# Examining query sentiment bias effects on search results in large language models

Alice Li[1], Luanne Sinnamon[1]

[1]School of Information, University of British Columbia, 1961 E Mall, Vancouver, BC, Canada, V6T 1Y3

## Abstract

Large Language Model (LLM)-based search systems that generate synthesized responses to queries have the potential to increase the efficiency of information retrieval compared to conventional web search systems. However, the benefits of efficiency must be considered against the loss of user agency and the increased potential for system and user bias in single response systems. As LLMs are trained on massive corpora of digital text and are largely opaque in their operations, LLM-based search systems may magnify bias impacts on users. Further, the relative novelty of LLM-based search systems and their rapid deployment to the general public creates a need to better understand their potential impact on human information seeking and learning. We conducted an algorithm audit to explore the nature and sentiment of responses from 3 LLM-based search systems (Bing Chat, ChatGPT, Perplexity) to a range of queries representing 4 socio-political topics (climate change, vaccination, alternative energy, trust in media) over the course of 7 days. Each topic was audited using 12 frequently searched queries ($N$ = 1,008). Initial findings indicate that the sentiment of query is positively correlated with the sentiment of LLM result in all three LLM-based search systems, and that sentiment of responses on these topics differs significantly across the three systems.

## Keywords

Large language models, Search engines, Audit, Sentiment bias, Confirmation bias

## 1. Introduction

Large language models (LLMs) can perform a wide variety of tasks, such as summarizing given content, organizing to-do lists, and finding and synthesizing online resources for users [1]. Open AI's ChatGPT reached 100 million users in two months after its public launch in December 2022, and was visited 590 million times in the single month of January 2023 [2]. Within the past year, LLMs have been rapidly deployed to power web search engines (SEs), despite calls for further testing and refinements of these complex technologies [3, 4]. Researchers have raised numerous concerns with LLM-based SEs, which can be categorized as intrinsic or usage-related [5]. Intrinsic issues are inherent to the systems [5], including 1) hallucination, where the generated content cannot be found in the cited content source or is summarized incorrectly [6, 7, 8], and 2) lack of currency [9]. Usage-related issues arise during user interaction with LLM-based SEs [5], and include 1) copyright infringement due to use of sources without permission or proper citation [5], 2) unethical usage, where authorship is falsely claimed by the tool [5],

*Symposium on Future Directions in Information Access (FDIA) co-located with The 2023 European Summer School on Information Retrieval (ESSIR), August 28–September 2, 2023, TU Vienna, Austria*

alicelii@mail.ubc.ca (A. Li); luanne.sinnamon@ubc.ca (L. Sinnamon)
0000-0002-3174-557X (A. Li); 0000-0001-9063-5699 (L. Sinnamon)

CEUR Workshop Proceedings (CEUR-WS.org)

and 3) over-reliance, where people depend on the generated content without critical assessment [5, 10]. Given the importance of SEs as a means of accessing information in the course of everyday life and their pivotal role in shaping public knowledge [11], the shift to LLM-based SEs requires scrutiny.

An area of potential usage-related concern is the amplification of confirmation bias, which is a human tendency to prefer and place more trust in evidence that supports existing expectations or beliefs [12]. As LLMs may generate text that reflects back users' own beliefs [13], confirmation bias could be reinforced by results designed to closely match the query features (e.g., sentiment). SEs that are heavily personalized or trained to mimic their users may reinforce this bias, thereby limiting the range and diversity of information encountered and reinforce social and political polarization. When people seek information, they are often motivated by learning, problem-solving, and decision-making tasks, in which case their goals are to make sense of the information, learn from it, and, in some cases, act upon it [14, 15]. LLM-based search system responses allow searchers to bypass several steps in the traditional search process, including selection and evaluation of listed resources, information extraction, and synthesis. As such, these systems bear a greater burden of responsibility for the impact of those (potentially biased) results on the searcher's knowledge, attitudes, and actions than traditional SEs. In particular, there is a need to understand the processes by which LLM-based search systems generate results and the impact of result features on users.

The primary aim of this study is to assess the extent to which these systems have the potential to reinforce confirmation bias, by investigating the sentiment (positive, negative, neutral) of LLM-based search system responses in relation to the sentiment of the query. Study data will also provide a comparative baseline of response features across several LLM-based SEs. The findings of this research may benefit SE designers seeking to improve the LLM and user interaction models, policymakers developing regulatory policies, and researchers investigating the implications of LLM-based SEs on human information interaction.

## 2. Literature Review

We first review research on cognitive biases in SE use as a basis for understanding the role of such biases in LLM-based SEs. Next, we summarize research on sentiment and bias in search results and user interactions. We provide an overview of algorithm auditing methods applied to SEs and present a rationale for the research method. We conclude by presenting the research questions guiding this study.

### 2.1. Cognitive bias in searching

Bias is any type of skew or preference that can cause harm or other undesired consequences [16]. Many sources and types of bias influence SEs, including those that arise from data inputs and user interactions [16, 17]. Prior research has identified SE bias that discriminates on the basis of race and gender, for example [18, 19, 20]. In this study, we focus on the interplay between presentation bias arising from design decisions regarding outputs, and interaction bias, arising from users' cognitive biases and perceptions of SEs [17]. Cognitive biases are inherent human prejudices and preconceptions that can influence perceptions and actions [21].

In the context of search, cognitive biases influence the selection and interpretation of results. Azzopardi [22] presented the major cognitive biases found in studies of information seeking and retrieval according to [23]'s four categories: 1) too much information, 2) not enough meaning, 3) needing to act fast, and 4) what to remember. Too much information suggests that people are overloaded with information [23], which can lead them to accept a particular viewpoint because it is easily accessible (availability bias) or to prefer results that are consistent with their beliefs (confirmation bias) [22]. Not enough meaning indicates people update their mental model based on the information they receive [23]. Thus, people who experience the same stimulus over and over again develop positive viewpoints toward it (reinforcement effects) [22]. Needing to act fast indicates people make decisions under time pressure [23], for example, leading people to choose trusted sources with reliable information rather than unknown sources (ambiguity effects) [22]. What to remember includes biases towards filtering out specifics [23], and priming effects [22]. While searchers have the opportunity to analyze the variety of search results and snippets in conventional SE results pages, results are synthesized for them in LLMs. Notwithstanding the claims of efficiency [24], these LLM-based responses may have unintended consequences, such as feeding into searchers' cognitive biases. This study focused on confirmation bias, where alignment in viewpoint (e.g., sentiment) between queries and results may limit the exposure of searchers to diverse perspectives and feed into an "echo chamber" effect.

## 2.2. Sentiment of search engine search results and user interactions

Studies investigating the sentiment of search results have found variation in sentiment by topic and impacts of sentiment on user behaviour. Health-related queries were found to retrieve more positive and confirmatory top search results [25, 26]; whereas, controversial queries (e.g., genetic cloning) retrieved results with more negative-oriented emotions (e.g., anger, fear) [27]. These findings raise questions regarding the extent to which search results may potentially reaffirm users' preexisting beliefs [26]. Moreover, individuals demonstrate a preference for selecting positive search results for both controversial [27] and medical topics [26]. A known bias in LLMs is sentiment bias, in which the emotion of generated texts is influenced by attributes of the prompt and/or by social biases embedded in the training data [28, 29]. As an example of sentiment bias, GPT-2 was found to produce more positive emotion in relation to the profession of "baker" than for "accountant" [28]. With the advancement of the GPT models (3.5 and 4 versions) and their deployment in SEs, audits are needed to examine the extent to which sentiment bias exists in LLM-based search results, and the potential for user impacts, including confirmation bias.

## 2.3. Algorithm audits

In response to concerns of algorithmic opaqueness and discrimination, Sandvig et al. [30] proposed algorithm audits, a method derived from traditional auditing practices in the social sciences. An algorithm audit systematically and repeatedly queries an algorithm, then observes the results to draw inferences about how the algorithm works [31]. Both traditional audits and algorithm audits are used to examine discrimination against people and groups and systemic bias, but algorithm audits focus on automated, rather than human, systems. [30] proposed five algorithm audit research designs, which can be seen in SE bias audit studies: 1) code audit, 2) non-invasive user audit, 3) scraping audit [32, 33], 4) sock puppet audit [34], and 5) crowdsourced audit [35].

The code audit directly examines the source code of the algorithm to increase transparency [30]. However, it is difficult to perform on SEs due to their black-box nature. The non-invasive user audit asks participants about their experience with SEs through questionnaires, but few studies have used this approach since it cannot establish causality in a non-experimental setting [30]. The scraping audit involves performing repeated queries to observe search results [30, 32]. [32] audited Google and Yandex results for disease and remedy queries, and found that between 32% to 44% of snippets confirmed misinformation. The sock puppet audit involves creating personas to test the system [30, 34]. [34] examined whether personalization (i.e., age, gender, geolocation, watch history) amplified misinformation for YouTube accounts, and found that watching videos containing misinformation increased misinformative video recommendations. Crowdsourced audits are similar to sock puppet audits but employ real human participants [30, 35, 36]. [35] used a crowdsourcing platform to audit SE results for queries about COVID-19 public health beliefs that varied in sentiment. The study found that search results varied by both sentiment (positive, negative) and query location. Drawing upon these examples, the current study utilized the best practices of algorithm audits as applied to SEs [31], adjusting for minor differences between conventional and LLM-based SEs.

### 2.4. The current study

Two research questions (RQs) guided this study. For topics of societal importance and public debate: **RQ1**: What is the relationship between the sentiment of a query (positive, negative, neutral) and the sentiment of the LLM result? **RQ2**: Is there variation in the sentiment of results across LLM-based search systems (ChatGPT, Bing Chat, Perplexity)?

## 3. Method

This study employed an algorithm audit approach to answer the research questions [31]. The methods are presented in the following sections: 3.1) study design, 3.2) study procedure, and 3.3) data analysis preparation.

### 3.1. Study Design

An audit study has two axes: the topic and the type of discrepancy [31]. We selected four topics identified as current global challenges [37, 38]: climate change, alternative energy, vaccination, and trust in the media. Views on these topics among the general population are expected to vary substantially. The type of discrepancy examined in this study is sentiment bias. Specifically, we examined whether queries expressing a particular sentiment (positive, negative, neutral) prompt results that are biased in relation to the sentiment of the query, as well as general variations in sentiment across topics and LLM-based SEs.

**Table 1**
Example queries and polarities for each topic

| Topic | Query * | Polarity ** |
|---|---|---|
| Climate Change | climate change benefit | 0.76 |
| | will climate change kill us | -0.57 |
| | will climate change ever stop | -0.06 |
| Vaccination | vaccination benefits | 0.58 |
| | why vaccine mandates are unethical | -0.61 |
| | vaccination in pregnancy | -0.08 |
| Alternative Energy | renewable energy is important | 0.84 |
| | why alternative energy is bad | -0.75 |
| | alternative energy advantages and disadvantages | 0.04 |
| Trust in Media | trust in media by country | 0.37 |
| | distrust in media | -0.72 |
| | trust in media all time low | -0.10 |

Note. * Queries were drawn from popular searches on Google via AnswerThePublic service [39]
** Polarity was calculated using a fine-tuned transformer model for sentiment analysis [40]

### 3.1.1. Search queries

Starting with the four identified topics, we gathered queries from AnswerThePublic [39], a fee-based service that scrapes Google autosuggestion data from popular searches. We targeted English queries from Canada; queries were collected in May 2023. Twelve of the most frequent search queries associated with each topic were selected with the goal of achieving substantial query sentiment variation within each set. The number of queries used is in alignment with other web SE algorithm audits [34, 35]. Some example queries and their sentiment polarity scores are presented in Table 1.

## 3.2. Systems

Free LLM-based SEs available to the public were chosen since the target audience of our audit is the general information-seeking public. We examined three freely available LLM-based SEs: ChatGPT [41] and Perplexity [42] powered by GPT-3.5, and Bing Chat powered by GPT-4[43]. We chose these GPT-powered systems to examine whether they have embedded different rules or logics in relation to the design of responses. Google Bard was not included in this study since it was not yet available in Canada as of May 2023.

### 3.2.1. Legal and ethical considerations

The terms of use of the automated systems were checked before the study began to see if the company permits algorithm audits [31]. While Bing Chat does not state information regarding data scraping [44], ChatGPT and Perplexity do not allow automatic data scraping [45, 46]. Thus, our study chose to audit the systems through the manual input of various search queries by the researchers. As the system responses could not be traced to specific human participants and all queries were submitted by the research team, this study was not considered human subjects research and behavioural research ethics approval has not been sought.

## 3.3. Study Procedure

Public computers (macOS and Microsoft Windows operating systems) and dummy accounts were used to conduct this study. The full set of search queries were submitted to the three LLM-based SEs (Bing Chat, Perplexity, and ChatGPT) daily, over the course of a week in June 2023, instead of a single trial. This was done to anticipate dynamic changes in the algorithms and outputs, as the purpose of this audit was not to target specific phenomena at a certain time [31]. The default options were selected for each system to mimic users who casually find answers through the systems. Each query was cleared after each interaction to reduce the impact of personalization of search results.

## 3.4. Data Preparation and Analysis

A total of 1,008 LLM results were collected in this study. The Python programming language (version 3.10) was used to pre-process the collected data and to conduct the sentiment analysis. NumPy (version 1.25.0; [47]) and Pandas (version 1.5.3; [48]) packages, and regular expression (version 3.10.12; [49]) were used to remove web source numbering for Bing Chat and Perplexity results, since inclusion of the web sources would have affected the word count and sentiment analysis.

Sentiment polarity scores for each result were computed using a fine-tuned transformer model [40]. A transformer model interprets the meaning of different words in a sentence [50], and has been extended and refined for sentiment analysis. The sentiment model provided confidence scores for positive, neutral, and negative sentiments. The overall sentiment polarity score was calculated as the difference between the positive and negative scores [51].

Among the three systems assessed, Bing Chat exhibited duplicate results for seven queries (i.e., identical content in the full paragraph output). These duplicates were retained in subsequent analyses due to their distinct collection timestamps and unique dummy accounts, indicating possible user exposure to the same outcomes.

### 3.4.1. Descriptive summaries of LLM-based search system results

The average length of results across all topics and systems is 234 words (Table 2). Results indicate little difference in response length across topics, but a notable difference across systems. Bing Chat produced the shortest responses overall, consisting of 35 words, and the lowest average word count ($M$ = 123). Perplexity results were in the mid-range ($M$ = 243), and ChatGPT results had the highest mean word count ($M$ = 336) and the longest overall responses, consisting of 635 words.

Bing Chat and Perplexity include source links in their results (Table 3). There is a high level of consistency in the number of sources provided by both systems, at $M$ = 5.13.

**Table 2**
Number of words for each topic and LLM-based search system

| | BingChat | | | ChatGPT | | | Perplexity | | | Total- All 3 systems | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Topic** | *M*(*SD*) | Min | Max | *M*(*SD*) | Min | Max | *M*(*SD*) | Min | Max | *M* | Min | Max |
| **Climate Change** | 139(50.56) | 56 | 291 | 347(85.25) | 189 | 500 | 283(95.59) | 156 | 597 | 256 | 56 | 597 |
| **Vaccination** | 117(42.06) | 35 | 224 | 297(115.36) | 67 | 493 | 225(71.12) | 88 | 428 | 213 | 35 | 493 |
| **Alternative Energy** | 117(34.76) | 49 | 227 | 346(89.61) | 114 | 528 | 214(50.15) | 90 | 426 | 226 | 49 | 528 |
| **Trust in Media** | 121(34.29) | 62 | 260 | 353(81.91) | 91 | 635 | 250(58.19) | 149 | 505 | 241 | 62 | 635 |
| **Total- All 4 topics** | 123(41.77) | 35 | 291 | 336(96.22) | 67 | 635 | 243(78.33) | 90 | 597 | 234 | 35 | 635 |

Note. Each topic for each system (*n* = 84)

**Table 3**
Number of source links for each LLM-based search system

| System | *n* | Min | Max | *M*(*SD*) |
|---|---|---|---|---|
| **Bing Chat** | 336 | 1 | 7 | 5.13(.05) |
| **ChatGPT** | 336 | 0 | 0 | 0(0) |
| **Perplexity** | 336 | 2 | 6 | 5.13(.05) |

# 4. Results

## 4.1. RQ1: What is the relationship between the sentiment of a query (positive, negative, neutral) and the sentiment of the LLM result?

Preliminary analyses ($N$ = 1,008) showed the relationship between query sentiment and LLM result sentiment to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test ($p$ > .05), and there were no outliers. Thus, a parametric Pearson's product-moment correlation was conducted to assess the relationship. There was a statistically significant, moderate positive correlation between query sentiment and LLM result sentiment, $r(1, 006) = .46, p < .001$, with query sentiment explaining 21% of the variation in LLM result sentiment.

In further analysis, Pearson's correlations were statistically significant between query polarity and result polarity for all three LLM-based SEs: Bing Chat ($r$ = .454, $p$ <.001), ChatGPT ($r$ = .435, $p$ <.001), and Perplexity ($r$ = .509, $p$ <.001). Namely, query sentiments are positively correlated with LLM result sentiments at a similar level in all three LLM-based SEs (Bing Chat, ChatGPT, and Perplexity).

## 4.2. RQ2: Is there variation in the sentiment of results across LLM-based search systems (ChatGPT, Bing Chat, Perplexity)?

Shapiro-Wilk's test demonstrated non-normal distribution of LLM result polarity scores for all three LLM systems ($p$ < .001). Thus, a non-parametric Kruskal-Wallis H test was used to test for differences in LLM result polarity between three groups of LLM systems: Bing Chat ($n$ = 336), ChatGPT ($n$ = 336), and Perplexity ($n$ = 336). Distributions of polarity scores were similar for all groups when analyzed using a boxplot. Median LLM result polarities were statistically significantly different between groups, $\chi^2(3) = 12.54, p = .002$. Pairwise comparisons were conducted using Dunn's procedure [52] with a Bonferroni correction for multiple comparisons. Adjusted p-values are presented in the following. The post hoc analysis revealed statistically significant differences in LLM result polarity between Perplexity (*Mdn* = -.19) and Bing Chat (*Mdn* = -.10) (*p* = .05), and between Perplexity and ChatGPT (*Mdn* = .01) (*p* = .002), but not between Bing Chat and ChatGPT. Thus, Perplexity generates results with significantly more negative sentiment compared to Bing Chat and ChatGPT for these topics.

## 5. Discussion and Implications

This audit study examined the output sentiment of LLM systems on four divisive socio-political issues (climate change, vaccination, alternative energy, and trust in media). For **RQ1**, findings reveal evidence of sentiment bias, in the form of a correlation between the sentiment polarities of queries and LLM results. This discovery aligns with previous studies that found sentiment bias in LLMs [28, 29], and highlights the need to manage this issue to avoid "echo chamber" effects in search. A challenge in this area is the lack of a baseline or gold standard for how LLM results should be constructed in relation to user queries. For **RQ2**, the data indicate that in addressing these four socio-political topics (climate change, vaccination, alternative energy, and trust in media), ChatGPT and Bing Chat responses tend towards neutral sentiment. Perplexity results are significantly more negative in sentiment than the other two systems, which suggests that it may be more impacted by sentiment bias arising from social bias in the data. As ChatGPT and Perplexity are both powered by GPT-3.5, this finding suggests that sentiment bias can arise from different implementations of the GPT models.

This study is subject to certain limitations. Notably, negative polarity output does not indicate a negative or critical stance towards the topic. For instance, the negative sentiment output for climate change queries likely reflects society-wide concerns regarding the impacts of the climate crisis, rather than disbelief in the science or critiques of climate action. To address these complexities, we intend to conduct a deeper qualitative analysis of the study data to understand the relationship between LLM-based SE queries and outputs, including aspects such as stance, rhetorical style and use of evidence, which will offer more comprehensive insights into these initial findings. Furthermore, we are planning a user study using materials from this audit, to better understand the relationship between sentiment bias and confirmation bias.

**Implications.** Theoretically, this study investigated the extent to which LLM-based SE responses may support confirmation bias through one form of sentiment bias: mimicry between queries and responses. Further, it explored topical sentiment bias in LLM system responses more generally. An important implication is the need for the information retrieval research community to develop standards and metrics for the evaluation of these new systems. Methodologically, this study applied the best practices of SE algorithm audits to the new generation of SEs, the operations of which are almost entirely opaque to system users and other stakeholders. A great deal more work is needed in this area.

**Acknowledgments.** We thank Shawn Ching-Chung Hsueh for his advising during the data collection and analysis stages. This work is supported in part by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC) CGS-D program, the Natural Sciences and Engineering Research Council of Canada (NSERC) CREATE program, and the Anne and George Piternick Student Research Award from the School of Information, University of British Columbia.

## References

[1] OpenAI, Examples, n.d. URL: https://platform.openai.com/examples.

[2] K. Hu, Chatgpt sets record for fastest-growing user base - analyst note, 2023. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[3] F. of Life Institue, Pause giant ai experiments: An open letter, 2023. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[4] R. Zhao, X. Li, Y. K. Chia, B. Ding, L. Bing, Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines, 2023. arXiv:2304.11076.

[5] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, W. Mansoor, The future of gpt: A taxonomy of existing chatgpt research, current challenges, and possible future directions, SSRN Electronic Journal (2023). doi:10.2139/ssrn.4413921.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. arXiv:2303.12712.

[7] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, p. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[9] N. Staudacher, What is chatgpt?, 2023. URL: https://help.openai.com/en/articles/6783457-what-is-chatgpt.

[10] C. Shah, E. M. Bender, Situating search, in: ACM SIGIR Conference on Human Information Interaction and Retrieval, ACM, Regensburg Germany, 2022, p. 221–232. doi:10.1145/3498366.3505816.

[11] J. Haider, O. Sundin, Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life, 1 ed., Routledge, London, 2019. doi:10.4324/9780429448546.

[12] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, Review of General Psychology 2 (1998) 175–220. doi:10.1037/1089-2680.2.2.175.

[13] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, 2021, p. 610–623. doi:10.1145/3442188.3445922.

[14] G. Marchionini, Information seeking in electronic environments, Cambridge series on human-computer interaction, Cambridge University Press, Cambridge ; New York, 1995.

[15] L. Freund, J. Berzowska, The goldilocks effect: Task-centred assessments of e-government information: The goldilocks effect: Task-centred assessments of e-government information, Proceedings of the American Society for Information Science and Technology 47 (2010) 1–10. doi:10.1002/meet.14504701261.

[16] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Transactions on Information Systems 14 (1996) 330–347. doi:10.1145/230538.230561.

[17] R. Baeza-Yates, Bias on the web, Communications of the ACM 61 (2018) 54–61. doi:10.1145/3209581.

[18] D. Metaxa, M. A. Gan, S. Goh, J. Hancock, J. A. Landay, An image of society: Gender and racial representation and impact in image search results for occupations, Proceedings of the ACM on Human-Computer Interaction 5 (2021). doi:10.1145/3449100.

[19] M. Kay, C. Matuszek, S. A. Munson, Unequal representation and gender stereotypes in image search results for occupations, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 3819–3828. doi:10.1145/2702123.2702520.

[20] S. U. Noble, Algorithms of oppression, New York University Press, 2018. doi:10.18574/nyu/9781479833641.001.0001.

[21] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (1974) 1124–1131. doi:10.1126/science.185.4157.1124.

[22] L. Azzopardi, Cognitive biases in search: A review and reflection of cognitive biases in information retrieval, in: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, ACM, Canberra ACT Australia, 2021, p. 27–37. doi:10.1145/3406522.3446023.

[23] B. Benson, Cognitive bias cheat sheet, 2016. URL: https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18.

[24] S. Pichai, An important next step on our ai journey, 2023. URL: https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/.

[25] G. Demartini, S. Siersdorfer, Dear search engine: What's your opinion about...? sentiment analysis for semantic enrichment of web search results, in: Proceedings of the 3rd International Semantic Search Workshop, SEMSEARCH '10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1–7. doi:10.1145/1863879.1863883.

[26] R. White, Beliefs and biases in web search, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 3–12. doi:10.1145/2484028.2484053.

[27] G. Kazai, P. Thomas, N. Craswell, The emotion profile of web search, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1097–1100. doi:10.1145/3331184.3331314.

[28] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, P. Kohli, Reducing sentiment bias in language models via counterfactual evaluation, 2020. arXiv:1911.03064.

[29] J. Tian, S. Chen, X. Zhang, X. Wang, Z. Feng, Reducing sentiment bias in pre-trained sentiment classification via adaptive gumbel attack, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 13646–13654. doi:10.1609/aaai.v37i11.26599.

[30] C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, Auditing algorithms: Research methods for detecting discrimination on internet platforms, Data and discrimination: converting critical concerns into productive inquiry 22 (2014) 4349–4357. URL: https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

[31] D. Metaxa, J. S. Park, R. E. Robertson, K. Karahalios, C. Wilson, J. Hancock, C. Sandvig, Auditing algorithms: Understanding algorithmic systems from the outside in, Foundations and Trends® in Human–Computer Interaction 14 (2021) 272–344. doi:10.1561/1100000083.

[32] A. Bondarenko, E. Shirshakova, M. Driker, M. Hagen, P. Braslavski, Misbeliefs and biases in health-related searches, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ACM, Virtual Event Queensland Australia, 2021, p. 2894–2899. doi:10.1145/3459637.3482141.

[33] R. W. White, A. Hassan, Content bias in online health search, ACM Transactions on the Web 8 (2014) 1–33. doi:10.1145/2663355.

[34] E. Hussein, P. Juneja, T. Mitra, Measuring misinformation in video search platforms: An audit study on youtube, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–27. doi:10.1145/3392854.

[35] B. Le, D. Spina, F. Scholer, H. Chia, A crowdsourcing methodology to measure algorithmic bias in black-box systems: A case study with covid-related searches, in: L. Boratto, S. Faralli, M. Marras, G. Stilo (Eds.), Advances in Bias and Fairness in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 43–55. doi:10.1007/978-3-031-09316-6_5.

[36] R. E. Robertson, S. Jiang, K. Joseph, L. Friedland, D. Lazer, C. Wilson, Auditing partisan audience bias within google search, Proceedings of the ACM on Human-Computer Interaction 2 (2018). doi:10.1145/3274417.

[37] Edelman, 2023 edelman trust barometer: Navigation a polarized world, 2023. URL: https://www.edelman.com/trust/2023/trust-barometer.

[38] G. of Canada, The next generation of emerging global challenges, 2018, October 19. URL: https://horizons.gc.ca/en/2018/10/19/the-next-generation-of-emerging-global-challenges/.

[39] N. Digital, Discover what people are asking about..., n.d. URL: https://answerthepublic.com.

[40] L. X. Yuan, distilbert-base-multilingual-cased-sentiments-student, 2023. URL: https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student.

[41] OpenAI, Welcome to chatgpt, n.d. URL: https://chat.openai.com.

[42] Perplexity, Ask anything..., n.d. URL: https://www.perplexity.ai.

[43] M. Bing, Ask me anything..., n.d. URL: https://www.bing.com/.

[44] Bing, Bing conversational experiences and image creator terms, 2023. URL: https://www.bing.com/new/termsofuse?FORM=GENTOS.

[45] OpenAI, Terms of use, 2023. URL: https://openai.com/policies/terms-of-use.

[46] P. AI, Terms of service, 2023, March 14. URL: https://www.perplexity.ai/tos.

[47] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362. doi:10.1038/s41586-020-2649-2.

[48] W. McKinney, Data structures for statistical computing in python, in: S. van der Walt, J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (SCIPY 2010), volume 445, 2010, pp. 56–61. doi:10.25080/Majora-92bf1922-00a.

[49] P. S. Foundation, re — regular expression operations, release 3.10.12, 2023. URL: https://docs.python.org/3.10/library/re.html.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[51] Y. Zhu, A. G. Hoepner, T. K. Moore, A. Urquhart, Sentiment analysis methods: Survey and evaluation, Available at SSRN 4191581 (2022). doi:10.2139/ssrn.4191581.

[52] O. J. Dunn, Multiple comparisons using rank sums, Technometrics 6 (1964) 241–252. doi:10.1080/00401706.1964.10490181.