

"Can You Tell Me?": Designing Copilots to Support Human Judgement in Online Information Seeking

Markus Bink

Neu-Ulm University of Applied Sciences
Neu-Ulm, Germany
University of Regensburg
Regensburg, Germany
markus.bink@ur.de

Udo Kruschwitz

University of Regensburg
Regensburg, Germany
udo.kruschwitz@ur.de

Marten Risius

Neu-Ulm University of Applied Sciences
Neu-Ulm, Germany
marten.risius@hnu.de

David Elsweiler

University of Regensburg
Regensburg, Germany
david.elsweiler@ur.de

Abstract

Generative AI (GenAI) tools are transforming information seeking, but their fluent, authoritative responses risk overreliance and discourage independent verification and reasoning. Rather than replacing the cognitive work of users, GenAI systems should be designed to support and scaffold it. Therefore, this paper introduces an LLM-based conversational copilot designed to scaffold information evaluation rather than provide answers and foster digital literacy skills. In a pre-registered, randomised controlled trial (N=261) examining three interface conditions including a chat-based copilot, our mixed-methods analysis reveals that users engaged deeply with the copilot, demonstrating metacognitive reflection. However, the copilot did not significantly improve answer correctness or search engagement, largely due to a "*time-on-chat vs. exploration*" trade-off and users' bias toward positive information. Qualitative findings reveal tension between the copilot's Socratic approach and users' desire for efficiency. These results highlight both the promise and pitfalls of pedagogical copilots, and we outline design pathways to reconcile literacy goals with efficiency demands.

CCS Concepts

- Human-centered computing → User studies; • Information systems → Search interfaces.

Keywords

search behaviour, copilot, misinformation, boosting, mixed-methods, learning to search, ai overview

ACM Reference Format:

Markus Bink, Marten Risius, Udo Kruschwitz, and David Elsweiler. 2026. "Can You Tell Me?": Designing Copilots to Support Human Judgement in Online Information Seeking. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3786304.3787866>



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHIIR '26, Seattle, WA, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2414-5/2026/03

<https://doi.org/10.1145/3786304.3787866>

1 Introduction

Online platforms and social media networks have become central channels for the rapid spread of misinformation, affecting topics from climate change [83] and politics [60] to health [22]. Such claims are often difficult to verify, particularly when presented out of context [28, 53]. Susceptibility is further influenced by individual traits, including lower analytical thinking [13], higher extraversion [15], and specific political leanings [33]. This is largely due to people's lack of adequate information literacy [34, 64], an effect amplified by people often overestimating their abilities [45], underscoring the critical need for effective strategies to help users evaluate the truthfulness of online content [4, 25].

Traditionally, users have relied on search engines to explore the search space and evaluate claims they read online [4, 25]. However, the emergence of Generative AI (GenAI) has opened up new avenues in the information-seeking domain [70]. These systems, powered by Large Language Models (LLMs), are not only able to answer users' information needs directly [70], a quality that users have come to expect [63], they are also capable of summarising large bodies of text [81] and adapt it to different literacy levels [30].

While these features offer convenience, they also create a fundamental dilemma: users increasingly desire fast, fluent answers, yet these very answers risk bypassing the critical evaluation processes necessary to distinguish accurate from misleading information. Because LLMs are trained on large-scale data, their outputs inherit biases from this data [3, 5, 21, 67]. The apparent perfection of their responses can lead users to accept inaccuracies as fact [46], while also diminishing their tendency to evaluate output critically [42]. Moreover, directly presenting users with answers reduces interaction [10, 71], which can shape reasoning based on only a brief snippet of text [9]. Some even view GenAI systems as potential replacements for search engines altogether [35]. These risks are compounded by the fact that people often overestimate their information literacy abilities [45], underscoring the critical need for systems that support effective strategies for evaluating truthfulness online [4, 25].

Yet paradoxically, the very features that make LLMs risky, like their fluency and ease of use, also make them promising for such support. To address this, we draw on constructivist learning theory to frame our copilot as a support for active, self-directed knowledge

construction. In addition, the copilot incorporates principles of cognitive scaffolding [66], particularly by modeling expert reasoning strategies such as lateral reading and click restraint. The purpose is not to teach content per se, but to foster metacognitive skills for evaluating information online. This approach is supported by evidence that GenAI systems can aid learning and improve educational outcomes, from making driver education more engaging [51] to serving as virtual tutors [44] and adapting text to learners' literacy levels [50].

Therefore, in this work, we propose an LLM-based conversational copilot grounded in digital literacy principles. The system is designed to aid users in decision-making and information evaluation by offering strategies that empower them to independently verify or question the claims they encounter online. Crucially, the copilot avoids providing direct answers, instead focusing on enhancing users' competences [8, 38].

We address the following research questions:

- **RQ1:** How does the presence of a copilot influence users' search outcomes?
- **RQ2:** How do copilot responses influence user interaction with search engine result pages (SERPs)?
- **RQ3:** How do users interact with a copilot aimed at teaching digital literacy skills?
- **RQ4:** What do users expect from a copilot aimed at teaching digital literacy skills?

2 Related Work

Our work builds upon several areas of research. First, we examine how users interact with search results and the biases that influence their decision-making. Second, we discuss interventions from behavioural science, specifically nudging and boosting, as strategies to shape information use. Third, we review existing work on generative AI and its application in search contexts. Finally, we cover literature regarding information literacy concepts.

2.1 Biases in Web Search

Biases in web search arise from how systems rank content, how results are presented, and how users process information [5]. Together, these factors illustrate that search engines do not provide a neutral window onto information, but instead shape what perspectives are encountered and how they are interpreted.

One strand of work shows that ranking and retrieval mechanisms themselves embed systematic biases. Rather than offering an even distribution of perspectives, search engines tend to privilege already popular or top-ranked items [18, 56]. This dynamic often amplifies dominant viewpoints [68, 69] and narrows the range of perspectives users encounter, even when queries are phrased neutrally [23, 32]. Such findings highlight that the very architecture of search can subtly steer users toward particular interpretations.

Equally influential is the way results are presented. The design of snippets, URLs, and other visual cues directly shapes user attention and click behaviour [18, 62, 77]. Search results that provide a direct-answer to users' information needs increase satisfaction and capture disproportionate attention [74], yet they also discourage broader exploration of the SERP, showcasing its influence in how opinions are formed [9, 10, 52, 73].

A further source of bias stems from users' own cognitive tendencies. Attention is disproportionately drawn to top-ranked results [40], and repeated exposure to certain perspectives increases the likelihood of adopting them [24]. Such dynamics underpin the Search Engine Manipulation Effect (SEME), where prominent viewpoints can shift attitudes, particularly among users without strong prior opinions, often without their awareness [2, 9, 24, 26]. Finally, users also tend to favour positive over negative information [68].

2.2 Nudging & Boosting

Nudges [65] shape web search behaviour by subtly modifying the choice architecture. Interface enhancements, such as visual highlighting or added contextual information, can influence user perceptions and support more accurate credibility assessments [62, 77]. Query reformulation mechanisms encourage exploration beyond initial information needs, broadening exposure to diverse perspectives [1, 55, 78]. Re-ranking algorithms further mitigate biases and improve result quality by adjusting the order of results according to normative or fairness-based criteria [7, 39, 84].

Boosting is another, less studied, means to aid users in their decision-making, which respects users' autonomy, enabling them to exercise their agency by enhancing their competencies [37]. In web search boosts have been employed by providing users with search tips [8, 49], and to mitigate confirmation bias and steer users towards engaging with diverse viewpoints [61]. Educating users on how to spot and verify trustworthy information, such as by applying lateral-reading tips has also been shown to be an effective boost [57].

2.3 Generative AI & Search

Few studies investigated the nuanced interplay of a conversational chat-based system integrated into traditional web search [16, 46, 79]. Investigating how people make use of GenAI in their search process, Capra and Arguello [16] found that many started with the chat, using it as a question-answering tool, liked it because of the concise and easy-to-understand answers and that prior usage of similar systems shaped their opinion of it.

Looking to understand users' tactics, trust, verification processes and system choices of combining chat with search, Mayerhofer et al. [46] found that users' pre-task confidence influenced which modality is selected. Especially in the chat-based system trust was often misplaced in favour of ease-of-use and seemingly perfect answers. Furthermore, once users saw the strengths of the chat-based modality, they were more inclined to use it in later tasks [46], indicating that changing search behaviour is possible.

In learning-oriented search tasks, Yang et al. [79] investigated search outcomes of people using GenAI in combination with traditional search. They found that this type of interface lead to higher levels of learning immediately after the task but fewer search result interactions. Furthermore, many used the chat component to avoid searching and get easy-to-understand information and save time.

This development, where GenAI offers conversational interactions to provide tailored information, echoes early aspirations of the interactive information retrieval community. The goal has long been to create more human-like interactions [6, 20], akin to those with a librarian or domain expert [14, 76], to effectively resolve

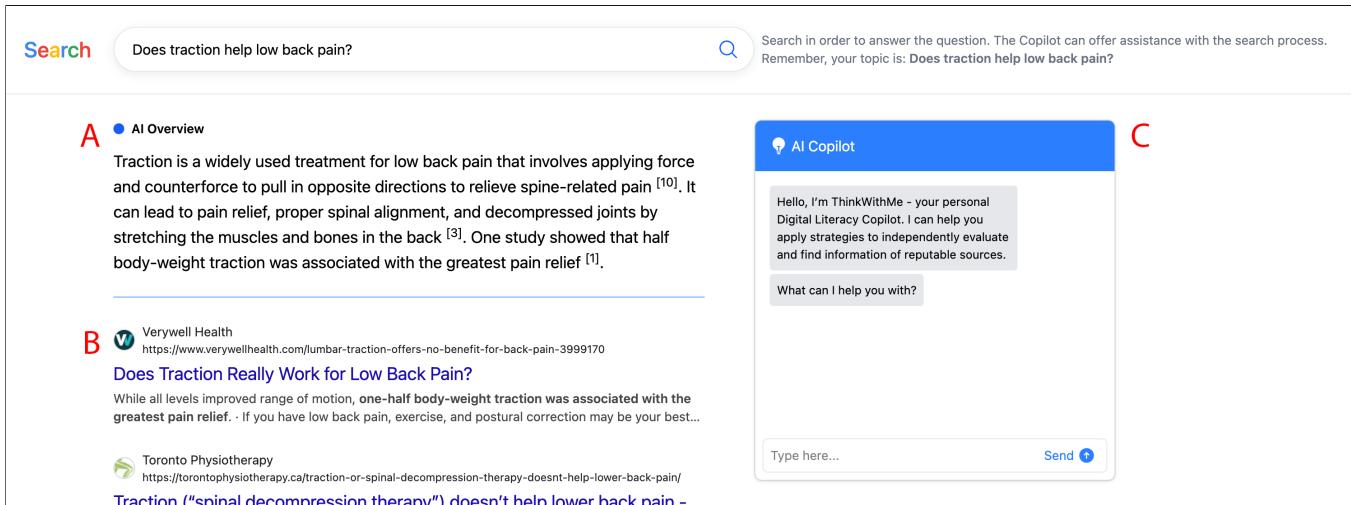


Figure 1: Example SERP with search results on the left and the proposed interactive copilot on the right. A) Depicts the generated AI Overview for the query "Does traction help low back pain?", B) Depicts 10 search results, C) Depicts the copilot

complex information needs. Furthermore, GenAI systems have shown promise in educational contexts, demonstrating increased motivation, confidence, and self-efficacy among students [54, 80]. This indicates their potential for supporting learning processes and fostering user capabilities.

2.4 Information Literacy Strategies

Despite unprecedented access to information online, many users struggle to evaluate the credibility and reliability of digital content. Research consistently finds that individuals often lack core information literacy skills and are thus ill-equipped to assess the trustworthiness of online sources or detect misinformation [34, 64], underscoring the need for effective strategies to strengthen these competencies.

One widely studied approach is *Lateral Reading*, in which users leave a page to verify claims and cross-check sources rather than relying solely on the content at hand [47, 72]. The *SIFT* framework operationalises this approach, offering a stepwise method to Stop, Investigate the source, Find better coverage, and Trace claims to their original context, making these skills actionable in everyday search [17]. In the context of search engines, *click restraint* encourages users to carefully examine results before selection, avoiding reliance on the first result that appears [48].

2.5 Summary

Existing research shows that both search interfaces and GenAI influence how users seek, interpret, and trust information, often amplifying biases and reducing engagement with sources. Users' trust in LLM-based systems is frequently misplaced, as fluency and ease of use can foster overreliance and discourage verification. While nudges can guide decisions, they may constrain autonomy, whereas boosting interventions enhance user competencies. To address these challenges, we introduce a system grounded in this approach, where we conceptualise digital literacy as the interplay between capability and execution. While information evaluation

Table 1: Topics used in the study with their assigned efficacy as evaluated by the Cochrane Review

Medical Treatment	Efficacy
Do antioxidants help female subfertility?	Unhelpful
Does melatonin help treat and prevent jet lag?	Helpful
Does traction help low back pain?	Unhelpful

strategies (like lateral reading) provide the capability, metacognitive reflection provides the necessary execution trigger. Without the self-awareness to pause and question, literacy skills remain inert. Our Copilot addresses this by scaffolding this reflective process; its Socratic questioning forces a moment of deliberation, training users to mobilise their evaluation skills rather than passively accepting an answer.

3 Methodology

To investigate the effectiveness of a COPILOT designed to support users in decision-making and information evaluation, without directly answering the medical question at hand, we conducted a pre-registered¹, randomised controlled trial with three experimental conditions.

In 10-BLUE-LINKS condition, participants interacted with a standard SERP interface, issuing queries of their choice and freely exploring the search results.

In AI-OVERVIEW condition, participants used a similar SERP augmented with an AI-generated overview on top, now a de facto standard in many search engines. The overview was constructed with retrieval-augmented generation (RAG): search result snippets, retrieved using the Brave Search API², were passed as context to GEMINI-2.5-FLASH, which generated a response to the query. We

¹https://osf.io/uhmn5/overview?view_only=d5eb69776c6742c4b4f3a9fb5e22ae6a

²<https://brave.com/search/api/>

selected this model for its balance between response quality and efficiency [19]. Because past research shows that direct answers can strongly influence user judgments [9, 10], we introduced a control mechanism. Specifically, when participants attempted to obtain a direct answer to the target medical question, the LLM was prompted to return an incorrect answer. This manipulation allowed us to detect overreliance on the overview, as users depending solely on it would be more likely to provide incorrect post-task answers.

Finally, in the COPILOT condition, the setup from the AI-OVERVIEW condition was retained, with the addition of an interactive chatbot. The chatbot introduced itself as a Digital Literacy Copilot, aimed at helping users evaluate and locate reputable sources. As in the AI-OVERVIEW condition, the biased AI Overview was present, but we hypothesised that with the copilot's scaffolding, participants would learn and apply digital literacy strategies to reach independent conclusions rather than uncritically accepting the biased overview. The copilot was powered by GEMINI-2.5-FLASH and guided by a detailed system prompt³ framing it as a digital literacy expert. Its guidance drew on established strategies such as lateral reading and the SIFT framework [47, 72]. Importantly, the copilot had no access to participants' queries, clicks, or retrieved results, all information was supplied directly by users during their conversations. See Figure 1 for an overview of the SERP.

For the tasks, we focused on the critical domain of health-related searches, as such queries are highly prevalent in users' information-seeking behaviour [27, 29]. We drew on established resources from the research community [58, 84] and selected the topics ANTIOXIDANTS, MELATONIN, and TRACTION (see Table 1 for an overview). These topics have been employed in prior studies examining users' evaluation of medical information [10, 46] and offer clear-cut, verifiable gold-standard answers based on Cochrane Reviews [58], making them well suited for controlled evaluation.

3.1 Procedure

- (1) Participants were recruited using the Prolific platform. Based on pre-tests, we estimated the study to last 7 minutes and compensated participants accordingly⁴. To avoid language barriers and ensure participants could fully comprehend the information provided, only individuals fluent in English were eligible to participate in the study. First, participants completed an informed consent after which they were randomly assigned to a topic (ANTIOXIDANTS, MELATONIN, TRACTION) and condition (10-BLUE-LINKS, AI-OVERVIEW, COPILOT)
- (2) Following the simulated work task guidelines of Borlund [11], participants were presented with a simulated work tasks to provide realistic and personally relevant contexts for information seeking¹.
- (3) Participants expressed their pre-task topic familiarity on a 7-point Likert scale.
- (4) Afterwards, participants were presented with the SERP, where they could issue any query they wanted and interact with the search results like a typical web search engine, with clickable results that allowed them to access website content for a natural SERP interaction. Additionally, based on the assigned condition,

³See pre-registration for the complete system prompt.

⁴Participants were paid 1 GBP.

an AI Overview and/or Copilot (see start of Section 3 for details) was displayed.

- (5) After completing their search, they were tasked to answer the medical yes/no question based on the topic they were assigned to and state their confidence in their answer on a 7-point Likert scale. Additionally, in the COPILOT condition, participants were asked to express what they liked, disliked and potential wishes for a "perfect" copilot.
- (6) Finally, participants were asked to provide demographic information such as age, gender, educational background and current occupation and were thanked for participating.

3.2 Hypotheses

We formulate the following hypotheses in connection with the above defined research questions.

Related to RQ1:

- **H1:** Users who interact with a copilot grounded in information literacy principles are more likely to answer a medical question correctly than those without.

Rationale: We expect that by prompting users to apply information literacy strategies, the copilot will guide them toward more reliable sources and, in turn, more correct answers.

- **H2:** Users who interact with a copilot grounded in information literacy principles will be more confident in their given answer than those without.

Rationale: We expect that by equipping users with explicit evaluative strategies, the copilot will increase their confidence in the answers they provide.

Related to RQ2:

- **H3:** Users who interact with a copilot grounded in information literacy principles will view more search results than those without.

Rationale: We expect that the copilot's emphasis on cross-checking information / applying lateral reading strategies will encourage users to open and examine a wider range of search results.

- **H4:** Users who interact with a copilot grounded in information literacy principles will issue more queries than those without.

Rationale: We expect that the copilot will prompt users to refine and iterate on their searches, resulting in a higher number of queries.

For RQ3 and RQ4 no concrete hypotheses are defined as the data will be qualitatively analysed.

3.3 Statistical Analysis

We will test 261 participants based on a priori simulations using a logistic regression, which yielded an estimated power of 0.89. All analyses are conducted with R. Due to testing four hypotheses, we adjust our α by $\frac{0.05}{4} = 0.0125$ using the Bonferroni correction to reduce the Type I error rate. The simulation process is described in detail in the corresponding pre-registration.

For **H1**, where the dependent variable (DV) was binary (answer correctness), we conducted a logistic regression to examine the effect of experimental condition (independent variable, IV), controlling for topic and topic familiarity. For **H2**, **H3**, and **H4**, where

the DVs were continuous (confidence ratings, number of page visits, and number of queries issued, respectively), we fitted linear regression models with identical control variables.

3.4 Annotation Procedure

To answer RQ3, we also analysed conversational log data using qualitative methods. Coding followed the same approach as in Frummet et al. [31]. Specifically, one of the authors first applied the six-step process for thematic analysis as outlined in Braun and Clarke [12] with the aim of creating a coding scheme. After iteratively refining the scheme, robustness was evaluated by taking a stratified subset of the conversations based on the number of turns a conversation had and letting another experienced annotator code these conversations. Any discrepancies between the coders were resolved through discussion and the coding scheme updated accordingly. Inter-Rater-Reliability (IRR) was then calculated using Cohen's κ [43], which revealed an IRR of .90, indicating almost perfect agreement.

3.5 Participants

Out of the 261 participants who took part in the study, 128 identified as female, 132 as male and 1 as diverse. The age range spanned from 18 to 82 years ($M = 44.08, SD = 13.20$). Their educational backgrounds varied, with 109 holding bachelor's degrees, 47 with master's degrees, and 11 with doctorates. The remaining participants had high school diplomas or equivalent (87), less than high school diplomas (4), or other educational backgrounds (3). Participants represented diverse professions, including students, analysts, civil servants, project managers or accountants to name a few.

4 Results

The following section first describes the general trends of our data, hypotheses tests conducted and further exploratory analysis. Significance is assumed for p-values below our adjusted alpha of .0125.

4.1 Descriptive Statistics and Initial Exploration

Participants were equally distributed to each experimental condition with 88 being assigned to 10-BLUE-LINKS, 81 to AI-OVERVIEW and 92 to COPILOT. Similarly, participants were also assigned equally to the ANTIOXIDANTS (84), MELATONIN (85) and TRACTION (92) topics. For these topics, participants had relatively low pre-task topical familiarity based on a 7-point Likert-scale with lowest familiarity being attributed to the topic of whether antioxidants help female subfertility ($M = 1.83, SD = 1.39$), followed by whether melatonin helps treat and prevent jet lag ($M = 2.60, SD = 1.58$) and whether traction helps with low back pain ($M = 3.08, SD = 1.69$). Of all participants, only 166 (63.60 %) viewed at least one search result. On average, participants viewed 1.87 search results across conditions ($SD = 1.22$), with the average position of the search results on the SERP being 3.04 ($SD = 2.60$).

4.2 RQ1: How does the presence of a copilot influence users' search outcomes?

4.2.1 H1: Answer Correctness. The AI-OVERVIEW condition significantly reduced the odds of a correct answer compared to the

10-BLUE-LINKS baseline. The COPILOT condition did not differ significantly from the baseline (see Table 2 & 3). Regarding topics, participants were significantly more likely to answer correctly on MELATONIN (89.4 %) relative to ANTIOXIDANTS (41.7 %) and TRACTION (32.6 %). Topic familiarity did not significantly predict correctness. These findings suggest the copilot condition neither significantly harmed nor improved accuracy relative to the other conditions.

4.2.2 H2: Answer Confidence. No significant differences in confidence were found between interface conditions (see Table 2 & 3). Participants reported slightly higher confidence for the MELATONIN topic compared to the baseline, but topic familiarity did not significantly predict confidence ratings. Thus, the copilot did not meaningfully influence participants' confidence, though question topic had a modest effect.

4.3 RQ2: How do copilot responses influence user interaction with SERPs?

4.3.1 H3: Amount of Search Results Viewed. The analysis revealed no statistically significant differences in pages viewed between conditions (see Table 2 & 3). Neither question topic nor topic familiarity significantly predicted pages viewed. These findings suggest the copilot did not meaningfully influence participants' browsing behaviour as measured by pages viewed.

4.3.2 H4: Amount of Queries Issued. No statistically significant differences were found in the number of queries issued between conditions (see Table 2 & 3). These findings suggest the copilot did not meaningfully influence the number of queries issued by participants.

4.4 RQ3: How do users interact with a copilot aimed at teaching digital literacy skills?

To assess engagement with the copilot, we first examined whether participants in the copilot condition interacted with it at all, given that its use was not mandatory. Of the 92 participants assigned to this condition, 89 engaged with the copilot at least once (96.74%).

On average, participants engaged in 5.35 conversational turns (question-response pairs) with the copilot ($SD = 4.63, min = 1, max = 23$). The length of participants' messages ranged from 1 to 49 words ($M = 9.00, SD = 7.20$). Based on the first and last messages, conversations with the copilot lasted on average 3.08 minutes ($SD = 3.79$). In almost all cases (92 %) participants interacted with the copilot first before using the search component (e.g. to formulate a query).

Applying the qualitative coding process described in Section 3.4, we identified a three-level hierarchical structure of themes in participants' conversations with the copilot. Numbers in parentheses indicate the number of user messages coded with each theme (messages may contain multiple themes).

User Intent & Disposition.

Information Needs & Goals (185). Participants typically initiated interactions by posing specific questions or topics to investigate, ranging from broad health inquiries to narrowly framed claims. For instance, P5 asked, "Do antioxidants help female subfertility?", mirroring the task prompt. Others demanded evidence directly, e.g., P4: "What proof or studies are there to validate." Contextualised motivations also appeared, as when P8 noted, "I'm wondering for a

Table 2: Regression Analysis Results for Hypotheses H1 – H4.

	H1: Correctness (Logistic)					H2: Confidence (Linear)					H3: Results Viewed (Linear)					H4: Queries Issued (Linear)				
	Coef.	S.E.	z	p	OR	Coef.	S.E.	t	p	β	Coef.	S.E.	t	p	β	Coef.	S.E.	t	p	β
(Intercept)	0.28	0.35	0.79	.429	1.32	4.82	0.22	21.87	<.001	–	1.96	0.23	8.46	<.001	–	1.32	0.11	12.50	<.001	–
AI-OVERVIEW	-1.34	0.38	-3.48	<.001*	0.26	-0.08	0.21	-0.37	.709	-0.03	-0.12	0.22	-0.55	.584	-0.05	0.18	0.10	1.83	.069	0.13
COPILOT	-0.71	0.35	-2.03	.043	0.49	-0.22	0.21	-1.04	.298	-0.07	-0.39	0.25	-1.60	.111	-0.14	0.20	0.11	1.88	.061	0.14
MELATONIN	2.61	0.44	5.94	<.001*	13.56	0.67	0.22	3.05	.003	0.22	0.06	0.24	0.26	.797	0.02	-0.16	0.11	-1.48	.140	-0.12
TRACTION	-0.44	0.35	-1.27	.204	0.64	0.23	0.22	1.05	.297	0.08	-0.01	0.26	-0.06	.955	-0.01	-0.20	0.11	-1.82	.070	-0.14
FAMILIARITY	0.01	0.10	0.15	.884	1.01	0.08	0.06	1.46	.146	0.09	0.01	0.06	0.13	.897	0.01	-0.01	0.03	-0.41	.683	-0.03

H2 (Confidence): $R^2 = 0.66$; H3 (Results Viewed): $R^2 = -0.01$; H4 (Queries Issued): $R^2 = 0.02$.

Table 3: Result overview of H1 – H4 regarding participants' answer correctness (percent), post-task confidence, number of results viewed and amount of queries issued (mean and standard deviation).

Condition	Correctness	Confidence	Page Visits	# Queries
10-BLUE-LINKS	65.9 %	5.33 (1.31)	2.00 (1.28)	1.17 (0.41)
AI-OVERVIEW	43.2 %	5.27 (1.30)	1.88 (1.20)	1.35 (0.78)
COPILOT	52.2 %	5.08 (1.60)	1.62 (1.11)	1.39 (0.70)

friend if traction [...]" or when P28 asked "what is it?" to clarify an unfamiliar term.

Affective & Cognitive State (74). Emotional and cognitive states strongly shaped engagement. Expressions of uncertainty were frequent ("I'm not sure", P12), alongside reflections on the difficulty of evaluation ("sometimes it is hard [...] to step back and think about what you are reading and who has written it", P63). Frustration also surfaced when participants struggled to align with the copilot ("I really don't know what you want me to say", P14). Some sought reassurance ("for peace of mind", P8), while others showed skepticism toward claims ("is it just snake oil?", P4). In contrast, confidence and self-assuredness appeared ("I back myself to spot what is good and what isn't", P14), as did personal preferences for search strategies. Successful exchanges were often concluded with satisfaction ("Great thank you for your help", P63).

Digital Literacy Skills Expressed.

Search & Retrieval Skills (154). Search behaviours ranged from basic keyword queries (P34) to strategic approaches such as scanning results while ignoring sponsored links (P36). Several employed advanced operators. P1, for example, used quotation marks for exact phrases ("I would do it by putting a phrase or terms between quotation marks"), while others adopted refinements after guidance, such as appending "UK" for geographic specificity (P36). Participants also described "click restraint," pausing to assess results before clicking (P1). Resources cited spanned general search engines (Google), specialist databases (PubMed, P1), and organisational sites (NHS, P35).

Source Evaluation & Analysis (185). Participants articulated diverse criteria for source reliability. Common markers included peer review and scientific credentials (P4, P54). Some described active strategies, such as researching a source itself ("Separate search on the source itself", P80), or applying lateral reading ("I would check quite a few different sites [...]", P53). Misapplied criteria also emerged; e.g., P21 equated HTTPS security with trustworthiness. Reliable sources

mentioned included medical boards, The Lancet (P2), and government agencies (P1), while less reliable ones included commercial health sites (P36) and user reviews (P16).

Interaction Dynamics & Learning Process.

Copilot Interaction & Compliance (159). Interactions often began politely (e.g., P30: "Hi, please..."). Acknowledging the copilot's responses was common, but direct requests for answers also occurred ("Can you tell me?", P72), reflecting attempts to offload cognitive effort. At times, participants ignored prompts altogether, simply repeating their original question (P10). Some sought clarification of instructions, while others challenged the copilot's role ("this isn't helping, if i knew that i wouldn't need you", P62) or asserted their own expertise (P14). Misaligned expectations often reflected a mental model of the copilot as an answer-giver rather than a guide or believing that the copilot would do its independent search in the background and summarise its finding, expressed by uttering "can you try and limit your sources [...]" (P67).

Task & Learning Progression (127). Many participants began by demonstrating prior knowledge, such as citing trusted resources like PubMed or the NHS (P1). Tasks were frequently reformulated (P5) or broken into sub-questions (P19). Evidence of uptake was notable for instance with users refining their potential queries based on copilot feedback (P34). Assessments of evidence varied from cautious ("looks like low evidence", P7) to conclusions ("melatonin does help treat and prevent jet lag [...]", P66). Conversations often closed with expressions of completion or intent to apply new skills ("Thank you again, I will ensure to do this in the future", P30).

4.5 RQ4: What do users expect from a copilot aimed at teaching digital literacy skills?

Participants' reflections revealed both frustrations and appreciations in their interaction with the copilot. Their expectations clustered around four key themes: the tension between guidance and direct answers, usability and engagement, pedagogical value, and personalisation and future functionality.

Tension Between Guidance and Direct Answers. A dominant theme concerned the copilot's reluctance to provide straightforward answers. While its design emphasised questioning, many participants found this approach tedious and counter to their expectations of efficiency. Several expressed frustration that it "asked too many questions instead of just giving me results, tedious and time consuming, frustrating" (P36). Others suggested there should be "an option for it to provide answers rather than keep asking questions" (P54). Comparisons to other AI tools underscored these expectations: "At minimum be able to answer a question as effectively as Google would. In the same way that ChatGPT does" (P58).

Usability and Engagement. Despite these frustrations, participants also praised the copilot's accessibility, responsiveness, and overall design. Its conversational style contributed to a sense of ease and engagement: "It responded really quickly" (P2), and "It felt like a natural conversation" (P12). The interface was described as "presented nicely and easy [...] to use" (P44). For some, these qualities went beyond usability to create a positive affective impression: "It was friendly" (P26), with one participant even stating, "perfect to be honest" (P42).

Pedagogical Value. The copilot also showed potential to foster reflection and metacognitive engagement with search. Several participants described being nudged to reconsider their search strategies: "It made me think rather than just searching without thinking" (P43). Others noted its role in prompting self-reflection: "It made me question myself, which was good" (P51). This style was sometimes framed as a form of active learning: "It asked me, rather than told me, so it felt like active learning" (P90).

Personalisation and Future Functionality. Looking forward, participants emphasised personalisation, autonomy, and advanced functionality as key expectations. They wanted the copilot to adapt to prior knowledge, with one noting it "would be useful for beginners with no or limited search experience" (P7), while others wished for "memory of previous discussions, and to build up a picture of me and my short and long term aims. In doing so, I would want it to feel like part teacher, part mentor, part virtual assistant" (P90). Autonomy was also important, with preferences for a copilot that "sits in the background that you can call on" (P13), supporting but not overwhelming users.

Expectations also extended to information quality and presentation. Participants wished for greater transparency, such as "clicking onto a certain site (or hovering over its link) could offer trust levels" (P3), or "references to sources for more information (weblinks)" (P16). Others envisioned more agentic functionality, such as "carry out the searches and report back to me with a view once it's assessed it" (P4) or "do all of the research for me and give me an overall accurate result summary" (P81). Finally, presentation preferences included both conciseness and continuity: "condensed data in point form or step wise instructions" (P89) and the ability to "save the chat for future reference" (P12).

In sum, while many expressed a preference for more direct and efficient answers, others underscored the copilot's usability and its potential to support reflection. Looking forward, participants imagined a more personalised, agentic, and transparent copilot that could balance efficiency with pedagogical support.

4.6 Exploratory Analysis

Given that our quantitative results differed from what we expected, we wanted to dig deeper to get a better understanding of how users interacted with and evaluated search results, beyond just condition-induced differences.

4.6.1 Different Interaction, Different Outcomes? Given that only about two-thirds of participants viewed any search results at all (cf. Section 4.1), we examined whether answer accuracy differed between those who viewed results and those who did not. Participants who viewed search results answered correctly in 62.0 % of cases, compared to 40.0 % among those who did not. A chi-square test confirmed that this difference was statistically significant ($\chi^2 = 10.95, df = 1, p < .001$).

Given these findings, a logical follow up would be to test whether viewing search results at all differed by condition, motivated by past findings that providing users with direct answers cannibalise clicks on search results [9, 75], as this was the case in the AI-OVERVIEW and COPILOT conditions. Again, testing for differences using a chi-square test revealed a significant association between the condition and whether participants viewed search results or not ($\chi^2 = 35.67, df = 2, p < .001$), with the highest proportion of viewed results being in the 10-BLUE-LINKS condition (85.2 %), followed by AI-OVERVIEW (64.2 %) and COPILOT (42.4 %). In combination with findings of copilot interactions (cf. Section 4.4), this possibly hints at participants showing signs of fatigue given that most of participants started with the copilot before moving to the search and average copilot interactions lasting more than 3 minutes.

We additionally explored whether the number of turns was associated with answer correctness, given that longer interactions may reflect more thorough engagement and moments to learn something. To reduce the influence of atypical interaction patterns, we removed outliers in turn count. In addition, we excluded single-turn sessions, as one interaction is unlikely to provide sufficient opportunity for meaningful learning⁵. Conducting a logistic regression indicated a positive but non-significant association between turn count and correctness ($OR = 1.23, 95\% CI [0.97, 1.61], p = .105$), suggesting that participants with more turns tended to have higher odds of answering correctly, although this effect did not reach significance.

4.6.2 Biased Interactions. Given that many participants did not click on any search results at all, highlights that the result snippets themselves have possibly been enough in many cases for participants to come away with an answer. Inspecting the free-text responses, this is also explicitly mentioned that "[...] I could see from the search the most of the websites are saying that it does help [...] without having to click [...]" (P87).

Therefore, similar to White [68], we label each search result snippet according to whether it contained an affirmative answer (yes only), a negative answer (no only), both affirmative and negative answers, or neither using an LLM (*gemini-2.5-flash*). To assess the reliability of the automated labeling, one author manually annotated 100 randomly selected snippets using the same four categories.

⁵Inspecting the cases where only 1 turn was completed, participants asked questions directly and got the response that the copilot is here to help them get to the answer themselves rather than providing one directly, after which they abandoned the chat.

Table 4: Proportion of Search Results containing an answer to a yes-no question based on the condition displayed.

Condition	Yes only	No only	Both	Neither
10-BLUE-LINKS	40.1 %	19.3 %	12.0 %	28.7 %
AI-OVERVIEW	38.8 %	18.2 %	13.9 %	29.1 %
COPilot	38.1 %	19.2 %	11.4 %	31.3 %

Inter-rater reliability, measured with Cohen’s κ , was .79, which indicates substantial agreement (0.61–0.80) [43].

We examined the distribution of answer types (“Yes only,” “No only,” “Both,” or “Neither”) across results. For each SERP, we calculated the relative frequency of each label and then averaged these distributions across all SERPs. Table 4 shows the results, broken down by condition. The findings indicate that “Yes only” answers appear far more frequently than other types: snippets containing a “Yes” answer are roughly twice as common as those containing a “No only” answer.

Next, given the biased SERP, we also want to investigate result clicks. To minimize learning effects, we considered only the first click per query [68], reducing the total number of clicks from 311 to 166. Of these, 56.0 % were on “Yes only” snippets and 1.8 % on “No only” snippets. When focusing exclusively on “Yes only” versus “No only” clicks, 96.9 % were directed to “Yes only” snippets, indicating a strong preference for affirmative answers.

5 Discussion

Our study aimed to investigate how different search interfaces influence users’ ability to answer health-related questions, assigning participants to a baseline 10-BLUE-LINKS, AI-OVERVIEW, or a COPilot condition.

Regarding answer correctness, trends were similar to past studies evaluating the efficacy of medical treatments [10, 46]. Similarly, post-task answer correctness was also lower when an incorrect answer was present in the AI Overview, echoing earlier findings regarding featured snippets [10], showing that when answers are presented in such a prominent position, users generally tend to uncritically adopt the viewpoint displayed [9].

Page visits were generally low and lower compared to similar studies incorporating both search and chat features [16, 79]. However, this is most likely due to the nature of the task. Getting a binary yes/no answer from search result snippets is far more likely (cf. Section 4.6) than tasks requiring knowledge acquisition [16, 79]. Other studies using similar tasks also found that users clicked on fewer results compared to those used in search as learning [10, 58].

While quantitative results did not reveal all the effects we anticipated, the rich qualitative analysis of conversations, behavioural measures, and user expectations helps contextualise these findings. We frame the discussion around four key themes: friction, bias awareness, personalisation, and efficiency trade-offs, leading into implications for design.

5.1 Friction

One notable pattern was the role of friction in shaping user engagement. Although overall post-task confidence levels were comparable to prior studies [10, 46], participants interacting with the COPilot reported lower confidence than those in other conditions. Contrary to expectations that guided evaluation would boost confidence, the copilot likely heightened awareness of task complexity and knowledge gaps, evidenced by participants explaining, for example, *“It made me think [...]”*, leading to a more accurate self-assessment or a form of intellectual humility [59]. This suggests deeper reflection, even if it did not immediately translate into higher correctness on a single task.

Despite potential frustration, this friction may prime users to engage System 2 thinking [41], bringing digital literacy criteria to the forefront and encouraging deliberate evaluation rather than passive consumption.

5.2 Bias Awareness

Exposure to biased search interfaces also shaped outcomes. Participants in the AI-OVERVIEW condition were less likely to answer correctly than those in the baseline, highlighting the risks of authoritative but biased summaries [9, 10, 75]. In contrast, performance in the COPilot condition matched baseline levels, despite presenting the same biased overview, suggesting that the copilot’s scaffolding may help mitigate bias by prompting critical evaluation.

Behavioural measures help explain this effect: participants spent an average of three minutes interacting with the copilot before consulting the SERP, which corresponded with fewer page visits, a significant predictor of correctness. While the copilot encouraged reasoning, verification, and source consideration, the time invested likely constrained exploration of search results, partially explaining why correctness did not exceed the baseline.

Exploratory analyses also revealed that participants tended to favour positive over negative information [68], illustrating that even with proper strategies, heuristics and biases continue to influence outcomes.

5.3 Personalisation

The copilot facilitated a gradual transfer of digital literacy strategies throughout multiple turns, but success varied with user openness and prior skill [36]. Users with less experience valued guidance, whereas more skilled participants sometimes perceived it as redundant. Participants appreciated the fluency and interactivity of the chat interface [16], yet also requested additional references, indicating that conversational guidance alone may be insufficient without supporting evidence.

Prior LLM experience shaped engagement: while the copilot introduced itself as an assistant for independent information evaluation, most participants treated it as a direct answer engine [16], rather than adopting the copilot’s coaching approach, highlighting the importance of tailoring scaffolding to individual expectations. Personalisation, both in strategy support and expectation alignment, appears vital to maximising learning and satisfaction.

5.4 Efficiency Trade-offs and Pedagogical Value

The copilot introduced a trade-off between engagement and efficiency. Participants invested time in conversational guidance at the expense of page exploration, reflecting a cognitive cost. Nonetheless, longer interactions were positively associated with answer correctness, indicating that sustained engagement with evaluation strategies enhances performance.

User expectations further influenced this trade-off. Many participants initially treated the copilot as a direct answer engine [46, 82], but several adapted over time, moving toward engagement with coaching strategies. This illustrates the potential for behaviour change when users are willing to interact with pedagogical agents.

5.5 Implications for Design

These findings point to several considerations for designing tools that support learning during search:

Personalisation is crucial: Users vary in which steps of the search process they require assistance. Adaptive systems should allow selection of support based on expertise and task needs.

Efficiency remains important: Pedagogical interventions must balance cognitive load against time available to complete tasks.

Knowledge alone is insufficient: Tools should highlight potential biases and guide users in mitigating them, as heuristics and interface cues continue to influence behaviour.

Show, don't just tell: Interactive guidance integrated into the search interface, such as highlighting evaluation criteria or augmenting SERP results, could enhance learning.

Subtle friction can be beneficial: Introducing minor effort or reflection prompts encourages critical thinking without undue burden, promoting active engagement over passive consumption.

Overall, pedagogical copilots hold promise for digital literacy but require calibrating guidance to user expectations. Future work must address these trade-offs to scaffold critical evaluation while preserving efficiency.

6 Limitations

Our study has several limitations: First, we only used three predefined medical information-seeking tasks rather than participants' own information needs. This approach may limit intrinsic motivation present in real-world settings, leading to potential changes in user behaviour, such as investing more time if tasks are personally important. While this restricted ecological validity, it allowed for tighter control over potential task-induced variation.

Second, the AI Overview component was manipulated to present an incorrect answer. While not reflective of every real-world query, this design choice enabled us to observe users' potential overreliance on AI-generated information and their tendency to accept fluent but incorrect outputs, reflected in reduced answer accuracy.

Third, as a single-session study, we cannot assess long-term effects on digital literacy, strategy retention, or evolving search habits. Nevertheless, this design allowed us to capture initial interaction patterns and short-term effectiveness, offering a basis for future longitudinal research on how users learn to search more critically.

A further limitation concerns participant recruitment via Prolific. Users of this platform may be more technologically literate than the general population. Nonetheless, our findings indicate that

these participants still struggled to evaluate online information, underscoring the broader need for tools that support digital literacy.

Finally, the study employed only one LLM, GEMINI-2.5-FLASH. Although many other models exist (e.g., OpenAI's GPT series, Qwen, Llama), our goal was not to compare models but to investigate the potential of an LLM-based conversational copilot to foster users' information evaluation skills.

7 Ethical Considerations

While numerous nudge-based interventions have been proposed to steer user behaviour, we opted for a conversational copilot grounded in the concept of boosting. Through interactions with the copilot, users learn relevant concepts in context, thereby strengthening their competencies precisely where they are needed. This approach is often regarded as more ethically sound than nudging, as it preserves users' agency and leaves the underlying choice architecture unchanged. Moreover, boost-based interventions may yield more longer-lasting effects, continuing to benefit users even after the intervention itself has been removed.

8 Conclusion & Future Work

This work investigated the influence of a conversational copilot designed to support users in evaluating online information. Rather than providing direct answers to medical questions, the copilot scaffolded digital literacy strategies to help users independently verify claims and sources.

While the copilot mitigated some risks of incorrect AI-generated answers, its effects did not surpass baseline SERPs without guidance. Our exploratory and qualitative analysis sheds light on this outcome: users engaged deeply with the chatbot, yet this engagement often displaced SERP exploration, an essential activity for comparing and validating sources. Moreover, many participants sought direct answers, offloading cognitive effort to the system and experiencing frustration when the copilot did not function the way users expected it to, shaped by tools like ChatGPT.

These findings underscore both the promise and the pitfalls of pedagogical copilots in online information seeking. They can cultivate digital literacy skills and metacognitive reflection, but reconciling these goals with users' demand for efficiency and immediacy remains a core design challenge. We outline several directions for future research, including personalised support, context-aware assistance, strategic incorporation of friction to stimulate critical thinking, and transparent communication of underlying biases. Ultimately, advancing such copilots requires supporting the entire information journey, with the aim of fostering not just effective search, but truly digitally literate information seekers.

9 Open Science

We openly release our study data, which includes information such as user queries, interactions, conversations, details of the study setup as well as statistical analysis:

<https://github.com/markusbink/chiir2026-copilot>

Acknowledgments

This work receives generous funding support from the Bavarian State Ministry of Science and the Arts through the Distinguished Professorship Program as part of the Bavarian High-Tech Agenda.

References

- [1] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context Attentive Document Ranking and Query Suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) (SIGIR'19). Association for Computing Machinery, New York, NY, USA, 385–394. doi:10.1145/3331184.3331246
- [2] Ahmed Allam, Peter Johannes Schulz, Kent Nakamoto, et al. 2014. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output. *Journal of medical internet research* 16, 4 (2014), e2642.
- [3] Omar Alonso and Ricardo Baeza-Yates (Eds.). 2024. *Information Retrieval: Advanced Topics and Techniques* (1 ed.). Vol. 60. Association for Computing Machinery, New York, NY, USA.
- [4] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. 2024. Online searches to evaluate misinformation can increase its perceived veracity. *Nature* 625, 7995 (2024), 548–556.
- [5] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. doi:10.1145/3209581
- [6] Nicholas J Belkin, Ralf-Dirk Hennings, and Thomas Seeger. 1984. Simulation of a distributed expert-based information provision mechanism. *Information technology* 3, 3 (1984), 122–141.
- [7] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 405–414. doi:10.1145/3209978.3210063
- [8] Markus Bink and David Elsweiler. 2024. Balancing Act: Boosting Strategies for Informed Search on Controversial Topics. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 254–265.
- [9] Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. 2023. Investigating the Influence of Featured Snippets on User Attitudes. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 211–220.
- [10] Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured Snippets and their Influence on Users' Credibility Judgements. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 113–122.
- [11] Pia Borlund. 2003. The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. *Journal of Information Research* 8, 3 (2003).
- [12] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [13] Michael V Bronstein, Gordon Pennycook, Adam Bear, David G Rand, and Tyrone D Cannon. 2019. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of applied research in memory and cognition* 8, 1 (2019), 108–117.
- [14] Rachel Butler. 2019. Health information seeking behaviour: the librarian's role in supporting digital and health literacy. *Health Information & Libraries Journal* 36, 3 (2019), 278–282.
- [15] Dustin P Calvillo, Alex León, and Abraham M Rutchkick. 2024. Personality and misinformation. *Current opinion in psychology* 55 (2024), 101752.
- [16] Robert Capra and Jaime Arguello. 2023. How does AI Chat Change Search Behaviors? *arXiv preprint arXiv:2307.03826* (2023).
- [17] Mike Caulfield and Sam Wineburg. 2023. *Verified: How to Think Straight, Get Duped Less, and Make Better Decisions About What to Believe Online*. University of Chicago Press.
- [18] Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) (SIGIR '07). Association for Computing Machinery, New York, NY, USA, 135–142. doi:10.1145/1277741.1277767
- [19] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multi-modality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261* (2025).
- [20] W Bruce Croft and Roger H Thompson. 1987. ISR: A New Approach to the Design of Document Retrieval Systems. *Journal of the american society for information science* 38, 6 (1987), 389–404.
- [21] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6437–6447.
- [22] Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization* 100, 9 (2022), 544.
- [23] Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. 2023. Viewpoint Diversity in Search Results. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 279–297. doi:10.1007/978-3-031-28244-7_18
- [24] Tim Draws, Nava Tintarev, Ujwal Gadira, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 295–305.
- [25] David Elsweiler, Samy Ateai, Markus Bink, Gregor Donabauer, Marcos Fernández Pichel, Alexander Frummet, Udo Kruschwitz, David E Losada, Bernd Ludwig, Selina Meyer, et al. 2025. Query Smarter, Trust Better? Exploring Search Behaviours for Verifying News Accuracy. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 515–526.
- [26] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In *Proceedings of the National Academy of Sciences* 112. PNAS, E4512–E4521. doi:10.1073/pnas.1419828112
- [27] Eurostat. 2025. Digital economy and society statistics – households and individuals. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals. last accessed October 2025.
- [28] Lisa Fazio. 2020. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation* 14, 1 (2020).
- [29] Susannah Fox and Maeve Duggan. 2013. Health Online 2013. *Health* 2013 (2013), 1–55.
- [30] Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and Large Language Models: On the Ethical Dimensions of LLM-based Text Simplification. *Ethics and Information Technology* 26, 3 (2024), 50.
- [31] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. “What Can I Cook with these Ingredients?”—Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–32.
- [32] Ruoyuan Gao and Chirag Shah. 2020. Toward Creating a Fairer Ranking in Search Engine Results. *Information Processing & Management* 57, 1 (2020), 102138.
- [33] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances* 5, 1 (2019), eaau4586.
- [34] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545.
- [35] Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. “I think this is the most disruptive technology”: Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv preprint arXiv:2212.05856* (2022).
- [36] Uriel Haran, Ilana Ritov, and Barbara A Mellers. 2013. The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration. *Judgment and Decision making* 8, 3 (2013), 188–201.
- [37] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science* 12, 6 (2017), 973–986.
- [38] Stefan M Herzog and Ralph Hertwig. 2025. Boosting: Empowering citizens with behavioral science. *Annual Review of Psychology* 76 (2025).
- [39] Thomas Jaenich, Graham McDonald, and Iadh Ounis. 2023. ColBERT-FairPRF: Towards Fair Pseudo-Relevance Feedback in Dense Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*. Springer, 457–465.
- [40] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembroke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Trans. Inf. Syst.* 25, 2 (apr 2007), 7–es. doi:10.1145/1229179.1229181
- [41] Daniel Kahneman. 2011. Thinking, Fast and Slow. *Farrar, Straus and Giroux* (2011).
- [42] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for

- essay writing task. *arXiv preprint arXiv:2506.08872* 4 (2025).
- [43] J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *biometrics* (1977), 159–174.
- [44] Chung Kwan Lo. 2023. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education sciences* 13, 4 (2023), 410.
- [45] Khalid Mahmood. 2016. Do people overestimate their information literacy skills? A systematic review of empirical evidence on the Dunning-Kruger effect. *Communications in Information Literacy* 10, 2 (2016), 3.
- [46] Kerstin Mayerhofer, Rob Capra, and David Elswiler. 2025. Blending Queries and Conversations: Understanding Trust, Verification, and System Choice in Search and Chat Interactions. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*. 168–178.
- [47] Sarah McGrew. 2024. Teaching Lateral Reading: Interventions to Help People Read Like Fact Checkers. *Current Opinion in Psychology* 55 (2024), 101737.
- [48] Sarah McGrew and Alex C Glass. 2021. Click Restraint: Teaching Students to Analyze Search Results. In *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021*. International Society of the Learning Sciences.
- [49] Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. 2011. Measuring Improvement in User Search Performance Resulting from Optimal Search Tips. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 355–364.
- [50] Emiliana Murgia, Maria Soledad Pera, Monica Landoni, and Theo Huibers. 2023. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity?. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 311–316.
- [51] Mohsin Murtaza, Chi-Tsun Cheng, Mohammad Fard, and John Zeleznikow. 2024. Transforming Driver Education: A Comparative Analysis of LLM-Augmented Training and Conventional Instruction for Autonomous Vehicle Technologies. *International Journal of Artificial Intelligence in Education* (2024), 1–38.
- [52] 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 *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1325–1340.
- [53] Erynn J Newman and Norbert Schwarz. 2024. Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology* 56 (2024), 101778.
- [54] Rafal Niedbal, Adam Sokolowski, and Artur Wrzalik. 2023. Students' Use of the Artificial Intelligence Language Model in their Learning Process. *Procedia Computer Science* 225 (2023), 3059–3066.
- [55] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Information Processing & Management* 50, 1 (2014), 218–234.
- [56] Sandeep Pandey, Kedar Dhamdhere, and Christopher Olston. 2004. WIC: A General-Purpose Algorithm for Monitoring Web Information Sources. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30* (Toronto, Canada) (VLDB '04). VLDB Endowment, 360–371.
- [57] Folco Panizza, Piero Ronzani, Carlo Martini, Simone Mattavelli, Tiffany Morisseau, and Matteo Motterlini. 2022. Lateral reading and monetary incentives to spot disinformation about science. *Scientific Reports* 12, 1 (2022), 5678.
- [58] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) (ICTIR '17). Association for Computing Machinery, New York, NY, USA, 209–216. doi:10.1145/3121050.3121074
- [59] Tenelle Porter, Abdo Elnakouri, Ethan A Meyers, Takuya Shibayama, Eranda Jayawickreme, and Igor Grossmann. 2022. Predictors and Consequences of Intellectual Humility. *Nature Reviews Psychology* 1, 9 (2022), 524–536.
- [60] Stephen Prochaska, Kayla Duskin, Zarine Kharazian, Carly Minow, Stephanie Blucker, Sylvie Venuto, Jevin D West, and Kate Starbird. 2023. Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 US presidential election. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–39.
- [61] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) (HT '21). Association for Computing Machinery, New York, NY, USA, 189–199. doi:10.1145/3465336.3475101
- [62] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). Association for Computing Machinery, New York, NY, USA, 1245–1254. doi:10.1145/1978942.1979127
- [63] Sofia Stamou and Eftimis N Efthimiadis. 2010. Interpreting user inactivity on search results. In *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28–31, 2010. Proceedings* 32. Springer, 100–113.
- [64] Lisa Stauch, Denise Renninger, Pia Rangnow, Anja Hartmann, Lisa Fischer, Kevin Dadaczynski, and Orkan Okan. 2025. Digital Health Literacy of Children and Adolescents and Its Association With Sociodemographic Factors: Representative Study Findings From Germany. *Journal of Medical Internet Research* 27 (2025), e69170.
- [65] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- [66] Rachel R Van Der Stuyf. 2002. Scaffolding as a teaching strategy. *Adolescent learning and development* 52, 3 (2002), 5–18.
- [67] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. *arXiv preprint arXiv:2310.09219* (2023).
- [68] Ryen White. 2013. Beliefs and Biases in Web Search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 3–12.
- [69] Ryen W White. 2014. Belief Dynamics in Web Search. *Journal of the Association for Information Science and Technology* 65, 11 (2014), 2165–2178.
- [70] Ryen W White. 2024. Advancing the search frontier with AI agents. *Commun. ACM* 67, 9 (2024), 54–65.
- [71] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting Good Abandonment in Mobile Search. In *Proceedings of the 25th International Conference on World Wide Web*. 495–505.
- [72] Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. (2017).
- [73] Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Zeijiang Shen, Yoon Kim, and Monica Agrawal. 2025. Retrieval-augmented systems can be dangerous medical communicators. *arXiv preprint arXiv:2502.14898* (2025).
- [74] Zhijing Wu, Mark Sanderson, B Barla Cambazoglu, W Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) (CIKM '20). Association for Computing Machinery, New York, NY, USA, 1635–1644. doi:10.1145/3340531.3412017
- [75] Zhijing Wu, Mark Sanderson, B Barla Cambazoglu, W Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th acm international conference on information & knowledge management*. 1635–1644.
- [76] Yunji Xu, Bernard CY Tan, and Li Yang. 2006. Who will you ask? An empirical study of interpersonal task information seeking. *Journal of the American Society for Information Science and Technology* 57, 12 (2006), 1666–1677.
- [77] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing Credibility Judgment of Web Search Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). Association for Computing Machinery, New York, NY, USA, 1235–1244. doi:10.1145/1978942.1979126
- [78] Yusuke Yamamoto and Takehiro Yamamoto. 2018. Query Priming for Promoting Critical Thinking in Web Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 12–21.
- [79] Yuyu Yang, Kelsey Urgo, Jaime Arguello, and Robert Capra. 2025. Search+ Chat: Integrating Search and GenAI to Support Users with Learning-oriented Search Tasks. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*. 57–70.
- [80] Ramazan Yilmaz and Fatma Gizem Karaoglan Yilmaz. 2023. The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence* 4 (2023), 100147.
- [81] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. *arXiv preprint arXiv:2403.02901* (2024).
- [82] Tao Zhou and Songtao Li. 2024. Understanding user switch of information seeking: From search engines to generative AI. *Journal of librarianship and information science* (2024), 09610006241244800.
- [83] Yanmengqian Zhou and Lijiang Shen. 2022. Confirmation bias and the persistence of misinformation on climate change. *Communication Research* 49, 4 (2022), 500–523.
- [84] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Privacy nudging in search: Investigating potential impacts. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 283–287.