Answer Bubbles: Information Exposure in AI-Mediated Search

Anonymous ACL submission

Abstract

Generative search systems are increasingly replacing link-based retrieval with AI-generated summaries, yet little is known about how these systems differ in sources, language, and fidelity to cited material. We examine responses to 11,000 real search queries across four systems—vanilla GPT, Search GPT, Google AI Overviews, and traditional Google Search—at three levels: source diversity, linguistic characterization of the generated summary, and source-summary fidelity. We find that generative search systems exhibit significant *source-selection* biases in their citations, favoring certain sources over others. Incorporating search also selectively attenuates epistemic markers, reducing hedging by up to 60% while preserving confidence language in the AI-generated summaries. At the same time, AI summaries further compound the citation biases: Wikipedia and longer sources are disproportionately overrepresented, whereas cited social media content and negatively framed sources are substantially underrepresented. Our findings highlight the potential for *answer bubbles*, in which identical queries yield structurally different information realities across systems, with implications for user trust, source visibility, and the transparency of AI-mediated information access.

1 Introduction

"Google users are less likely to click on a link when they encounter search pages with AI summaries."
—Pew Research Center, July 2025 (Chapekis and Lieb, 2025)

Generative search systems are increasingly mediating how users search for and access information, shifting the landscape from "traditional" link-based search toward AI-generated overviews that select, synthesize, and cite sources on users' behalf (Gao et al., 2023; Li et al., 2024b). Rather than returning ranked lists of links for users to evaluate themselves, these AI systems collapse the search process into a single synthesized output that implicitly determines what information the users receive, how it is framed, and which sources are attributed.

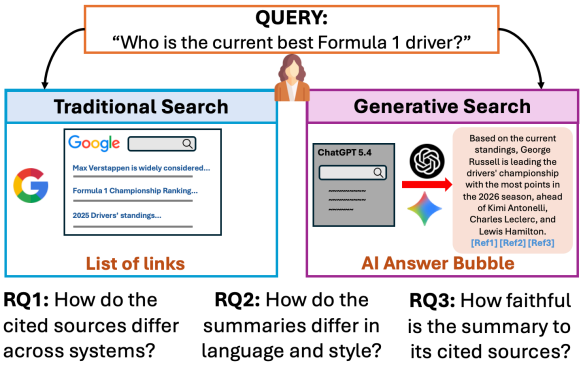

Figure 1: **Paper Overview.** Traditional search returns a ranked list of links for users to evaluate, while generative search produces an *answer bubble* containing AI-generated summaries synthesized from multiple sources. We study these answer bubbles along three dimensions: the sources they cite (RQ1), the linguistic and epistemic qualities of their summaries (RQ2), and how faithfully those summaries represent cited content (RQ3).

Recent evidence suggests that this shift is already reshaping user behavior at scale. A Pew Research study found that when Google's AI Overview appeared in search results, users clicked on traditional result links at roughly half the rate compared to searches without a summary, and clicked on sources cited within the AI summary itself in only 1% of visits (Chapekis and Lieb, 2025). As users increasingly rely on these synthesized overviews rather than navigating to original sources, the design choices embedded in generative search systems, i.e., which sources they retrieve, how they synthesize them, and what language they use, can have great impact on how they shape public information exposure (Sharma et al., 2024).

Despite this growing influence, systematic empirical evaluations across generative search systems remain limited. Prior work has examined informational accuracy in individual large language model (LLM)-based systems (Vu et al., 2024), but how different systems utilize various types of cited

sources, incorporate epistemic attributes in outputs, and maintain fidelity to the original cited information remains understudied. This gap is consequential: just as *filter bubbles* describe how personalized algorithms restrict the diversity of content users encounter (Pariser, 2011), generative search systems risk creating what we term *"answer bubbles"*: self-contained, system-specific information realities in which the same query issued to different search platforms could yield structurally different answers, without the user's knowledge or ability to compare across systems.

Furthermore, information-seeking is also influenced by *how* it is presented. Affective and epistemic cues in language can signal trustworthiness, curiosity, or authority (Vogl et al., 2020; Krakowska, 2020). When synthesizing text, these attributes are no longer properties of the original sources alone, but of the system's own voice. Whether and how these attributes vary across systems remains an open question with direct implications for user trust and information quality.

In this paper, we present the first comparative study of generative and traditional search systems. We query 11,000 real user search queries across 11 topics through four systems—vanilla GPT, Search GPT, Google AI Overviews, and traditional Google Search—and ask three research questions (RQs):

RQ1: How do the sources cited by generative search systems differ from each other and from traditional search in terms of diversity, concentration, and domain composition?

RQ2: How do the epistemic, psycholinguistic, and stylistic qualities of generative search responses vary across systems and topics?

RQ3: How faithfully do generative search summaries represent information from their sources?

Summary of Findings: Our analysis yields three key findings. First, generative search systems draw from fundamentally different source pools, with Search GPT's top-100 domains overlapping only 24–25% with traditional Google Search, favoring encyclopedic and news sources while nearly excluding social media. Second, search grounding systematically reshapes the epistemic character of responses, reducing hedging language by up to 60% while preserving confidence markers, with effects varying significantly by topic. Third, biases in source selection compound during synthesis: Wikipedia is both the most-cited and most over-represented source, whereas social media and discussion forums, despite being cited, are substantively under-used by up to 22 percentage points.

2 Related Work

Auditing Generative Search Systems: The shift from link-based retrieval to LLM-synthesized search responses has prompted growing scrutiny of how these systems access and present information. Liu et al. (2023) audited citation verifiability across four generative search engines, finding that only 51.5% of generated statements were fully supported by their citations. Li and Sinnamon (2024) conducted a smaller-scale audit of ChatGPT, Bing Chat, and Perplexity, documenting sentiment and source-authority biases across public-interest topics. Vu et al. (2024) showed that augmenting LLMs with search improves factual accuracy on time-sensitive queries, and Li et al. (2024b) surveyed LLM-powered search and recommendations. However, these studies focus on individual system properties—accuracy, verifiability, or bias—rather than systematically comparing user-relevant properties across systems. *Our work fills this gap with a large-scale cross-system study spanning source selection, linguistic characterization, and source-summary fidelity.*

Source Bias in Retrieval and Generation: Search engines have long shaped source visibility through ranking concentration (Trielli and Diakopoulos, 2019). Generative search deepens this dynamic, as systems now actively select which sources to incorporate and suppress (Dai et al., 2024b). Prior work has shown that dense retrievers exhibit systematic biases toward short, surface-level documents (Coelho et al., 2024; Fayyaz et al., 2025). Our work extends these findings beyond the retrieval stage, showing that source-selection biases compound during synthesis. *We operationalize this using Equal Coverage and Coverage Parity metrics adapted from Li et al. (2025), quantifying source-summary fidelity as a measurable property of generative search.*

Epistemic and Psycholinguistic Dimensions: Psycholinguistic factors and epistemic emotions shape how users seek and evaluate information (Krakowska, 2020; Vogl et al., 2020), while linguistic register influences trust (Zhao and Liu, 2021; Choo, 2023). Furthermore, behavioral evidence shows users increasingly rely on AI summaries over original sources (Chapekis and Lieb, 2025), amplifying the influence of system-generated language. Research on human-AI interaction further suggests users tend to over-rely on AI outputs lacking explicit uncertainty signals (Buçinca et al., 2021). *We operationalize these constructs through LIWC-based analyses, show-*

ing that search grounding selectively suppresses hedging while preserving confidence markers.

3 Data & Response Collection

We now outline our process for curation of the query dataset and subsequent response collection.

Query Dataset: We drew queries from Google's Natural Questions (NQ) corpus, a collection of real user search queries issued to Google with an information-seeking intent (Kwiatkowski et al., 2019). From this corpus, we randomly sampled 50,000 queries and classified each into one of 11 topical categories using GPT-4o-mini in a few-shot setting with temperature 0. The 11 topics (listed in Table A1) were adapted from prior work (Huynh et al., 2025) through a combination of zero-shot topic modeling using BERTopic (Grootendorst, 2022) and manual inspection of queries categories. To validate the automated classification by GPT-4o-mini, two authors independently labeled a random subset of 120 queries, achieving 90.00% and 90.83% agreement with GPT-4o-mini with an inter-annotator $\kappa = 0.9910$ between the two human raters. To construct the final dataset, we randomly sampled 1,000 queries per topic, yielding a balanced set of 11,000 queries. See §A for descriptive statistics and details on key terms in each topic.

Response Collection: We collected responses to each of the 11,000 queries from three sources, yielding 33,000 query–response pairs:

(1) Vanilla GPT: We queried GPT-4o-mini via the OpenAI API without web search. These responses reflect the model's parametric knowledge alone.

(2) Search GPT: We queried GPT-4o-mini with the *"web_search"* tool call, which retrieves and cites web sources in its generated responses. These responses represent search-grounded LLM output.

(3) Google AI Overview (Google AIO): We programmatically issued each query to Google Search via SerpAPI and extracted the AI Overview (a generative summary powered by a Gemini model) if produced. SerpAPI has been used extensively by prior work for scraping Google search results (Li et al., 2024a; Vu et al., 2024; Sun et al., 2025).

The temperature for GPT generation was set to 0, and the location for querying SerpAPI was a fixed [anonymized] location in Central United States.

4 RQ1: Where Do Various Search Systems Get Their Information?

In this section, we analyze what types of sources generative and traditional search systems synthesize information from to answer user queries. Across query topics, we aggregate and compare counts of domains and specific websites that were cited by GPT and Google AIO.

4.1 Findings

Citation volume, coverage, and overlap: The three systems differ in number of sources cited. Google Organic returns a near-constant 9–10 results per query (the first page of Google search). Search GPT cites a median of 3 sources, with 97% of queries containing at least one citation. Google AIO, however, has the most variable coverage, with only 57.8% of queries triggering a generative overview with a mean of 7.0 references. See Table A2 for more details. We also find that coverage varies by topic: business (84%), history (81%), and education (74%) most frequently produce AIOs, while sports (34%), technology (32%), and travel (32%) rarely do. To quantify overlap, we computed Jaccard similarity on each pair's top-100 most-cited domains. While Google AIO and Organic share 68% of their top-100 domains, Search GPT's top-100 overlaps only 24% with AIO and 25% with Organic Google Search, indicating that it retrieves from a largely distinct set of sources.

Domain-level source patterns: Wikipedia is the most cited domain across all systems: Organic Google Search (81% of queries), Google AIO (28%), and Search GPT (49%). Search GPT also surfaces other encyclopedic and news-wire sources including Britannica (9%), Reuters (5%), and AP News (4%). In contrast, Google AIO and Organic prominently feature user-generated platforms: Facebook (10% and 36%), Quora (8% and 24%), and Reddit (7% and 42%).

Grouping domains into categories, we observe the largest contrast between social media and forums. First, Search GPT draws almost none of its citations from social platforms (0.1%), while Google AIO (8.5%) and Organic (13.4%) rely on them substantially more. Conversely, Search GPT draws disproportionately from encyclopedic and reference sources (27.3% vs. ∼10% for both Google systems) and from news/media outlets (7.2% vs. 2.4% and 1.7%). Government and institutional sources are roughly comparable across all three (∼7–9%). These patterns indicate fundamentally different source-selection philosophy: while Search GPT seems to favor authoritative, editorially-curated content, both of Google's systems surface a broader range of user- and community-driven sources.

Topic-level source distributions: Domain prefer-

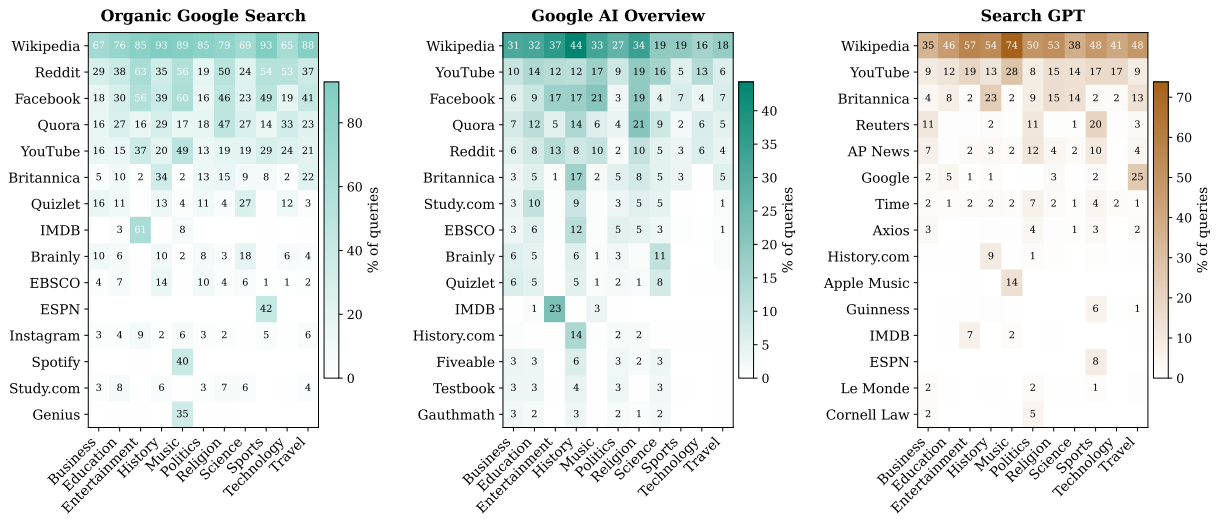

Figure 2: Top-15 cited domains by query topic for each source (% of queries citing each domain). Cell values ≥1% are shown. Domain preferences are strongly topic-dependent: IMDB dominates entertainment, ESPN dominates sports, and Spotify/Genius dominate music, but only in Google's systems and not Search GPT.

ences are strongly topic-dependent (Figure 2). Certain domains concentrate almost exclusively in specific topics. For example, IMDB dominates entertainment queries (61% Organic, 23% AI Overview, 7% Search GPT), ESPN dominates sports (43% Organic, 8% Search GPT), and Spotify (40%) and Genius (35%) appear almost exclusively in music queries within Organic results. Search GPT shows distinctive topic-specific information sourcing patterns, with Reuters and AP News concentrated in sports and politics/news (20% and 12% respectively), while google.com appears in 25% of travel queries (likely Google Maps/Travel content). Google AIO surfaces more educational platforms for factual topics—study.com (11% education), brainly.com (11% science), and quizlet.com (8% science)—while Quora reaches 21% for religion queries. Wikipedia itself varies substantially by topic, ranging from 65% (technology) to 93% (history/sports) in Organic results and from 35% (business) to 74% (music) in Search GPT.

5 RQ2: Linguistic Characterization of Generative Responses

Having established *where* generative search systems obtain their information (RQ1), we now examine *how* they present it. To characterize the linguistic qualities of the 33,000 generated responses, we computed a range of lexical, stylistic, and model-based features spanning three categories:

(1) LIWC categories: We scored each response on 73 Linguistic Inquiry and Word Count (LIWC) categories (Pennebaker et al., 2015). We focus on three conceptual groupings: *Affect* (positive and negative emotion), *Cognition* (insight, tentative, comparative, and overall cognitive process language), and *Analytical* (causal, certainty, discrepancy, and differentiation language). Scores are expressed as proportions of total word count.

(2) Verbosity and readability: We measured *word count*, *average sentence length*, *average word length*, and *type-token ratio* (TTR), which is the ratio of unique word types to total tokens, indicating lexical diversity. We assessed readability through the Coleman-Liau index (Ward et al., 2026).

(3) Model-based scores: We computed four feature-based scores using transformer-based models. *Politeness* was measured with a fine-tuned XLM-RoBERTa classifier[1] (Srinivasan and Choi, 2022), outputting the probability of a text being polite (0–1). *Formality* was measured with a fine-tuned RoBERTa classifier[2], outputting a score from 0 (informal) to 1 (formal). *Polarity* quantifies sentiment from -1 (negative) to $+1$ (positive), via TextBlob (Loria et al., 2026). *Subjectivity* measures the extent to which text reflects personal opinion (0–1), also via TextBlob.

To compare the three response sources (GPT-Vanilla, GPT-Search, and Google AIO), we used Kruskal-Wallis H tests for each feature, followed by pairwise Mann-Whitney U tests with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) for multiple comparisons to determine statistical significance. Effect sizes are reported as Cohen's d (pooled SD).

[1] `Genius1237/xlm-roberta-large-tydip`

[2] `s-nlp/roberta-base-formality-ranker` trained on the GYAFC (Rao and Tetreault, 2018) and Online Formality (Pavlick and Tetreault, 2016) corpora

Feature	Mean			KW	Cohen's d		
	GPT_V	GPT_S	$Google_{AI}$		GPT_V vs GPT_S	GPT_V vs $Google_{AI}$	GPT_S vs $Google_{AI}$
Verbosity & readability							
Word Count	87	140	83	***	-0.65 ***	0.05 ***	0.64 ***
Avg. Sent. Length	16.3	18.0	11.4	***	-0.17 ***	0.58 ***	0.59 ***
Avg. Word Length	5.07	5.25	4.17	***	-0.34 ***	1.06 ***	1.31 ***
Coleman-Liau Index	11.5	12.8	11.7	***	-0.32 ***	-0.06 **	0.27 ***
Type-Token Ratio	0.73	0.64	0.75	***	0.73 ***	-0.14 n.s.	-0.65 ***
Affect							
affect	0.054	0.053	0.031	***	0.03 n.s.	0.57 ***	0.58 ***
posemo	0.045	0.044	0.025	***	0.04 n.s.	0.61 ***	0.61 ***
negemo	0.008	0.009	0.006	***	-0.02 ***	0.14 ***	0.17 ***
Cognition							
cogproc	0.087	0.067	0.044	***	0.39 ***	0.77 ***	0.48 ***
insight	0.021	0.018	0.010	***	0.15 ***	0.55 ***	0.44 ***
tentat	0.025	0.015	0.010	***	0.43 ***	0.63 ***	0.27 ***
compare	0.028	0.023	0.016	***	0.19 ***	0.51 ***	0.39 ***
Analytical							
cause	0.018	0.018	0.011	***	0.02 ***	0.34 ***	0.35 ***
certain	0.010	0.007	0.005	***	0.21 ***	0.41 ***	0.24 ***
discrep	0.006	0.003	0.002	***	0.26 ***	0.37 ***	0.14 ***
differ	0.021	0.013	0.011	***	0.36 ***	0.45 ***	0.13 ***
Feature-based model scores							
Politeness	0.52	0.45	0.41	***	0.39 ***	0.65 ***	0.30 ***
Formality	0.93	0.95	0.84	***	-0.16 ***	0.80 ***	0.99 ***
Polarity	0.10	0.09	0.09	***	0.05 n.s.	0.09 ***	0.05 ***
Subjectivity	0.39	0.39	0.40	***	0.00 *	-0.05 n.s.	-0.06 ***

Table 1: Psycholinguistic and lexico-semantic differences across response sources (Benjamini-Hochberg corrected). KW: Kruskal-Wallis test. Pairwise significances determined by Mann-Whitney U. ***$p < .001$, **$p < .01$, *$p < .05$.

5.1 Findings

In comparing linguistic characteristics of the generative responses, the key design variable is search grounding: Vanilla GPT generates from parametric knowledge alone, whereas Search GPT and Google AIO both retrieve web content before generation. Table 1 reports all features across these systems. Three major patterns emerge:

Structural divergence: Despite all being LLM-based, the three systems produce characteristically different responses. The largest contrast is between the two search-grounded systems: as compared to Google AIO, Search GPT responses are 69% longer ($M = 140$ vs. 83 words; $d = 0.64$), use longer words ($d = 1.31$), are substantially more formal ($d = 0.99$), and infer a higher reading level (Coleman-Liau: $M = 12.8$ vs. 11.7 grade level; $d = 0.27$). In contrast, Google AIO favors shorter sentences ($M = 11.4$ vs. 18.0 words; $d = 0.59$), simpler vocabulary, and higher lexical diversity ($d = 0.65$). These differences suggest completely different design philosophies. While Search GPT produces elaborate, paragraph-style information, Google AIO produces compressed, content-heavy information optimized for quick consumption within a search interface. Vanilla GPT falls between the two on verbosity ($M = 87$ words), shares Search GPT's syntactic complexity and formality, and has TTR comparable to Google AIO ($d = -0.14$) but much higher than Search GPT ($d = 0.73$). Therefore, Vanilla GPT achieves conciseness through vocabulary variety rather than structural compression.

Epistemic and emotional attenuation: The systems exhibit significant epistemic differences.

(1) Hedging and uncertainty: Tentative language (*"maybe," "perhaps"*), the primary marker of epistemic caution, drops by 40% from Vanilla GPT to Search GPT ($0.025 \rightarrow 0.015$; $d = 0.43$) and by 60% from Vanilla GPT to Google AIO ($0.025 \rightarrow 0.010$; $d = 0.63$). Overall cognitive process language also drops by 23% and 49%, respectively, indicating that hedging is reduced at a faster rate than other types of cognitive language.

(2) Reasoning depth: Insight language (*"think," "know"*) is halved from Vanilla GPT to Google AIO ($0.021 \rightarrow 0.010$; $d = 0.55$), and comparative language (*"versus," "rather"*) follows a similar pattern ($d = 0.51$). Analytically, causal language (*"because," "therefore"*) and certainty markers (*"always," "never"*) also decline, though with smaller effects ($d = 0.34$ and $d = 0.41$, respectively). Notably, certainty drops slower than hedging, where the ratio of certainty-to-tentative language (which measures assertiveness) rises from 0.39 in Vanilla GPT to 0.49 in both search systems. Therefore, search grounding does not just compress all epistemic language equally, but rather selectively removes the markers of deliberation while preserving markers of confidence.

(3) Emotional flattening: Positive emotion drops by 44% from Vanilla GPT to Google AIO ($0.045 \rightarrow 0.025$; $d = 0.61$), while negative emotion drops by only 25% ($d = 0.14$). This asymmetry means Google AIO's affect is not just reduced, but

shifted to a lower positive-to-negative emotion ratio. The ML-based polarity and subjectivity scores, by contrast, show negligible differences ($d \leq 0.09$), indicating that the emotional flattening operates at the lexical level, rather than the semantic level, by using fewer emotion words. Google AIO's politeness is also markedly lower than Vanilla GPT's ($d = 0.65$), suggesting a more impersonal style.

Topic-wise variation: These effects are not uniform across topics. Technology and travel/geography show the largest divergence (mean $|d| > 1.0$), with Google AIO reducing 64% of hedging and 74% of positive emotion relative to Search GPT ($d = 1.37$ for cogproc, $d = 1.30$ for posemo) in technology queries. Notably, these two topics also least frequently generated AIOs (32% availability per RQ1). At the other extreme, education, history, and business/finance show near-zero epistemic divergence ($d = -0.03$ to 0.12 for hedging and cogproc), indicating that the two search-grounded systems produce effectively indistinguishable output for these high AI-overview-availability topics (74–84%). Interestingly, for politeness, politics/news and religion show nearly no differences across search systems ($d = 0.02$ and $d = 0.00$), meaning Google AIO matches Search GPT's courtesy for more socially sensitive queries. This contrasts with technology ($d = 0.84$) and sports ($d = 0.94$), where AIOs tend to adopt a much terser tone. Formality, in contrast, remains high across nearly all topics ($d = 0.82$–1.76), suggesting very low topical sensitivity. See Table A3 in §A for detailed results.

6 RQ3: Source-Summary Fidelity

We now shift focus from how generative search systems *cite* sources (RQ1) and *express* information (RQ2) to how faithfully they *synthesize* cited content. Specifically, we examine whether generative summaries draw evenly from sources or systematically favor certain sources over others.

6.1 Methods

We adapt the coverage-based fairness framework of Li et al. (2025) to the generative search setting. The pipeline consists of three steps:

(1) Atomic Content Unit (ACU) decomposition: Each generative response is decomposed into ACUs (single, self-contained factual statements) using GPT-4o-mini with few-shot prompting. We generated ACUs for 1,100 queries per source (100 per topic), producing 17,644 ACUs for Search GPT (mean 16.0 per summary) and 16,898 for Google AIO (mean 15.4).

(2) Extract source context: For each query, we scraped the full textual content of every cited URL using `trafilatura` (Barbaresi, 2021) and `BeautifulSoup4`, filtering for core article content excluding ads and boilerplate. Each source document is chunked into overlapping blocks of approximately 100 words, following Li et al. (2025).

(3) Entailment probability: We compute entailment probabilities $P(\text{chunk entails ACU})$ for all (source chunk, ACU) pairs using RoBERTa-large fine-tuned on MNLI (Liu et al., 2019). ACUs with mutual probability > 0.95 are deduplicated via connected-component clustering (Li et al., 2025). We then compute two coverage metrics:

(i) Equal Coverage (EC): For a single query, EC measures whether the summary draws evenly from all K cited sources:

$$\text{EC}(q) = \frac{1}{K} \sum_{k=1}^{K} |p(q,k) - \bar{p}(q)| \quad (1)$$

where $p(q,k)$ is the mean entailment probability between source k's chunks and the summary's ACUs, and $\bar{p}(q) = \frac{1}{K}\sum_k p(q,k)$ is the overall mean. EC $= 0$ indicates perfectly even coverage, with higher values indicating selective reliance.

(ii) Coverage Parity (CP): CP tests whether sources with a particular attribute a are systematically over- or under-represented across the corpus:

$$\text{CP} = \frac{1}{|A|} \sum_{a \in A} |\mathbb{E}_q\left[p(q,k) - \bar{p}(q) \mid \text{attr}(k) = a\right]| \quad (2)$$

Lower CP values indicate less systematic bias. We adapt CP to the generative search setting by using attribute schemes that capture structural, topical, and linguistic properties of cited sources. We assess statistical significance via bootstrap resampling (1,000 iterations) and use the Mann-Whitney U test for cross-source comparison.

6.2 Findings

Table 2 reports per-topic EC and coverage probability, and Table 3 reports CP across attribute schemes. We now expand upon key findings below.

Google AIO covers sources less evenly than Search GPT: Google AIO exhibits significantly higher EC than Search GPT ($M = 0.177 \pm 0.091$ vs. 0.164 ± 0.113; $p < .001$), indicating that it distributes attention across cited sources less evenly. Google AIO also achieves lower overall coverage (0.447 vs. 0.500; $p < .001$), meaning its summaries capture less of the total content in cited sources despite citing nearly twice as many (mean=5.0 vs.

	Search GPT		Google AIO	
Topic	EC	Coverage	EC	Coverage
Business/Fin./Econ.	0.162	0.500	0.162	0.431
Education/Lit.	0.189	0.481	0.175	0.465
Entertainment	0.185	0.480	0.217	0.459
History	0.160	0.460	0.169	0.445
Music	0.236	0.449	0.177	0.407
Politics/News	0.152	0.506	0.168	0.444
Religion	0.138	0.523	0.165	0.465
Science	0.141	0.543	**0.149**	0.453
Sports	0.164	0.542	0.204	0.455
Technology	**0.131**	0.535	0.189	0.433
Travel/Geography	0.162	0.477	0.177	0.421
Overall	*0.164*	*0.500*	*0.177*	*0.447*

Table 2: Equal Coverage (EC) and mean coverage probability (Coverage) by topic. EC measures how unevenly a summary draws from its cited sources (**lower** = more even). Coverage indicates the fraction of source content entailed by summary ACUs.

	Search GPT Mean Diff (pp)	Google AIO Mean Diff (pp)
Document Length		
Short (<200 words)	-0.134^{***}	-0.177^{***}
Medium (200–800)	$+0.000$	$+0.015^{**}$
Long (>800 words)	$+0.039^{***}$	$+0.046^{***}$
Source Sentiment		
Positive	$+0.004$	$+0.011^{**}$
Neutral	-0.008	-0.010
Negative	-0.037	-0.138^{***}
Domain Category		
Encyclopedia	$+0.022^{***}$	$+0.050^{***}$
Government	-0.042^{*}	-0.035^{**}
Social/Forum	–	-0.221^{***}
Other	-0.015^{*}	$+0.001$
Wiki vs. Non-Wiki		
Wikipedia	$+0.026^{***}$	$+0.054^{***}$
Non-Wikipedia	-0.017^{**}	-0.009^{**}
Citation Position		
Position 0 (first)	$+0.015^{*}$	$+0.032^{***}$
Position 1	$+0.001$	$+0.020^{*}$
Position 2	-0.021^{*}	$+0.012$
Source Subjectivity		
High subjectivity	-0.023^{*}	-0.032^{***}
Low subjectivity	$+0.004$	$+0.008^{**}$

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.
Positive = over-represented; negative = under-represented.

Table 3: Coverage Parity (CP) analysis across attribute schemes. Higher CP values indicate more systematic bias. Significance via bootstrap resampling (n=1,000).

2.7). This suggests a trade-off: Google AIO cites more sources but synthesizes each one less thoroughly. Topic-wise, for Search GPT, music shows the highest coverage inequality (EC=0.236) and technology the lowest (0.131). For Google AIO, entertainment shows the highest inequality (0.217) while science shows the lowest (0.149).

Longer sources are the most over-represented: The largest CP effect for both systems is for source document length (Search GPT=0.058; Google AIO=0.079). Short sources are strongly under-represented (Search GPT: -13.4pp; Google AIO: -17.7pp; both $p < .001$), while long sources are over-represented (3.9pp and 4.6pp). This suggests that these systems place more emphasis on the amount of extractable text than on relevance or quality, exhibiting a length bias that parallels brevity biases in dense retrievers (Fayyaz et al., 2025).

Wikipedia, encyclopedia, and government sources: Encyclopedic sources are over-represented in both Search GPT ($+2.2$pp) and Google AIO ($+5.0$pp). For Wikipedia specifically (Search GPT: $+2.6$pp; Google AIO: $+5.4$pp; both $p < .001$), we find a significant compounding effect where it is not only cited most often (RQ1), but its content also receives disproportionate weight in summaries. In contrast, government sources are under-represented in both systems (-4.2pp and -3.5pp, respectively).

Negative emotion and social media forums are under-represented: Source sentiment shows markedly different patterns across systems. Sentiment effects are non-significant for Search GPT (CP $= 0.016$) but large and significant for Google AIO (CP $= 0.053$): negative-sentiment sources are under-covered by 13.8pp ($p < .001$), while positive-sentiment sources are slightly over-represented ($+1.1$pp, $p < .001$). This asymmetry suggests that Google AIO's synthesis implicitly filters against negative content.

Domain-category analysis further shows that social media and forum sources (Reddit, Quora) are under-covered by 22.1pp in Google AIO ($p < .001$), suggesting that when Google AIO draws far less content from Reddit when cited alongside other sources. Combined with our finding from RQ1 that Google AIO *does* cite social platforms (8.5% of domains), this highlights a gap between citation and synthesis: the model lists user-driven sources in its references but draws disproportionately little content from them. Reddit ($n = 149$, diff $= -0.221$, $p < .001$) and Khan Academy ($n = 21$, diff $= -0.311$, $p < .001$) are the most under-covered in the synthesized summaries.

Sources with epistemic markers receive more coverage: Source-level LIWC analysis reveals that sources containing causal language ("because," "therefore") and certainty markers ("always," "definitely") are systematically over-covered in both systems. Causal language shows CP $= 0.071$ (Search GPT) and 0.070 (Google AIO). Conversely, highly subjective sources are under-covered in both systems (-2.3pp and -3.2pp). Together, these patterns suggest that the models favor sources with assertive, explanatory prose that is easier to extract factual claims from.

7 Discussion

Our study reveals that generative search systems do not simply retrieve and present information, but they actively reconstruct it through source selection, linguistic transformation, and selective synthesis. We organize our discussion around three interconnected themes that emerge from our findings.

The Compounding Effect From Citation to Synthesis: A central finding of this work is that biases in source selection (RQ1) are amplified rather than mitigated during synthesis (RQ3). For example, Wikipedia is not only the most frequently cited domain across all systems, but its content is disproportionately over-represented in generated summaries. Conversely, social media and forum sources present an inverse pattern: Google AIO cites Reddit and Quora in 7–8% of queries, yet draws 22.1% less content from social platforms than from other sources. This creates a *citation-synthesis gap*: the model references user-generated sources lending an appearance of source diversity, while relying on a narrower set of encyclopedic or authoritative texts. This compounding dynamic resonates with concerns raised by Dai et al. (2024a,b) about systematic biases in how models prioritize sources, and extends them by showing that the bias operates not only at the retrieval stage but deepens during generation. For users, this gap is largely invisible, and the cited sources in an AI Overview may suggest breadth that the synthesized text does not reflect. This also has implications for source visibility and the economic viability of smaller publishers and community-driven platforms.

Epistemic Reshaping of Summaries from Uncertain Sources: Our linguistic analysis (RQ2) reveals that search grounding does not uniformly compress language but selectively reshapes its epistemic character. The ratio of certainty-to-tentative language rises from 0.39 in Vanilla GPT to 0.49 in both search-grounded systems, connecting directly to observations from RQ3 findings that sources containing causal and certainty language are systematically over-covered, while highly subjective sources are under-covered. Together, these results suggest that generative search systems preferentially extract from assertive, explanatory prose and then present the resulting synthesis in a style that is more assertive and less hedged than the underlying sources warrant. This epistemic flattening is particularly concerning as it could potentially present uncertain information to users with unwarranted confidence. This aligns with broader concerns in information-seeking literature about how linguistic style influences perceived credibility (Zhao and Liu, 2021; Choo, 2023) and how epistemic emotions shape knowledge exploration (Vogl et al., 2020). If users interpret the absence of hedging as a signal of factual reliability, the epistemic reshaping we find could miscalibrate the trust-reliance association in educational settings (Pal et al., 2026).

Implications and Future Directions: Our findings surface important implications for different stakeholders. *For NLP researchers*, we offer a reusable framework for auditing source-summary fidelity in any retrieval-augmented generation system. Our finding that document length is the strongest predictor of over-representation parallels recent work on brevity biases in dense retrievers (Fayyaz et al., 2025) and suggests that length bias may persist through the full RAG pipeline. Future work should investigate whether debiasing at the retrieval stage propagates to more equitable synthesis, or whether generation introduces its own independent biases. *For system designers*, the divergence between Search GPT and Google AIO demonstrates that these systems have fundamentally different design choices about what constitutes a good search response. Making these design philosophies more transparent to users through source diversity indicators or confidence signals could help users make more informed judgments about the information they receive. *For users and policymakers*, the key takeaway is that the same query yields different "answer bubbles" that the end user is unaware of. We propose that generative search systems should be subject to transparency requirements analogous to those for algorithmic recommendation systems (Kowald, 2024).

8 Conclusion

We presented a large-scale comparative study of generative and traditional search systems across 11,000 real user queries, revealing that the same query produces fundamentally different information realities depending on the system used. Generative search systems diverge sharply in which sources they cite, how they linguistically frame responses, and how faithfully they synthesize cited material, with biases in source selection compounding during generation. As these systems increasingly mediate public information access, our findings underscore the need for greater transparency in how AI-generated search responses are constructed and the sources they prioritize.

Limitations

Our study has limitations that outline interesting directions for future work.

Temporal and geographic scope: Our audit captures behavior of generative search systems at a single point in time from a single location. Generative search systems are updated frequently, and source selection strategies, linguistic patterns, and synthesis behaviors may shift with model updates or policy changes. Similarly, Google search results and AI Overview availability vary by geography, language, and user profile. Longitudinal and cross-regional studies are needed to determine how stable the answer bubble patterns we find are.

Query representativeness: Our queries are drawn from Google's Natural Questions (NQ) corpus, which is primarily focused on information-seeking intent and was originally collected around 2018. Other types of queries may elicit different source selection and linguistic behaviors. Furthermore, our 11-topic list might collapse heterogeneous subtopics into single categories (e.g., "Science" spans biology, physics, and healthcare). Finer-grained topic analysis and more diverse query types could reveal additional insights that our current design did not capture.

Methodological proxies: Our RQ3 pipeline relies on GPT-4o-mini for atomic content unit decomposition and RoBERTa-large for entailment scoring, both of which may introduce noise and domain-specific errors. Since we utilized the work of Li et al. (2025), we did not conduct additional human validation of ACU quality or entailment accuracy. Additionally, not all cited URLs were successfully scraped (e.g., paywalled content), potentially biasing coverage estimates if non-scrapable sources differ systematically from scrapable ones.

No user-study or causal claims: While we provide detailed characterization of outputs, we do not measure user perception or behavior. Therefore, we can not make claims about whether users would notice or be affected by the epistemic and source-diversity differences we find. Controlled user studies measuring how answer bubbles influence trust calibration and downstream decision-making would directly extend this work. Moreover, our study is correlational, drawing associations between search grounding and linguistic attenuation. We therefore cannot establish causal mechanisms or disentangle the contributions of model architecture, training data, and retrieval pipeline design on the effects we find.

System coverage and confounds: We audit four systems, but the generative search landscape includes many other systems like Perplexity and open-source search-augmented models that may show different results. Additionally, we treat Google AIO as a black box without access to its specific model version, retrieval pipeline, or post-processing steps, limiting mechanistic interpretation of our findings. Future work incorporating a broader set of systems would strengthen the generalizability of our findings.

Acknowledgments

We utilized Claude Opus 4.6 and GPT Codex 5.3 through Cursor and Claude Code to develop part of the codebase for our experiments.

Ethical Considerations

Our study analyzes system-generated outputs from publicly accessible search platforms and does not involve human subjects, personal data, or user interaction. All queries are drawn from the publicly available Natural Questions (NQ) corpus. We followed the terms of service and standard usage policies for both SerpAPI and OpenAI API.

We also want to highlight that we do not intend to single out any particular system with our findings, but aim to characterize broad patterns across the generative search landscape that have implications for information access. We believe that transparent, independent auditing of AI-mediated search systems serves the public interest.

References

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21.

Athena Chapekis and Anna Lieb. 2025. Google users are less likely to click on links when an ai summary appears in the results.

Chun Wei Choo. 2023. Climate change information seeking. *Journal of the Association for Information Science and Technology*, 74(9):1086–1099.

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, Bangkok, Thailand. Association for Computational Linguistics.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024a. 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*, pages 6437–6447.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024b. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9136–9152, Vienna, Austria. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Tuan-Luc Huynh, Thuy-Trang Vu, Weiqing Wang, Yinwei Wei, Trung Le, Dragan Gasevic, Yuan-Fang Li, and Thanh-Toan Do. 2025. Promptdsi: Prompt-based rehearsal-free continual learning for document retrieval. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 383–401. Springer.

Dominik Kowald. 2024. Transparency, privacy, and fairness in recommender systems. *ArXiv*, abs/2406.11323.

Monika Krakowska. 2020. Affective factors in human information behavior: a conceptual analysis of interdisciplinary research on information behavior. *Zagadnienia Informacji Naukowej–Studia Informacyjne*, 58(1A (115A)):75–95.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Alice Li and Luanne Sinnamon. 2024. Generative ai search engines as arbiters of public knowledge: An audit of bias and authority. *Proceedings of the Association for Information Science and Technology*, 61(1):205–217.

Haoyuan Li, Yusen Zhang, Rui Zhang, and Snigdha Chaturvedi. 2025. Coverage-based fairness in multi-document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9801–9819.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024a. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024b. A survey of generative search and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven Loria, pre-commit-ci[bot], dependabot[bot], dependabot-preview[bot], dependabot-support, Roman Yankovsky, Jeff Schnurr, pyup.io bot, Jamie Moschella, Emmy Zero, John Franey, Evan Dempsey, Jeff Kolb, Daniel Ong, Jhon Eslava, Romain Casati, Wesley Childs, Tim Gates, Robin Bürgi, and 11 others. 2026. sloria/TextBlob. https://github.com/sloria/TextBlob.

Olivia Pal, Veda Duddu, Agam Goyal, Drishti Goel, and Koustuv Saha. 2026. Do we know what they know we know? calibrating student trust in ai and human responses through mutual theory of mind. *arXiv preprint arXiv:2601.16960*.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4:61–74.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 129–140.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2025. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*.

Daniel Trielli and Nicholas Diakopoulos. 2019. Search as news curator: The role of google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pages 1–15.

Elisabeth Vogl, Reinhard Pekrun, Kou Murayama, and Kristina Loderer. 2020. Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion*, 20(4):625.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Fresh-LLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Alex Ward, Sean Hastings, LKirst, Shivam Bansal, Lukas Schoebel, Alonso Cisneros, Christian Clauss, Guillem García Subies, John Giorgi, massvoice, Caleb, MKaras, Hugo van Kemenade, Takeshi Kishiyama, Rinaldo Rex, Andrew McNutt, Aether, Gérard Laras, Lane Clark, and 11 others. 2026. textstat/textstat. https://github.com/textstat/textstat.

Shuguang Zhao and Yiming Liu. 2021. The more insufficient, the more avoidance? cognitive and affective factors that relates to information behaviours in acute risks. *Frontiers in Psychology*, 12:730068.

A Appendix

In this section we extend the description of the query dataset by providing simple descriptive statistics and highlighting key terms across topics.

The queries have a mean length of 9.32 ± 1.91 words. The majority of queries are phrased as questions (78%), with the remainder being declarative statements or keyword fragments. The five most frequent key terms across all queries are *world* (377), *India* (346), *name* (338), *time* (327), and *United States* (249). Table A1 presents the top key terms per topic, computed after lemmatization and aggregating counts across morphological variants.

Table A2 reports extended description of our findings from RQ1.

B Compute Resources

All experiments on open-source models were run on internal organization CPU and GPU servers equipped with 3xNVIDIA A40.

Topic	# Queries	Top 5 Key Terms
Business, Finance, & Economy	1,000	company (53), India (51), United States (48), world (43), bank (43)
Education & Literature	1,000	book (107), wrote (95), school (77), name (50), summary (40)
Entertainment	1,000	play (193), movie (116), season (110), episode (90), film (56)
Geography & Travel	1,000	locate (148), map (90), world (69), state (61), river (60)
History	1,000	war (92), name (58), India (58), world (38), United States (34)
Music	1,000	song (239), sing (216), write (88), love (72), lyrics (47)
Politics & News	1,000	state (147), India (90), court (68), president (65), united (61)
Religion	1,000	Bible (235), book (90), church (81), God (70), catholic (45)
Science	1,000	earth (33), locate (32), cell (24), function (23), name (23)
Sports	1,000	time (142), world cup (95), game (75), team (73), NBA (54)
Technology	1,000	use (111), difference (55), computer (53), Xbox (42), Windows (40)
All Queries	**11,000**	**world (377), India (346), name (338), time (327), United States (249)**

Table A1: **Query Dataset Summary.** Table representing all 11 topics and the number of queries per topic in our final dataset. We also report the top 5 most frequently appearing key terms per topic with frequencies in parentheses.

	Search GPT	Google AIO	Organic
Queries w/ $\geq$1 citation	97.0%	57.8%	99.8%
Mean citations/query	3.2	4.0	9.3
Median citations/query	3	4	10
Unique domains	7,606	14,279	26,891
Domain category (% of cited domains)			
Encyclopedia/Reference	27.3%	10.3%	11.0%
News/Media	7.2%	2.4%	1.7%
Gov./Institutional	7.0%	8.5%	7.9%
Social Media/Forums	0.1%	8.5%	13.4%
Video/Streaming	6.9%	3.7%	3.3%
Organization (.org)	8.4%	11.9%	10.9%
Other	43.0%	54.0%	51.0%

Table A2: **Source summary.** *Domain categories* show each category's share of all cited domains.

Topic	cogproc	tentat	posemo	formality	politeness
Technology	**1.37**	0.69	**1.30**	**1.19**	**0.84**
Travel/Geography	**0.88**	0.64	**1.01**	**1.76**	0.79
Sports	0.78	0.40	**0.85**	**0.95**	**0.94**
Science	0.75	0.43	**0.82**	**1.28**	0.38
Politics/News	0.56	0.37	0.76	**1.21**	0.02
Music	0.40	0.17	0.62	**1.26**	0.34
Religion	0.41	0.18	0.51	**1.00**	0.00
Entertainment	0.28	0.16	0.32	**0.82**	-0.06
Education/Lit.	0.11	-0.03	0.37	**0.86**	0.04
History	0.20	0.07	0.29	0.76	0.09
Business/Fin./Econ.	0.12	0.03	0.39	0.68	0.11
Overall	*0.48*	*0.27*	*0.61*	*0.99*	*0.30*

Table A3: Cohen's d for Search GPT vs. Google AIO by topic. Positive d indicates higher values in Search GPT. **Bold** := large effect ($|d| \geq 0.8$). Topics are sorted by mean $|d|$ across the five features.