RESEARCH-ARTICLE

"Although Powerful, it's not Infallible": Investigating Academic Researchers' Verification Challenges with LLMs

MONICA VISANI SCOZZI, City St George's, University of London, London, U.K.

STEPHANN MAKRI, City St George's, University of London, London, U.K.

PRANAVA MADHYASTHA, City St George's, University of London, London, U.K.

Open Access Support provided by:

City St George's, University of London

Published: 22 March 2026

CHIIR '26: 2026 ACM SIGIR Conference on Human Information Interaction and Retrieval
March 22 - 26, 2026
Seattle, USA

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3786304.3787865
ISBN: 9798400724145

"Although Powerful, it's not Infallible": Investigating Academic Researchers' Verification Challenges with LLMs

Monica Visani Scozzi*
Centre for HCI Design
City St George's, University of London
London, United Kingdom
monica.visani-scozzi@citystgeorges.ac.uk

Stephann Makri
Centre for HCI Design
City St George's, University of London
London, United Kingdom
Stephann@citystgeorges.ac.uk

Pranava Madhyastha
City St George's, University of London
London, United Kingdom
Computer Science
The Alan Turing Institute
London, United Kingdom
Pranava.Madhyastha@citystgeorges.ac.uk

Abstract

LLMs have great potential for shaping how people find and understand information. However, current tools can struggle to provide authoritative sources, fabricate plausible references, and present obstacles to assessing truthfulness of their outputs. Understanding how users verify LLM outputs is particularly important in scholarly disciplines where information produced becomes the foundation of future knowledge. We investigated the factors that influence academic researchers' decisions to verify LLM responses, their verification strategies, and the effectiveness of those strategies. We conducted a naturalistic think-aloud study, followed by a semi-structured interview, where we observed 16 researchers across disciplines using LLMs of their choice to conduct a research information-seeking task. Our findings highlight that prevailing LLM design can hamper users' ability to satisfy their information needs for several reasons, such as lack of transparency about sources used in LLM outputs and lack of faithfulness of LLM outputs to the source. Based on these findings, we discuss how future LLMs can better support users in effective verification.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

HCI, LLM, Generative AI, Verification, Information seeking, Information behaviour, Trust

ACM Reference Format:
Monica Visani Scozzi, Stephann Makri, and Pranava Madhyastha. 2026. "Although Powerful, it's not Infallible": Investigating Academic Researchers' Verification Challenges with LLMs. 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/10.1145/3786304.3787865

1 Introduction

Large Language Models (LLMs) have become increasingly adopted for information retrieval tasks [11], yet concerns persist about the accuracy and reliability of their responses [20]. While LLM outputs often appear credible and authoritative, closer examination reveals systematic inaccuracies. Recent research acknowledges that these fabrications cannot be fully eliminated through current technical approaches [25, 26]. Recent work has also demonstrated that attempts to mitigate inaccuracies through increased model parameters, contextualised training data, or integration of external knowledge bases remain insufficient [59, 60].

In parallel, research in the fields of Human-Computer Interaction and Information Science has identified significant risks associated with the recent rise of LLM generated misinformation. Misleading content produced by LLMs can reinforce false beliefs and potentially lead to harmful decision-making, particularly when users lack the knowledge or tools to verify outputs [e.g., 42, 55].

While extensive research has examined how people verify information found on the web and social media [e.g., 17, 28, 40], there is limited empirical evidence on their verification approaches when interacting with LLMs, especially for scholarly tasks. Addressing this gap is important because academic researchers represent a population with high stakes in information accuracy and established practices for evaluating sources, yet face novel challenges when these established practices meet opaque systems.

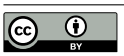
In this paper, we investigate how academic researchers verify LLM responses when using them to conduct research-related information-seeking tasks. We examine the factors that influence their decision whether or not to verify information, and the strategies they employ to assess the veracity (and relevance) of responses. Through a naturalistic observation of 16 researchers conducting self-selected scholarly tasks, we provide empirical evidence of verification behaviour in real academic contexts.

We address two key research questions: RQ1) What factors influence academic researchers' decisions to verify or accept LLM outputs during information-seeking tasks?; and RQ2) What verification strategies do academic researchers employ when interacting with LLMs, and what difficulties do they experience when employing them? Our findings contribute a detailed characterisation of verification practices with LLMs, highlighting both the strategies researchers employ and the barriers they encounter. These findings have implications for designing LLM-based tools that better support rigorous information verification.

2 Background

Even pre-internet, verifying the accuracy of information was deemed essential to human-information interaction, particularly in a scholarly research context [16]. With the advent of web search, concerns

*Main author to this research



CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/26/03
https://doi.org/10.1145/3786304.3787865

grew about whether and how users verified online information: Rieh [39] found that scholars made predictive and evaluative judgments shaped by domain and system knowledge, author credentials, source reputation, URL type, and presentation. This was extended into a three level credibility framework (construct, heuristics, interaction) [19]. Despite remaining sceptical, users often do not verify information [17], partly perceiving web search as factual retrieval [22]. Users relied instead on heuristics [35], while concerns over misinformation [28] and studies of fact-checking strategies [57] highlighted the need for better verification practices. Recent research [5, 49] shows whether people verify depends on trust, perceived effort, awareness, and truth-seeking motives. At the same time, research has worryingly demonstrated that information found online was subject to confirmation bias and memory distortions that could reinforce misinformation [29].

These longstanding issues with online verification set the stage for even more complex and qualitatively different challenges with LLMs. Firstly, LLM responses predominantly lack transparency in their generation process. Unlike search engines that return distinct sources with explicitly visible provenance, LLMs synthesise information from their training data, obscuring the origins of specific claims while presenting outputs in coherent, authoritative prose. This opacity makes it difficult for users to trace information back to verifiable sources, while perceiving the responses as confident [41]. Secondly, because LLMs are trained on vast amounts of web-sourced data [20], much of which is unverified, they may potentially inherit and propagate the same misinformation challenges that plague web content, but in a form that appears more authoritative and is harder to trace. Finally, the natural language conversational interface, long advocated as ideal for information seeking [3], paradoxically introduces new verification challenges. Given the challenges associated with both verification in general and verification of LLM outputs, it is critical to examine how academic researchers - who rely heavily on accurate information - approach verification in LLMs.

Even before LLMs' public release, Shah and Bender [41] warned that users might perceive LLM responses as confident, but not have sufficient opportunity for verification. Surfacing original sources was suggested as a potential way to mitigate the risk of false or misleading responses [42]. Furthermore, a call was made for user studies to determine whether and how users actually verify LLM responses [42]. Our study addresses this call.

Albeit with caution, researchers are increasingly adopting these tools for various research tasks [1, 37, 54]. Many researchers recognise the risks of accepting LLM-generated information without verification [33, 44], as inaccuracies in foundational knowledge can compound throughout the research process [29]. Researchers bring established verification strategies from their experience with web-based information: cross-checking multiple sources, assessing source authenticity and authority, and evaluating internal coherence, all of which remain relevant tools, as demonstrated [34].

It hence remains unclear whether these theories and pre-LLM strategies transfer to LLM interactions, and whether current tools provide adequate support for verification. Academic researchers also vary in their domain knowledge, research experience, and familiarity with LLM capabilities—factors that may influence verification behaviour. Understanding how these factors interact with LLM verification practices can guide their design to better support researchers of varying expertise, as prior studies suggest [8].

In this work, we provide empirical evidence of academic researchers' verification practices, from a human-information interaction perspective. Our aim is to understand verification behaviours to inform the design of LLM tools that better support the acquisition of accurate and trustworthy information, especially for research purposes. Prior research has highlighted the importance of accounting for user characteristics which includes domain knowledge, tool knowledge, task experience, and task complexity when studying information-seeking behaviour [9, 52]. In scholarly research, these factors are particularly salient: domain knowledge is essential to academic contribution, and both cognitive and affective states influence the information-seeking process [4, 27]. To observe verification behaviour in its most faithful form, we adopted a naturalistic observation approach rather than assigning artificial tasks to participants. In our study, the participants performed self-selected, research-related information-seeking tasks that they were already planning to complete using LLMs. This approach allowed us to observe verification behaviours that emerge naturally during scholarly work, rather than behaviours that might be prompted by experimental demands or artificial task constraints.

We now present our method and findings, which shed light on these verification behaviours.

3 Method

In this section, we present our naturalistic observation approach, which aimed to examine how academic researchers verify information when using LLMs for scholarly tasks.

3.1 Participants

We recruited 16 academic researchers through adverts distributed via university mailing lists (PhD programmes and departmental lists) and through affiliated research groups. All participants had prior experience using LLMs for information-seeking and retrieval tasks related to their research. Participants represented diverse academic disciplines and used various LLM tools (such as: ChatGPT, Gemini, Copilot, Perplexity and specialised tools like ConsensusGPT). Their academic roles ranged from PhD students to postdoctoral researchers and faculty members. This diversity in disciplinary background, academic experience, and tool familiarity enabled us to observe how different contexts and expertise levels influenced verification behaviour. We based our sample size on previous related work for exploratory qualitative research involving observation and interviews, such as [12, 32].

3.2 Eligibility and Screening

To ensure participants were likely to perform authentic research tasks, we established inclusion and exclusion criteria. Our inclusion criteria were: a) prior experience using LLMs for research-related information seeking; b) a planned information-seeking research task where they would have use an LLM; c) a suitable information-seeking task where there was the potential for verification practices to be demonstrated. We excluded three categories of tasks to focus on contexts where verification behaviour would be observable and

meaningful. These were: 1) programming tasks – which are fundamentally different tasks than information seeking; 2) language translation tasks – where verification depends on bilingual fluency and subjectivity; and 3) creative writing tasks – where assessment is subjective and does not involve factual verification. Inclusion and exclusion criteria were decided based on the authors' reflexive discussions about their own experiences when using LLMs for academic research, existing studies [50, 56] and verification opportunities: tasks had to be self-assigned by the participants, authentic to the participant's research, and planned for completion using an LLM regardless of study participation. This screening strategy was designed to ensure that we observed tasks with genuine stakes and realistic verification needs.

To ensure participants' self-selected tasks involved a substantial amount of information-seeking and had potential for verification, we asked them to send the researcher 2 or 3 possible tasks they might undertake ahead of the study. To illustrate the range of possible suitable tasks and support their task choice, we sent them a list of examples beforehand (Table 1).

3.3 Study Approach

Each session lasted one hour, consisting of: (1) the participant spending 30 minutes conducting the self-selected task(s) while thinking-aloud, with the researcher asking opportunistic questions about their behaviour and rationale [31]; (2) the researcher completing a post-task questionnaire to obtain contextual information about familiarity with the topic, Generative AI used, demographic information like age, gender, years topic researched; (3) the researcher conducting a 15 minutes semi-structured interview with the participant, to clarify the behaviour observed.

Sessions were conducted via Zoom, with both audio and screen activity recorded. The think-aloud protocol captured affective, behavioural, and cognitive processes, surfacing challenges and decision-making that might otherwise remain implicit.

3.4 Transcription and Analysis

We analysed the data through a three-stage process, combining inductive Thematic Analysis [7] with focused coding aligned to our research questions. The first author conducted the coding, while all authors collaboratively discussed codes, refining their definition and boundaries. The authors reviewed the final codes together to ensure cohesion and consistency.

Stage 1: Familiarisation and enrichment. We transcribed all think-aloud recordings and interviews verbatim. While reviewing transcripts alongside video recordings, we enriched the data by: (a) inserting descriptions of observed actions in square brackets; (b) capturing screenshots of salient interactions; and (c) adding preliminary analytical notes about observed verification behaviours. This process allowed deep familiarisation with the data while preserving contextual details that audio transcription alone would miss.

Stage 2: Inductive Thematic Analysis. We conducted an initial inductive Thematic Analysis [7] to identify general patterns in verification behaviour. During this phase, we generated codes describing observed behaviours without imposing predetermined categories. For example, we coded instances where participants checked alignment between responses and prompts, verified claims against sources, or assessed reference credibility. This stage remained open and exploratory.

Stage 3: Focused coding aligned to research questions. The richness of the qualitative data necessitated a more structured approach in the third stage. While maintaining the ethos of Thematic Analysis, we reorganised codes to directly address our research questions. We structured codes into categories descriptive of the verification experience: situations of explicit or implicit omitted verification, verification triggers and strategies adopted to verify.

We iterated through each participant transcript again, applying this focused coding scheme. Codes were then reviewed individually to assess consistency, merging or splitting them where necessary to capture distinct behaviours. For example, we initially coded various source-checking behaviours together, but later distinguished between checking source authenticity versus source authority, as these reflected different verification concerns. The information collected through the screening survey and the post-task questionnaire complemented the data in the format of tables, offering possible interpretations of the think-aloud and interview.

3.5 Participants' Choice of LLM tools

The LLM tool(s) used by each participant are reported in Table 2. Although many of the tools had different interfaces and underlying models, they were not very different from each other from a user interaction perspective and therefore we did not note substantial differences in verification behaviour across tools, apart from the following notable points: no Gemini users interacted with the *double-check functionality*; ChatGPT and Perplexity listed more sources than were referenced in context and ChatGPT responses were inconsistent in how they presented references (sometimes no references at all, only links, only list, list and in-context references).

3.6 Ethical considerations

The study received ethical approval from our departmental ethics committee. All participants provided informed consent prior to participating. While we did not initially disclose that the study focused specifically on verification behaviour (to avoid influencing naturalistic behaviour), participants were fully debriefed during the post-task interview phase. The consent process informed participants that the study was investigating 'information-seeking practices with LLMs' without specifying verification, and participants were informed they could withdraw at any time. No participants expressed concerns about this approach during debriefing.

Table 1: Task types participants undertook, categorised by their primary information need, with participant identifiers (three participants performed two tasks).

Task type	Information need	No. of tasks	Participants
Literature exploration	Identify reliable and authoritative sources	4	P1, P3, P7, P15
Ideation	Explore ideas for a research proposal	3	P2, P4, P16
Q&A	Expand own knowledge on a topic	5	P3, P6, P8, P11, P12
Summarisation	Understand a paper content to gain knowledge	4	P5, P10, P13, P14
Operational (How-to)	Find a method for a research task	3	P2, P14, P9

4 Findings: Factors influencing Verification

We identified seven factors that influenced whether and when participants chose to verify LLM responses during their information-seeking tasks. These factors often interacted, with multiple elements simultaneously shaping verification decisions. We present each factor with illustrative examples from our observations.

Table 2: Participant demographics, research context, and LLM tools used (Yrs = years on research topic; LLM Expertise = self-assessed 1–5 scale; * = free version; ** = ConsensusGPT).

Participant	Role	Tasks	Yrs	Discipline	Age range	Gender	LLM Expertise	LLM
P1	Post-doc	Literature exploration	<1	HCI	35-44	Female	4	ChatGPT-4o*
P2	PhD student	Ideation, Operational	1	Management	35-44	Female	4	ChatGPT-4o
P3	PhD student	Literature exploration, Q&A	1	Law	35-44	Male	4	ChatGPT-4o, Gemini
P4	PhD student	Ideation	<1	Accounting	45-54	Male	2	Perplexity
P5	PhD student	Summarisation	2	Management	45-54	Female	4	ChatGPT-4o
P6	PhD student	Q&A	<1	Finance	25-34	Male	5	ChatGPT-4o*
P7	Post-doc	Literature exploration	1	HCI	35-44	Male	5	ChatGPT-4o**
P8	PhD student	Q&A	<1	Journalism	45-54	Male	5	Gemini
P9	PhD student	Operational	1	Health Services	65-74	Female	3	ChatGPT-4o*
P10	PhD student	Summarisation	2	Data Visualisation	18-24	Female	4	ChatGPT-4o
P11	PhD student	Q&A	1	Data Visualisation	55-64	Female	3	ChatGPT-4o*, ChatPDF
P12	PhD student	Q&A	1	Psychology	35-44	Not declared	3	Copilot Cloud
P13	Post-doc	Summarisation	4	Data Science	35-44	Female	4	ChatGPT-4o
P14	Post-doc	Summarisation, Operational	6	HCI	35-44	Female	4	Copilot Teams
P15	PhD student	Literature exploration	5	Games studies	35-44	Female	4	Perplexity
P16	Post-doc	Ideation	4	Artificial Intelligence	35-44	Female	4	ChatGPT-4o*

4.1 Influence of task type

We classified participants' tasks according to the information need stated by the researcher as their goal when initiating the task: *Literature exploration* (acquiring new sources on a topic), *Ideation* (generating ideas in a context), *Question & Answer* (referred hereby as Q&A, expanding own knowledge on a topic), *Summarisation* (summarising a paper to refresh or distil knowledge), and *Operational* (finding a prescriptive method to perform another research task). Out of the sixteen participants, three performed two tasks during their session (P2, P3, P14). See Table 1 for a summary of the task distribution and to Table 2 for participant demographics.

Researchers approached **Literature Exploration** tasks with a strong focus on verifying relevance and credibility of suggested sources. This was driven by a general distrust in AI responses, prompting them to cross-check by following embedded links or using search engines. In all **Ideation** tasks, participants assessed response relevance, but verification varied: two researchers did not verify sources surfaced by the LLM and seemed to value the output of the ideas disregarding their origin (*"It's the idea that counts"*, P4). However, another researcher applied stricter verification, influenced by prior journalism experience and a highly structured ideation process oriented to funding opportunities. **Question & Answer** tasks (Q&A) highlighted a tendency to prioritise satisfying the information need over immediate verification, with source-checking often deferred. For **Summarisation** tasks, verification was driven by topic familiarity - known papers prompted selective verification (cognitive dissonance, prior AI inaccuracies), while new ones were often accepted without checking, reflecting perceived tool proficiency in summarisation and cost-benefit reasoning. In **Operational** tasks, participants relied on prior knowledge and postponed verification until implementation the steps suggested by the LLM, with only P14 proactively seeking additional sources.

4.2 Perceived importance of verification

Participants emphasised the importance of verification for research versus everyday life tasks, placing it at the heart of academic rigour and intellectual integrity (*"the way I've framed genAI is that you're using it as a tool...and that also comes with its own level of academic rigour and understanding" (P15)*, subjecting LLM output to critical scrutiny.

Several participants (P3, P5, P9, P10, P11, P14) postponed verification until they needed to incorporate the information found into their research (writing a paper or thesis, finalising a systematic literature review protocol or a data visualisation dataset): *"If I'm gonna use that information in my work, I go for the verification, because I'm not going to report something that I'm not sure if is accurate or not"*, (P3).

However, when researchers lacked subject matter expertise, their approaches to verification varied: P12 admitted accepting the information from their Q&A task if it sounded plausible, in contrast to others (P1, P3, P7, P16), who stressed verification as essential, disregarding topic knowledge or task type. Two researchers highlighted the heightened risk of accepting information at face-value when unfamiliar with the topic, highlighting the need for rigorous evaluation and critical thinking (P11, P15). Another participant (P10) observed that they applied strict verification when conducting core tasks such as writing, while they were softer on ChatGPT responses on side tasks. As P15 posed it: *"You can't just be like 'oh, this is gospel'. It's like: 'no, it's not gospel'. It's one nugget of a wider process that you have to leverage."* This shows that information acquired through LLMs represents only part of a more complex research process, as reflected in Byström and Hansen [8].

4.3 Cost-benefit evaluation

Researchers weighed the cost of verifying information versus perceived benefits, represented by the value of the information or the additional insight gained through verification, in line with Information Foraging theory [38]. Their decisions on whether to verify relied on prior knowledge and importance of the information: if the response resonated and induced low or no cognitive dissonance, verification was perceived as outweighing the benefit, and skipped. Six participants explicitly reported inaccuracies but accepted them without further verification, such as P10, who said *"I've noticed they've got something wrong, which is because I have looked at this paper a little bit. It was GPT-4v, not just GPT-4. But it kind of doesn't matter because it amounts to the same thing"*, (P10). In this example, the perceived benefit of verification was low, and the error inconsequential. However, P10 *did* verify information related their core objective, understanding the dataset: *"I need to know when it says 'prior research' what does it mean. Like, where is the data from.."* (P10). In summary, when the information sought was at the core of the task and the response lacked accuracy, the benefit was perceived as outweighing the cost and verification was initiated.

4.4 Perceived response accuracy

Several aspects of LLM responses influenced their perceived accuracy and amplified or attenuated researchers' need for verification: **prior knowledge** about the domain or source, **references**, **language dissonance** and **structure**.

Prior knowledge. All participants referenced prior knowledge of the topic or the source when assessing response accuracy, though it functioned in complex ways. **Prior knowledge** triggered verification when the response generated a cognitive dissonance versus what known about the topic or the source. For example, both P10 and P15 decided to verify when they thought that LLM misattributed claims in the responses. For both participants, prior knowledge raised suspicion and triggered verification. In other cases participants accepted the responses when encountering familiar authors and concepts. Other factors contributed to participants

accepting the response: **perceived cost-benefit of verifying**, **perceived trust in the LLM** and the **task's role in the wider research process**. For example, in P2's case the task was part of a broader ideation process about race and feminist movements: *"They suggested 'Intersectionality', which I'm familiar with. So I would take this and build upon Kimberlé Crenshaw"*. P2 did not further verify, stating they would later consult traditional digital libraries.

When **prior knowledge was not available**, participants found some responses useful but exercised caution to information presented, recognising their lack of knowledge. As P11 stated while looking for information about participatory process evaluation, *"This is just giving me information about what process evaluation is because I am literally just on the outside of this and I don't know what I'm doing."* Their lack of task knowledge prompted them to shift their information-seeking task from understanding the topic to its sources. Yet, in another situation the same participant decided not to verify due to the novelty of the information encountered: *"Okay. Outcomes, several methodologies. Okay. Participatory impact. I have never heard of this. So this is, I guess, useful"*. This shows how prior knowledge can swing the decision towards or away from verification, depending on other factors.

References. The presentation of **references** in plausible academic format reassured some participants (*"These references are now coming up and I like the look of that"*, P9), while made some others sceptical (*"There have been a couple of occasions where I've seen Chat produce what phantom references is"*, P2). This was in general due to prior experience with reference inaccuracies. During the study, LLMs provided participants with links leading to deprecated or obsolete pages (P3, P11) and fabricated references (P1).

In another set of examples, three participants noticed that lengthy LLM responses included a maximum of five references, while the full list of sources cited was much larger (in one case 33 sources). This triggered suspicion in two participants, who then investigated the full list of citations, but not in the third who had uploaded the selected paper for an ideation task.

Language dissonance. At times, participants questioned if the **language** in the LLM's response was a reflection of the source or a fabrication, and decided to go back to the original source to verify. For example, when reading an LLM-generated summary of a paper, P5 stated *"it seems to me that it doesn't use the language the authors used"*.

In another situation, unclear verbose responses confused P6, who concluded *"I need to ask to my professor why. [..] It's just wording but.. useless. It's just words without any meaning"*. P6 then moved to Google Search to find a better source of information. Boldness of language also triggered verification in both P10 and P15: (*"when ChatGPT gives me some bold statement and usually doesn't give you any reference, I go and look at that through publication"*, P15).

Mismatch between response and source structure. Sometimes LLM tools presented information closely mirroring the source (e.g. reproducing a complete list of items). However, other times they **structured** information not entirely matching the source. When this happened, participants (e.g. P5, P10) perceived a discrepancy and this triggered verification. For example, when examining a list of deceptive verification practices, P10 noticed that the LLM had merged two items in the list. The mismatch created an increased cognitive load and sparked verification.

4.5 Perceived trustworthiness of the LLM

Participants' trust in the LLM worked both ways: on the one hand, inaccurate information encountered in prior interactions influenced researchers to distrust response elements, like references (phantom references, links to a different paper, or a paper being misrepresented as something it was not) and quotes. On the other hand, prior positive experience with AI induced lower willingness to verify.

We observed that successful verification within a single interaction could induce a 'state of trust': P13 verified a second quote (response 4), since having previously received inaccurate quotes. The verification was successful – two parts of the quote appeared in two separate sentences in the same paragraph. P13 did not verify the two successive quotes, nor the first: *"the quote is a little bit.. Let me see. It seems to be a little bit outside. But anyway..[..] So it's correct. I just felt that it was maybe hallucinating"*, (P13). Nine researchers expressly stated they did not trust LLMs (e.g. P2, *"although it is powerful, it's not infallible"*), yet stated using it because of its usefulness. Some researchers demonstrated awareness of stochastic word generation in LLMs, and approached interactions with LLMs cautiously. For example, P16 stated *"the verification there is important when I'm building my own knowledge about certain topic so I want to make sure that I'm not just adapting certain random word that has been concatenated based on some stats"* .

4.6 Illusion of Relevance

LLM responses sometimes created an *Illusion of Relevance* by echoing the prompt. This resulted in many participants believing the content aligned to the prompt due to two key cues: 1) the opening paragraph often paraphrased the prompt (e.g. P1, P4, P10, P12), setting the expectation of alignment; 2) listed sources appeared relevant based on their descriptions or titles, but closer inspection revealed they did not actually support the implied claims (e.g. P1, P2, P7).

As P2 noted, *"maybe because I'd used a certain word it surfaced a certain paper, but it didn't pertain to the question that I had asked or the prompt...given. But I wouldn't know that until I really checked the paper, because...the title was quite relevant"*. P2 had learnt the necessity of verification due to the apparent relevance, which increased participants' cognitive cost of verification: they had to inspect sources not relevant, and at times felt compelled to revisit their prompts and previous responses, initially assuming an oversight on their part as the cause for irrelevant sources, only to realise it was actually the LLM's behaviour.

4.7 Conversation integrity

When the integrity of the conversation was compromised due to inconsistencies across responses or the *Illusion of Relevance*, some researchers (P1, P10, P7) initiated verification of the interaction history, to assess the cause of the inconsistency: was it the result of an inaccurate prompting, or was the response contradictory or inaccurate in relation to the prompt? This inconsistency manifested as cognitive dissonance, resonating with Borlund's concept of a

'cognitive relevance check' [6]: users noticed conversation misalignments and violations of conversational principles, as in Grice's pragmatics [18]. This intriguing behaviour was induced by LLM responses.

As an example, P7 was using an LLM to identify papers about the use of storytelling in participatory design for health conditions. They seemed confused because response 2 did not focus on papers about 'storytelling'. They revisited their first prompt, realised it included the word 'narratives', momentarily considered this as the originating issue, but then noticed that the first paragraph of response 1 did indeed mention 'storytelling'. They concluded about both responses: *"oh, so it says here 'storytelling'...but it didn't suggest, I think, storytelling papers as such"*, (P7). P7 thought that their request for 'storytelling'-related sources had propagated through the interaction, but realised only upon an effortful verification of the chat history that it had been missed from the outset of the conversation. P7 detected this AI sycophancy and realised their misunderstanding.

4.8 Summary of factors

Our findings demonstrate that task complexity, prior knowledge and problem structure interact to influence academics' decisions on whether or not to verify LLM responses. We found that inaccuracies in LLM responses triggered the need to verify for source authenticity, while also strengthening the need to check for authority and relevance. Academics determined the importance of verification based on situational attributes in Byström and Hansen [8]. In particular, the research task goal and the role of information-seeking in the wider research process dictated academics' verification priorities. Also, researchers' perceived usefulness of the LLM tool strongly influenced their perception of credibility of the responses. When academics identified misalignment between successive LLM responses or between references and responses, the *illusion of relevance* was broken, triggering the perceived need for verification. Finally, the simplistic interfaces and lack of transparency of LLMs increased the influence of tool knowledge on meta-cognitive awareness, as found in prior research [4, 9, 52].

5 Findings: Verification strategies and their effectiveness

Researchers adopted various strategies to verify information found, often deciding to leave the LLM tool and use search engines, digital libraries or other AI tools. Verification strategies were often applied in parallel, complementing each other. Some of their verification strategies included **finding sources** to support claims, **assessing authenticity** and **authority** of sources, **interacting with sources** to check them in detail, **re-prompting** with focus, **comparing responses to prompts** to understand the root cause of response misalignment, or **using other LLMs** to cross-check information. We present each of these strategies in the following subsections.

5.1 Finding sources

Among the most adopted strategies, researchers looked for sources to support the claims in the response. This strategy was triggered by the need for cross-verification, or when sources were missing in the LLM's response. For example, when looking at a response with no embedded references, P11 stated *"these are nice ideas which I can look at, but I need to know where they've come from"* and re-prompted for sources.

Researchers would either re-prompt in the LLM tool (*"give me the list of some accurate and reliable resources"*, P3), or employ Google, Google scholar or other search tools, as with P8: *"'give me the verifiable sources' and I go to their papers. Because I don't differentiate...the fabricated response from the credible one"*, (P8).

This re-prompting strategy failed when the LLM was not capable of providing sources aligned to a claim in the response. For example, the LLM tool struggled to find relevant prior research in a novel research area for P1. This led P1 to re-prompt three times for sources. P1 received references that were fabricated references, yet formatted in an academic style, and the same two links were repeated across the responses, despite being only partially relevant.

5.2 Assessing source authenticity

A few researchers distrusted the LLM tool due to prior experiences of fabricated references, and purposely checked authenticity of sources by opening the references, copying a reference and searching it in Google, Google Scholar or other digital libraries: P16 searched project names in Google, distrusting the LLM: *"So here it gives me some names...This is I'm assuming one project. So I'm going to look at it, to make sure it's not hallucinating"*. P16's verification strategy was external to the LLM, and similarly P7 searched for a reference directly in Google Scholar, remarking *"if I go here, then that's Google Scholar, and I will verify the publication exists"*.

Participants' behaviour reflected their perception of the tool as not capable to offer the reliability and accuracy required for their research task. Yet, their strategy failed when a link in the response led to a missing or different web page (P3 and P11 respectively), or when references were fabricated, as in P1's example, where the reference could not be found in the page linked in the LLM's response, nor in Google Scholar.

5.3 Assessing source authority

Those participants who engaged in literature exploration, Q&A and some ideation tasks explicitly stated the need for authoritative sources, discarding blogs, Wikipedia and other non-academic sources. As P15 stated, *"some of the stuff where it's like they've written a blog. I'm like, okay, like there's not going to be anything too empirical"*. They frequently discarded sources they did not consider academically reliable.

Authority assessment was performed first using the information in the response (year, authors), and then by opening the source to conduct authenticity and further authority checks. After looking at a source in a publishing journal, P7 discarded it since *"it's in Japanese Psychological Research, so I probably won't look at this because it doesn't seem like a journal that would be highly ranked"*. As with other participants (e.g. P11, P16), P7 sought publications deemed particularly reputable in academia. In P7's example, source authority could not be fully assessed in the LLM's response, due to missing publisher information. Another participant (P16) used an articulated process, by looking for the author in a research directory (DBLP) to assess their publications, academic activity and affiliations. Other researchers also mentioned this approach, stating

they would use it at a later stage, when examining sources more closely. As P8 put it, *"I got names and papers suggested by Gemini. I go to Google Scholar and I look for their papers, I read a little bit of their papers and then I go to the profile...Google Scholar gives you in fact the name and where that individual works and how renowned he is and things like that"*. In this case, P8 implied they would perform authority verification more thoroughly later. What was evident across examples was that source authority could not be verified within the LLM tool, so participants used external resources.

5.4 Interacting with sources

Cognitive dissonance often triggered doubt in our participants, either due to nuances versus prior knowledge about the topic or the source, or due to an intuition of relevance misalignment, as explained by Sperber and Wilson [46]. Participants adopted the most common verification strategy: they would find the original source and see if the claim was aligned to the source. This strategy was laborious, leading them outside of the LLM tool. Only occasionally did the LLM provide a precise indication about the location of the claim. For example, P14 uploaded a PDF for a summarisation task and the LLM provided the page number of the quote mentioned in the response. In general, researchers interacted with sources either to verify claims, relevance and/or authority. When the source was an academic paper, researchers would often consult the abstract - also achieving authority verification.

When opening links directly from the response, researchers assessed the content of the web page and discarded sources if deemed not authoritative or relevant. For example, P7 opened a link in Consensus, which erroneously labelled the paper as a 'systematic review'. However, P7's strategy failed, since they trusted Consensus's label and believed the paper was a systematic review, despite distrusting ConsensusGPT itself, which referred to the paper as a 'study'. P7 discarded the paper as they said they were *"not interested in systematic reviews"*. When verification was triggered by specific claims, researchers opened the source and looked for the point mentioning the claim: P13 copied the text 'mind perception theory' from ChatGPT, moved to the source paper and pasted it in the search function using CTRL+F. As no results were found, they removed 'theory' from the search query and found where 'mind perception' was mentioned. The strategy of finding text in the source by using the search function failed for other participants, either due to difficulty committing the exact wording used by the LLM to memory, or due to the LLM rephrasing concepts that were mentioned with nuanced difference in the source.

5.5 Re-prompting with focus

Re-prompting was used as a general strategy to address insufficient accuracy in responses: participants re-prompted to obtain *"verifiable, credible journals and sources"* (P8), to ask for 'academic' sources (P1), or to address inadequate response depth or tone. P12 was conducting research on inferentialism, re-prompted Copilot to ask why it defined an example in their previous prompt as a 'deductive inference,' and P14 explicitly asked the LLM tool if a paper did *"include the word <impairment X>"*: both participants needed more precise information, and the re-prompt aimed to address information needs not satisfied in the previous interaction. This strategy was not always successful. In these examples, while P14 achieved their goal, P12 did not achieve clarity and sought a known paper as a source of reliable information.

5.6 Comparing responses to prompts

Participants were confused by responses that, at first, seemed aligned to their prompt (paraphrasing it in the first paragraph), but later proposed information that was not actually entirely related to the prompt (see P7's example in section 4.7). In these cases, participants inspected their prompts for accuracy, assuming an ommision on their part. When they found their prompt sufficiently accurate, they engaged in a second type of verification - comparing parts of the response to the prompt. This form of verification was laborious, due to the initial apparent alignment. It was only by continuing to read the response that they realised the paragraphs that followed did not align to it, revealing AI sycophancy, as P1 observed: *"it actually talks about <impairment 1> at the top. But the result is for blind people, so it doesn't connect it"*.

5.7 Using other LLMs

P6 used two LLMs in parallel throughout the observation, and some other participants (P8, P9, P15) explained that they would perform their information-seeking task across multiple LLM tools. This strategy was adopted to optimise the outcome of the task, but also as a form of cross-verification. P3 commented: *"Gemini is telling me: 'Factors favoring a strong regulation'...Okay, so we have some common answer"*. Other participants decided not to rely on the LLM's description of a source, which they deemed insufficiently accurate, and instead used ChatPDF to summarise the original source. This strategy was successful but convoluted, since it first involved downloading the PDF file from the publication web page, then uploading it into ChatPDF. P6 deserves particular mention as, in order to understand a formula, they asked ChatGPT an example. But as they did not trust the AI calculations, they used a development environment to verify the values in the example.

5.8 Summary of strategies

In this section, we reported the verification strategies most used by academic researchers. These included: locating cited sources (either through the LLM tool or independently), verifying source authenticity and/or authority, examining original sources to verify specific claims, re-prompting to obtain more accurate responses, and comparing the prompt with the response. Other (less common) strategies observed but not reported involved comparing responses, comparing sources to prior knowledge and consulting other people.

6 Discussion

In this section, we discuss limitations in current LLM design that we identified through our findings: **faithfulness to sources, task-specific verification needs, apparent conversation integrity, lack of transparency about interaction actions-consequences**. We outline future research directions and reflect on limitations.

6.1 Need for user-centered sources

Academic tasks require interaction with original sources and researchers consistently verified information directly from these

sources rather than relying solely on the LLM tool. Participants experienced several barriers during source verification: *technical barriers*, such as links leading to incorrect pages or responses listing more references than actually used; *information accuracy barriers*, (references that were fabricated or not deemed authoritative); and *trust-related barriers*, as responses sometimes cited references that appeared relevant but were ultimately misleading, breaking the user's trust. Even research-specific tools (i.e. ConsensusGPT) failed to reliably reflect underlying sources by citing sources that, upon scrutiny, were not relevant.

In academic research, where terminology precision is critical, generic LLMs often fall short and research has examined alternatives, such as context-based LLMs and using external knowledge. In both cases researchers reported improved performances [10, 21, 36], yet studies also showed that even under those conditions LLMs generated credible but inaccurate responses [10, 21, 30, 59] and unreliably summarised information [60], aligning with our findings.

These barriers suggest deeper problems in how LLMs represent and reference source material, rooted in the design and training of LLMs and exacerbated by outdated or incomplete training data – something participants did not seem consciously aware of. This suggests the need for LLMs aimed at supporting knowledge work to provide accessible original metadata when referencing sources, in alignment with Information Foraging Theory [38]. The information sample would be faithful to the patch, better supporting trust in the LLM and efficiency, rather than undermining it.

6.2 Better supporting verification

We observed that research tasks were broken down in sub-tasks, each with its own verification needs, based on task type and domain knowledge. Our findings align with Shah et al.'s model of search tasks as trees rather than linear [43], where tasks on each level could belong to any of Byström and Järvelin's task complexities [9]. A researcher may follow a defined strategy for assessing the authenticity of a paper, but may operate on a higher level when deciding if a paper is worthy of reading. LLMs could follow a researcher from macro-tasks to actions, and adapt how information seeking and verification are supported. While task structure informs verification needs, our findings highlight the critical role of user knowledge – specifically domain knowledge, research expertise, and tool expertise - in shaping verification behaviour, as seen elsewhere [54].

Domain knowledge changes during tasks, especially Q&A ones [4], but also constantly supports users in acquiring accurate information [9]. While it generally aids verification, we observed highly knowledgeable users noticing nuances, which triggered cognitive dissonance and sparked verification, whereas researchers new to a topic often did not verify at all. When users do not have sufficient domain knowledge, they risk taking responses at 'face value'. In this scenario, LLM tools could be designed to encourage and support verification, and provide explanatory cues for the responses, as suggested by prior work on Explainable AI [15, 45, 51].

Furthermore, **familiarity with the tool** itself shaped user expectations of the quality of its outputs. These expectations align with Subramonyam et al.'s [47] concept of the 'gulf of envisioning' and Belkin's Anomalous states of Knowledge [4]: users must bridge the cognitive gap between their intention to prompt and its formulation. Only by possessing sufficient knowledge about the domain and the tool (its capabilities, training data, and prompting techniques), can they prompt effectively. This knowledge helps form an expectation about the response, which users then use to assess its quality (see P3, *"I have more than basic information about the probable answers"*).

Finally, **research expertise** also influences academic verification rigour. Although our study did not isolate this factor, prior research has shown how Information presumed true when first encountered persists [14] and continues to influence reasoning. For this reason, LLMs for academics should, by design, provide structured verification support, especially to new researchers who trust LLMs for understanding novel concepts and postpone verification until later - a moment which may never come.

Given these findings, we recommend that LLMs should actively support users, especially those who are not yet experienced in research practice, subject matter, or tool usage, by respectively adapting verification support to the sub-task, encouraging claim verification and guiding academics on how to use LLM tools responsibly rather than only reinforcing their limitations [54].

6.3 Combatting the illusion of relevance

Our findings showed that participants had to verify both the authenticity *and* the relevance of the information provided by LLMs. Relevance judgement is recognised as a key activity in interactive information retrieval [8, 52], yet is not as straightforward as it seems, as the user may not always have a clear idea of their information needs, as these often evolve and become clearer during the search process [4]. In traditional web-based information seeking, we would have expected to observe complex relevance verification approaches in Q&A tasks and more nuanced ones in e.g. literature exploration. However, we observed intense relevance verifications across all tasks. An explanation can be found in the concept of *pragmatics*, where relevance lays the foundation of the conversation between individuals [46].

As explained in our findings, LLMs created an *Illusion of Relevance* in their responses, reinforced by an authoritative confident tone and by the tendency to rephrase the user's prompt in the first paragraph. At first glance, this behaviour mirrors the human conversational strategy of restating a question to confirm understanding. As Thomas et al. pointed out, the Cooperative Principle in Grice's Pragmatics suggests that effective conversation relies on participants making contributions that are appropriate to the purpose and direction of the exchange [18, 48].

When LLMs repeat the prompt in their first paragraph, they appear to follow Grice's principle and maxims of Relation (staying relevant) and Manner (being clear). However, this behaviour in LLMs breaks down when they restate the prompt at the beginning of the response: in human conversation, this has the intention of ensuring that the interpretation of the question is accurate before proceeding, and implies two important assumptions: first, the speaker is offering the listener the opportunity to correct them, and second, the speaker successive utterance would be in alignment to their understanding. LLMs violate both assumptions. They do not allow for real-time correction and their follow-up content often

diverges from the apparent understanding they signalled. This can confuse users and undermine trust.

Perhaps unsurprisingly, participants developed preventive strategies, like ignoring the first paragraph (P4), and initiating a new chat in an attempt to regain control of the consistency of the conversation (P10, P11). They did this out of concern that the LLM would start confusing prompts and lose focus. Essentially, participants were trying to mitigate for the LLM's conversation shortcomings and lack of transparency.

6.4 Prompting the prompters

Our participants varied in tool expertise, from occasional users to highly knowledgeable researchers working in the AI field. By design, in our study all participants had prior experience with LLMs. Some aspects of the opacity of LLM tools surfaced through the observations - related to prompting strategies, tool settings and the impact of training data versus online retrieval and context window limitations. Prompting difficulties deserve particular attention, since they are at the heart of task formulation; in our study, prompts often generated unsatisfactory responses that triggered lengthy verification approaches. As designing LLMs for underspecified prompts is a fragile approach [58], it may be more fruitful to design tools that support users in prompt construction (e.g. [53]).

6.5 Reflections on the naturalistic approach

The think-aloud protocol elicited participants' thoughts and surfaced the slow and fast thinking theorised by Kahneman [24], but also demonstrated seamless transition between the two [13], based on sub-task relevance and cognitive dissonance triggered. With the think-aloud method, the majority of fast thinking decisions *not* to verify were invisible to the researcher. Only more conscious, stated decisions to verify (or not) were observable. Future research could investigate thought processing when verifying LLM responses and how cognitive dissonance arises (see for e.g. [23]).

Due to the naturalistic nature of the study, some participants experienced unexpected events during the session, introduced either by design (choice between answers, new functionality like generating charts) or by programming error (e.g. links in ConsensusGPT intermittently broken). Users had to alter their process to adapt to the event. On the one hand, this provided an opportunity to observe users' recovery strategies. On the other hand, this may have meant that their verification strategies were altered due to tool unpredictability (e.g. with P2, P5, P7). As an example, P7 had to choose between two possible responses provided by ConsensusGPT. Both were, in fact empty responses, since the custom GPT expected to be granted permission to retrieve each response from Consensus. P7 randomly selected a response, and then granted the permission.

6.6 Design implications and future research

Based on our findings, we propose several implications for the design of future LLMs, especially for academic use. Future conversational LLM systems for information retrieval must provide a faithful representation of authentic sources – a description aligned to the paper content, rather than to the prompt, and reliable data sources, as comprehensive research already suggests [59].

Their design should also return agency to users, by promoting a clearer distinction between the accountability and ownership of system retrieval versus user retrieval. As Bates [2] argues, the role of user and system in information retrieval should be a design choice: LLMs should be designed with how information retrieval is facilitated by the system and how information seeking is conducted by the user in mind. LLM design currently blurs the line between the role of the human versus the role of the LLM in interactive information retrieval, and can leave the user confused or conflicted - between high perceived usefulness and low trust.

Designers could consider introducing explicit alignment steps for addressing non-relevant responses, through a tighter feedback loop built into the interactive conversation. This loop could support users in providing and LLM systems in incorporating useful relevance feedback - already an established information retrieval technique [8]. Similarly, LLMs could provide greater transparency about which parts of responses are based on parametric information (from the training dataset) versus retrieved information (from the Web).

Finally, our study surfaced aspects of the interaction that deserve further investigation: the intersection of domain knowledge, research expertise and tool understanding plays a key role in the decision to verify information in LLMs, as seen in traditional information behaviour literature [16]. Further research could consider how each of these factors interplay with the others when interacting with opaque systems and responses. In this regard, we observed unconscious misconceptions about how LLMs operate, skewing interactions even in LLM expert users. Research could investigate further how the conversational interaction may influence those misconceptions, beyond anthropomorphism.

7 Conclusion

Our naturalistic study highlighted a critical tension in academic information seeking with LLMs: the need to rely on original sources versus the limitations of current LLMs in supporting users in doing so. To investigate this tension, our study examined how researchers engaged with LLMs during academic information-seeking tasks, focusing on their verification behaviours and trust in the responses. Two key aspects of verification were examined: (1) factors that influenced their decision on whether or not to verify — such as perceived accuracy of the response relative to the user's domain knowledge, and overall trust in the LLM tool and (2) their verification strategies — which included consulting original sources, comparing prompt to response and assessing authenticity and authority.

Our findings illustrate that transparent source selection, improved source faithfulness, accurate alignment between responses and their sources and support for verification and prompting are not merely desirable technical enhancements, but *essential design requirements*. To usefully support academic workflows, LLMs must rely on accurate representation of sources through a corpus of information, enabling users to make informed decisions about whether or not to pursue further exploration. Furthermore, conversational information-seeking tools should engage in a meaningful exchange by empowering and supporting users to articulate their information needs and verify that the information provided is meeting them, rather than making assumptions about the user's intent and making it difficult to verify their outputs. Only by addressing these limitations can LLMs evolve from the current tools that risk undermining scholarly rigour to trustworthy partners in academic research.

References

[1] Abdulrahman M Al-Zahrani. 2024. The impact of generative AI tools on researchers and research: Implications for academia in higher education. *Innovations in Education and Teaching International* 61, 5 (2024), 1029–1043.

[2] Marcia J Bates. 1990. Where should the person stop and the information search interface start? *Information Processing & Management* 26, 5 (1990), 575–591.

[3] Nicholas J. Belkin, Pier Giorgio Marchetti, and Colleen Cool. 1993. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information processing & management* 29, 3 (1993), 325–344.

[4] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for information retrieval: Part I. Background and theory. *Journal of documentation* 38, 2 (1982), 61–71.

[5] Sarbottam Bhagat, Russell R Torres, and Dan J Kim. 2025. Curbing Dissemination of Fake News: Role of Information Verification Cost, Trust in Information, Truth-Seeking, and Fake News Awareness on Information Verification Behavior. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 56, 2 (2025), 31–58.

[6] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for information Science and Technology* 54, 10 (2003), 913–925.

[7] Virginia Braun and Victoria Clarke. 2021. *Thematic analysis: A practical guide*. SAGE publications Ltd.

[8] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information science and Technology* 56, 10 (2005), 1050–1061. doi:10.1002/asi.20197

[9] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management* 31, 2 (1995), 191–213.

[10] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip Yu, and Lichao Sun. 2025. A survey of ai-generated content (aigc). *Comput. Surveys* 57, 5 (2025), 1–38.

[11] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt*. Technical Report. National Bureau of Economic Research.

[12] Mira Crouch and Heather McKenzie. 2006. The logic of small samples in interview-based qualitative research. *Social science information* 45, 4 (2006), 483–499.

[13] Wim De Neys. 2023. Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences* 46 (2023), e111.

[14] Ullrich KH Ecker, Stephan Lewandowsky, Briony Swire, and Darren Chang. 2011. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic bulletin & review* 18, 3 (2011), 570–578.

[15] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The who in XAI: how AI background shapes perceptions of AI explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–32.

[16] David Ellis. 1989. A behavioural approach to information retrieval system design. *Journal of documentation* 45, 3 (1989), 171–212.

[17] Andrew J Flanagin and Miriam J Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society* 9, 2 (2007), 319–342.

[18] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.

[19] Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information processing & management* 44, 4 (2008), 1467–1484.

[20] Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data intelligence* 6, 1 (2024), 201–239.

[21] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* (2024).

[22] Isto Huvila. 2013. In Web search we trust? Articulation of the cognitive authorities of Web searching. *Information Research* 18, 1 (March 2013). https://informationr.net/ir/18-1/paper567.html

[23] Kaixin Ji, Danula Hettiachchi, Flora D Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing information seeking processes with multiple physiological signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1006–1017.

[24] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[25] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *arXiv preprint arXiv:2509.04664* (2025). doi:10.48550/arXiv.2509.04664

[26] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. https://openai.com/index/why-language-models-hallucinate/

[27] Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science* 42, 5 (1991), 361–371.

[28] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[29] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[30] Subhankar Maity and Manob Jyoti Saikia. 2025. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* 12, 6 (2025), 631.

[31] Stephann Makri, Ann Blandford, and Anna L Cox. 2010. This is what I'm doing and why: reflections on a think-aloud study of dl users' information behaviour. In *Proceedings of the 10th annual joint conference on Digital libraries*. Association for Computing Machinery, New York, NY, USA, 349–352. doi:10.1145/1816123.1816177

[32] Kerstin Mayerhofer, Rob Capra, and David Elsweiler. 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.

[33] Lisa Messeri and Molly J Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 8002 (2024), 49–58.

[34] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology* 58, 13 (2007), 2078–2091.

[35] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.

[36] Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, Vol. 11. MDPI, 57.

[37] Helen Pearson. 2024. Can AI review the scientific literature—and figure out what it all means? *Nature* 635, 8038 (2024), 276–278.

[38] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[39] Soo Young Rieh. 2002. Judgment of information quality and cognitive authority in the Web. *Journal of the American society for information science and technology* 53, 2 (2002), 145–161.

[40] Mirjam Seckler, Silvia Heinz, Seamus Forde, Alexandre N Tuch, and Klaus Opwis. 2015. Trust and distrust on the web: User experiences and website characteristics. *Computers in human behavior* 45 (2015), 39–50.

[41] Chirag Shah and Emily M Bender. 2022. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 221–232.

[42] 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.

[43] Chirag Shah, Ryen White, Paul Thomas, Bhaskar Mitra, Shawon Sarkar, and Nicholas Belkin. 2023. Taking search to task. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 1–13.

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

[45] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551.

[46] Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Vol. 142. Harvard University Press Cambridge, MA.

[47] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[48] Paul Thomas, Mary Czerwinksi, Daniel McDuff, and Nick Craswell. 2021. Theories of conversation for conversational IR. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–23.

[49] Russell Torres, Natalie Gerhart, and Arash Negahban. 2018. Epistemology in the era of fake news: An exploration of information verification behaviors among social networking site users. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems* 49, 3 (2018), 78–97.

[50] Johanne R Trippas, Damiano Spina, and Falk Scholer. 2024. Adapting generative information retrieval systems to users, tasks, and scenarios. In *Information Access in the Era of Generative AI*. Springer, 73–109.

[51] Chuan-Ching Tsai, Jin Yong Kim, Qiyuan Chen, Brigid Rowell, X Jessie Yang, Raed Kontar, Megan Whitaker, and Corey Lester. 2025. Effect of artificial intelligence helpfulness and uncertainty on cognitive interactions with pharmacists: Randomized controlled trial. *Journal of Medical Internet Research* 27 (2025), e59946.

[52] Pertti Vakkari. 1999. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information processing*

& management 35, 6 (1999), 819–837.

[53] Ben Wang, Jiqun Liu, Jamshed Karimnazarov, and Nicolas Thompson. 2024. Task supportive and personalized human-large language model interaction: A user study. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 370–375.

[54] Jiyao Wang, Chunxi Huang, Song Yan, Weiyin Xie, and Dengbo He. 2025. When young scholars cooperate with LLMs in academic tasks: the influence of individual differences and task complexities. *International Journal of Human–Computer Interaction* 41, 8 (2025), 4624–4639.

[55] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[56] Ryen W White. 2024. Advancing the search frontier with AI agents. *Commun. ACM* 67, 9 (2024), 54–65.

[57] Sam Wineburg and Sarah McGrew. 2019. Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record* 121, 11 (2019), 1–40.

[58] Chenyang Yang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu. 2025. What Prompts Don't Say: Understanding and Managing Underspecification in LLM Prompts. *arXiv preprint arXiv:2505.13360* (2025).

[59] Wan Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics* 13, 5 (2025), 856.

[60] 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).