven by selection behavior, and to identify which citation mechanisms account for differences observed at aggregated level.

Table 3. Citation Probabilities by Document Type

Metric	AI	Human	Unclassified	$\Delta_{\text{AI}-\text{H}}$	Rel. diff. (AI/H)
$P_{\text{cited}}(t)$	0.2123	0.1548	0.0973	+0.0574	**+37.10%**
$P_{\text{rcited}}(t)$	0.0741	0.0776	0.0405	-0.0035	-4.55%
$P_{\text{ncited}}(t)$	0.1382	0.0772	0.0568	+0.0610	**+78.98%**

Note: For each t, $P_{\text{cited}}(t) = P_{\text{rcited}}(t) + P_{\text{ncited}}(t)$ holds

The probabilities in Table 3 are computed according to Equations 3.5–3.7 as conditional frequencies over query–document pairs $(q, d) \in \Omega$. For example, the overall citation probability for AI-generated documents $P_{\text{cited}}(\text{AI})$, according to Equation 3.5, is defined as the ratio between the number of cited AI query–document pairs and the total number of AI query–document pairs. In our data, 4,102 of the 19,326 AI pairs are cited, yielding $P_{\text{cited}}(\text{AI}) = 4{,}102/19{,}326 \approx 0.2123$. Similarly, 29,545 of 190,834 Human pairs are cited, giving $P_{\text{cited}}(\text{Human}) \approx 0.1548$, while 6,026 of 61,920 Unclassified pairs are cited, giving $P_{\text{cited}}(\text{Unclassified}) \approx 0.0973$.

Retrieved and non-retrieved citation probabilities are computed in the same manner using however Equations 3.6 and 3.7. For retrieved citations, 1,432 of the 19,326 AI pairs correspond to retrieved citation events, yielding $P_{\text{rcited}}(\text{AI}) \approx 0.0741$, while the corresponding values for Human and Unclassified documents are $14{,}814/190{,}834 \approx 0.0776$ and $2{,}510/61{,}920 \approx 0.0405$, respectively. For non-retrieved citations, 2,670 AI pairs are cited outside the retrieved list, yielding $P_{\text{ncited}}(\text{AI}) \approx 0.1382$, compared to $14{,}731/190{,}834 \approx 0.0772$ for Human documents and $3{,}516/61{,}920 \approx 0.0568$ for Unclassified documents. Note that for each t, $P_{\text{cited}}(t) = P_{\text{rcited}}(t) + P_{\text{ncited}}(t)$ holds.

The absolute magnitudes of the citation probabilities in Table 3 should be interpreted as conditional quantities rather than global citation likelihoods. All probabilities are defined over the observable query–document space Ω, which consists only of highly query-relevant documents appearing in the top-ranked retrieval results or among AI Overview citations. Given that AI Overviews cite approximately 15 documents per query from a consideration pool of roughly 100–110 documents, baseline citation probabilities within this restricted space are inherently non-negligible. Accordingly, the most informative signals are the relative differences in citation probabilities across document types under identical conditioning. These differences reveal systematic variation in citation selection behavior, which we summarize in the following findings.

Finding 3 (Higher citation likelihood for AI-generated documents)**.** *Within the observable query–document pool, AI-generated documents are significantly more likely to be cited than Human-authored documents. As shown in Table 3, the overall citation probability for AI-generated documents exceeds that of Human documents by 37.10%, indicating that, conditional on document availability, AI-generated sources are preferentially selected by the AI Overview.*

Finding 4 (Non-retrieved citations drive AI–Human disparity)**.** *The observed citation advantage for AI-generated documents is driven primarily by non-retrieved citation events. While retrieved citation probabilities are nearly identical for AI and Human documents (relative difference of −4.55%), the non-retrieved citation probability for AI-generated documents exceeds that of Human documents by 78.98%. Because non-retrieved citations account for approximately half of all citation events, this disparity directly accounts for the aggregate difference in citation probabilities.*

4.4. **Rank-Conditioned Citation Selection**

In practice, documents appearing at higher ranks receive substantially greater exposure and are more likely to be examined and cited, independent of document type. To investigate provenance effects from rank effects, we further condition citation behavior on retrieval position. In this section, we analyze how citation probabilities vary with rank cutoffs for AI-generated and Human-authored documents, enabling a controlled comparison of citation selection under comparable exposure levels.

Applying Equation 3.8, citation probabilities are computed by aggregating over all observable query-document pairs $(q, d) \in \Omega$ that satisfy the rank cutoff $\text{rank}_q(d) \leq k$. For each document type t (AI-generated or Human-authored), we evaluate the fraction of such pairs that are cited by the AI-generated summary, resulting $P_{\text{top}}(t, k)$, the rank-conditioned citation probability.

Consider for example $k = 10$. For AI-generated documents, the numerator of $P_{\text{top}}(\text{AI}, 10)$ counts all $(q, d) \in \Omega$ in which the document is classified as AI-generated, appears within the top-10 retrieved positions, and is cited, while the denominator counts all AI-generated documents appearing within the top-10 regardless of citation. In our data, this yields 589 cited pairs out of 1,337 eligible pairs, resulting in $P_{\text{top}}(\text{AI}, 10) \approx 0.4405$. Applying the same computation to Human-authored documents yields 8,000 cited pairs out of 20,366 eligible pairs, corresponding to $P_{\text{top}}(\text{Human}, 10) \approx 0.3928$. The difference is $\Delta = 0.048$.

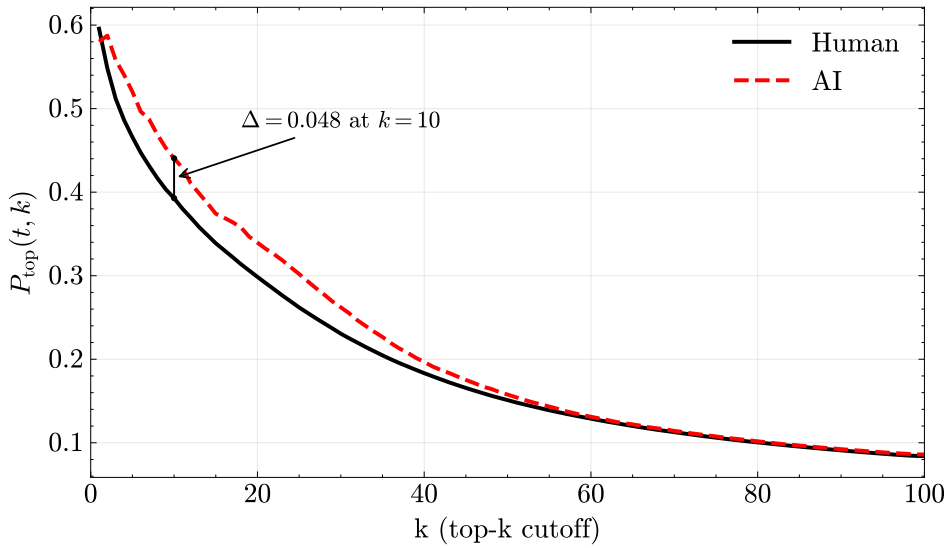

Figure 2. Probability $P_{\text{top}}(t, k)$ as a function of the top-k retrieval cutoff.

Figure 2 reports these probabilities across increasing rank cutoffs k, plotting $P_{\text{top}}(t, k)$ for values of k ranging from 1 to 100. Two separate curves are shown for AI-generated and Human-authored documents, illustrating how the rank-conditioned citation probability varies with retrieval depth for each document type.

It is observed that, at $k = 1$, the probability is slightly higher for Human documents than for AI documents (59.57% vs. 58.02%). When $k = 2$, the relationship is reversed, with AI documents having a higher conditional citation probability (58.74%) than Human documents (54.81%). For all $k \geq 2$, the conditional citation probability remains higher for AI documents than for Human documents. The difference reaches its maximum at $k = 7$, where the probabilities are 48.72% for AI documents and 43.15% for Human documents, corresponding to a gap of 5.57 percentage points. At $k = 10$, the probabilities are 52.02%

for AI documents and 46.60% for Human documents. As k increases, the difference between the rank-conditioned citation probabilities for the two document types decreases. Beyond approximately $k = 50$, the difference becomes negligible.

Finding 5 (Early-rank amplification of AI-generated documents)**.** *At small rank cutoffs, AI-generated documents exhibit a substantially higher conditional citation probability than Human-authored documents. The gap emerges immediately after the top-ranked position and reaches its maximum within the top-10 results, indicating that citation selection amplifies provenance differences most strongly at high-exposure ranks.*

Finding 6 (Long-tail convergence under rank expansion)**.** *As the rank cutoff increases, the conditional citation probabilities for AI-generated and Human-authored documents converge. Beyond approximately the top-50 retrieved positions, citation probabilities become nearly identical, suggesting that provenance-driven differences are primarily a head-ranking phenomenon rather than a long-tail effect.*

5. Conclusion

In this paper, we studied citation behavior in AI-generated search summaries using an observational auditing approach. We proposed a system-agnostic framework that formalizes retrieval and citation as observable processes over query–document pairs and enables rank- and provenance-conditioned analysis of citation selection without requiring access to internal system mechanisms. We instantiated this framework in a large-scale empirical study of Google AI Overviews, focusing on YMYL queries, and analyzed how citation probabilities vary with document provenance, retrieval status, and rank position.

In particular, our analysis yields six empirical findings (Findings 1-6) that collectively reveal previously unmeasured patterns in citation behavior within generative search systems. While human-authored documents continue to dominate both retrieval and citation overall, AI-generated documents are systematically over-represented in citations relative to their retrieval presence, largely due to citations originating outside the retrieved ranking. At the level of individual query–document pairs, AI-generated documents exhibit higher citation probabilities than human-authored documents, an advantage that persists across nearly all retrieval depths and is most pronounced at highly ranked positions. These findings provide the first rank-conditioned, provenance-aware evidence that citation selection in AI-generated search summaries is not solely explained by retrieval exposure, introducing a new dimension for auditing and evaluating generative search behavior.

As a measurement-oriented approach for auditing citation behavior in generative search systems, the framework is observational and system-agnostic, relying only on external inputs and outputs rather than assumptions about internal mechanisms. Its application to Google AI Overviews shows that meaningful and interpretable patterns in citation selection can be uncovered despite the opacity of commercial systems. The framework decomposes citation behavior into simple rank- and provenance-conditioned components, thus it can be readily applied to other search engines, and extended to additional document attributes. As generative search systems continue to evolve, to our belief, such transparent and repeatable auditing tools will be increasingly important for understanding how information is retrieved and then endorsed through citations. Finally, we remark that combining observational auditing with complementary qualitative or user-centered studies may help clarify how citation behavior in generative search systems influences user trust and decision-making, particularly in high-stakes domains.

Acknowledgments

We gratefully acknowledge the support for this research provided by the Mitacs Business Strategy Internship (BSI) program, in partnership with Originality.ai. Rustem Kakimov and Xing Tan also acknowledge support from a Discovery Grant of NSERC, as well as a research grant from the Faculty of Science & Environmental Studies at Lakehead University.

References

[1] Z. Sun, Z. Zhang, X. Shen, Z. Zhang, Y. Liu, M. Backes, Y. Zhang, and X. He. "Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 22975–23005.

[2] Z. Chen, J. Ye, B. Tsai, E. Ferrara, and L. Luceri. "Synthetic politics: Prevalence, spreaders, and emotional reception of AI-generated political images on X". In: *Proceedings of the 36th ACM Conference on Hypertext and Social Media*. 2025, pp. 11–21.

[3] L. La Cava, L. M. Aiello, and A. Tagarelli. "Machines in the Crowd? Measuring the Footprint of Machine-Generated Text on Reddit". In: *arXiv preprint arXiv:2510.07226* (2025).

[4] D. H. Spennemann. "Delving into: the quantification of AI-generated content on the Internet (synthetic data)". In: *arXiv preprint arXiv:2504.08755* (2025).

[5] A. Akram. "Quantitative analysis of AI-generated texts in academic research: A study of AI presence in Arxiv submissions using AI detection tool". In: *arXiv preprint arXiv:2403.13812* (2024).

[6] Y. Bang, Z. Ji, A. Schelten, A. Hartshorn, T. Fowler, C. Zhang, N. Cancedda, and P. Fung. "HalluLens: LLM Hallucination Benchmark". In: *arXiv preprint arXiv:2504.17550* (2025).

[7] C. Dilmegani and A. Daldal. *AI Hallucination: Comparison of the Popular LLMs*. Accessed Oct. 2025. URL: https://research.aimultiple.com/ai-hallucination/.

[8] Z. Xu, S. Jain, and M. Kankanhalli. "Hallucination is Inevitable: An Innate Limitation of Large Language Models". In: *arXiv preprint arXiv:2401.11817* (2024).

[9] S. Banerjee, A. Agarwal, and S. Singla. "LLMs Will Always Hallucinate, and We Need to Live With This". In: *Intelligent Systems Conference*. Springer. 2025, pp. 624–648.

[10] Google. *Creating helpful, reliable, people-first content*. Accessed Oct. 2025. URL: https://developers.google.com/search/docs/fundamentals/creating-helpful-content.

[11] Q. Chen, X. Geng, C. Rosset, C. Buractaon, J. Lu, T. Shen, K. Zhou, C. Xiong, Y. Gong, P. Bennett, et al. "MS MARCO Web Search: A large-scale information-rich web dataset with millions of real click labels". In: *Companion Proceedings of the ACM Web Conference 2024*. 2024, pp. 292–301.

[12] *Originality.ai*. Accessed Oct. 2025. URL: https://originality.ai/.

[13] Google. *Find information in faster & easier ways with AI Overviews in Google Search*. Accessed Oct. 2025. URL: https://support.google.com/websearch/answer/14901683.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[15] *GPTZero*. Accessed Oct. 2025. URL: https://gptzero.me/.

[16] *ZeroGPT*. Accessed Oct. 2025. URL: https://www.zerogpt.com/.

[17] *Winston AI*. Accessed Oct. 2025. URL: https://gowinston.ai/.

[18] L. Dugan, A. Hwang, F. Trhlík, A. Zhu, J. M. Ludan, H. Xu, D. Ippolito, and C. Callison-Burch. "Raid: A shared benchmark for robust evaluation of machine-generated text detectors". In: *Proceedings of the 62nd ACL (Volume 1: Long Papers)*. 2024, pp. 12463–12492.

[19] A Akram. "An Empirical Study of AI-Generated Text Detection Tools". In: *Adv Mach Lear Art Inte* 4.2 (2023), pp. 44–55.

[20] J. Q. Liu, K. T. Hui, F. Al Zoubi, Z. Z. Zhou, D. Samartzis, C. C. Yu, J. R. Chang, and A. Y. Wong. "The great detectives: humans versus AI detectors in catching large language model-generated medical writing". In: *IJEI* 20.1 (2024), p. 8.

[21] J. Gillham. *Which AI Detection Model Should I Use?* Accessed Oct. 2025. URL: https://originality.ai/blog/which-ai-detection-model-to-use.