Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-based Search

Pranav Narayanan Venkit
Pennsylvania State University
University Park, USA
pranav.venkit@psu.edu

Philippe Laban
Salesforce AI Research
Palo Alto, USA
phillab@berkeley.edu

Yilun Zhou
Salesforce AI Research
Palo Alto, USA
yilun.zhou@salesforce.com

Yixin Mao
Salesforce AI Research
Palo Alto, USA
y.mao@salesforce.com

Chien-Sheng Wu
Salesforce AI Research
Palo Alto, USA
wu.jason@salesforce.com

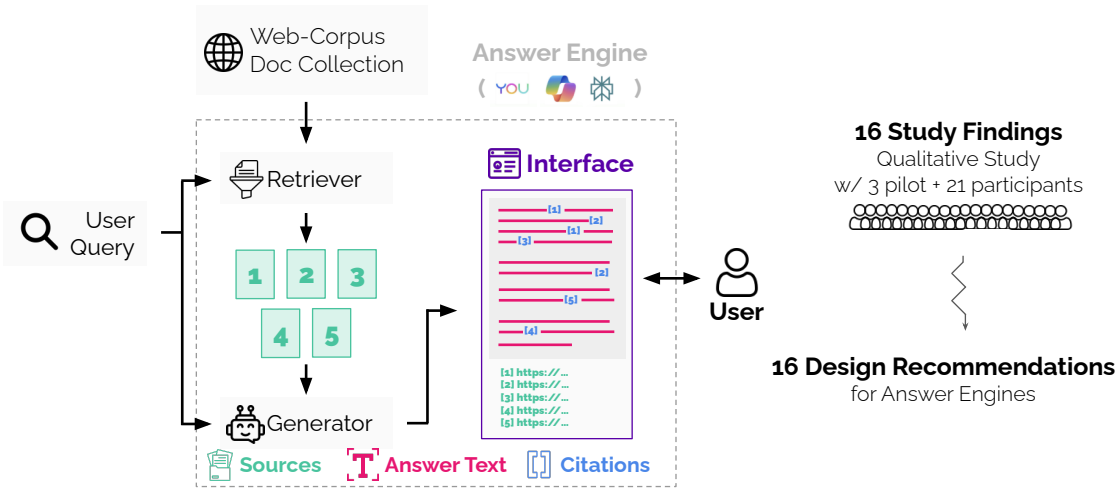


Figure 1: The design pipeline of an answer engine and the study framework used to audit these systems. The figure shows the key components of an answer engine, including how it generates answers based on user queries, with a focus on outputs such as sources, answer text, and citations.

Abstract

As Large Language Model (LLM) applications transition from research to widespread sociotechnical products, they are fundamentally changing how people access and process information. Answer engines - LLM-powered search tools that generate source-cited summaries rather than just returning relevant links - represent a particularly significant shift from traditional search engines. Through a qualitative study of 21 expert users comparing answer engines to traditional search, we identified 16 ethical and societal limitations in current answer engine implementations. We showcase how current LLM-based search engine still lack the neccessary features to be treated as a safe sociotechnical system for public consumption. Based on these findings, we propose a framework of 16 corresponding design recommendations to guide the development of more trustworthy answer engine systems. We showcase of this application is still nascent and how we need to be more sensitive to their social ramifications, to create a better and safer experience for individuals.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732089

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **User studies**; **Empirical studies in HCI**; **Natural language interfaces**; • **Computing methodologies** → **Natural language generation**.

Keywords

Answer Engines, Generative Search Engine, RAG systems, Ethical Audit, Fairness and Ethics

ACM Reference Format:

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-based Search. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732089

1 Introduction

LLMs have recently become part of daily life for many, with services such as ChatGPT and Claude offering AI-based conversational assistance to hundreds of millions of customers [17, 46, 52]. In doing so, such systems have graduated from research tools that were evaluated from a technical standpoint to *sociotechnical systems* [9] that have both technical and social impact, and that require more nuanced evaluation, as they can influence various facets of society, including communication, information dissemination, and decision-making [46, 57].

A prominent example of an LLM-based sociotechnical system is the *Answer Engine*, also known as a Generative Search Engine. Answer Engines are marketed as replacements for traditional search engines – such as Google or Bing – and work in the following retrieval-augmented way: a user with an information need formulates a search query [42]. The system first retrieves relevant source documents that likely contain answer elements to the user's query, leveraging a retrieval system (which can be a traditional search engine). The Answer Engine then composes a textual prompt that contains the user's query, and the retrieved sources, and instructs an LLM to generate a long and self-contained answer for the user, based on the content of the sources. Crucially, citations are inserted into the answer, with each citation linking to the sources that support each statement within the answer. This citation-enriched answer is provided to the user in a user interface: the citation forms the semantic glue between the generated answer and the sources, with a click on a citation allowing the user to navigate to the source or sources that support any statement. In essence, the answer engine promises a streamlining of a user's information-seeking journey [57].

However, there are several well-known limitations to Answer Engines, primarily stemming from the use of LLMs as part of answer generation. First, LLMs are known to hallucinate information and cannot detect factual inconsistencies [30, 62], even when authoritative sources are provided. Second, prior work [33, 38] has also shown limitations of answer engines' ability to assess the accuracy of citations within an answer. Third, LLMs accumulate knowledge within their internal weights during pre-training, and prior work has shown limited success at enforcing that an LLM generates information based solely on documents provided in a prompt rather than based on pretraining information which can be noisy or outdated [32]. Yet prior work has predominantly evaluated LLMs and their output primarily from a technical perspective [16, 33]. Since Answer Engines are used by millions daily, it is important to evaluate them from a social perspective, to understand how users perceive Answer Engines, and how they navigate limitations.

To address this need, we conduct an audit-centric usability study (Section 3) involving 24 participants[1] *with expertise in technical domains (e.g., sociology, economics).* Participants interact with Answer Engines and traditional Search Engines on two query types: *expertise* and *debate* queries. Expertise queries are technical queries that participants self-report being experts on. Participants' familiarity with the answer allows us to evaluate how Answer Engines perform on deeply technical questions. Debate queries are queries related to a debate topic, formulated either to be pro or against the debate (e.g., "Why should we abolish Daylight Savings"). By initially asking participants if they support one side of the debate, we evaluate how participants interact with the answers that support or refute their opinions.

The usability study enables us to obtain two main kinds of insights: (i) quantitative insights on how users interact with answers, citations, and sources in both Answer Engines and traditional Search Engines, (ii) qualitative feedback from participants which we group using inductive reasoning [22, 23] followed by a qualitative coding method [2, 60] into *16 limitations of Answer Engines.* With the study completed, we then propose actionable *16 Design Recommendations* that can measure whether systems make progress progress towards safer interaction of these systems.

2 Related Works

2.1 The Lack in Understanding AI of Today

As AI becomes more embedded in daily life, their role has evolved from simple technical tools to complex sociotechnical systems. These systems involve an intricate interplay between social actors and technological components that together shape goal-oriented behaviors [9, 46, 61, 62]. AI systems in domains like education [29], healthcare [1], and policymaking [6, 53] are thus deeply entwined with the social practices and institutional contexts in which they operate [19]. However, despite the recognition of AI as inherently sociotechnical, current research often adopts a *technocentric* perspective, focusing on algorithmic and computational aspects while neglecting broader societal implications [13, 18, 46, 65]. This gap [5] is evident in the conceptualization of terms like harm [4, 67], sentiment [61], and hallucination [62]. As Venkit et al. [61] argues, understanding AI through a sociotechnical lens is crucial to fully grasp its impacts, biases, and potential harms. Similarly insights from Human-Centered Explainable AI (HCXAI) Ehsan and Riedl [14] advocates for positioning "the human" at the core of technology design, leveraging the social dynamics and context of AI systems to bridge gaps for non-technical users to create safer and trustworthy systems— emphasizing the need for a human-centric approach to technology, focusing on how AI systems can better align with human understanding and accessibility [14, 15]. Adopting this tenet of HCXAI and sociotechnical lens shows how the development of these technologies must account for human factors, moving beyond traditional human-computer interactions, which is the primary motivation of our work.

[1]Pilot study with 3 participants and a final expert oriented usability study with 21 participants

2.2 The Ethical and Societal Shortcomings of Answer Engines

Answer engines are marketed as efficient tools for simplifying information retrieval by reducing the need for users to manually sift through data repositories [54, 56]. However, recent developments, such as Google AI Overview and Perplexity, have exposed new ethical challenges and negative user experiences. For example, Google's answer engine erroneously advised users to "put glue on their pizza," revealing how the system misinterpreted sarcastic content from the internet, presenting it as fact with undue authority[2]. Such cases of misinformation highlight the risks associated with automating information retrieval, especially under the guise of 'Google doing the Googling' for users [54, 56, 66].

The release of OpenAI's 'SearchGPT,' marketed as a 'Google search killer' [56], further exacerbates these concerns. As reliance on these tools grows, so does the urgency to understand their impact. Lindemann [36] introduces the concept of Sealed Knowledge, which critiques how these systems limit access to diverse answers by condensing search queries into singular, authoritative responses, effectively decontextualizing information and narrowing user perspectives [27, 44]. This "sealing" of knowledge perpetuates selection biases and restricts marginalized viewpoints [36]. Building on this, Sharma et al. [58] argues that answer engines foster echo chambers, where exposure to diverse opinions is minimized, reinforcing existing beliefs and reducing the visibility of minority perspectives. This is particularly problematic given the established Western-centric bias in text generation models [20, 45, 47, 64]. When integrated into search engines, these models further propagate a predominantly Western viewpoint or an 'automated colonial impulse' [10, 11, 34], underscoring the need for comprehensive studies on their risks.

Another key concern surrounding answer engines is their inherently 'black-box' nature [3, 50] due to the opacity of their decision-making processes and the hidden biases within them [3, 28, 41]. Answer engines intensify this problem by merging two opaque systems: a search engine [12, 24, 59] and a generative AI model [3], resulting in compounded opacity and reduced user autonomy [21, 51]. This dual-layered opacity leads to problematic outcomes, such as those identified by Li and Sinnamon [35], who revealed sentiment bias based on query content, along with commercial and geographic biases in the sources answer engines use. The over-reliance on uneven quality sources, heavily skewed toward news, media, and business, further illustrates the need for transparency [35]. Without such scrutiny, these systems risk perpetuating biases and misinformation with significant societal implications.

3 Answer Engine Usability Study

3.1 Study Design

Our study protocol is designed as a 90-minute one-on-one session via video conference, which was recorded and transcribed with participants' consent. The study protocol was reviewed and approved by the institution's Ethics Office.

3.1.1 Participant Recruitment. Motivated by Kang et al. [31], Venkit et al. [63] on the use of expert insights in user study and to enable expertise audit of the model results, we recruit participants with technical expertise on specific topics, to allow the evaluation of system responses on complex queries participants have expertise in. Our recruitment criteria targeted experts from diverse academic and professional backgrounds. This recruitment strategy was intentional, allowing us to study how Answer Engines respond to highly technical queries that our participants would have in-depth knowledge about, allowing study insights to reveal how such systems perform on in-depth query and evaluate them accurately. Participants were recruited through a combination of academic channels (via email invitation and LinkedIn), and social media platforms (via Twitter and Reddit). The study was conducted using the *User Interviews platform*[3], and Google Meet for video-conferencing.

We recruited 24 participants aged 22 to 38 years (M=29.3, SD=2.99), with a gender distribution of 66.67% female (n=16), 33.34% male (n=7), and 4.16% non-binary (n=1). Participants' occupations distribution were 45.83% Ph.D. students that are at least three years into their graduate program (n=11), 16.67% research professionals including postdoctoral researchers (n=4), 33.34% industry experts (n=8), and 4.17% other professionals (n=1). Regarding participant expertise, 25.00% were experts in Human-Computer Interaction (n=6), 25.00% in Artificial Intelligence and Computational Research (n=6), 20.83% in Healthcare and Medicine (n=5), 16.67% in Applied Sciences (e.g., Transportation, Meteorology) (n=4), and 12.50% in Education and Social Sciences (n=3).

An initial pilot study with *3 participants* helped refine our methodology and develop a codebook. The final study was conducted with the remaining *21 participants*. All participants were compensated with a $60 gift card. The anonymized participant description and the answer engines interacted with are described in Appendix A.3.

3.1.2 Study Part 1: Pre-Study Questionnaire (5 minutes). Participants completed a questionnaire (exact questions in Appendix A.2) that asked for (1) high-level demographic information, (2) participants' familiarity with answer engines, and (3) a list of 3-4 specific technical questions in their area of expertise that they could query in a search engine.

3.1.3 Study Part 2: Expertise Information Retrieval (40 minutes). During this part, participants go over one technical question at a time from the list they provided in the pre-study questionnaire and alternate between querying it in an answer and a traditional search engine. Participants are asked to "think aloud" as they review results [40, 49], articulating their thoughts and reactions. This approach captures detailed insights of successes and failures of each engine on a concrete query. Participants are encouraged to interact in-depth with the results, including by clicking on external hyperlinks. Once they are done, they proceed with the next question on their list. Participants typically spent 5-10 minutes per query, going through an average of 6 queries in the 40-minute time frame.

3.1.4 Study Part 3: Debate Information Retrieval (40 minutes). Part 3 follows a similar structure to Part 2 but uses opinion-based queries, a common use case for search engines [26, 58]. Participants start with answering a series of questions measuring their agreement with various socially and politically charged statements collected

[2]https://www.theverge.com/2024/5/23/24162896/google-ai-overview-hallucinations-glue-in-pizza

[3]https://www.userinterviews.com/

from the ProCon debate website[4]. Based on their responses, we constructed questions that reflected both supportive and opposing viewpoints. For example, if a participant agreed with the statement "Zoos should exist", a supportive query is: "Why should zoos exist?" and an opposing query is: "Why should zoos not exist?". For each issue where a participant expressed a non-neutral opinion, we prepared either a supportive or opposing query, which the participant ran through both the answer engine and the traditional search engine. We alternated between supportive and opposing queries, allowing the study of how participants interact with responses that align or diverge from their viewpoints.

The 21 participants were divided into three groups of 7. Each group interacted with a single answer engine: YouChat, Bing Copilot, or Perplexity AI. These platforms were chosen due to their public accessibility as AI-as-a-Service (AIaaS) systems and their marketing as answer engines[37]. Google Search was selected for the traditional search engine.

3.2 Pre-Survey Questionnaire Analysis and Results

Regarding frequency of use, 37.5% (9/24) of participants use Generative AI (GAI)-based applications daily, 29.1% (7/24) weekly, 25% (6/24) monthly, and 8.3% (2/24) a few times per month. Regarding the use of answer engines specifically, 70.83% (17/24) of participants were familiar with these systems, 41.16% use them multiple times per week, and 58.84% at least once a month. Participants utilize answer engines to conduct research, brainstorm, plan, learn new skills, and obtain faster results compared to traditional search engines.

3.3 Qualitative Analysis and Results: A Sociotechnical Audit of Answer Engine Shortfalls

We employed a constructivist grounded theory using a qualitative coding approach [8] to analyze our user interview data. Following Charmaz [7], the authors individually coded the transcripts line-by-line, employing inductive reasoning to develop theories [22, 23]. Using the qualitative coding platform Taguette[5], we generated themes from the transcript snippets. These themes were then synthesized and refined through collaborative discussions between the authors, and insights were categorized with respect to the four components of an answer engine: (1) **answer text**, (2) **citation**, (3) **sources**, and (4) **user interface**. The 16 primary findings, with the number of participants who explicitly identified and expressed concerns for each component, is explained as follows (and are summarized in Table 1).

Theme 1: Answer Text Findings

A.I. Need for Objective Detail in Generated Answers *(21/21)*: A pervasive issue across all three answer engines was the lack of detail and contextual depth in generated responses. This shortfall affected both expertise and debate queries. Participants repeatedly found the summaries to be overly generic and insufficient, often driving them to seek more comprehensive information through traditional search engines like Google. Participant P1 noted, *"It was just trying to answer without actually giving me a solid answer or a more thought-out answer, which I am able to get with multiple Google searches."* P10 emphasized, *"It's too short and just summarizes everything a lot. [The model] needs to give me more data for the claim, but it's very summarized."* These reflections highlight a common issue: a desire for answer conciseness leads to *frequently omitted critical details that would substantiate claims.* As a result, responses are perceived as superficial, **lacking the necessary specificity and nuance**.

A.II. Lack of Holistic Viewpoint *(19/21)*: Our study revealed a limitation in the behavior of answer engines when participants engaged with *opinionated queries.* The answer engines frequently aligned with the bias implied in the question, **neglecting to present diverse perspectives** available from the retrieved sources. The responses often appeared to support only the side of the argument the model inferred the participant was "looking for," thereby **reinforcing user biases**. Figure 2 illustrates this by showing an example of a one-sided answer (left, Perplexity.ai) and a comparably more nuanced answer (right, You.com). Participant P19 noted, "*I want to find out more about the flip side of the argument... this is all with a pinch of salt because we don't know the other side and the evidence and facts.*" Similarly, P1 stated, "*It felt like it was trying to just conform to my question, even though the sources that it was citing actually spoke about all the negatives and positives*," indicating a mismatch between the source content and the generated answer.

A.III. Confident Language While Presenting Claims *(16/21)*: Another issue identified is the generation of **overly confident responses**. All three systems frequently used terms of affirmation and certainty, *even when addressing subjective opinions or claims.* This approach can lead participants to trust the generated content without the necessary context, with the problem being highlighted for both debate and expertise queries. Figure 2 illustrates the issue: in **[A]** the answer engine confidently represents information without stating the nuances. Participants highlighted this issue through their interactions with the models. P3 observed that *"the model uses a very magnetic or authoritative voice while making claims,"* which *"can definitely convince someone that this is 'the answer' instead of actually giving them the opportunity to see the issue."* Similarly, P4 noted that the model employs the 'God Voice', likening it to articles that *"make you think that it's the ultimate truth."*

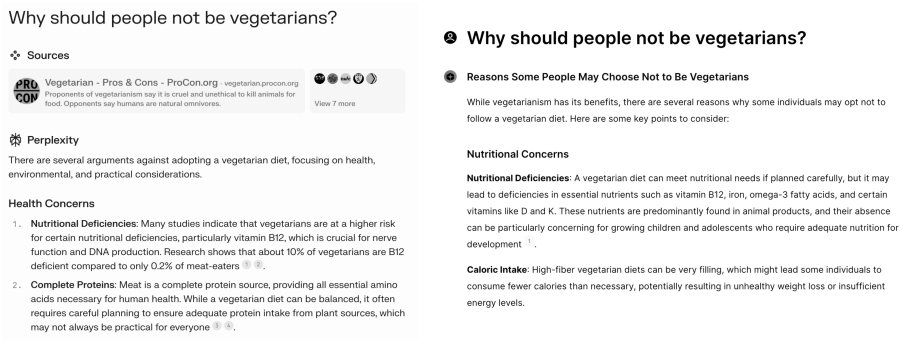
A.IV. Overly Simplistic Writing Form and a Lack of Critical Thinking *(14/21)*:

The fourth finding highlighted by many participants is the simplistic nature of the language used in responses. Participants noted that this simplicity in language reflects a **lack of critical analysis and thinking**. For example, P13 found the answers to be *"kind of childish,"* noting they did not match the scientific level required. P2 described the text as *"fluffy"* and *"similar to what a fifth grader might write without consulting sources"*. Additionally, some participants perceived the systems as **'people pleasers'**, presenting information in a manner that was agreeable or comforting rather than providing an accurate response. P1 noted, *"It's being a people pleaser and only trying to validate me."*

[4]https://www.procon.org/
[5]https://app.taguette.org/

Answer Text	Citation	Sources	User Interface
A.I Need for objective details in generated answers (21/21)	**C.I** Misattribution and misinterpretation of sources cited (21/21)	**S.I** Low Frequency of Sources Used for Summarization (19/21)	**U.I** The lack of selection, and filtering of sources (17/21)
A.II Lack of holistic viewpoints for opinionated or charged questions (19/21)	**C.II** Cherrypicking information based on assumed context (19/21)	**S.II** More sources retrieved than used for generating the actual answer (13/21)	**U.II** Lack of human input in generation and source selection (17/21)
A.III Overtly confident language while presenting claims (16/21)	**C.III** Missing citations for claims and information generated (18/21)	**S.III** Lack of trust in sources used by the answer engine (12/20)	**U.III** Answer engines take additional work to verify and trust (14/21)
A.IV Simplistic language and a lack of creativity and critical thinking (14/21)	**C.IV** Transparency of source selection in model responses (15/21)	**S.IV** Redundancy in source citation and duplicate content retrieved (12/21)	**U.IV** Citations formats are not a normalized interaction (12/21)

Table 1: Summary of key findings, organized thematically by answer engine components, with the number of participants who explicitly identified and expressed concerns for each component.



[A] [B]

Figure 2: Comparison of outputs from [A] Perplexity, which reflects the bias inherent in the question by presenting only a one-sided response, and [B] YouChat, which acknowledges multiple perspectives, avoiding presenting incomplete information.

Theme 2: Citation

C.I. Misattribution of Sources: Correct Summaries, Incorrect Citations *(21/21)*:

A common issue identified in this theme was the **misattribution of sources**. This occurs when the answer engine cites a source that does not factually support the cited statement, misrepresenting the source content. For instance, P12 noted, "*It has cited irrelevant parts of the paper for this question.*" P15 commented, "*But this statement doesn't seem to be in the source. I mean the statement is true; it's valid... but I don't know where it's even getting this information from.*" Participants felt that the systems were **using citations to legitimize their answer**, creating an illusion of credibility. This facade was only revealed to a few users who proceeded to scrutinize the sources. P4 expressed concern, "*I mean it's just like you see a citation, you assume it's a valid source...I'll just see that there is a source. That's it. I don't verify it.*"

C.II. Cherrypicking Information Based on Assumed Context *(19/21)*:

When participants posed expert or opinion-based questions, they noticed that the system often selectively presented information from retrieved sources, highlighting **a particular perspective instead of a comprehensive view of the article**. For instance, P7 noted, "*I feel [the system] is manipulative. It takes only some information and it feels I am manipulated to only see one side of things.*" Similarly, P8 observed, "*[The source] actually has both pros and cons, and it's chosen to pick just the sort of required arguments from this link without the whole picture.*" Participants felt that the model does not accurately represent the full scope of the source material. The tendency to reinforce assumed user biases further contributes to an **echo chamber effect**, limiting exposure to a broader range of perspectives and potentially distorting the original intent of the source material.

C.III. Missing Citation for Claims and Information Generated *(18/21)*:

The absence of citations in many of the sentences generated by *all three answer engines* emerged as a critical issue, especially when **key claims or facts are presented without necessary factual support from the sources**. P16 highlighted the inconvenience caused, noting, "*Not having the references for each sentence is annoying...you want to know what's the resource retrieved is. Now we'll have to actually go to the website and compare notes, which is an additional step which no one wants to do. I would have gone to the website in the first place instead.*" These comments reveal a clear demand for citations, particularly for sentences that are important to answering the question. Figure 3 shows an example of an uncited statement in Perplexity, compared to a similar Copilot cited answer.

C.IV. Lack of Transparency of Source Selection in Model Responses *(15/21)*:

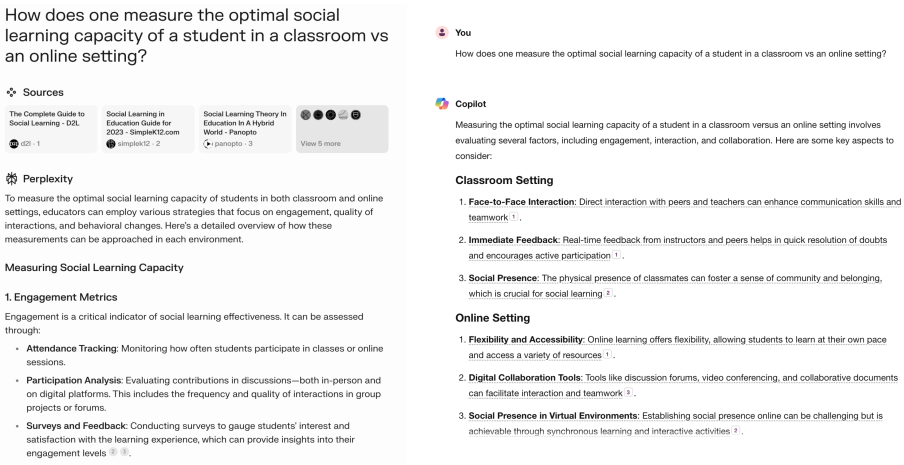


Figure 3: Comparison of outputs from [A] Perplexity, which lacks citations for the points generated, causing confusion on the actual source of each sentence, and [B] Copilot, which effectively indicates the sources for each statement.

Participants raised significant concerns about the **lack of transparency in how the system selects and prioritizes citation**, indicating the need for a clearer explanation of its decision-making process. This "*black box*" nature of the system makes it difficult for users to trust the outputs, as they cannot discern why certain sources are cited over others. Participants frequently noted that the system did not adequately prioritize important or factual sources, leading to a general de-emphasis of critical references. For instance, P4 questioned, "*Where is it getting this thing from? Why is it getting it from these particular sources is what I'm curious about.*" Additionally, P2 expressed frustration with the system's opaque process , stating, "*It's very easy to just cough out sources and be like, 'this is where I took all this information from.' But which part of the information did you take from? That kind of explanation doesn't exist.*"

Theme 3: Sources

S.I. Low Number of Sources *(19/21)*:

For both expert and opinionated questions, participants encountered experiences where they **needed more sources to address the question** at hand. Participants highlighted this issue with specific feedback. For instance, P16 remarked, "*It feels like it's pulling all of this from one source*," while P1 noted, "*Again, everything is extracted from the same source, which is very weird.*" This indicated a pattern where the answer engine heavily relied on a limited number of sources, *averaging to three sources used*, often leading to incomplete answers. Interestingly, this issue also caused a phenomenon where the model **overtly emphasized certain sources for generation**. Participants flagged this as a consequence of the models using very few sources for their answer. P5 mentioned, "*If it's like citing the right review paper, it can get away with citing only one [source], but it isn't doing that and citing one weak article throughout.*"

S.II. More Sources Listed Than Used (13/21):

Participants using *Bing Copilot and Perplexity* noted that these systems often **listed multiple retrieved sources without actually citing them in the generated answer**, a behavior described as "buffing"—creating an impression of thoroughness without substance. This practice led to confusion and eroded trust, as users saw citations that were not integrated into the generated response. For example, P12 remarked, "*It's giving the impression of multiple sources, but it's just citing something that has a blurb citing to the other source. So it's really just coming off of this one source.*" This selective use of sources caused frustration, as participants believed the models ignored more reliable or relevant articles, diminishing the perceived quality of the generated content. Notably, however, YouChat did not display this specific weakness and consistently cited all listed sources in its responses.

S.III. Lack of Trust in Source Types (12/20):

The answer engines retrieve content from varying sources, including forums, blogs, opinion pieces, and research articles. However, participants expressed general distrust toward certain sources due to perceived biases, or lack of authority. This distrust was evident in feedback like P1's remark, "*Who knows who has written that post [...] it's a LinkedIn post when the question is a scientific one,*" and P2's observation, "*The sources are not convincing. They seem to be these non-factual sources from where this answer is kind of drawn.*" Additionally, participants noted issues with **outdated or misinformed content** being used to generate answers. For example, P10 mentioned, "*I think the citation is outdated... because it says 'Windows 10', and we've already switched to better OS, which is what the answer needed*".

S.IV. Different Sources but Duplicate Content (12/21):

Participants identified instances where the answer engine retrieved and cited multiple sources that, upon closer inspection, contained identical or highly similar content. While these sources were presented as distinct entities with different citation numbers, they **ultimately contributed redundant information**. For example, P3 observed, "*Source 1 and 2 are just the same content in different formats. This is funny as it's using them differently,*" and P9 stated, "*I think this is a big problem! I should have checked the citations on the other answer as well because it's basically just giving*

me one website but citing it differently multiple times." As a result, the system provides a misleading appearance of a well-rounded answer while simply recycling the same content. Figure 4 provides an example of duplicate sourcing.

Theme 4: User Interface

U.I. Autonomy over Source Validation and Selection *(17/21)*: Participant feedback underscored that answer engines often offer narrow perspectives, due to users having little to no control over the information presented to them, leading to concerns about a lack of autonomy and the inability to evaluate the credibility of sources independently. P11 elaborates, "*[The answer engine] is taking from these sources, but are they even legit? I cannot choose or remove the ones it chose. At least with Google, I have a lot more choices to click.*" The inability to select, filter, or assess sources independently not only limits the breadth of information but also undermines user confidence in the accuracy and reliability of the answers provided.

U.II. Lack of Human Input in Generation and Source Selection *(17/21)*: Our study identified a key issue with answer engines: the **tendency to lose context when generating responses** due to the answer engine assuming the most probable context rather than accurately interpreting or clarifying the specific context. Participant P7 highlighted this issue: "*I think the sources were right, but the context is lost. It did not manage to go to the expected answers at all.*" Participants suggested that the single-interaction design contributes to this issue. In cases where context is critical, participants recommended that **the system should adopt a more interactive, conversational approach by asking questions before answering**. By asking clarification questions, answer engines could better address ambiguity and provide more accurate responses. P1 also remarked, "*Having [the system] ask further questions to alleviate any potential ambiguity on the generation would have helped, instead of assuming the context.*"

U.III. Additional Work to Verify and Trust Answers *(14/21)*: As described in previous findings, participants often felt compelled to independently verify the provided sources due to distrust. Therefore, contrary to their intended purpose, some participants found that **the answer engines often resulted in more work for users, undermining their marketed goal of simplifying information retrieval**. P17 echoed this, stating, "*We have to go to the website and compare notes and all of that [for verification], which is an additional step which no one wants to do. I would have just gone to the website in the first place.*" The necessity to cross-check each source individually added complexity, extending the time required to complete tasks that might have been faster with a traditional search engine. P5 emphasized the inefficiency: "*The [answer engine] does not speed up the process. I don't feel like I would trust it enough to just go off that.*"

U.IV. Citation Formats are Not a Normalized Interaction *(12/21)*: The format of citations affects user experience. The most common method involves numerical citations (e.g., [1], [1][2]), where numbers correspond to references listed below the content. While familiar to those accustomed to academic writing, this format can be **less intuitive or effective for individuals who do not regularly engage with such reference systems**. Participant P10 highlighted this issue: "*I actually really don't understand the citations because in my job and daily life, it isn't used.*" This concern suggests that numerical citations, while clear to some, may not be meaningful to a general audience unfamiliar with academic or research practices. However, participants who interacted with Bing Copilot, which uses a hover-over feature to display source information, reported a better understanding of where the information came from. This feature **encouraged direct interaction with the sources in a more accessible manner**. Participant P2 commented, "*[The on-hover] is useful to have. Sources here, I think it will help more people to feasibly check for sources every time they're taking something from a system like this as they are forced to interact with it.*" We have identified the most agreed upon findings from our participant interactions in this section. Additionally, minor findings are included in Appendix A.4.

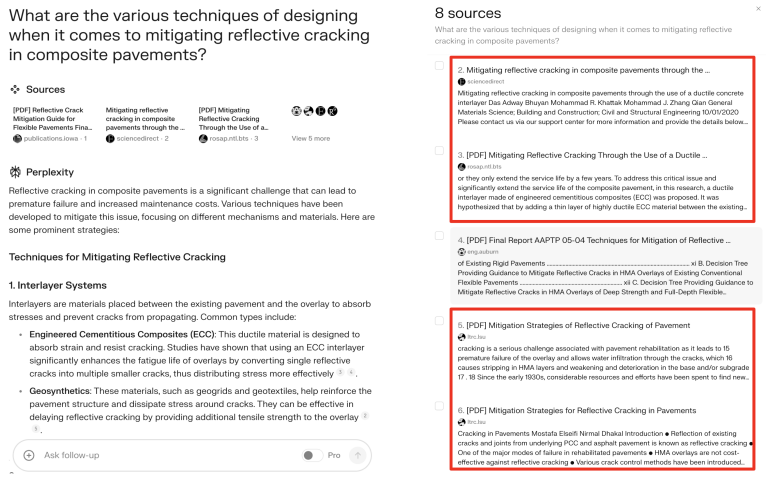
3.4 Quantitative Analysis and Results: Source and Citation Interaction Evaluation

To complement qualitative insights, we investigate the number of sources displayed and cited by the three answer engines and participants' interaction with these sources. Our study revealed that the chosen answer engines cite a limited subset of sources displayed. Table 2 illustrates the disparity between sources retrieved **(mean: 4.31)** and sources cited **(mean: 3.0)** in the final answers, varying across answer engines. Out of the three answer engines, Perplexity displays the most sources (mean: 5.00), but cites the least (mean: 2.58), while YouChat cites all the sources that are displayed (mean: 3.57). With *an average of only three sources available*, users have limited autonomy in selecting and verifying information.

Further investigation into user interaction with sources reveals a stark contrast in how individuals **hover and click sources** when using answer engines versus traditional search engines. As shown in Table 3 and Figure 5, participants display a much narrower scope of inquiry when leveraging answer engines, **hovering** an average of only 2 sources *(median: 2; SD: 1.39)*. This limited engagement likely results from the fewer sources provided by answer engines, which restricts the breadth of information users can explore. In contrast, participants using traditional search engines adopt a more comprehensive approach, hovering an average of 12 sources *(median: 11; SD: 3.56)*. Additionally, there is a notable difference in the number of **sources clicked** to thoroughly analyze to find answers. With answer engines, users tend to click on a single source and trust the model's selection. However, when using traditional search engines, participants engage with a wider range of sources, clicking and analyzing an average of 4 sources.

We conducted independent two-sample t-tests to statistically assess our findings, comparing user interactions during traditional search engine use with those observed when using each answer engine. The significance level was set to $\alpha = 0.01$. As shown in Table 4, the results reveal statistically significant differences in user interactions, allowing us to reject the null hypothesis that answer engines foster similar interactions as traditional engines.

Finally, our analysis in Table 5 examined how individuals interact with answer engines when asking questions that align with or challenge their acknowledged stance. When faced with contradictory questions, participants engaged in more analysis, hovering (mean = 1.72) and clicking (mean = 2.95) a higher number of sources.



[A] [B]

Figure 4: Results generated by Perplexity [A] and the corresponding sources retrieved [B]. The image illustrates how the model retrieved 8 sources, many of which are duplicates of the same source. Despite this, the model cites them differently, creating an illusion of varied content when it is actually the same.

	Sources Displayed by Answer Engine				Sources Cited by Answer Engine			
	All	Perplexity	Bing Copilot	YouChat	All	Perplexity	Bing Copilot	YouChat
Mean	**4.31**	**5.00**	**4.46**	**3.57**	**3.00 (69%)**	**2.58 (51%)**	**2.80 (62%)**	**3.57 (100%)**
Median	5.00	5.00	4.00	4.00	3.00	3.00	3.00	4.00
SD	1.32	0.00	1.56	1.25	1.15	0.65	1.18	1.25
Max	8.00	5.00	8.00	6.00	6.00	4.00	6.00	6.00
Min	1.00	5.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Sources Retrieved vs Cited by the three answer engines evaluated in the study. The percentage of sources cited is mentioned in '( )' to identify the subset of sources actually cited or used for the generated answer, for each answer engine.

	Citations Hovered by Users				Citations Clicked by User			
	Perplexity	Bing Copilot	YouChat	Google	Perplexity	Bing Copilot	YouChat	Google
Mean	2.12	2.10	2.00	**11.81**	1.29	0.76	1.07	**3.80**
Median	2.00	2.00	2.00	**11.00**	1.00	1.00	0.50	**4.00**
SD	1.39	1.21	1.72	**3.56**	0.90	0.89	1.41	**0.96**
Max	5.00	4.00	6.00	**24.00**	3.00	4.00	5.00	**6.00**
Min	0.00	0.00	0.00	**6.00**	0.00	0.00	0.00	**1.00**

Table 3: Citations Hovered versus Clicked by a participant in a traditional search engine and the chosen answer engines. The tradiditional search engine results are highlighted to differentiate the system and its result.

	Citation Hovered by Users				Citations Clicked by User			
	T vs All	T vs Perp	T vs Copilot	T vs You	T vs All	T vs Perp	T vs Copilot	T vs You
t-statistic	22.94	13.00	14.59	14.00	17.13	11.40	15.06	11.43
pvalue	8.14e-53	1.56e-23	1.73e-27	5.06e-26	3.21e-28	5.08e-20	1.66e-28	2.50e-20

Table 4: Independant two sample t-test between the citations hovered and clicked between traditional search engine (T) and each of the answer engines (All Answer Engine (All), Perplexity (Perp), Bing Copilot (Copilot) and You Chat (You))

In contrast, when asking aligned questions, participants relied on fewer sources (hovering: mean = 1.08; clicking: mean = 0.48) and often trusted information with lesser verification. The two sample t-tests confirmed a significant difference in verification behaviors between the two conditions (hovering; p-value = 4.93e-06, clicking; p-value = 4.88e-05).

4 The Greater Impact of LLM-based Search on Society: A Discussion

We now reflect on the societal impacts of answer engines, based on the findings and data collected from user interaction and feedback. The main themes of discussion are as follows:

Lack of autonomy and choice afforded to users by these systems: Shah and Bender [57] argues that information retrieval is

	Citation Hovered		Citations Clicked	
	Aligned Q	Disaligned Q	Aligned Q	Disaligned Q
Mean	1.08	2.95	0.48	1.72
Median	1.00	3.00	0.00	2.00
SD	1.25	1.21	0.77	1.12
Max	4.00	6.00	3.00	5.00
Min	0.00	0.00	0.00	0.00

Table 5: Citations hovered versus clicked by a participant in settings where they asked question aligned/disaligned from their acknowledged biases, across all three systems combined.

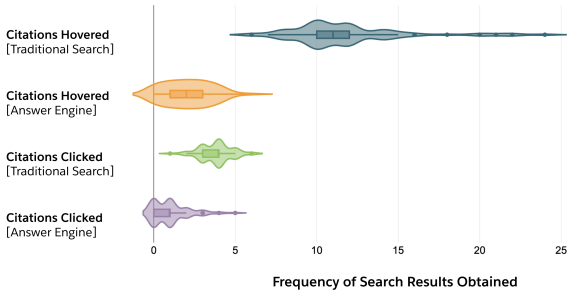


Figure 5: Violin plot showcasing the distribution of number of sources hovered and clicked on by participants for Traditional Search versus Answer Engines.

fundamentally a human-centered activity, underscoring the importance of human agency in seeking information, noting that "information access is not merely an application to be solved by the so-called 'AI' techniques du jour." Our participants echoed this, expressing concerns over the erosion of autonomy when using answer engines. They felt compelled to rely on a single, system-generated answer, limiting their ability to explore multiple sources. While answer engines may simplify information retrieval, they do so at the cost of user autonomy and choice, which raises broader societal concerns. Preserving human agency must be central despite the convenience of AI-driven solutions.

Biased Viewpoints and the Automation of Echo Chambers: Our study revealed a significant bias in responses to opinion-based or debate-oriented queries. These systems often reduce topics with multiple viewpoints to one-sided summaries, reflecting biases from the participant or the system itself. This reductionist approach automates and intensifies the *echo chamber effect*, prioritizing biased viewpoints over factuality and nuance. P2 noted, *"Even if I don't have those opinions, it's going to think I have those opinions,"* highlighting the risk of reinforcing biases that do not align with the user's beliefs. This highlights a fundamental challenge in the design of answer engines: the heavy reliance on user input, which can inadvertently lead to biased outputs.

Self-regulating Populism in Information Retrieval: A significant broader implication identified is the *populistic approach in generative search engines*, which reflects *'sealed knowledge'* [36]. This approach relies mainly on popular or highly ranked search results, often marginalizing less visible but valuable information [36]. Populism here refers to prioritizing widely accepted perspectives while sidelining alternative views [25, 64]. Our findings show that participants using traditional search engines often explore beyond the top five results (Table 3), a behavior not mirrored in answer engines, which prioritize content from the top two or three results (Table 3). This reliance on ranking algorithms, known to be biased [24, 39, 43], skews information by amplifying popular viewpoints and sidelining minority perspectives.

The Lack of Critical Thinking to Trust and Verify: The rise of answer engines, which provide instant, summarized responses, is reshaping how society interacts with information, raising concerns about the erosion of critical thinking skills. Traditionally, searching for information involves cognitive steps like identifying sources, parsing content, and rejecting unreliable information [57]. This process fosters control over information flow, discernment of quality, and critical engagement. Shifting from active search to passive reception of answers can undermine these cognitive processes. P2 expressed, *"If I start using this regularly, at some point, my heuristic is going to change where I won't check the sources at all. Then my writing and decisions, in fact, will not be mine anymore."* As these systems become more integrated into daily life, it is important to ensure they do not erode critical thinking and independent inquiry skills. This unnecessary need to automate such experiences is aptly captured by Salvaggio [55] stating, "the productivity myth suggests that anything we spend time on is subject to automation...implying that the goal of writing is merely to fill a page, rather than to engage in the deeper process of thought that a completed page represents" [55].

Lack of Interaction and Revenue to Actual News and Media Sources: A noteworthy concern identified in our findings is the potential economic impact on these sources, which rely heavily on web traffic as a key revenue stream. The development and increasing use of answer engines raise concerns that users might completely stop visiting the original websites, diminishing the revenue streams that support journalism and content creation, as seen with the very few sources that participants interact with [66]. Participant P9 voices out the issue: *"There is the problem of the websites,*

Design Recommendation	Associated System Weakness
Provide balanced answers	Lack of holistic viewpoints for opinionated questions [A.II]
Provide objective detail to claims	Overly confident language when presenting claims [A.III]
Minimize fluff information	Simplistic language and a lack of creativity [A.IV]
Reflect on answer thoroughness	Need for objective detail in answers [A.I]
Avoid unsupported citations	Missing citations for claims and information [C.III]
Double-check for misattributions	Misattribution and misinterpretation of sources cited [C.I]
Cite all relevant sources for a claim	Transparency of source selected in model response [C.IV]
Listed & Cited sources match	More sources retrieved than used [S.II]
Give importance to expert sources	Lack of trust in sources used [S.III]
Present only necessary sources	Redundancy in source citation [S.IV]
Differentiate source & LLM content	More sources retrieved than used for generation [S.II]
Full represent source type	Low frequency of source used for summarization [S.I]
Incorporate human feedback	Lack of search, select and filter [U.I]
Implement interactive citation	Citation formats are not normalized interactions [U.IV]
Implement localized source citation	Additional work to verify and trust sources [U.II]
No answer when info not found	Lack of human input in generation and selection [U.I]

Table 6: Sixteen design recommendations for answer engines. The recommendations derive from the findings of our usability study, summarized in the middle column with corresponding findings [ID]. Appendix A.5 defines each recommendation.

which are probably human-created, not getting the advertising revenue... increasingly, there is no way that people will now visit these sites anymore." Recent developments further show that answer engines often avoid subscription-based content like that from The New York Times, limiting user access to high-quality, paywalled information. This exclusion places premium content providers at a disadvantage. In some cases, our study found that certain systems, like Perplexity, still retrieved content from subscription-based sites, effectively bypassing the paywall and providing users with information from these sources without requiring them to visit the site or engage with the paywall. This was actively seen in recent news [48] where Perplexity generated information from Forbes, a website that requires a subscription to access. As answer engines continue to evolve, there is a risk that the financial viability of high-quality journalism could be compromised, particularly for outlets that depend on subscriptions and advertising revenue.

Absence of Policy Governing How Generative Models Affect Society: Our findings finally culminate in a critical issue: the absence of robust policy governance for generative models like answer engines. The lack of clear regulations, especially for systems functioning as sociotechnical entities, poses significant risks to individuals and society. Participants frequently highlighted the need for better governance. The lack of policy also raises concerns about privacy and data security, with participants recognizing the risks but noting a gap in understanding how to mitigate them. Our study also revealed concerns about the environmental impact of generative models, especially regarding energy consumption and carbon footprint. P21 noted: *"The whole energy consumption and stuff behind it. It's a point of real concern...it is really scary."* This highlights a broader issue often overlooked: the sustainability of these technologies.

4.1 Answer Engine Design Recommendations

Based on the discussion and the study findings, we design ctionable set of design recommendations for answer engine development to address these issues. Table 6 provides the mapping between the study findings and the design recommendations. Appendix A.5 provides the detailed definition of each recommendation.

5 Conclusion

We presents a systematic audit of answer engines that bridges technical evaluation with critical societal considerations. Through our audit-based usability study, we uncovered fundamental limitations in how answer engines process and respond to queries, particularly concerning algorithmic bias, lack of transparency, and potential societal harms. Based on these findings, we developed 16 concrete design recommendations to guide the development of more equitable systems. Our analysis demonstrates that while answer engines show promise, their current implementation as social systems requires careful scrutiny. We emphasize the shared responsibility among users, researchers, and policymakers to critically examine these technologies and advocate for the integration of our recommendations in future developments.

6 Positionality Statement

As researchers with backgrounds in NLP, human-computer interaction, and socioinformatics, we approach this study with a deep commitment to ethical considerations in AI. Our work is grounded in the belief that AI systems must be designed to promote transparency, fairness, and societal well-being, taking into consideration the users using the system. We acknowledge that our perspectives are shaped by our prior experiences working in both academia and industry, particularly in designing trustworthy AI systems at our respective institutions.

7 Ethics Consideration and Adverse Impact Statement

A significant limitation of our work is its Western-centric focus. The answer engines we evaluated were primarily developed and deployed in U.S.-centric contexts, which influenced both the scope of our audit and the recruitment process. While we included diverse participants within this framework, the study inherently reflects Western perspectives due to the demographics of those who responded to our invitations.

Additionally, answer engines often rely on location-specific data, and the models we evaluated were no exception. Our findings may not fully apply to users in other regions, particularly the Global South, where cultural, social, and technological contexts differ. Future research should explore how these systems perform in non-Western settings.

We recognize that while our recommendations, derived from our human-centric usability study, serves as a valuable ethics sheet, it is not the definitive gold standard. There is ample room for growth and refinement, and we are committed to continually updating and improving the study, such as developing a usable evaluation benchmark to automate the recommendations to a usable artifact.

Finally, as the AI landscape rapidly evolves, model behaviors are frequently updated. Our findings reflect the state of answer engines as of October 2024, but these systems' behaviors may change. Therefore, our results should be viewed in the context of this dynamic and shifting field.

References

[1] Sabira Arefin. 2024. AI Revolutionizing Healthcare: Innovations, Challenges, and Ethical Considerations. *MZ Journal of Artificial Intelligence* 1, 2 (2024), 1–17.

[2] Carl Auerbach and Louise B Silverstein. 2003. *Qualitative data: An introduction to coding and analysis*. Vol. 21. NYU press.

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[4] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 948–958.

[5] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[6] Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, et al. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus* 3, 6 (2024).

[7] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[8] Kathy Charmaz. 2017. Constructivist grounded theory. *The Journal of Positive Psychology* 12, 3 (2017), 299–300.

[9] Robert Cooper and Michael Foster. 1971. Sociotechnical systems. *American Psychologist* 26, 5 (1971), 467.

[10] Dipto Das. 2023. Studying Multi-dimensional Marginalization of Identity from Decolonial and Postcolonial Perspectives. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 437–440.

[11] Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The"Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[12] Tim De Jonge and Djoerd Hiemstra. 2023. UNFair: search engine manipulation, undetectable by amortized inequity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 830–839.

[13] Sarah Dean, Thomas Krendl Gilbert, Nathan Lambert, and Tom Zick. 2021. Axes for sociotechnical inquiry in AI research. *IEEE Transactions on Technology and Society* 2, 2 (2021), 62–70.

[14] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*. Springer, 449–466.

[15] Upol Ehsan, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Sunnie SY Kim, Niels Van Berkel, Andreas Riener, and Mark O Riedl. 2024. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[16] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217* (2023).

[17] Emilio Ferrara. 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* (2024), 1–21.

[18] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 205–216.

[19] Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776* (2024).

[20] Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. *arXiv preprint arXiv:2407.14779* (2024).

[21] Talia B Gillis, Vitaly Meursault, and Berk Ustun. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 377–387.

[22] Barney Glaser. 1992. Basics of grounded theory analysis: Emergence vs forcing. (1992).

[23] Barney Glaser and Anselm Strauss. 1967. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[24] Eric Goldman. 2005. Search engine bias and the demise of search engine utopianism. *Yale JL & Tech.* 8 (2005), 188.

[25] David Grant. 2025. *Populism, Artificial Intelligence and Law: A New Understanding of the Dynamics of the Present*. Taylor & Francis.

[26] Jutta Haider and Olof Sundin. 2019. *Invisible search and online search engines: The ubiquity of search in everyday life*. Taylor & Francis.

[27] Donna Haraway. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*. Routledge, 455–472.

[28] Carol Mullins Hayes. 2023. Generative artificial intelligence and copyright: Both sides of the Black Box. *Available at SSRN 4517799* (2023).

[29] Wayne Holmes and Ilkka Tuomi. 2022. State of the art and practice in AI in education. *European Journal of Education* 57, 4 (2022), 542–570.

[30] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[31] Ruogu Kang, Wai-Tat Fu, and Thomas George Kannampallil. 2010. Exploiting knowledge-in-the-head and knowledge-in-the-social-web: Effects of domain expertise on exploratory search in individual and social search environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 393–402.

[32] Navreet Kaur, Monojit Choudhury, and Danish Pruthi. 2024. Evaluating Large Language Models for Health-related Queries with Presuppositions. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 14308–14331. https://aclanthology.org/

2024.findings-acl.850
[33] Philippe Laban, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems. *arXiv preprint arXiv:2407.01370* (2024).
[34] Alina Leidinger and Richard Rogers. 2023. Which Stereotypes Are Moderated and Under-Moderated in Search Engine Autocompletion?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1049–1061.
[35] Alice Li and Luanne Sinnamon. 2024. Generative AI Search Engines as Arbiters of Public Knowledge: An Audit of Bias and Authority. *arXiv preprint arXiv:2405.14034* (2024).
[36] Nora Freya Lindemann. 2024. Chatbots, search engines, and the sealing of knowledges. *AI & SOCIETY* (2024), 1–14.
[37] Sebastian Lins, Konstantin D Pandl, Heiner Teigeler, Scott Thiebes, Calvin Bayer, and Ali Sunyaev. 2021. Artificial intelligence as a service: classification and research directions. *Business & Information Systems Engineering* 63 (2021), 441–456.
[38] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 7001–7025.
[39] Patrick Maillé, Gwen Maudet, Mathieu Simon, and Bruno Tuffin. 2022. Are search engines biased? Detecting and reducing bias using meta search engines. *Electronic Commerce Research and Applications* (2022), 101132.
[40] Sharon McDonald, Helen M Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication* 55, 1 (2012), 2–19.
[41] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1699–1710.
[42] Shahan Ali Memon and Jevin D West. 2024. Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707* (2024).
[43] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring search engine bias. *Information processing & management* 41, 5 (2005), 1193–1205.
[44] Rainer Mühlhoff. 2018. Digitale Entmündigung und User Experience Design. *Leviathan* 46, 4 (2018), 551–574.
[45] Tarek Naous, Michael Joseph Ryan, and Wei Xu. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:258865272
[46] Pranav Narayanan Venkit. 2023. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 1004–1005.
[47] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 554–565.
[48] Casey Newton. 2024. How to stop perplexity and save the web from bad ai. https://www.platformer.news/how-to-stop-perplexity-oreilly-ai-publishing/
[49] Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive systems*. 209–218.
[50] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
[51] Brooke Perreault, Johanna Hoonsun Lee, Ropafadzo Shava, and Eni Mustafaraj. 2024. Algorithmic Misjudgement in Google Search Results: Evidence from Auditing the US Online Electoral Information Environment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 433–443.
[52] Sanjeev Pulapaka, Srinath Godavarthi, and Dr Sherry Ding. 2024. GenAI and the Public Sector. In *Empowering the Public Sector with Generative AI: From Strategy and Design to Real-World Applications*. Springer, 31–43.
[53] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 836–898.
[54] Kylie Robison. 2024. Google promised a better search experience - now it's telling us to put glue on our pizza. https://www.theverge.com/2024/5/23/24162896/google-ai-overview-hallucinations-glue-in-pizza
[55] Eryk Salvaggio. 2024. Challenging the myths of Generative AI. https://www.techpolicy.press/challenging-the-myths-of-generative-ai/
[56] Deepa Seetharaman. 2024. https://www.wsj.com/tech/ai/openai-search-engine-searchgpt-97771f86
[57] Chirag Shah and Emily M Bender. 2024. Envisioning information access systems: What makes for good tools and a healthy Web? *ACM Transactions on the Web* 18, 3 (2024), 1–24.
[58] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[59] Pedro Silva, Bhawna Juneja, Shloka Desai, Ashudeep Singh, and Nadia Fawaz. 2023. Representation online matters: Practical end-to-end diversification in search and recommender systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1735–1746.
[60] Elizabeth A St. Pierre and Alecia Y Jackson. 2014. Qualitative data analysis after coding. 715–719 pages.
[61] Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J Passonneau, and Shomir Wilson. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13743–13763.
[62] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. " Confidently Nonsensical?": A Critical Survey on the Perspectives and Challenges of'Hallucinations' in NLP. *arXiv preprint arXiv:2404.07461* (2024).
[63] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An Audit on the Perspectives and Challenges of Hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 6528–6548.
[64] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 116–122.
[65] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. 26–34.
[66] PJ Vogt. 2024. How much glue should be in your pizza? https://pjvogt.substack.com/p/how-much-glue-should-be-in-your-pizza
[67] 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.

A Appendix

A.1 Study Design Illustration

A.2 Pre-Study Questionnaire

This pre-study questionnaire was designed to assess participants' eligibility and gather additional information about their usage of answer engines and generative AI in their daily lives. The questions from the survey below:

(1) **Please Enter Your Name**
Short answer response
(2) **Please Enter Your Age**
Short answer response
(3) **Please Select Your Gender**
- Male
- Female
- Non-Binary
- Genderfluid
- Agender
- Bigender
- Other

(4) **What is your current profession or job title?**
Short answer response
(5) **What is your current or most recent educational qualification?**
- High School Degree or Equivalent
- Some College, No Degree
- Associate Degree
- Bachelor's Degree
- Master's Degree
- Doctorate Degree

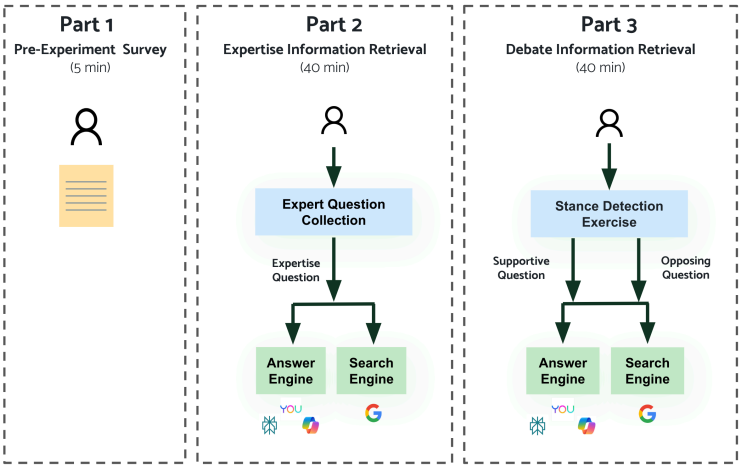


Figure 6: High-level diagram of the three parts to the 90-minute usability study we conducted, and the work that derives from study findings: design recommendations, and the Answer Engine Evaluation (AEE) framework.

- Other

(6) **What do you consider are your current field of expertise? Mention at least 3 topics. (Eg: Multilingual Text Summarization, STEM Education Enhancement, Geology, VR Gaming, etc.)**
Long answer response

(7) **You are familiar with the concept of RAG (Retrieval Augmented Generation) models or Advanced Search Engine or Generative Search Engine (like perplexity.ai or Bing chat).**
- Strongly Disagree
- Disagree
- Neutral
- Agree
- Strongly Agree

(8) **How often do you interact with Generative AI models? (ChatGPT, Stable Diffusion, DALL-E)**
- Every Day
- Several Times a Week
- Several Times a Month
- Rarely
- Never

(9) **How often do you interact with Generative Search Engine like perplexity.ai, Bing Copilot, Google AI Overview?**
- Every Day
- Several Times a Week
- Several Times a Month
- Rarely
- Never

(10) **How do you use or foresee using Generative Search Engine?**
Long answer response

(11) **Please provide any one of your public social media ID for verification purposes (Eg: LinkedIn, Google Scholar, Company or Institutional profile, etc.)**
Short answer response

(12) **To participate in this study, you will need a stable and good internet connection, a device that can share your screen, and a quiet environment for the entire session. Can you meet these requirements?**
- Yes
- No

A.3 Participant Data Information

We present an anonymized list of participants, in Table 7, highlighting the diversity of expertise and the answer engines utilized by each participant throughout the study.

A.4 Additional Community-Centric Themes of Answer Engine Shortfalls

L.V. Amplification of Western Context in Generation *(12/21)*: In some of the user studies, participants keenly observed that the contextual assumptions generated by all the three selected answer engine predominantly **reflect a Western perspective, even when such assumptions are not explicitly stated**. For instance, questions like "Should a government provide universal health care?" or "Should we be vegetarians?" were interpreted as references to the **US government and Western cultural norms**, regardless of the user's actual context. Participants found this problematic, noting that not all contexts should be assumed to align with a Western viewpoint.

Participants expressed concerns, with P9 stating, *"Yeah, it is interesting that it immediately in the second sentence goes to the US governmental context which I didn't specify."* Similarly, P19 noted, *"Even for the previous answer, we didn't mention the country in context and it just took the US context automatically."* P8 also voiced out with an example stating, *"For questions on agriculture, it does not give any geographic specifications of other countries like India, for example, where the answer is very cultural and different."*

While acknowledging that the models are primarily developed and used in the United States, participants emphasized the need for these models to recognize and explicitly address this bias, rather

Participant ID	Field of Work	Occupation	Model Used
P1	Human-Computer Interaction	Doctorate Student	Perplexity AI
P2	Human-Computer Interaction	Doctorate Student	Perplexity AI
P3	Healthcare and Medicine	Doctorate Student	Perplexity AI
P4	Human-Computer Interaction	Doctorate Student	Perplexity AI
P5	Meteorology and Climate Science	Postdoctoral Researcher	Perplexity AI
P6	Human-Computer Interaction	Postdoctoral Researcher	Bing Copilot
P7	Education and Social Sciences	Doctorate Student	Bing Copilot
P8	Transportation Engineering	Doctorate Student	Bing Copilot
P9	Education and Social Sciences	Doctorate Student	You Chat
P10	Information Science	Program Manager	You Chat
P11	Healthcare and Medicine	Doctorate Student	You Chat
P12	Human-Computer Interaction	Postdoctoral Researcher	You Chat
P13	Artificial Intelligence	Research Scientist	Bing Copilot
P14	Education and Social Sciences	Doctorate Student	Perplexity AI
P15	Healthcare and Medicine	Doctorate Student	Bing Copilot
P16	Healthcare and Medicine	Doctorate Student	You Chat
P17	Healthcare and Medicine	Doctorate Student	Bing Copilot
P18	Artificial Intelligence	Research Scientist	Perplexity AI
P19	Human-Computer Interaction	Postdoctoral Researcher	You Chat
P20	Artificial Intelligence	Research Scientist	Bing Copilot
P21	Public Services	Medical Practitioner	You Chat
P22	Artificial Intelligence	Research Scientist	Bing Copilot
P23	Artificial Intelligence	Research Scientist	Perplexity AI
P24	Artificial Intelligence	Research Scientist	Perplexity AI

Table 7: Overview of participants' anonymized information, including their professional field and occupation. The table also indicates the specific answer engine each participant used. Participants P22 to P24 took part in the pilot study, while the remaining participants were recruited for the primary study.

than presenting such assumptions with unwarranted confidence. This finding aligns with existing research on text generation models having a **Western or Global North alignment** and is further exacerbated by the type of sources these models select [3, 20, 46]. Prior works have established how text and image generation models can perpetuate biases and harms due to misalignment, however, the exacerbation of these biases can be seen translated in these answer engines as well. Participants therefore stressed the importance of the models providing contextually appropriate responses that consider diverse cultural perspectives and avoid defaulting to a Western-centric viewpoint.

U.V. Forces Answers When There is Not Enough Information *(10/21)*:

Another issue on user interaction identified by participants involved answering expert-level questions that needed more context or content for proper answer generation. These types of queries, which we call 'intractable questions,' often do not have clear or straightforward answers available in existing resources. When participants asked such questions, the answer engines tended to force the generation of whatever limited information they could find from search results, often leading to out-of-context or redundant responses. Participant P7 highlighted this issue by stating, "So it did go to the right domain, but it was not able to find the right answer because that doesn't exist. There is no one research paper that talks about combining these three elements yet. But it still generates an answer." Similarly, P12 noted the problematic approach of the answer engine, saying, "The [answer engine] has pulled out certain terms and then has tried to generate sentences from random parts of the paper that do not answer the question at all." This suggests that in the absence of adequate content, the answer engine's attempt to generate a response often results in misleading or irrelevant information.

Participants suggest that the answer engine would be more effective in such cases if it could recognize intractable questions and refrain from generating forced answers. Instead, the system should be capable of categorizing these queries as intractable, thereby indicating to the user that a direct answer may not be available. This approach would prevent the dissemination of irrelevant or misleading information, especially in cases where a user is not aware of this shortcoming.

A.5 Design Recommendation Explanation

S-I. Provide Balanced Answers for Leading Questions: To mitigate bias in responses, it is essential not to assume or reinforce the biases of the user. For topics or questions that are potentially leading or biased, participants in our study strongly indicated the need for neutral and balanced answers. The system should focus on addressing the broader context of the topic rather than providing an expected answer that aligns with any assumed biases.

S-II. Provide Objective Details to Claims: Participants frequently observed that the model often lacked objective details to substantiate its claims. It is crucial for answer engines to avoid

excessive summarization of sources. Instead, they should provide comprehensive information that supports the claims being made. Wherever necessary, responses should include objective details such as percentages, figures, or specific data points to strengthen the credibility of the answer.

S-III. Minimize Fluff Information in Answers: Many participants reported instances where the model generated simplistic answers containing irrelevant or extraneous information. Future answer engines should ensure that each sentence in the generated response is contextually accurate and directly relevant to the question posed. If a sentence does not contribute meaningfully to the response, it should be reconsidered or omitted to maintain clarity and precision.

S-IV. Reflect on Source's Thoroughness: A significant concern highlighted by participants was the lack of transparency regarding how the system selected and utilized sources. The black-box nature of current models creates distrust, as users are often unclear about the rationale behind the cited sources. To address this, an additional trust layer should be incorporated, providing users with insights into why specific sources were used and how they contribute to the generated answer. This transparency will enhance users' ability to critically evaluate the response.

C-I. Avoid Unsupported Citations: Participants observed that many statements generated by answer engines lacked proper citations, particularly when making claims that required supporting references. It is crucial to evaluate each statement's need for citation, ensuring that claims are backed by relevant sources retrieved by the model. If a statement cannot be properly cited, the system should either remove the statement or clearly indicate its relevance to the overall answer.

C-II. Double-Check for Misattribution: Misattribution was another common issue identified by participants, where sources were cited out of context or incorrectly attributed. To prevent this, answer engines should externally evaluate citations by considering the full content of the source rather than just a fragment. Additionally, revealing which part of the source contains the cited information can help reduce instances of misattribution, ensuring greater accuracy in the generated answers.

C-III. Cite All Relevant Sources for a Claim: Participants found it confusing when answer engines cited only one source for claims that clearly required multiple references. This practice hindered their ability to discern the importance of different points within the response. To address this, models should cite all relevant sources wherever necessary, helping users understand the breadth of support for a given claim. This approach reduces the likelihood of giving undue attention to non-important points or sources that mention irrelevant information.

C-IV. Retrieved Sources Must be Equal to Used Sources: Participants noted that some answer engines, like Bing Copilot and Perplexity, retrieved more sources than were actually used in the generated answers. This practice led to a confusion of trust, where users believed that many sources were used to construct the answer when, in reality, only a small percentage were utilized. To maintain transparency and trust, it is essential that the number of retrieved sources matches the number of sources actually used in the response. This alignment ensures users can accurately assess the reliability of the generated information.

S-I. Give Explicit Attention to Expert Sources: Participants observed that answer engines often fail to prioritize authoritative sources, such as research papers or government websites, even when these sources provide the most accurate information. Instead, the system tends to distribute attention equally among various types of content, including less reliable sources like blog posts and opinion pieces. It is crucial that the system recognizes and prioritizes expert sources, particularly when they offer definitive answers (e.g., CDC for COVID-19 updates). The source's authority should take precedence over search engine ranking, ensuring that the most reliable information forms the core of the response.

S-II. Retrieve and Use Only Necessary Sources: There were instances where the model retrieved sources that were either inaccurate or irrelevant to the question asked. Although these sources were marked as relevant by the search engine, they were not utilized in generating the final answer, sometimes limiting the response to a single source. To improve the accuracy and relevance of answers, the model should be more selective in retrieving sources, ensuring that only those necessary for constructing a precise and contextually appropriate response are used. Irrelevant sources should be discarded to make way for more suitable alternatives, or, if none exist, the system should acknowledge the lack of appropriate sources.

S-III. Differentiate Source-Based vs. Model-Generated Content: Answer engines are designed to retrieve and synthesize information from the web, minimizing the risk of hallucination—generating content not grounded in reality. However, participants noted several instances where significant claims or sentences lacked citations, leaving users uncertain of their origin. These uncited statements are likely generated from the model's training data rather than retrieved sources. While these statements may be factually correct, the inability to distinguish them from retrieved content undermines trust. To address this, the system should differentiate model-generated content from source-based content, perhaps through color coding or disclaimers, enhancing transparency and user trust in the system.

S-IV. Explicitly Mention and be Aware of Source Types: The origin and type of a source are critical factors in determining its reliability. Participants noted that, when using traditional search engines, they typically assess the credibility of a source before trusting its information. This behavior was less evident when using answer engines, where the source type and origin were not always transparent. Participants recommended that answer engines become more discerning about source types and their relevance to the question. The top search results are not always the most accurate; hence, the model should intelligently assess and prioritize source types, ensuring that the most credible and relevant sources are used to generate answers.

U-I. Incorporate Human Feedback on Sources and Text: One significant limitation identified by participants was the restricted interactivity within the answer engine's interface. Users were not given the option to modify sources or provide feedback on how the generated content could be improved. To enhance the quality and relevance of the generated answers, it is recommended to implement a human feedback system. This would allow users to contribute insights on the search results and suggest adjustments, leading to more accurate and contextually relevant responses.

U-II. Implement Interactive Citations (e.g., on Hover): Answer engines are increasingly used as sociotechnical systems across

various fields, including education, IT, and healthcare, where they are expected to provide quick and reliable answers. However, the use of citations—a familiar tool in academic contexts—is not as intuitive for many users in their daily lives. Participants suggested the development of interactive citations, such as on-hover pop-ups, which would display detailed source information when users hover over a citation. This feature could encourage users to verify the information and understand the source content more thoroughly, thereby increasing the reliability and usability of the system.

U-III. Incorporate Paragraph-Level Local Citations: Currently, answer engines often place citations at the end of sentences, which can create confusion about whether the entire sentence or just part of it was sourced from the cited reference. Participants expressed uncertainty when it was unclear which parts of the sentence were directly supported by the source. To address this issue, the system should implement paragraph-level local citations, clearly indicating exactly what information was cited and from where. This approach would improve transparency and help users better understand the relationship between the generated content and its sources.

U-IV. Avoid Forced Answers When Information is Insufficient: Participants observed that answer engines often generate responses even when there is insufficient information or when the question has no legitimate answer. This tendency can result in the dissemination of misinformation or fabricated content. For instance, when faced with a question about a non-existent concept, such as "Does the theorem of Law dispersion explain relative accentuation?" the system should recognize that no such theorem exists and explicitly state that no answer is available. Similarly, when information is insufficient, the system should refrain from generating an answer, thereby preventing the spread of inaccurate or misleading information.