Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking

NIKHIL SHARMA, Johns Hopkins University, USA
Q. VERA LIAO, Microsoft Research, Canada
ZIANG XIAO, Johns Hopkins University, USA

Large language models (LLMs) powered search systems have already been used by hundreds of millions of people, and are believed to bring many benefits over traditional search. However, while decades of research and public discourse interrogated the risk of search in increasing selective exposure and creating echo chambers—limiting exposure to diverging opinions and leading to opinion polarization, little is known about such risk of LLM-powered search. We conduct two experiments to investigate: 1) whether and how LLM-powered conversational search increases selective exposure compared to traditional search; 2) whether and how LLMs with opinon biases that either reinforce or challenge the user's view change the effect. Overall, we found that participants engaged in more biased information querying with LLM-powered search, and an opinionated LLM reinforcing their views exacerbated this bias. These results present critical implications for the development of LLMs and conversational search systems, and the policy governing these technologies.

CCS Concepts: • **Human-centered computing → Collaborative interaction**; **Empirical studies in HCI**; **User studies**; *Web-based interaction*; • **Information systems → Collaborative search**; **Search interfaces**; *Web searching and information discovery*.

Additional Key Words and Phrases: generative search, information retrieval, user behaviour studies

ACM Reference Format:
Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2023. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. In . ACM, New York, NY, USA, 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Exposure to diverse viewpoints is essential for critical thinking, balanced views, and informed decision-making, and at a collective level, preventing opinion polarization or even dangerous radicalization. However, such ideals are often challenging to achieve because people have a natural tendency of selective exposure [15], or confirmation bias [59], favoring consonant information and avoiding dissonant information. In the last two decades, much research and public discourse have expressed concerns regarding the exacerbating effect of information and web technologies on selective exposure. For example, by personalization and supplying only information people want to see, search engines and recommender systems such as news feeds may produce "filter bubbles" [42] of ideological and cultural isolation. By allowing people to easily select whom to connect with and what they want to see, social media can create "echo chambers" where people end up only interacting with like-minded others. In short, information technologies can have multiplex mechanisms to exacerbate people's selective exposure bias, from data and algorithmic biases, to biases induced by interaction affordances [5].

Manuscript submitted to ACM

With the recent rise of powerful large language models (LLMs) such as the GPT-4, a new generation of LLM-powered information technologies has emerged, from conversational search, open-domain or specialized chatbots, to various productivity tools such as writing support. These technologies can have profound effects on the information consumption of individuals and the society at large. First, LLMs are in essence "next token predictors" that optimize for giving expected and "likable" outputs, and thus can potentially be more biased towards providing consonant information than traditional information system algorithms. However, LLMs can be used to provide "synthesized" content (e.g., a concise summary) based on a collection of documents, which may help increase people's exposure to diverse information by removing biases that people may exhibit when selecting which document to read. Information technologies powered by LLMs also support natural and interactive conversational interactions. How these new affordances of interaction shape people's information-seeking behaviors and information consumption remains an open question. Last but not least, LLMs are known to encode biases from the training data [1, 48], and can be easily steered to exhibit certain opinion biases through model adaptation techniques such as fine-tuning or prompting [6]. Little is known whether the encoded opinion biases can exacerbate selective exposure of people with similar views, or be used to bring benefits to those with different viewpoints. Given the rapidly growing reach and usage frequency of LLM-powered information technologies, it is paramount for the research community to investigate these issues and LLMs' effect on information diversity to inform the development of LLMs, design of LLM-powered systems, as well as policy governing these technologies.

In this work, we take a formative step toward understanding the echo chamber effect of LLMs by focusing on LLM-powered conversational search systems. Since the public release of Microsoft Bing Chat and Google Bard in 2023, LLM-powered conversational search systems have already reached hundreds of millions of users in just a few months. While conversational search is believed to bring many benefits [47, 63] such as ease of interaction with natural language, support for complex queries through multi-turn interactions, and overall user engagement, little is known at present about how people actually interact with LLM-power search systems, let alone their drawbacks and potential harms. We conducted a critical investigation into LLM-powered search systems through two experiments. In the first experiment (N = 115), using information-seeking tasks on controversial topics, we compared people's information-seeking behaviors and their attitude change outcomes when using a traditional search system versus LLM-powered conversational search systems (versions with and without references to the information sources). In the second experiment (N = 213), we explored whether and how conversational search systems using LLMs with manipulated opinion biases that either reinforce or challenge the user's existing attitude on the given controversial topic change their selective exposure. In short, we ask the following research questions:

- How does interacting with an LLM-powered conversational search system affect selective exposure and opinion polarization compared to a web search system?
- How does an LLM-powered conversational search system that exhibits opinion bias either consonant or dissonant with the user's existing attitude affect people's selective exposure and opinion polarization?

Below, we first review related work that shaped our study, then present the methods and results of the two experiments. For each study, we will list hypotheses after introducing the measurements in the method section.

2 RELATED WORK

2.1 Selective Exposure, Confirmation Bias, and Echo Chamber Effect

Psychologists have extensively studied people's selective exposure bias [15, 21]—systematic preference towards information that is consonant with one's existing view over dissonant information, and the related concept of confirmation bias [59]—actively seeking or assigning more weights to consonant information. Both biases can be attributed to a fundamental desire to avoid or reduce cognitive dissonance [14]. Selective exposure and confirmation bias have been found to lead to opinion bias and polarization as well as suboptimal decision-making in many settings such as health, politics, and scientific research [21, 39]. Collectively, these biases can lead to an information environment or segregated group communication where only information of a certain belief or ideology is shared—this is often referred to as "echo chamber effect" in social and political sciences [7, 54].

The HCI and broader research and activist communities have had long-standing concerns over the negative effect of information and web technologies on the diversity of information that people consume. In particular, by coining the term "filter bubble", Eli Pariser [42] raised much public attention in the 2010s on the potentials of personalization and algorithmic filtering used by search engines, recommendation systems, and social networking platforms in reducing people's exposure to conflicting viewpoints and creating ideological and social fragmentation. Other researchers were concerned about the affordances of technologies for people to selectively curate their own information environment, such as by following only or mostly like-minded others on social media platforms [2, 8] or limiting the diversity of the sources for one's news feed [17]. However, others challenged these concerns, suggesting that the actual selective exposure is less prevalent than theorized [20], and that there are individuals who actively seek diverse perspectives [31, 36] and the high-choice environment made possible by information and web technologies can facilitate diversity-seeking [11].

Researchers have also explored various approaches to combat selective exposure and increase information diversity. In information retrieval and recommender systems, serendipity is studied as an optimization criterion to increase the exposure to novel and diverse information [50, 55]. Many systems were developed to help people encounter diverse perspective [13], learn about and deliberate on controversial topics [12, 28, 44], be aware of one's own information bubble and better control filtering mechanisms [16, 25, 37]. As Garrett and Resnick [18] argued, to increase people's consumption of attitude-challenging information, the key lies in presenting high-quality challenging items in the right context, and/or reducing people's cognitive dissonance. To this end, HCI researchers conducted experiments to study the effects of diversity-enhancing designs such as highlighting or presenting agreeable information first to reduce cognitive dissonance [36], and highlighting the expertise [33], focused aspect [34], or the common ground [32] of challenging information. Overall, these designs are shown to have a positive diversity-enhancing effect but often only for a sub-group of people who have a predisposition to be open to diverse views, highlighting the challenge in combating selective exposure.

Building on these prior works, our research aims to investigate whether and how LLM-powered search systems can exacerbate people's selective exposure and reduce information diversity. Our experimental design was informed by previous HCI research conducting laboratory studies on selective exposure [16, 31–33, 36], adopting an information-seeking task of writing an essay for a controversial topic, and measurements of biases in information seeking behaviors and post-task attitude changes.

2.2 Human-LM Interaction

HCI researchers have started exploring developing applications of generative language models (LMs) and studying human interactions with them before this wave of widely adopted LLMs. Popular applications of language models include code generation as programming assistance [51, 53, 61], various forms of writing assistance such as next sentence generation [23, 29], summarization of previous content [9], rewriting [64], and metaphor generation [19], as well as chatbots [40, 62] and social agents [43]. Research on LLM-powered search is only recently emerging. For example, Liu et al. [35] conducted a human evaluation to audit popular LLM-powered search systems including Bing

Chat, and found that while the responses are fluent and appear informative, they frequently contain unsupported statements and inaccurate references (i.e. URLs to original sources).

Our study is particularly informed by works that are concerned with the negative effects of language models on people's information consumption and production. In the context of co-writing with LMs, common concerns include over-reliance on AI and automation-induced complacency [41] that can lead to not only sub-optimal writing outcomes [3] but also loss of human agency and perceived ownership of the created content [10]. While over-reliance on AI's suggestions has been studied as a common issue in human-AI interaction, LMs can exert additional informational influence by exposing people to some views more often than others, or making it easier to express some views over others. Jakesch et al. [22] refer to this effect as "latent persuasion" by language models. Their experiment demonstrated that when writing with an LLM-powered writing assistant that was configured to have a certain bias on the given topic, not only did participants' writing exhibit more of the model's bias, but also their own opinion shifted towards that direction afterwards. These negative effects of LMs can be exploited by malicious parties to influence public opinion or spread misinformation [26, 67].

Our work contributes to the literature on human-LM interaction with insights about a new and popular application domain—conversational search, and explores whether and how a human bias—selective exposure—interacts with the properties and affordances of LLMs to impact people's information seeking and consumption.

2.3 Conversational Search

While LLM-powered search systems are a recent phenomenon, years of research has pursued the idea of "conversational search", which allows information retrieval through natural and flexible conversations [24, 47, 63, 65]. In a foundational paper, Radlinski and Craswell [47] laid out the desirable properties of conversational search, including supporting users to express complex information needs, revealing system capabilities through multi-turn interactions, supporting mixed-initiative interactions, and so on. The authors also argue that conversational search is especially suitable for complex information tasks, such as when searching for a set of items or referencing a set of criteria or relevant items. Others similarly argued that conversational search could better engage users to interact for multiple rounds and respond to system questions to form more complex and/or refined queries [27, 66]. However, the implementation of conversational search, especially in an open-domain context, still faced technical challenges before LLMs emerged. Empirical studies of human interaction with a conversational search system were relatively limited [4, 56, 58], and often relied on wizard-of-oz approaches. For example, using a human intermediary, Trippas et al. [56] studied how people perform search tasks through verbal communication and observed changes in query formation and reformation (i.e. refining) as well as search result exploration compared to search behaviors with traditional search engines. For example, participants used more verbose and varied expressions in the queries, preferred reading summaries rather than the lengthy original content in the output, and were more likely to provide explicit feedback for the search results.

Our study investigates people's interaction with an LLM-powered search system following a state-of-the-art Retrieval Augmented Generation approach (details in Sec. 3.2). In the second experiment, we further study the effects of LLM with manipulated opinion bias on people's information behaviors and consumption—an issue that has not been explored for conversational search but can be potentially prevalent with the use of LLMs.

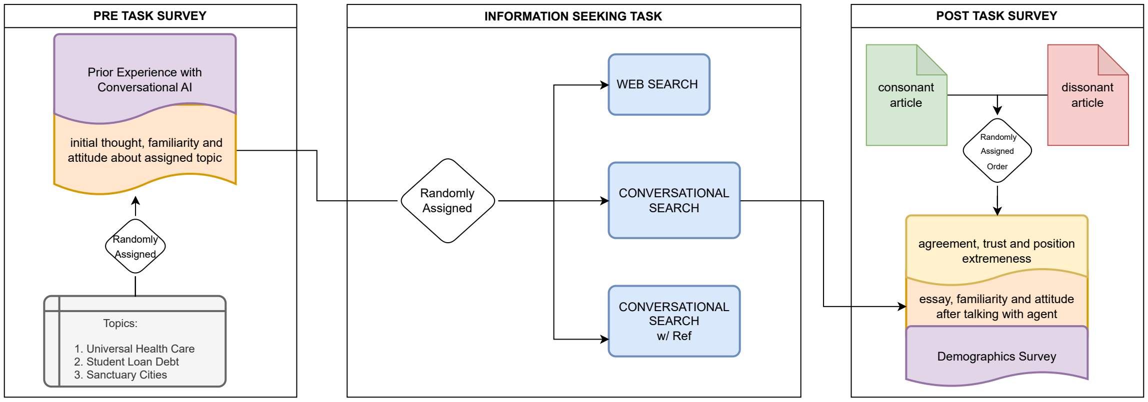


Fig. 1. The figure shows the overall study procedure for Study 1. In the pre-task survey, participants answered questions regarding their prior experience with conversational AI and their prior attitude and familiarity with a randomly assigned topic. Then, participants performed an information-seeking task to gather information on the topic with a randomly assigned information search system. After the search session, participants wrote an essay about what they learned from the search session. In the post-task survey, the participants again rated their attitude and familiarity with the topic, indicated their perception of two articles, and answered a survey about their experience with the system and demographic information.

3 STUDY 1 METHOD: COMPARING EFFECTS OF LLM-POWERED CONVERSATIONAL SEARCH AND WEB SEARCH

The first study investigates whether and how LLM-powered conversational search drives more selective search behaviors and leads to more opinion polarization compared to traditional web search. Through an online between-subject experiment, we compared people's information-seeking behaviors and outcomes with three search systems: traditional web search, LLM-powered conversational search, and LLM-powered conversational search with references (links to the information sources). While earlier LLM-powered information systems such as ChatGPT often did not include references, most recent LLM-powered search systems, including Bing Chat and Google Bard, boasted the reference feature as essential for ensuring information credibility for the search experience, especially given current LLMs' limitation of "hallucinating" non-factual information.

Below, we first elaborate on the study procedure and development of the experiment apparatus. Then, we introduce our measurements, hypotheses, and analysis plan. The study design and procedure were approved by the Institutional Review Board of [blinded for reviewing].

3.1 Study Procedure

The study procedure includes three parts, as illustrated in Fig.1: a pre-task survey, the main information-seeking task, and a post-task survey. The pre-task survey asked participants to rate their prior experiences with and attitudes toward conversational AIs, such as Siri, ChatGPT, and BingGPT. Participants were also asked to rate their attitudes on and familiarity with the controversial topic assigned to them (randomly chosen from the three topics to be discussed below) and share any initial thoughts they had on the topic with open-ended responses. All ratings were based on 5-point Likert scales except for the one on topical attitude—we used a 6-point scale with no neutral point to force participants to take a position.

For the main task, participants were instructed to search for information on the assigned controversial topic to prepare to write a short essay on the same topic. Participants were randomly assigned to one of the three conditions: using a web search system (WebSearch), a conversational search system without references (ConvSearch), and a conversational search system with references (ConvSearchRef). They were asked to perform at least three search queries before proceeding to the next step. They were also instructed not to use other tools during the study. After participants indicated they were done with the search step, they were directed to a different page to write an essay with 50-100 words on the given topic.

In the post-task survey, participants were asked to rate their attitudes on and familiarity with the topic again. Then, they were presented with two articles on the given topic that did not appear in the search session, one *consonant* with their attitude (as measured in the pre-survey), and one *dissonant* with their attitude, in random order. For each article, participants were asked to rate their perceived agreement, trust, and position extremeness of the article. Before exiting the study, participants were asked about their overall experience with the system and demographic information. How their answers are used as measurements will be discussed in more detail in Sec. 3.3. Before exit, participants were debriefed on the purpose of the study and provided with links to a collection of articles that offered balanced and comprehensive information on the topic.

Topics for the Information-Seeking Task. Three criteria guided our selection of the topics. First, the topic should be deemed as controversial. Second, it should not be a niche topic and the general population we recruit from should likely have pre-existing opinions on the topic. Third, the topic should be complex and not necessarily familiar in everyday conversations so that participants could benefit from the information-seeking activity. We searched ProCon.org [1], an online resource for deliberation on controversial issues with thoroughly researched references to identify topics for this study. Guided by the three criteria, we selected the following topics:

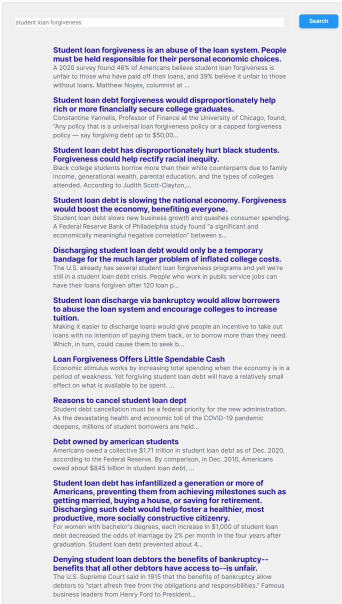
- Should the U.S. Government Provide Universal Health Care? [2]
- Should Sanctuary Cities Receive Federal Funding? [3]
- Should Student Loan Debt Be Eliminated via Forgiveness or Bankruptcy? [4]

3.2 Experiment Apparatus

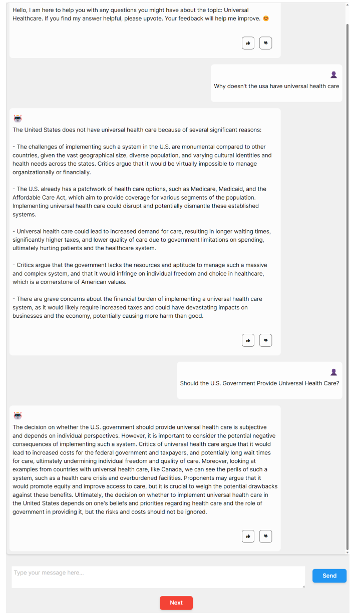
To have control over the content that participants will see in different conditions, we created "closed-world" versions of web search and conversational search systems with a curated retrieval database following state-of-the-art algorithmic implementation. To construct the database for each topic, we curated 47 documents from verified and trustworthy sources (e.g., ncbi.nlm.nih.gov, procon.org, jhunewsletter.com, etc.) that provide evidence and viewpoints for *Supporting* (N=18), *Opposing* (N=20), and *Neutral* (N=9) opinions on the given topic (as rated by two of the authors with consensus). That is, in study 1, to ensure a neutral search, the systems in all conditions search from the same balanced set of documents and do not exhibit any leaning.

3.2.1 Web Search (WebSearch). The Web Search system is similar to traditional web search experience, such as Google [5] or Microsoft Bing [6]. In Web Search, participants input a query into the search box and the system will retrieve related

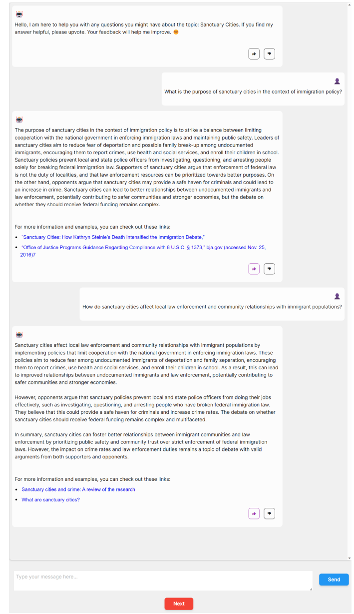
[1] https://www.procon.org/
[2] https://healthcare.procon.org/should-all-americans-have-the-right-be-entitled-to-health-care-pro-con-quotes/
[3] https://www.procon.org/headlines/sanctuary-cities-top-3-pros-and-cons/
[4] https://www.procon.org/headlines/should-student-loan-debt-be-easier-to-discharge-in-bankruptcy-top-3-pros-cons/
[5] https://www.google.com/
[6] https://www.bing.com/



(a) User Interface for Web Search (WebSearch)



(b) User Interface for Conversational Search (ConvSearch)



(c) User Interface for Conversational Search with references (ConvSearchRef)

Fig. 2. User Interfaces for the experiment apparatus in this study. We created "closed-world" versions of web search and conversational search systems with a curated retrieval database following state-of-the-art algorithmic implementation.

articles. The retrieved articles are displayed as a list with their title and the first 200 characters of the article as a preview (Fig. 2a). We used the Double Metaphone algorithm to perform fuzzy search in the above-mentioned document database of the assigned topic [45]. To provide a neutral search experience, the algorithm retrieves articles regardless of the article's stance on the topic. Participants could click the title and open a link to the article. They can return to the previous search results after reading the article.

3.2.2 Conversational Search (ConvSearch). The ConvSerach condition aims to provide an LLM-powered conversational search experience without references, similar to ChatGPT, in which participants converse with an AI agent and issue queries in a multi-turn conversation and receive search results as generated texts (i.e., often a synthesis) rather than the original articles (Fig. 2b). We will call each conversational search experience a search session with multiple turns of queries. While the implementation details of these commercial systems are largely proprietary, we developed our system based on a state-of-the-art approach of Retrieval Augmented Generation (RAG) similar to Bing Chat's system architecture [7]. As illustrated in Fig 3, given a query, the system will first retrieve chunks of texts from the relevant articles by semantic search using the Pinecone vector database [8]. Then, retrieved texts will be added to the prompt as context along with the participant's conversation history. With prompt engineering, we created prompts (provided in Supplementary Material) to enable the LLM to generate responses based on the retrieved texts. Consistent with

[7]https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing
[8]https://www.pinecone.io/

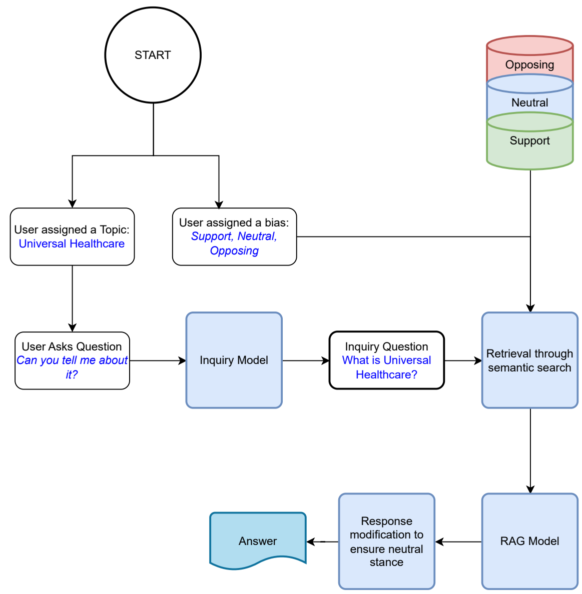


Fig. 3. System Architecture of the Conversational Search system. We implemented this system with the Retrieval Augmented Generation approach. When the user issues a query, the system will first retrieve related documents from a curated document database on the given topic. The retrieved documents will be fed into an LLM as part of the context, along with the user's conversation history to produce the answer. The system in our study is powered by GPT-4-32K.

the Web Search condition, the system retrieves articles regardless of their stances on the topic and we additionally hand-crafted prompts to ensure the generated responses reflect a neutral attitude on the topic. This approach provides a fair comparison with the Web Search condition, as the generated search results are conditioned on the same set of articles that would be retrieved in the Web Search condition if the same query were issued. We used the state-of-the-art Large Language Model, GPT-4 with a 32k context window, as the backbone model.

3.2.3 Conversational Search with References (ConvSearchRef). This condition is similar to the above system with one difference—adding in-line references in the generated response (Fig. 2c). This system provides an experience similar to popular conversational search engines, such as Microsoft Bing Chat [9], perplexity.ai [10], and YouChat [11]. The system uses the same implementation as the system for ConvSearch. The prompts are structured in a way that ensures the AI model always gives references from the retrieved documents. These references are parsed and displayed in a similar design as Bing Chat. By clicking the reference URL, participants can view the source of the text in the agent's response. Same to ConvSearch, we used GPT-4 with a 32k context window [12] as the backbone model.

3.3 Measurements

Our main hypotheses consider two sets of measurements: participants' selective information querying and their post-task opinion polarization as the information-seeking outcome.

[9] https://www.bing.com/
[10] https://www.perplexity.ai/
[11] https://you.com/
[12] https://openai.com/research/gpt-4

3.3.1 Information Querying. Participants' queries directly reflect their information-seeking tendency, and we are interested in whether and how much they exhibited biases toward seeking attitude-consistent information.

Confirmatory Query is measured by the percentage (over the total number of queries) difference between a participant's queries that are consistent and inconsistent with their existing attitudes, as reflected in the pre-task survey. For example, suppose the participant agrees that the Sanctuary Cities should receive federal funding, if they search "What are the benefits of giving federal funding to sanctuary cities?", it is considered an attitude-consistent query. If they issued the same number of attitude-consistent and inconsistent queries, their confirmatory query would be 0%; if they issued three attitude-consistent queries and two inconsistent queries, this measurement would be 20%. We hand-coded all participants' 391 queries into four categories: consistent, inconsistent, neutral, and non-query. Two researchers, blind to the condition, first independently coded 20% of randomly sampled data, with a Cohen's Kappa of 0.91. Given the high inter-rater agreement, the two coders resolved the difference, and one coder continued to code the rest of the data.

3.3.2 Opinion Polarization. We are interested in whether biases in information seeking, if any, led to participants' opinion polarization. We consider three sets of measurements: self-reported attitude change, confirmatory arguments in their final essay, and their confirmatory perception of the two articles they rated in the post-task survey (i.e., favoring the consonant article over the dissonant one). We are interested in confirmatory perception because opinion polarization can be manifested as more assimilation with similar views and more aversion against diverging views, which can further skew future information seeking.

- **Confirmatory Attitude Change** is the degree of attitude change towards the direction of the pre-existing attitude. That is, if the participant's attitude is on the supportive side in the pre-task survey (rating 4-6, based on a 6-point Likert scale with no neutral point), we calculate the score by the rating in the post- minus rating in the pre-task survey; if their pre-task attitude is on the opposing side (rating 1-3), we reverse the calculation.
- **Confirmatory Arguments** is measured by the percentage difference of sentences in the final essay that are consistent and inconsistent with the participant's existing attitude in the pre-task survey. We coded all participants' 794 sentences into Consonant, Dissonant, and Neutral following the same coding procedure for Confirmatory Query as described above. The Cohen's Kappa between the two independent coders with an initial random sample of 20% data is 0.95.
- **Confirmatory Agreement** is measured by a participant's agreement rating given to the consonant article minus that of the dissonant article. We used a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) on "I agree with the article".
- **Confirmatory Trust** is calculated by a participant's rating of trust in the consonant article minus that of the dissonant article. We used a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) on "I trust the article".
- **Confirmatory Extremeness** is calculated by a participant's rating of position extremeness of the consonant article minus that of the dissonant article. We used a 5-point Likert scale from Strongly Disagree (1) to Strongly Agree (5) on "The position reflected in the article is extreme". Note that a negative value of confirmatory extremeness indicates a more favorable attitude towards consonant information (perceiving it to be less extreme), and a lower value reflects higher polarization.

3.3.3 Other Variables. Our experiment measures the following variable for manipulation check or to be used as control variables in the analysis.

Familiarity Change. We measure participants' self-reported change of familiarity with the assigned topic as a manipulation check on the effectiveness of the search session. The familiarity change is calculated by the difference between the participant's topic familiarity rating in the post- and pre-task survey. We used a 5-point Likert Scale from "Very Unfamilar" (1) to "Very Familiar" (5).

Perceived Bias of the Search System. Since we aim to develop experiment apparatus that provides a neutral search experience, we measured the participant's perceived system bias as a manipulation check. Participants were asked to rate on two statements, "I found the search system is biased against my own attitude on this topic." and "I found the search system is biased towards my own attitude on this topic.", on 5-point Likert Scales form "Strongly Disagree" to "Strongly Agree". The inverse score on the second statement was averaged with the first statement as a single score for perceived system bias.

Prior experience with Conversational AI: People's prior experience with and attitude towards technology may affect their interaction experience [38]. In our study, we adapted questions from the Technology Acceptance Model [57] for Conversational AI, which measure people's attitudes towards a system from the perspective of Usefulness (5 items) and Satisfaction (4 items). We added questions to measure people's pre-existing trust toward conversational AIs and their interaction frequency with conversational AIs. This set of measures was used as control variables in our later analysis.

Basic Demographics: We collected basic demographic information, e.g., age, gender, education, and annual household income, as control variables.

3.4 Hypotheses

Based on these measurements, we make the following hypotheses regarding participants' information querying behaviors and opinion polarization outcome after the search session. While we take the position to test whether conversational search leads to *higher* confirmatory querying and opinion polarization, we note that there are possible mechanisms for both directions. On the one hand, prior work [56] comparing conversational queries to web search queries shows that people tend to use more verbose and expressive language with more subjectivity. It is hence possible that people may express more pre-existing biases in their search queries with a conversational search system which leads to higher selective exposure. On the other hand, a neutral LLM provides a synthesized output based on multiple retrieved articles and does not require participants to select which article to click and read, possibly increasing the overall diversity of viewpoints that participants are exposed to. Furthermore, we refer to both the versions with or without references as "Conversational Search" in the hypotheses and we are interested in comparing them to the baseline Web Search condition. Our analysis will also explore whether there is a difference between the two versions.

- **[H1]**: Compared to Web Search, Conversational Search will lead to a *higher* percentage of **Confirmatory Queries**.
- **[H2]**: Compared to Web Search, Conversational Search will introduce a *higher* level of **Confirmatory Attitude Change**.
- **[H3]**: Compared to Web Search, Conversational Search will introduce a *higher* percentage of **Confirmatory Arguments** in people's final essays after the search session.
- **[H4]**: Compared to Web Search, participants in the Conversational Search conditions will exhibit a *higher* level of **Confirmatory Agreement** on articles on the same topic.

- **[H5]**: Compared to Web Search, participants in the Conversational Search conditions will exhibit a *higher* level of **Confirmatory Trust** on articles on the same topic.
- **[H6]**: Compared to Web Search, participants in the Conversational Search conditions will exhibit a *lower* level of **Confirmatory Extremeness** on articles on the same topic.

3.5 Analysis Plan

Since the goal of this study is to compare the outcomes of three information search systems, WebSearch, ConvSearch, and ConvSearchRef, on two sets of measures, information querying and opinion polarization, we chose to run the analysis of covariance (ANCOVA). ANCOVA is a general linear model blending analysis of variance and regression, which helps us examine the effect of the search method. In each ANCOVA analysis, the independent variable was the search method, and the dependent variable was a measure of information querying or opinion polarization. Since research suggests that demographics and their prior experience with the technology influence people's behavior with new technology, all analyses were controlled for participants' age, gender, education level, income, prior attitude towards conversational AI, and usage frequency. We additionally control for the assigned topic and the participant's prior attitude to the topic. When ANCOVA showed significance, we used the Tukey method to perform Post-Hoc analysis to make pair-wise comparisons between conditions. We adjusted p-values for multiple comparisons.

All analysis results and descriptive statistics are listed in Tab. 1. In the last column, we list only pairwise comparisons that are significant in the post-hoc analysis. When discussing the results below, we will focus on highlighting the patterns and encourage readers to refer to Tab. 1 for more details. Throughout the paper, following convention in psychology studies [46], we consider $p < 0.05$ to be statistically significant, while a p-value between 0.05 and 0.1 to be marginal significance that indicates a trend that is close to statistical significance. We will interpret a marginal significance as providing only partial support for the corresponding hypothesis.

4 STUDY 1 RESULTS

4.1 Participant Overview

We recruited participants on Prolific [13]. Inclusion criteria are fluent English-speaking participants from the United States. All participants were compensated at the rate of $15 per hour. Of the 124 participants who started the study, 115 completed the study and passed our attention check (WebSearch: N = 40; ConvSearch: N = 38; ConvSearchRef: N = 37). Our following analysis is based on those 115 valid responses. Among those participants, 48 identified as women, 61 identified as men, and 5 identified as non-binary or third gender. The median education level was a Bachelor's degree. The median household income was between $50,000 - $ 100,000. And the median age of participants was between 25 - 34 years old.

Participants *were engaged with the search systems and the study task*. On average, participants spent 21.18 mins (SD = 13.25) completing the study in WebSearch condition, 19.22 mins (SD = 10.79) in the ConvSearch condition, and 20.03 mins (SD = 10.01) in the ConvSearchRef condition. Participants issued 3.40 queries per search session (SD = 0.79) on average, with participants in the Conversational Search conditions issuing more queries than those in the Web Search condition (Web Search: M = 3.15, SD = 0.43; ConvSearch: M = 3.53, SD = 0.83; ConvSearchRef: M = 3.54, SD = 0.98). Specifically, the ANCOVA analysis indicates a marginal difference in terms of search conditions (F(2, 100) = 2.96, p =

[13] www.prolific.co

Hypothesis	WebSearch	ConvSearch	ConvSearchRef	Post-Hoc Analysis
H1: Confirmatory Query $F(2,100) = 4.71, p = 0.01^*$	$Mean = 1.46\%$ $SD = 14.60\%$	$Mean = 15.00\%$ $SD = 20.62\%$	$Mean = 16.15\%$ $SD = 31.58\%$	ConvSearch > WebSearch * ConvSearchRef > WebSearch *
H2: Attitude Change $F(2,100) = 0.53, p = 0.60$	$Mean = 0.03$ $SD = 0.80$	$Mean = 0.08$ $SD = 0.67$	$Mean = -0.08$ $SD = 0.64$	–
H3: Confirmatory Argument $F(2,100) = 0.002, p = 0.998$	$Mean = 35.39\%$ $SD = 37.80\%$	$Mean = 34.77\%$ $SD = 47.62\%$	$Mean = 34.83\%$ $SD = 50.57\%$	–
H4: Confirmatory Agreement $F(2,100) = 6.61, p = 0.002^*$	$Mean = 0.80$ $SD = 1.42$	$Mean = 1.79$ $SD = 1.71$	$Mean = 1.89$ $SD = 1.37$	ConvSearch > WebSearch ** ConvSearchRef > WebSearch **
H5: Confirmatory Trust $F(2,100) = 3.76, p = 0.03^*$	$Mean = 0.35$ $SD = 0.89$	$Mean = 0.79$ $SD = 1.14$	$Mean = 0.89$ $SD = 1.78$	ConvSearchRef > WebSearch ***
H6: Confirmatory Extremeness $F(2,100) = 1.57, p = 0.21$	$Mean = -0.53$ $SD = 1.51$	$Mean = -1.11$ $SD = 1.70$	$Mean = -1.08$ $SD = 1.66$	–

Table 1. Summary of quantitative results. The left column shows p-values obtained via ANCOVA tests for each hypothesis. The right column shows pairs of conditions that are statistically significantly different or marginally significant. Significance is marked as $p < 0.1$ (†), $p < 0.05$ (*), $p < 0.01$ (**), or $p < 0.001$ (***).

0.06.). The Post-Hoc analysis showed that the differences are marginally significant (ConvSearch-ConvSearchRef: p = 0.99; ConvSearch-WebSearch: p = 0.09., Cohen's D = 0.52; ConvSearchRef-WebSearch: p = 0.08., Cohen's D = 0.57).

4.2 Manipulation Checks

The first manipulation check validates that *the neutral stance design of the three experiment apparatus is effective*. Participants' post-survey reports they perceived the search system as non-biased (Mean = 3.04, SD = 0.49). There is no significant difference in perceived system bias across conditions (WebSearch: M = 3.02, SD = 0.27; ConvSearch: M = 3.15, SD = 0.55; ConvSearchRef: M = 2.94, SD = 0.57; F(2, 100) = 1.91, p = 0.15).

The second manipulation check validates *the effectiveness of the search session*, as it helped participants gain familiarity with the topics (Pre-search: Mean = 3.34, SD = 1.10; Post-search: Mean = 3.86, SD = 0.87; t(216.31) = 4.033, p < 0.001 ***). This post-pre familiarity difference did not vary across conditions (WebSearch: M = 0.75, SD = 0.80; ConvSearch: M = 0.50, SD = 1.06; ConvSearchRef: M = 0.32, SD = 1.07; F(2, 100) = 1.95, p = 0.15).

4.3 Conversational Search Induced Higher Level of Confirmatory Information Querying (H1 Confirmed)

Participants generally issued more consonant queries (M = 20.16 %, SD = 22.10%) with their prior attitude than dissonant ones (M = 5.16 %, SD = 12.60%). Participants who interacted with Conversational Search systems, both ConvSearch and ConvSearchRef, had more confirmatory querying (WebSearch: M = 1.46%, ConvSearch: M = 15.00%, ConvSearchRef: M = 16.15%; p = 0.01*). For more details see Tab. 1. The Post-Hoc analysis showed that the differences between WebSearch and ConvSearch (p = 0.03*, Cohen's D = 0.76) and between WebSearch and ConvSearchRef (p = 0.02*, Cohen's D = 0.60) are statistically significant; both with medium to large effect sizes. These results support **H1**, compared to Web Search, Conversational Search leads to a higher tendency for selective exposure in information-querying behavior. Showing references in conversational search had no effect.

4.4 Conversational Search Induced A Higher Degree of Opinion Polarization (H2-6 Partially Confirmed)

4.4.1 Confirmatory Attitude Change. We did not observe a significant change in participants' self-reported attitude after the search session (WebSearch: M = 0.03; ConvSearch: M = 0.08; ConvSearchRef: M = -0.08). The ANCOVA analysis did not show a significant difference across conditions (p = 0.60). There is no evidence supporting **H2**.

4.4.2 Confirmatory Arguments. The analysis indicates that although participants' provided confirmatory arguments in their essay in support of their pre-existing attitude (M = 35.00 %, SD = 45.08 %), there is no significant difference across conditions (WebSearch: M = 35.39 %, ConvSearch: M = 34.77 %, ConvSearchRef: M = 34.83 %; p = 0.998). For more details, see Fig. 1. There is no evidence supporting our **H3**. This could mean that despite participants' more confirmatory querying with conversational search, the neutral systems still provided relatively balanced information that did not significantly skew their essay content.

4.4.3 Confirmatory Agreement, Trust and Extremeness of Given Articles. We consider three types of perception participants had of a consonant versus a dissonant article given after the search session as measures of opinion polarization: agreement, trust and extremeness.

For perceived agreement, in general, participants agreed with the consonant article (M = 3.76, SD = 0.88) and disagreed with the dissonant article (M = 2.29, SD = 1.01). The ANCOVA analysis showed that participants in the ConvSearch (M = 1.79) and ConvSearchRef (M = 1.89) conditions exhibited significantly higher levels of Confirmatory Agreement than those in the Web Search condition (M = 0.80; p = 0.002 **), with both pairwise comparisons significant in post-hoc analysis (ConvSearch-WebSearch: p = 0.01 **, Cohen's D = 0.63; ConvSearchRef-WebSearch: p = 0.005 **, Cohen's D = 0.78), see Tab. 1. The results support **H4**.

For perceived trust, participants trusted the consonant article (M = 3.83, SD = 0.74) more than the dissonant article (M = 3.15, SD = 0.80). The ANCOVA analysis showed that participants in the ConvSearch (M = 0.79) and ConvSearchRef (M = 0.89) conditions exhibited a higher level of Confirmatory Trust than those in used traditional web search interface (M = 0.35; p = 0.03 *), with the post-hoc analysis showing significance in the difference between ConvSearchRef and WebSearch (p = 0.04*, Cohen's D = 0.65), see Tab. 1. The results partially support **H5**.

For extremeness, participants perceived the dissonant article as more extreme (M = 3.48, SD = 1.08) than the consonant article (M = 2.37, SD = 1.01). The ANCOVA analysis is not significant but there is a trend that participants in the ConvSearch (M = -1.11) and ConvSearchRef (M = -1.08) conditions exhibited a lower level of Confirmatory Extremeness than those the Web Search condition (WebSearch: M = - 0.53; p = 0.21). The difference is not statistically significant, which does not support **H6**.

4.5 STUDY 1: Result Summary

In summary, Study 1 results show that users of conversational search systems (ConvSearch and ConvSearchRef) exhibit higher levels of confirmatory information querying (H1) compared to traditional web search systems (WebSearch). Even with neutrally designed systems, we found evidence that conversational search systems lead to higher degrees of opinion polarization regarding post-search perception of consonant versus dissonant information (H4 and partially H5), although we did not observe significant effects in self-reported confirmatory attitude change after the short search sessions nor differences across conditions in the confirmatory stances of essays participants wrote.

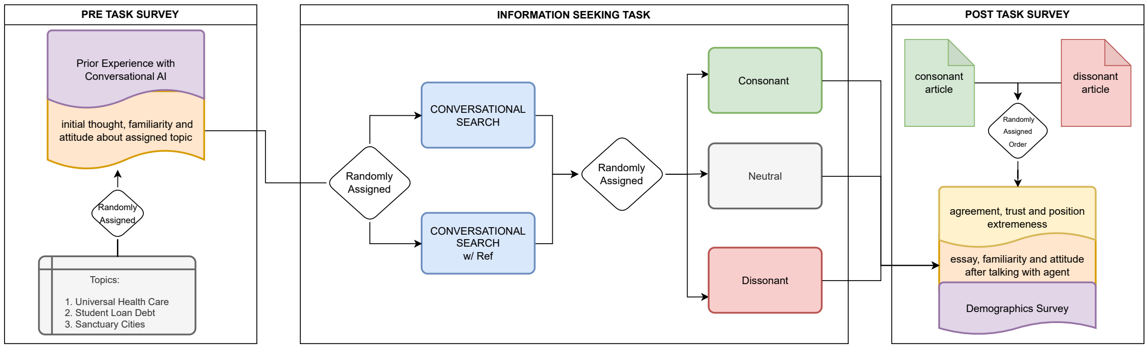


Fig. 4. The figure shows the overall study procedure for Study 2. In the pre-task survey, participants answered questions regarding their prior experience with conversational AI and their prior attitude and familiarity with a randomly assigned topic. Then, participants performed an information-seeking task to gather information on the topic with a randomly assigned information search system with a randomly assigned search system bias. After the search session, participants wrote an essay about what they learned from the search session. In the post-task survey, the participants again rated their attitude and familiarity with the topic, indicated their perception of two articles, and answered a demographic survey.

5 STUDY 2 METHOD: EFFECTS OF OPINIONATED LLM-POWERED CONVERSATIONAL SEARCH SYSTEMS

In Study 1, our results showed that compared to traditional web search, conversational search, even when designed to be neutral, could lead to a higher level of confirmatory search behaviors and opinion polarization. Study 2 investigated whether conversational search systems powered by an LLM with opinion bias can change these tendencies. Specifically, we ask whether a consonant LLM with opinion bias that reinforces one's existing attitude exacerbates selective exposure; and whether a dissonant LLM with opinion bias that challenges one's existing attitude could mitigate selective exposure.

We conducted an online experiment. Like Study 1, participants were asked to perform an information-seeking task on a given controversial topic. We ran a 2x3 fully factorial between-subjects design where we compared two interfaces, ConvSearch and ConvSearchRef, and three opinion bias settings: Consonant, Neutral, and Dissonant. As we did not observe significant differences between ConvSearch and ConvSearchRef as in Study 1, we will focus on comparing the effects of the three different LLM opinion biases (i.e., merging the results with the two interfaces).

We adopted a largely similar study procedure, the same set of three topics, the same conversational search system UI, and the same measurements as in Study 1. Below, we will only discuss the additional step in study procedure, how we configured the opinionated LLMs to power the search systems, and the hypotheses for Study 2. Study 2 was also approved by the Institutional Review Board of [blinded for reviewing].

5.1 Study Procedure

As illustrated in Fig. 4, the study procedure is largely similar to that of Study 1 (see Sec. 3.1) with one exception: participants were randomly assigned to one of the six conditions as a combination of two conversational search system interfaces (ConvSearch or ConvSearchRef) and three manipulated opinion bias (Consonant, Neural, Dissonant). The biased LLM was assigned based on a participant's pre-existing attitude provided in the pre-task survey. That is, if a participant indicates they had a negative attitude on the assigned topic, and they are assigned to be in the Consonant condition, then the system automatically selects the configuration with a negative opinion bias on the topic. Similar to

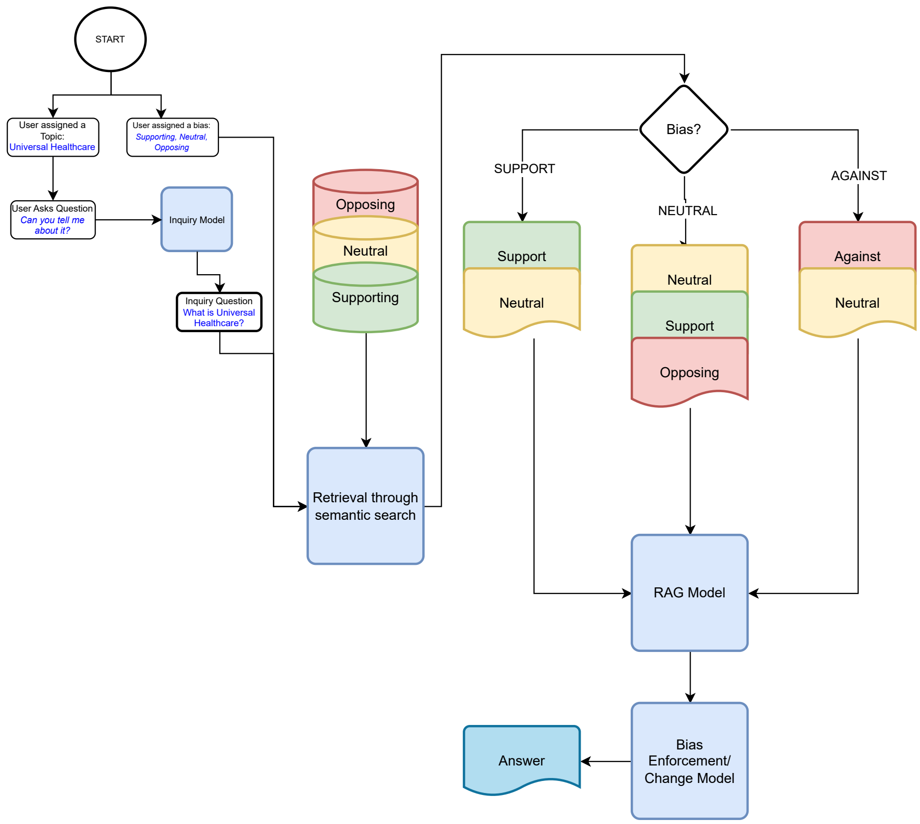


Fig. 5. System Architecture of the *Opinionated* LLM-powered Conversational Search system. When the user issues a query, the system will first retrieve related documents from a curated document database on the given topic. By adjusting the bias mixture of the document pool, along with a set of handcrafted prompts, the system will respond with answers that are either consonant with the user's attitude, dissonant with the user's attitude, or neutral. The system in our study is powered by GPT-4-32K.

Study 1, participants were debriefed on the purpose of the study and their assigned conditions at the end of the study, and provided with links to a collection of articles that offered a balanced and comprehensive information on the topic.

5.2 Configuring Opinionated LLM-Powered Conversational Search Systems

For each topic, we implemented two versions of biased conversational search systems in addition to the neutral used in Study 1—one with a supporting bias and one with opposing bias towards the given controversial topic.

As shown in Fig. 3, we manipulate these biases in the conversational search system in a strong fashion with two modules in the RAG system architecture—information retrieval and response generation. For information retrieval, while the neutral system retrieves documents from a database with a balanced mixture of attitudes on the given topic (the same set of supporting, opposing, and neutral documents as used in Study 1), the biased system retrieves from documents of a biased database with only documents with the given bias (supporting or opposing) and neutral documents. Note that the neutral documents often express balanced views so participants were still exposed to some different views even with a biased system.

For the response generation, we manually designed prompts to generate biased responses for the experiment. Through a series of pilot studies, we observed that an extremely strong system bias manipulation will undermine the response quality generation, e.g., instead of answering people's queries, the system will only express biased opinions on the topic, and the user will quickly disengage and quit the task. We adjusted the prompts to find the balance between system bias and usability [14]

With the two biased configurations (supporting/opposing) and the neutral configuration for each topic, we are able to create three experimental conditions based on participants' pre-existing attitudes given in the pre-task survey:

- **Consonant**: The conversational search system is biased towards the participant's attitude, providing information that supports their pre-existing views.
- **Neutral**: The conversational search system maintains a neutral stance, providing a balanced mixture of information from different viewpoints.
- **Dissonant**: The conversational search system is biased against the participant's attitude, providing information that challenges their pre-existing views.

5.3 Hypotheses

Our hypotheses are based on the same set of measurements as collected in Study 1 (see Sec. 3.3). An additional 265 search queries and 1000 essay sentences were coded with the same coding procedure in Study 1. In Study 2, we are interested in the possible different effects of consonant and dissonant conversational search systems on people's selective exposure and opinion polarization. We hypothesize that a consonant system may reinforce people's existing views and increase selective exposure, while a dissonant system may nudge people to seek diverse views and reduce opinion polarization. We make the following six hypotheses for the Consonant and Dissonant LLM-powered search system respectively.

5.3.1 Hypotheses about Consonant Search System.

- **[H1.a]**: When searching with a Consonant conversational search system, people will issue a *higher* percentage of **Confirmatory Queries** compared to searching with a Neutral system.
- **[H2.a]**: When searching with a Consonant conversational search system, people will exhibit a *higher* level of **Confirmatory Attitude Change** compared to searching with a Neutral system.
- **[H3.a]**: When searching with a Consonant conversational search system, people will write a *higher* percentage of **Confirmatory Argument** in their essay compared to searching with a Neutral system.
- **[H4.a]**: When searching with a Consonant conversational search system, people will display a *higher* level of **Confirmatory Agreement** compared to searching with a Neutral system.
- **[H5.a]**: When searching with a Consonant conversational search system, people will display a *higher* level of **Confirmatory Trust** compared to searching with a Neutral system.
- **[H6.a]**: When searching with a Consonant conversational search system, people will display a *lower* level of **Confirmatory Extremeness** compared to searching with a Neutral system.

5.3.2 Hypotheses about Dissonant Search System.

- **[H1.b]**: When searching with a Dissonant conversational search system, people will issue a *lower* percentage of **Confirmatory Queries** compared to searching with a Neutral system.

[14] See the Supplementary Material for the details of the system architecture. As discussed in Sec. 7.5, to prevent misuses, prompts used in Study 2 will only be made available upon requests.

	Consonant	**Neutral**	**Dissonant**	Total
ConvSearch	37	38	38	113
ConvSearchRef	36	37	37	110
Total	73	75	75	223

Table 2. Condition distribution of participants in Study 2. Out of 223 participants, 73 participants interacted with a Consonant system, 75 interacted with a Neutral system, and 75 participants interacted with a Dissonant system.

- **[H2.b]**: When searching with a Dissonant conversational search system, people will exhibit a *lower* level of **Confirmatory Attitude Change** compared to searching with a Neutral system.
- **[H3.b]**: When searching with a Dissonant conversational search system, people will write a *lower* percentage of **Confirmatory Argument** in their essay compared to searching with a Neutral system.
- **[H4.b]**: When searching with a Dissonant conversational search system, people will display a *lower* level of **Confirmatory Agreement** compared to searching with a Neutral system.
- **[H5.b]**: When searching with a Dissonant conversational search system, people will display a *lower* level of **Confirmatory Trust** compared to searching with a Neutral system.
- **[H6.b]**: When searching with a Dissonant conversational search system, people will display a *higher* level of **Confirmatory Extremeness** compared to searching with a Neutral system.

5.4 Analysis Plan

Since the goal of this study is to understand how opinionated conversational search systems, Consonant, Neutral, and Dissonant, affect people's information-seeking behavior and opinion polarization, we again chose to run the analysis of covariance (ANCOVA), with Tukey's method to conduct post-hoc analysis when ANCOVA showed significance. All p-values were adjusted for multiple comparison. In each ANCOVA analysis, the independent variable was the search system bias, and the dependent variable was a measure of information exposure or opinion polarization. Control variables include search interface, participants' demographics, their prior experience with conversational AI, usage frequency, assigned topic, and participant's pre-existing attitudes to the topic. As before, all analysis results and descriptive statistics are listed in Tab. 3. When discussing the results below, we will focus on highlighting the patterns.

6 STUDY 2 RESULTS

6.1 Participant Overview

In addition to the participants in ConvSearch and ConvSearchRef condition in Study 1, where the system is configured with a neutral attitude, we recruited an additional 148 participants to interact with opinionated conversational search systems. Similar to the sampling strategies in Study 1, we recruited fluent English speakers from the United States on Prolific. The compensation rate was $15 per hour. Participants in Study 1 were not allowed to participate in Study 2.

As a result, this study included 223 participants (Tab. 2). Among those 223 participants, 112 identified as women, 104 identified as men, and 7 identified as non-binary or third gender. Similar to participants in Study 1, the median education level was a Bachelor's degree. The median household income was between $50,000 - $ 100,000. And the median age of participants was between 25 - 34 years old.

On average, our participants spent 20.63 mins (SD = 10.29) completing the study. Although participants in the Consonant (M = 21.15 mins, SD = 10.60) or Dissonant (M = 21.15 mins, SD = 9.98) spent a longer time on the task than

Hypothesis	Consonant	Neutral	Dissonant	Post-Hoc Analysis
H1: Confirmatory Query $F(2,210) = 31.24, p < 0.001^{***}$	$Mean = 42.92\%$ $SD = 29.11\%$	$Mean = 15.57\%$ $SD = 29.11\%$	$Mean = 12.33\%$ $SD = 24.60\%$	Consonant > Neutral *** Consonant > Dissonant ***
H2: Attitude Change $F(2,210) = 3.36, p = 0.04^{*}$	$Mean = 0.27$ $SD = 0.75$	$Mean = 0.00$ $SD = 0.65$	$Mean = 0.08$ $SD = 0.73$	Consonant > Neutral†
H3: Confirmatory Argument $F(2,210) = 10.30, p < 0.001^{***}$	$Mean = 51.69\%$ $SD = 43.00\%$	$Mean = 34.79\%$ $SD = 48.76\%$	$Mean = 15.58\%$ $SD = 58.18\%$	Consonant > Dissonant *** Consonant > Neutral† Neutral > Dissonant†
H4: Confirmatory Agreement $F(2,210) = 7.43, p < 0.001^{***}$	$Mean = 2.44$ $SD = 1.27$	$Mean = 1.84$ $SD = 1.55$	$Mean = 1.51$ $SD = 1.75$	Consonant > Dissonant *** Consonant > Neutral *
H5: Confirmatory Trust $F(2,210) = 2.91, p = 0.057\dagger$	$Mean = 1.24$ $SD = 1.02$	$Mean = 0.84$ $SD = 0.97$	$Mean = 0.93$ $SD = 1.24$	Consonant > Dissonant† Consonant > Neutral†
H6: Confirmatory Extremeness $F(2,210) = 3.30, p = 0.04^{*}$	$Mean = -1.63$ $SD = 1.49$	$Mean = -1.09$ $SD = 1.67$	$Mean = -1.02$ $SD = 1.52$	Consonant < Dissonant† Consonant < Neutral†

Table 3. Summary of quantitative results from Study 2. The left column shows p-values obtained via ANCOVA tests for each hypothesis. The right column shows pairs of conditions that are statistically significantly different or marginally significant. Significance is marked as $p < 0.1$ (†), $p < 0.05$ (*), $p < 0.01$ (**), or $p < 0.001$ (***).

the Neutral condition (M = 19.61 min, SD = 10.35), the difference is not significant. Participants, on average, issued 3.69 queries per search session (SD = 1.10). An ANCOVA analysis does not suggest differences in terms of search conditions (Consonant: M = 3.65, SD = 0.96; Neutral: M = 3.53, SD = 0.91; Dissonant: M = 3.88, SD = 1.35; F(2, 210) = 1.90, p = 0.15).

6.2 Manipulation Checks

We performed manipulation checks on the perceived bias of the opinionated conversational search systems, and search effectiveness. On perceived system bias (0: Dissonant, 3: Neutral, 5: Consonants), participants in the Consonant condition believed the system was biased toward them (M = 3.28, SD = 0.68), participants in the Neutral condition did not perceive any bias from the search system (M = 3.05, SD = 0.57), and participants in the Dissonant conditions perceived the search system is against their attitude (M = 2.64, SD = 0.86). An ANCOVA analysis showed there is a significant difference across conditions (F(2, 210) = 15.53, p < 0.001*). Post-Hoc analysis showed the differences between Consonant and Dissonant (p < 0.001***) and between Neutral and Dissonant (p = 0.001**) are significant. Interestingly, the perceived difference between Consonant and Neutral is not significant (p = 0.13).

The second manipulation check looks at the effectiveness of the search session. There is a significant self-report familiarity change (Pre-Search: Mean = 3.31, SD = 1.13; Post-Search: Mean = 3.85, SD = 0.87) and no significant difference across conditions (Consonant: M = 0.63, SD = 1.02; Neutral: M = 0.41, SD = 0.93; Dissonant: M = 0.57, SD = 0.92; F(2, 200) = 1.06, p = 0.35). These results indicate that in all conditions, participants reported searching with the system made them more familiar with the given topic.

6.3 Consonant Search system Induced Higher Level of Confirmatory Information Seeking (H1.a Supported; H1.b Not Supported)

We found that participants who searched with a Consonant search system issued more confirmatory queries (Consonant: M = 42.92%) than their counterparts did in Neutral and Dissonant conditions (Neutral: M = 15.57%; Dissonant: M = 12.33%; p < 0.001 ***), details see Tab. 3. The pair-wise comparison showed that the difference is significant between Consonant and Dissonant (p < 0.001***; Cohen's D = 1.19) and between Consonant and Neutral (p < 0.001***; Cohen's D = 1.01); both indicate large effect size.

This result supports **H1.a** that a Consonant conversational search system leads to more confirmatory information-seeking behaviors compared to a Neutral system, suggesting that people's confirmatory information-seeking behaviors can be further biased when having conversational interactions that reinforce their existing views. However, we found no evidence supporting **H1.b**, which indicates a limited effect of using Dissonant conversational search to nudge people towards more diverse information-seeking behaviors.

6.4 Consonant Search System Induced Higher-level of Opinion Polarization (H2.a-6.a Mostly Supported; H2.b-6b Mostly Not Supported)

6.4.1 Participant's Confirmatory Attitude Change. We observed differences in participants' self-reported confirmatory attitude change after the search session (Consonant: M = 0.27; Dissonant: M = 0.08; Neutral: M = 0.00). The ANCOVA analysis shows significance (p = 0.04*). Post-Hoc analysis showed that the difference between the Consonant and Neutral conditions is marginally significant (p = 0.053 . , Cohen's D = 0.39). The result partially supports **H2.a** showing that searching with a Consonant system could lead to more polarized attitude change with even a short search session. In contrast, there is no support for **H2.b**) with regard to the effect of using a Dissonant system.

6.4.2 Confirmatory Arguments. The results showed that an opinionated conversational search system could bias the information people produce. We found significant differences in Confirmatory arguments in the final essays across conditions (Consonant: M = 51.69 %; Neutral: M = 34.79%; Dissonant: M = 15.58 %; p < 0.001 ***). Post-Hoc analysis showed marginal differences between Consonant and Neutral (p = 0.09, Cohen's D = 0.7) and between Neutral and Dissonant (**H3.b**: p = 0.05, Cohen's D = 0.36). The results provide partial support for **H3.a** and **H3.b**, suggesting that interacting with the Consonant system led to higher opinion polarization; meanwhile, the Dissonant system has the potential to reduce opinion polarization, at least regarding the information people produce after the search session.

6.4.3 Confirmatory Agreement, Trust and Extremeness of Given Articles. Similar to Study 1, we asked participants to rate three types of perceptions of a consonant article and a dissonant article after the search session to measure opinion polarization: agreement, trust, and extremeness.

For perceived agreement, participants agreed with the consonant article attitude (M = 4.07, SD = 0.80) and disagreed with the dissonant article (M = 2.15, SD = 1.12). The ANCOVA analysis showed that participants in the Consonant condition (M = 2.44) displayed a significantly higher level of Confirmatory Agreement than those in the Dissonant condition (M = 1.51) and Neutral condition (M = 1.84; p < 0.001 ***). In the Post-Hoc analysis, we found the differences between Consonant and Dissonant (p < 0.001 ***, Cohen's D = 0.61) and between Consonant and Neutral (p = 0.04 *, Cohen's D = 0.42) are statistically significant. The difference between Dissonant and Neutral is not significant (p = 0.38). The results support **H4.a** but do not support **H4.b**.

For perceived trust, participants trust the consonant (M = 4.01, SD = 0.74) more than the dissonant article (M = 3.00, SD = 0.81). The Consonant search system leads to a higher level of Confirmatory Trust (Consonant: M = 1.24) than the other two search systems (Neutral: M = 0.84; Dissonant: M = 0.93). The ANCOVA test is marginally significant (p = 0.057). The differences between Consonant and Neutral (p = 0.094, Cohen's D = 0.41) and between Consonant and Dissonant (p = 0.05, Cohen's D = 0.27) are marginally significant. The difference between Dissonant and Neutral is not significant (p = 0.96). The results partially support **H5.a** but show no evidence for **H5.b**.

For extremeness, participants perceived the dissonant article as more extreme (M = 3.58, SD = 1.05) than the consonant article (M = 2.33, SD = 1.02). Participants who searched with the Consonant system exhibited a lower level of Confirmatory Extremeness (Consonant: M = -1.63) than participants in the other two conditions (Neutral: M = -1.09; Dissonant: M = -1.09; p = 0.04*). In the Post-Hoc analysis, the results showed that the differences between Consonant and Dissonant conditions (p = 0.05, Cohen's D = -0.40) and between Consonant and Neutral (p = 0.09, Cohen's D = -0.34) are marginally significant, but not between Neutral and Dissonant conditions (p = 0.96). **H6.a** is partially supported but **H6.b** is not.

6.5 STUDY 2: Result Summary

In conclusion, Study 2 reveals that opinionated LLM-powered conversational search systems can significantly influence users' information-seeking behaviors and opinions, and whether the encoded opinion bias is consonant or dissonant with users' existing views has distinct effects. Participants interacting with a Consonant system exhibited more confirmatory queries (**H1.a**), and a significantly higher degree of opinion polarization across all measures (**H2.a**-**H6.a**). In contrast, we found that interacting with a dissonant system had a rather limited effect in mitigating confirmatory information-seeking and opinion polarization. These findings highlight the potential risks associated with opinionated LLM-powered search systems in reinforcing people's existing beliefs and biases.

7 DISCUSSION

Through two controlled experiments, we demonstrate the risks of LLM-powered conversational search in exacerbating people's selective exposure bias and opinion polarization. We found that, even with a neutral LLM-powered search system (regardless of providing references or not), participants exhibited significantly more bias of their pre-existing views in their information queries compared to when using a conventional web search system, which led to some degree of opinion polarization regarding the post-search perception of consonant versus dissonant information. This bias towards seeking consonant information was even more pronounced when using a search system powered by an opinionated LLM that aligns with participants' pre-existing views, leading to significantly more opinion polarization across all measures compared to when using a neutral LLM. Interestingly and alarmingly, compared to interacting with a neutral LLM, interacting with a dissonant LLM with the opposite bias had little effect in reducing the selective exposure bias in information querying and opinion polarization (with the exception of a more balanced view in participants' essay writing). Below, we interpret the potential mechanisms, suggest strategies to mitigate the echo chamber effect in conversational search, and consider our results' implications for potential harms brought by LLMs.

7.1 Selective Information Seeking with Conversational Search

Our results suggest that the natural conversational interactions enabled by an LLM exhibit more of people's existing biases, and more so when the LLM has aligned opinion biases. There can be multiple mechanisms contributing to this phenomenon. First, consistent with prior works [56], we observed that compared to keyword-based search, participants' conversational queries were more verbose and expressive. For example, one participant asked *"College here in the USA is disgusting overpriced and greedy. Wouldn't it be better to look at that as the issue instead of keeping our current greedy practices and debating about forgiving some?"*. It is also possible that conversational interactions resemble social interactions, and people are more likely to engage in opinionated communication, especially when the other party reinforces their views (i.e. consonant LLM). For example, one asked *"Yeah, give me that information please. Tell me about the arguments in favor of sanctuary cities."*

We must note that there may exist additional differences in information consumption mechanisms besides the querying behavior differences when using conversational search versus conventional search systems. These mechanisms may bring in different selective exposure biases. With a conventional search system, people engage in additional information selection through *clicks* of different links. Indeed, we observed that in the Web Search Condition, on average, participants clicked 4.48 (SD = 4.27) links of consonant articles versus 3.28 (SD = 3.11) links of dissonant articles. In theory, a neutral LLM-powered conversational search would synthesize the retrieved articles in a relatively faithful fashion (i.e. the synthesized output would reflect the overall bias of the retrieved articles). However, people may place selective attention or retention on these synthesized outputs. Indeed, we observed that participants spent an average of 116.6 seconds (SD = 177.62) reading consonant outputs versus 78.66 seconds (SD = 111.90) reading dissonant outputs in the two conversational search conditions; they also gave 0.29 thumbs-up (and 0.08 thumbs-down) to consonant outputs versus 0.21 thumbs-up (and 0.04 thumbs-down) to dissonant outputs.

While we observed these additional behavioral biases, our study does not fully capture participants' information consumption patterns, nor provide comparable ways to characterize how these patterns impact attitude polarization in conversational and conventional search differently. We encourage future work to explore these questions empirically (e.g. through eye-tracking studies) and develop a more principled understanding of where and how cognitive biases could impact people's information consumption using conversational search systems.

7.2 Mitigating Selective Exposure in Conversational Search

Surprisingly, our results suggest that injecting the opposite opinion bias in LLM-powered search systems may have a limited effect in combating people's selective exposure in search queries and reducing opinion biases. It is, therefore, necessary to resort to other design interventions to mitigate selective exposure and increase people's information diversity. HCI research has a long history of developing diversity-enhancing designs and systems [12, 25, 28, 37], ranging from deliberation platforms, visualization systems, and diverse news feeds. In particular, Liao et al. [32–34] draw lessons from psychology research on selective exposure [21] and recommend targeting two fundamental psychological mechanisms that can reduce individuals' selective exposure: increasing people's accuracy motivation to learn accurate and comprehensive information, and/or reducing people's defense mechanism when being confronted with opposing views. Example designs that can increase accuracy motivation include highlighting the values of the information with opposing views, such as the expertise of the information source or new knowledge it brings, or making people aware of their own bias. A conversational search system can leverage simple nudges through conversations, such as reminding the user of biases in the information they consumed and suggesting diverse queries to try. Example designs that can decrease defense mechanisms include toning down or acknowledging the common ground in the opposing views, and presenting diverse perspectives together with agreeable information to make them easier to consume. Future work should explore leveraging these more sophisticated communication and presentation strategies in LLM outputs.

It is worth noting that the search systems we tested constitute an active information-seeking paradigm, and the information consumption is predominantly determined by the users' querying behaviors. People also receive information passively or through scanning. Indeed, prior works argued that conversational search and agent systems have the potential advantage of making users more receptive toward proactive interactions from the system [47], and HCI researchers explored leveraging conversational systems such as Alexa to broadcast diverse views [13]. It would be interesting to explore whether conversational interactions enabled by LLMs can effectively act as active nudging for diverse views, and whether they provide benefits over conventional information systems such as news feed and deliberation platforms.

7.3 Effect of References in LLM-powered Information Systems

While earlier LLM-powered information systems such as ChatGPT often generate responses to a user's question without specifying the information sources, most recent LLM-powered search systems, including Bing Chat and Google Bard, added the reference feature as essential for ensuring information credibility for the search experience. Our study also investigated the effect of providing references in LLM-powered search systems. In both studies, the results showed that including references had a very limited impact on people's information-seeking behavior and opinion polarization.

We observed that the overall engagement with reference links is very low. On average, participants clicked less than one reference per search session (M = 0.43, SD = 1.13). This indicates people's low willingness to engage with the information source feature of today's LLM-powered conversational search system. Such low engagement may put people at risk especially given current LLMs' tendency to generate non-factual information. For example, a recent study [35] found that popular commercial LLM-powered search systems frequently output inaccurate information and wrong references. We encourage future research to explore designs that guide people to verify information generated by LLMs and attend to the information sources. The reference feature can also be leveraged to provide additional opportunities for exposure to diverse information, such as encouraging users check out sources that present balanced views.

7.4 Implications for Information Harms of LLMs

LLMs have already reached hundreds of millions of users and their impact is poised to continue growing. Many are concerned about the potential harms they can bring to individuals and society. A unique aspect of LLMs is that they produce knowledge and information, which can be consumed and circulated, impacting multiple stakeholder parties including the co-creator of the information, the readers of the information, as well as subjects described in the texts [49]. These processes can lead to multiple types of harm [52, 60], from discrimination and exclusion, to disinformation and misinformation, as well as creating information hazards such as leaking sensitive information or compromising privacy. Our results further imply that the conversational interaction affordance of LLMs may have a reinforcing effect on their information harms, and are especially likely to harm those already with a predisposition towards the biases, misinformation, disinformation or other information hazards produced by LLMs. In other words, potential harms of LLMs should be approached as sociotechnical problem, considering not only the limits of the technology, but also people's interaction behaviors with LLMs and their personal and social contexts.

As large-scale information and knowledge systems, we must also consider LLMs' societal risks through "subjugation" [52]—how they can proliferate dominant views and languages and foreclose alternative ones. A previous study by Jakesch et al. [22] highlights that LLM-powered writing support can exert latent persuasion with biases in its generated content. Our results imply an additional mechanism for such a risk by creating echo chambers. We must consider both dominant views and biases encoded in widely used LLMs, as well as the danger of targeted opinion influence by political or commercial groups that exploit the echo chamber effect. As our Study 2 shows, it can be extremely easy to steer LLMs to exhibit certain biases through adaptation techniques. These biased LLMs can be used for not only conversational search but also writing support, chatbots, and social media bots. Our results suggest that opinion biases in LLMs present risks that may outweigh the potential benefits. We encourage developing technical guardrails to detect opinion biases of LLMs and prevent malicious manipulation of such biases. Policymakers and society at large must also grapple with how to establish norms and regulations to restrain such manipulation and make transparent LLMs' possible opinion biases. We also encourage future work to explore how the embedding of LLM-powered information systems and agents can shape the dynamics of different political and opinion groups and public opinions.

7.5 Limitations and Ethical Considerations

We acknowledge several limitations of our study. First, we adopted a closed-world version of search systems and a highly focused task of essay writing. The results may not be fully generalizable to diverse real-world information-seeking settings with search systems. Second, we acknowledge that there can be alternative implementations of LLM-powered conversational search systems and ways to create opinionated LLMs, and they may exhibit different biases and influences. For example, the RAG architecture utilizes prompt engineering to ensure the synthesized outputs to reflect the views in the retrieved articles in a relatively faithful and balanced fashion. It is possible that without this layer, LLMs are more likely to generate more catering responses that can introduce further biases. In study 2, we manipulated the LLM's bias in both the retrieval and the response generation modules. This approach may create opinionated search systems with strong biases, which may not be as pronounced in other settings especially when biases are "unintended". Second, participants engaged in relatively short search sessions (an average of 3.40 queries per session), and the results may not be generalizable to information-seeking settings that require longer or continuous learning. However, we emphasize the differences in querying behavioral patterns and opinion polarization that can be observed even in such short interaction sessions. Third, as discussed in Section 7.1, our study focuses on information querying behaviors without capturing other information consumption mechanisms such as selective attention, perception, and retention that can further affect opinion polarization. Lastly, it is known that there are individual differences in diversity-seeking tendencies [31, 36] and tendencies to engage in human-like conversational interactions with machines [30]. Our study only looked at participants' behaviors at an aggregate level without considering these individual differences.

We also acknowledge that our results and system design may incur intentional misuses. Overall, we caution against the use of LLMs with opinion biases to power search systems without appropriate risk assessment and oversight, which should be enforced not only through technical guardrails but also norms and policy. We hope our results can inform such guardrails. To prevent misuses, we decide to not make public the prompts we used to generate biased LLM response, but will only make them available for requests that we can verify for safe usage (e.g. scientific and non-commercial purposes).

8 CONCLUSION

The recently developed powerful LLMs have experienced exponential growth in user population and companies have adopted them for various information systems, such as conversational search, open-domain, or specialized chatbots, to various productivity tools such as writing support. These applications have been found to have a profound impact on the way people search and consume information. In this study, through two controlled experiments, we empirically showed: 1) LLM-powered conversational search could lead to increased selective exposure and opinion polarization compared to traditional web search, by inducing more confirmatory querying behaviors; 2) an opinionated LLM that reinforces the user's view could exacerbate the effect, together suggesting the risk of "generative echo chambers". Our study also suggests the limitations of interventions such as providing references and leveraging an LLM that challenges one's existing view, both of which had little effect in reducing selective exposure. With millions of people already being exposed to LLM-powered information technologies, these results call for actions to regulate the use of LLM-powered search systems, develop technical guardrails against misuses of LLMs for opinion influence, and explore mitigation strategies for selective exposure in conversational search.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[2] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. 36–43.

[3] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.

[4] Sandeep Avula, Bogeum Choi, and Jaime Arguello. 2022. The effects of system initiative during conversational collaborative search. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30.

[5] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 2–2.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[7] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[8] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, Vol. 5. 89–96.

[9] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[10] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2023. The AI Ghostwriter Effect: Users Do Not Perceive Ownership of AI-Generated Text But Self-Declare as Authors. *arXiv preprint arXiv:2303.03283* (2023).

[11] Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, communication & society* 21, 5 (2018), 729–745.

[12] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1175–1184.

[13] Tom Feltwell, Gavin Wood, Phillip Brooker, Scarlett Rowland, Eric PS Baumer, Kiel Long, John Vines, Julie Barnett, and Shaun Lawson. 2020. Broadening exposure to socio-political opinions via a pushy smart home device. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[14] Leon Festinger. 1962. Cognitive dissonance. *Scientific American* 207, 4 (1962), 93–106.

[15] Dieter Frey. 1986. Recent research on selective exposure to information. *Advances in experimental social psychology* 19 (1986), 41–80.

[16] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–16.

[17] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

[18] R Kelly Garrett and Paul Resnick. 2011. Resisting political fragmentation on the Internet. *Daedalus* 140, 4 (2011), 108–120.

[19] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[20] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. Avoiding the echo chamber about echo chambers. *Knight Foundation* 2, 1 (2018), 1–25.

[21] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin* 135, 4 (2009), 555.

[22] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[23] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[24] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[25] Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. Chamberbreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

[26] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[27] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *Comput. Surveys* 55, 6 (2022), 1–40.

[28] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 265–274.

[29] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[30] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What can you do? Studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*. 264–275.

[31] Q Vera Liao and Wai-Tat Fu. 2013. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2359–2368.

[32] Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 184–196.

[33] Q Vera Liao and Wai-Tat Fu. 2014. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2745–2754.

[34] Q Vera Liao, Wai-Tat Fu, and Sri Shilpa Mamidi. 2015. It is all about perspective: An exploration of mitigating selective exposure with aspect indicators. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*. 1439–1448.

[35] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. arXiv:2304.09848 [cs.CL]

[36] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.

[37] Sayooran Nagulendra and Julita Vassileva. 2014. Understanding and controlling the filter bubble through interactive visualization: a user study. In *Proceedings of the 25th ACM conference on Hypertext and social media*. 107–115.

[38] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[39] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[40] Changhoon Oh, Jinhan Choi, Sungwoo Lee, SoHyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding user perception of automated news generation system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[41] Raja Parasuraman, Robert Molloy, and Indramani L Singh. 1993. Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23.

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

[43] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442* (2023).

[44] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.

[45] Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ users journal* 18, 6 (2000), 38–43.

[46] Laura Pritschet, Derek Powell, and Zachary Horne. 2016. Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological science* 27, 7 (2016), 1036–1042.

[47] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.

[48] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. CoRR arXiv:2112.11446.

[49] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, et al. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems* 35 (2022), 24720–24739.

[50] Urbano Reviglio. 2019. Serendipity as an emerging design principle of the infosphere: challenges and opportunities. *Ethics and Information Technology* 21, 2 (2019), 151–166.

[51] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 491–514.

[52] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2022. Sociotechnical harms: scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791* (2022).

[53] Jiao Sun, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating explainability of generative AI for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*. 212–228.

[54] Cass R Sunstein. 1999. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper* 91 (1999).

[55] Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. 2013. Escape the bubble: Guided exploration of music preferences for serendipity and novelty. In *Proceedings of the 7th ACM conference on Recommender systems*. 335–338.

[56] Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 32–41.

[57] Jinke D Van Der Laan, Adriaan Heino, and Dick De Waard. 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies* 5, 1 (1997), 1–10.

[58] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2187–2193.

[59] Peter C Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology* 12, 3 (1960), 129–140.

[60] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[61] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection not required? Human-AI partnerships in code translation. In *26th International Conference on Intelligent User Interfaces*. 402–412.

[62] Ziang Xiao, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2023. Inform the Uninformed: Improving Online Informed Consent Reading with an AI-Powered Chatbot. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[63] Ziang Xiao, Q Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2–18.

[64] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[65] Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval* 17, 3-4 (2023), 244–456.

[66] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*. 177–186.

[67] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.