Query Smarter, Trust Better? Exploring Search Behaviours for Verifying News Accuracy

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

Samy Ateia
University of Regensburg
Regensburg, Germany
samy.ateia@ur.de

Markus Bink
University of Regensburg
Regensburg, Germany
markus.bink@ur.de

Gregor Donabauer
University of Regensburg
Regensburg, Germany
gregor.donabauer@ur.de

Marcos Fernández-Pichel
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Santiago de Compostela, Spain
marcosfernandez.pichel@usc.es

Alexander Frummet
University of Regensburg
Regensburg, Germany
alexander.frummet@ur.de

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

David E. Losada
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Santiago de Compostela, Spain
david.losada@usc.es

Bernd Ludwig
University of Regensburg
Regensburg, Germany
bernd.ludwig@ur.de

Selina Meyer
University of Regensburg
Regensburg, Germany
selina.meyer@ur.de

Noel Pascual-Presa
University of Santiago de Compostela
Santiago de Compostela, Spain
noel.pascual.presa@usc.es

Abstract

While it is often assumed that searching for information to evaluate misinformation will help identify false claims, recent work suggests that search behaviours can instead reinforce belief in misleading news, particularly when users generate queries using vocabulary from the source articles. Our research explores how different query generation strategies affect news verification and whether the way people search influences the accuracy of their information evaluation. A mixed-methods approach was used, consisting of three parts: (1) an analysis of existing data to understand how search behaviour influences trust in fake news, (2) a simulation of query generation strategies using a Large Language Model (LLM) to assess the impact of different query formulations on search result quality, and (3) a user study to examine how 'Boost' interventions in interface design can guide users to adopt more effective query strategies. The results show that search behaviour significantly affects trust in news, with successful searches involving multiple queries and yielding higher-quality results. Queries inspired by different parts of a news article produced search results of varying quality, and weak initial queries improved when reformulated using full SERP information. Although 'Boost' interventions had limited impact, the study suggests that interface design encouraging users to thoroughly review search results can enhance query formulation. This study highlights the importance of query strategies in evaluating news and proposes that interface design can play a key role in promoting more effective search practices, serving as one component of a broader set of interventions to combat misinformation.

CCS Concepts

• **Information systems** → **Users and interactive retrieval**; • **Human-centered computing** → **Empirical studies in HCI**.

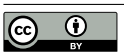
Keywords

search behaviour, misinformation, mixed methods

ACM Reference Format:
David Elsweiler, Samy Ateia, Markus Bink, Gregor Donabauer, Marcos Fernández-Pichel, Alexander Frummet, Udo Kruschwitz, David E. Losada, Bernd Ludwig, Selina Meyer, and Noel Pascual-Presa. 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 (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3726302.3730067

1 Introduction

The proliferation of false and misleading information through online media and social networks is a complex, global phenomenon influenced by large online platforms and individual and collective behaviours [52, 54]. The recent announcement by Meta that it will cease fact-checking [47] highlights a troubling shift in the fight against misinformation. Facebook's efforts had been shown to be



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730067

effective [4] and studies suggest that fact-checking helps users identify accurate information [75, 90] and boosts trust in social media platforms [1]. The removal of these efforts just underlines the need to enhance users' information literacy skills.

While much of the literature focuses on the role of social networks in spreading misinformation, the dominant influence of search engines in shaping the information environment remains under-explored [6]. This is particularly significant, since it is widely assumed that searching online to evaluate misinformation would reduce belief in it; this behaviour is common among professional fact-checkers [88] and is often taught as part of information literacy interventions [19, 57]. However, a recent study challenges this assumption. Aslett and colleagues [6] report on the results of five large-scale studies demonstrating that online search, when used to evaluate the truthfulness of false news articles, can actually increase the likelihood of believing them. These authors argue that people often use vocabulary from the source article when formulating their queries, which leads to confirmatory results from low-quality information spaces referred to as "data voids". They posit that this occurs because specialist terminology is typically not shared between high- and low-quality sources, causing search engines to return corroborating but misleading information.

In this paper, we investigate the impact of various query generation strategies on verifying news articles. We begin by analysing Aslett et al.'s data in greater detail (Sec 3), revealing that the way people search plays a larger role in determining trust in a fake news article than the mere act of searching itself. Next, we simulate different query generation strategies to examine their impact on search result quality (Sec 4). Our findings show that both the parts of the original article used to inspire the initial query and the methods and frequency of query reformulation significantly influence the quality of retrieved results. Finally, we conduct a user study to examine how interface design can help users adopt effective query strategies to improve the accuracy of their information evaluations (Sec 5). The findings suggest that while boost strategies had limited success in influencing user behaviour, reading search listings in full appeared to help participants create better queries and achieve improved outcomes, likely by encouraging the use of vocabulary not directly sourced from the article. This suggests several potential avenues for future research that could enhance querying behaviour and, ultimately, more effective news verification.

2 Related Work

We review three relevant bodies of literature: theories and information literacy interventions from the social sciences, contributions from the IR community on misinformation detection, and studies on how people evaluate and trust web-based information, and how this can be positively influenced.

2.1 Misinformation Research in Social Sciences

Research on misinformation in the behavioural and social sciences has introduced various means of improving users' competences and behaviours. Disciplines such as cognitive science, political and social psychology, and education research have inspired interventions including debunking false claims [42, 53], enhancing digital media literacy [10, 37], building resilience against manipulation [73, 82], slowing the spread of misinformation via the interface design [34], subtly encouraging people to think about the accuracy of articles [65], and highlighting the trustworthiness of information [27]. The effectiveness of these interventions has been tested using diverse methodologies ranging from controlled experiments (e.g., [65, 88]) to naturalistic field studies (e.g., [10, 65, 73]). Kozyreva et al. compiled a comprehensive toolbox of behavioural and cognitive interventions to combat online misinformation, synthesising evidence from 81 studies conducted worldwide across the social sciences [50]. The categories of interventions include *nudges*, which subtly influence people's behaviour by altering the environment or context in which decisions are made without restricting their options [81] as well as *boosts* and educational interventions, which aim to enhance individuals' skills and knowledge, empowering them to make better-informed decisions [43].

While Large Language Models (LLMs) can hallucinate and amplify the spread of misinformation [77], they have also been used in personalised conversations to help users reduce their belief in conspiracy theories [28, 84], showcasing their potential benefits to the field. Given the influence of misinformation, the exploration of new evidence-based approaches to improve users' abilities to recognise such content and limit its spread continues to be a highly relevant research area [20, 32].

2.2 Misinformation in IR

IR has predominantly focused on the detection of misinformation (see [83] for a review). Prominent research includes the work of Castillo et al. [22], who developed a method to classify tweets as credible or not based on features such as the number of reposts. Sondhi et al. [79] explored link- and content-based features to create a supervised method for identifying unreliable medical webpages, finding that combining all features yielded the best results. Similarly, Shim et al. [78] proposed embedding web search results into vectors and using traditional machine learning classifiers, which outperformed standard fake news detection models. Mazzeo et al. [56] showed that extracting URL features enhanced the detection of COVID-19 fake news in web search engines. Recent work has shown that utilising the graph structure of the information ecosystem via graph neural networks (GNNs) can help identifying fake news, e.g. [29, 30].

Initiatives like the *TREC Health Misinformation (HM)* Track, the *CLEF eHealth Consumer Health Search (CHS)* Task, and the *CLEF Check That! Lab* aim to develop retrieval methods that prioritise credible and accurate information over misinformation [25, 26, 59, 80]. For example, Pradeep et al. [71], as part of their *TREC HM* participation, proposed a multistage retrieval system with a final supervised re-ranker based on a fine-tuned T5-3b model. In a subsequent study, they integrated LLMs to estimate correct answers to health-related queries and generate query reformulations that improved performance [70]. Similarly, Bevendorff et al. [14] used the ChatNoir search engine [13] for initial candidate retrieval, followed by custom query-based re-ranking.

Events like the *Reducing Online Misinformation through Credible Information Retrieval* workshop (ROMCIR) further advance research in this area [66–68, 74]. These efforts, coupled with increasing interest from the IR community and beyond, underscore

that misinformation and its impact on end-users remain unresolved challenges. Addressing this multi-faceted problem extends beyond detecting misinformation, encompassing the study of its effects on searchers, improving information presentation, and empowering users to critically inspect and verify information [3].

2.3 (Changing) Search Behaviour

Research shows that user interaction with search results is often biased [5, 11]. One key bias is *position bias*, where users click results based on their placement rather than relevance [46]. Factors like missing or short snippets, absent query terms, and complex URLs further reduce a result's visibility [24]. Users also tend to favour results that confirm their existing beliefs [49, 60, 72, 85].

Search result composition influences users' beliefs after searching. Results biased toward incorrect information reduce a user's chance of correctly answering a question, while those favouring correct information improve it [69]. Position bias worsens this issue, evidenced by inaccurate featured snippets significantly affecting credibility judgments [17, 18]. Users are typically inconsistent in judging which search results to trust, with individuals relying on different cues or interpreting the same cue differently [48]. This aligns with Prominence-interpretation theory [35], which emphasises the role of subjective perception in trust formation. Users often trust the majority viewpoint in search results, a phenomenon called the *Search Engine Manipulation Effect* [31].

To address these issues, search systems have been developed to enhance user decision-making, often drawing on ideas relating to nudge concepts from the social sciences [81]. For example, result re-ranking algorithms tackle algorithmic biases and improve search quality by rearranging results based on specific criteria [7, 15, 23, 44, 91]. Other nudge-like strategies focus on query refinement, encouraging users to explore diverse perspectives or generate alternative ideas [2, 45, 61]. Additionally, some systems provide extra information about results in the snippet to help users make more informed credibility judgments [76, 89].

Boost-style interventions have also been applied in search contexts. For instance, search tips enhance user effectiveness [58], while tools designed to counter confirmation bias encourage users to engage with diverse viewpoints [16, 72]. In the area of query generation, interactive examples of high-quality queries during search sessions help users identify key attributes, craft effective queries, and align with expert-level standards [41].

This section demonstrates how social science approaches can complement IR and Interactive IR (IIR) research. Here, we integrate methods from these three fields by building on a social science study of search behaviour. First, we conduct a deeper analysis of Aslett et al's data, before applying query simulation techniques commonly used in the IR community to evaluate hypothetical behavioural strategies. Finally, we return to a social science-inspired IIR approach to test whether boost interventions can encourage users to adopt the beneficial behaviours identified in the simulations.

3 Analysing Aslett et al's data

Aslett et al. [6] model two conditions in their experiments: with and without search. The main goal of their work was to study the change on beliefs about news articles in users who conducted search sessions to evaluate the truthfulness of the news, and compare this effect with that of users who did not perform any search. We analyse their publicly available data[1] focusing on the search condition, specifically using study 5 data, the only one with query logs. This dataset includes queries, search results, metadata (e.g., NewsGuard scores measuring media quality[2]), and participant demographics[3]. In this first step, we conduct an exploratory analysis to test whether *how* people search is more important than *if* they searched.

The dataset contains 765 queries relating to news articles that were misleading. At the end of the search session the participant made a clear judgement with respect to the veracity of the article. 411 of the queries led to the participant correctly identifying the article as fake news (*Misl*) and 354 (46.3%) resulted in the participants trusting the content of the article (*True*).

Queries were slightly longer when participants got it wrong ($\text{mean}_{\text{Misl}} = 6.08$, $\text{sd}_{\text{Misl}} = 4.44$ vs $\text{mean}_{\text{True}} = 6.66$, $\text{sd}_{\text{True}} = 4.23$, $t = -1.8511$, $df = 755.19$, $p = 0.06$). This was initially surprising since longer queries are typically associated with search experience and expertise [8, 86, 87]. A second observation that contradicts some prior findings is that successful search sessions —where participants correctly identified fake news— involved more queries on average ($\text{mean}_{\text{Misl}} = 1.82$, $\text{sd}_{\text{Misl}} = 1.21$) than unsuccessful ones, where articles were misclassified as true ($\text{mean}_{\text{True}} = 1.53$, $\text{sd}_{\text{True}} = 0.87$, $t = 3.3283$, $df = 565.38$, $p < 0.001$). While prior work shows advanced searchers typically issue fewer queries but interact more deeply with results [87], this finding aligns with evidence that domain experts submit more queries than non-experts [64, 86].

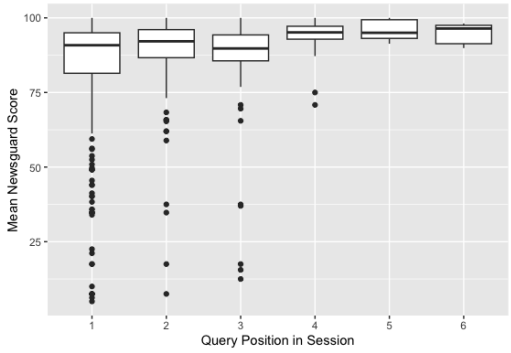


Figure 1: NewsGuard Scores for Search Results against Query Position in the Session

Figure 1 shows one potential reason for the performance improvement when sessions contain more queries. It shows how the average NewsGuard score for the SERP results varies based on the query position in the session. Later queries tend to return results of higher quality ($\rho = 0.15$, $p < 0.001$). Moreover, the average NewsGuard scores for search results were higher when participants correctly identified the fake news (mean = 87.32, sd=16.41), compared to when they did not (mean = 82.92, sd=19.09,

[1] available at https://github.com/SMAPPNYU/Do_Your_Own_Research
[2] https://www.newsguardtech.com evaluates the credibility of news websites based on nine criteria, focusing on transparency and journalistic practices, with reviews conducted by trained journalists and editors.
[3] Unfortunately, no click-through or interaction data are available.

$t = 3.413, df = 694.8, p < 0.001$). This suggests that the quality of media sources in the search results predicts participants' ability to identify fake news accurately after a search session.

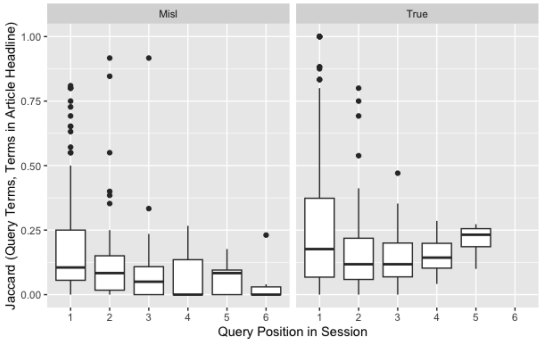


Figure 2: Word Overlap between Query and Article's Headline measured with the Jaccard Coefficient. The left plot represents the cases where users identified the article as false news, while the right plot represents the cases where users believed the article was truthful.

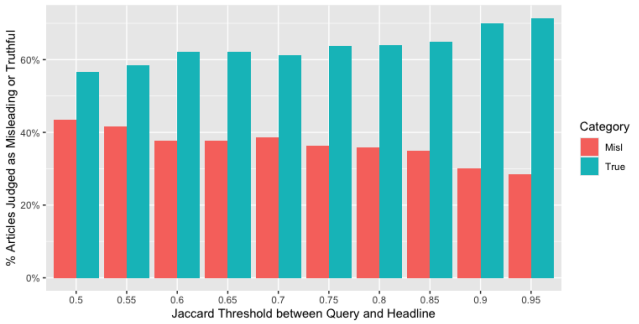


Figure 3: Percentages of Misleading and Truthful Decisions for Queries with Varying Levels of Overlap with the Article's Headline

Aslett et al. suggested that using query terms from the source article leads to confirmatory results. Figures 2 and 3 confirm that using the article vocabulary is problematic and that the amount of overlap is also important. This explains why longer queries were not always better and why queries sometimes improved as the session progressed. Figure 2 shows that when participants believed the article was truthful after searching (right plot), the overlap between query terms and the article's headline was high and remained consistent throughout the session. In contrast, when the article was identified as fake news (left plot), the overlap started lower and decreased over consecutive queries. Figure 3 shows varying levels of overlap between the query and headline when participants trusted or distrusted the article after searching. The percentage of query terms from the headline likely predicts whether participants classify the article as truthful.

These initial analyses suggest that the problem with validating news articles does not lie in the act of searching itself but in the way people conduct their searches. The findings indicate that examining querying strategies more closely could provide valuable insights. In the next section, we address this by simulating user query sessions using different querying strategies to assess how these impact the quality of search results, measured by the mean NewsGuard score.

4 Simulation

Our analysis of Aslett et al.'s data suggests that the source of query terms significantly affects search result quality. To investigate this further, we address two research questions: **(RQ.S1)** Can we identify effective strategies for generating initial queries by leveraging different sections of the source document? and **(RQ.S2)** Can we determine effective strategies for reformulating queries to improve performance throughout a search session?

4.1 Method

To address these questions, we performed a simulated study that evaluates search performance based on query generation from different parts of the source article and tests alternative reformulation strategies that simulate how users derive inspiration from search results. This reflects established patterns of user interaction with search engines (see review above). The IR community has a strong tradition of employing simulated query generation studies, which enable systematic, controlled, and efficient testing of query strategies [9, 33, 55], including reformulating queries throughout a session [38–40], without human users.

After preliminary experiments with traditional and newer simulation approaches, including methods oriented to sample queries from classic language models [21] and neural techniques based on docT5query [62] and keyBERT [36], we decided to use a Generative-AI approach powered by Llama3 (see repository for specific technical details)[4]. This decision was based on the fact that traditional methods tended to produce unrealistic queries, often drifting away from the topic of the source article. The problem of topic drift in query session simulation has been discussed in the literature [39] and, thus, we manually inspected the simulated queries to ensure that topic drift was not an issue in the final configuration of the simulations.

For our simulation process, we started with the 17 articles labelled as fake news in Aslett et al. [6] study 5[5]. We used Bing search API for searching and Llama3 (8B parameters) to generate the synthetic queries. We hypothesised that focusing on different aspects of the initial article and the search results during query generation can impact the quality of the results. To test this, we developed a series of query generation variants. A key feature of these approaches is their interpretability, allowing them to be translated into concrete behavioural strategies that human users could easily implement.

- **Initial query generation strategies**: We instructed the LLM to build queries after reading the article headline (**H** variant), after reading the headline and the first paragraph (**H 1P**), or after reading the full text of the article (**FT**). The

[4]Full details of the simulation process, including code, prompts and the queries themselves can be found here: https://github.com/MarcosFP97/sim-sigir
[5]Since 9 of these articles were no longer available online at the time of our study, we used the Wayback machine to recover these.

respective parts of the article were fed to the LLM together with instructions for the target search task (news verification). The generated query was run against the Bing search API, obtaining the first top 10 results of the search session.

- **Query reformulation strategies**: For the follow-up searches within the session, we tested two reformulation approaches. The first approach instructs the LLM to consider the previous queries in the session and the title and snippets from the top 10 search results of the last search (**TS** reformulation variant). The second approach also forces the LLM to consider the previous queries, the title & snippets from the top 10 results but additionally feeds the first paragraph of the top two search results (**TS 1P TOP2** reformulation variant). This approach simulates typical user behaviour (the first two results are much more likely to receive user clicks [46])[6].

In total, we tested six different search variants (three initial query generation * two query reformulation strategies). We simulated search sessions of lengths from one to five, based on the lengths in the sessions in Aslett et al's study. Each variant was executed ten times to minimise randomness in the results. We instructed the LLM to generate queries of approximately 3 to 5 words, based on the query lengths in the real data. Only search results published on the same day of the article or earlier were considered. All other results were removed as to simulate web pages available online at the date of publication of the article.

As a quality measure, we used the mean NewsGuard score of the search results since we previously demonstrated its correlation with users making better decisions. We observed that the SERPs of the simulated sessions showed good coverage of pages that have a NewsGuard score assigned (over 50% of the retrieved webpages, which is higher than the percentage of NewsGuard-scored pages in the original user study, 30%). This guarantees that the SERPs of the simulation can be assessed with sufficient confidence.

4.2 Results

The following subsections present the simulation study results:

4.2.1 Initial Query Generation RQ.S1 examines the best strategy for generating initial queries. Notably, all three tested strategies produced high NewsGuard scores for the SERPs of the first search. The H variant achieved an average score of $\text{mean}_{\text{H}} = 94.58$ (SD = 4.6), while H 1P and FT scored $\text{mean}_{\text{H1P}} = 95.26$ (SD = 3.9) and $\text{mean}_{\text{FT}} = 94.10$ (SD = 5.2), respectively. These results are significantly higher than the scores observed in Aslett's data. We believe this reflects a dual process: fake news articles are often removed from publication (the articles were over two years old by the time of our simulation), and search engines adapt by demoting or removing low-quality content as declining clicks lower their rankings. This likely also explains the higher percentage of results with NewsGuard scores in our simulated study.

We compared the initial query generation strategies and found that using the headline and the first paragraph (H 1P) resulted in search results with a significantly higher average NewsGuard score than those obtained with the full text (FT) strategy. Specifically, the mean for H 1P was 95.26 (SD = 3.9), compared to 94.10 (SD = 5.2) for FT, with a U value of 59355, $df = 658$, and $p = 0.04$[7]. The full text (FT) strategy was the least effective, and using only the headline (H) produced slightly lower quality search results. However, the difference between H and H 1P was not statistically significant: the mean for H 1P was 95.26 (SD = 3.9) and for H it was 94.58 (SD = 4.6), with a U value of 48802, $df = 648$, and $p = 0.09$. The comparison between FT and H also showed no significance, with a mean of 94.10 (SD = 5.2) for FT and 94.58 (SD = 4.6) for H, a U value of 51587, $df = 648$, and $p = 0.61$. From these results, we can conclude that the most effective strategy is to generate queries from both the headline and the first paragraph (H 1P), suggesting that the first paragraph may contain more neutral language, while later sections of the document might introduce misinformation cues —unique terms or phrases that can lead users toward data voids.

Table 1: Average NewsGuard Score in the SERPs (first and fifth query in the session)

Simulation variant	First query	Fifth reformulation
H - TS	94.93	95.38
H - TS 1P TOP2	95.00	95.65
H 1P - TS	94.94	95.48
H 1P - TS 1P TOP2	96.07	95.79
FT - TS	95.00	95.61
FT - TS 1P TOP2	94.43	95.22

4.2.2 Query Reformulation Strategies RQ.S2 explores the most effective reformulation strategy for search queries. We compared different strategies by evaluating the average NewsGuard scores from all queries in the simulated sessions.

The comparison between FT-TS and FT-TS 1P TOP2 revealed a significant difference ($U = 307816$, $df = 1644$, $p = 0.001$), with FT-TS having a mean of 94.9 (SD = 5.1) and FT-TS 1P TOP2 a mean of 94.2 (SD = 5.2). No significant differences were found for other pairwise comparisons between reformulation strategies. These results suggest that if the initial query generation strategy is strong (H 1P or H), the specific reformulation strategy has little impact. However, when the initial query is poor (FT), reformulations inspired only by the entire search engine results page (TS) yield significantly better results than those achieved by TS 1P TOP2. This makes sense as a weak query might lead us to poor top results and, thus, if we reformulate the query guided by the top 2 results (TS 1P TOP2) we might be getting to even poorer results. Observe also that the most effortful reformulation strategy (TS 1P TOP2, which involves not only inspecting the SERP but also reading a couple of paragraphs from the top 2 results) is not the most effective.

We analysed the evolution of NewsGuard scores during search sessions and found that query reformulation reduces performance differences, with the most variation in the first query. By the fifth reformulation, differences between approaches are minimal (see Table 1). While this might seem reassuring, Aslett's data show us that web users rarely generate that many queries in a single session.

[6]We restricted the input to the LLM to the first paragraphs from the top 2 pages because, otherwise, the context becomes too lengthy and noisy. Furthermore, these leading paragraphs arguably reflect the parts of these pages that are more likely read by web users.

[7]We used a non-parametric Mann-Whitney U test for these comparisons

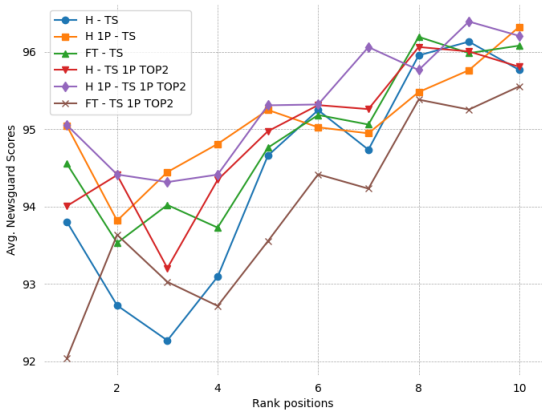


Figure 4: NewsGuard Scores at different rank positions

These simulation findings align with Aslett et al.'s data, emphasising the importance of query reformulation. While most users do not reformulate extensively, effective reformulation becomes critical when the initial query is weak, enabling performance improvements with a few reformulations.

We also analysed how NewsGuard Scores vary at different rank positions. As shown in Figure 4, these scores tended to increase with the ranking positions, at least up to position 10. This trend occurred for all variants and suggests that the most reputed results (at least according to NewsGuard assessment) are not necessarily at the top positions. Web users should, therefore, inspect the full SERP and not only the top positions. This effect could be due to the tendency of search engines to promote popularity (e.g., with link-based metrics), but popularity does not equal reputation.

This plot also confirms that FT-* methods are inferior to their counterparts. H1P-* methods yield the highest proportion of high quality docs at top positions, where users more likely click. Again, we ran statistical tests that confirmed that the H1P-* variants were superior for the top entries in the rank.

Table 2: Jaccard overlap between queries from different simulation variants and article headline (H) or first paragraph (1st)

Variant	Mean J(Q,H)	SD J(Q,H)	Mean J(Q,1st)	SD J(Q,1st)
FT TS	0.12	0.11	0.04	0.04
FT TS 1P TOP 2	0.10	0.10	0.05	0.04
H 1P TS	0.10	0.08	0.04	0.04
H 1P TS 1P TOP 2	0.08	0.08	0.03	0.04
HTS	0.09	0.08	0.03	0.03
HTS 1P TOP 2	0.09	0.09	0.02	0.03

Given our findings in Section 3, it was surprising that the headline conditions performed the best. To better understand why this was the case, we examined the Jaccard overlap between the generated queries and both the source article's headline and first paragraph (see Table 2). The results reveal that the best-performing variants had, on average, the least overlap with the source article. In other words, supplying the model with text does not guarantee that the model will merely extract words from that text to generate the query. Indeed, when more of the source article was provided, the query terms were more likely to be sourced from there. This may also apply to human users.

5 Empowering Users to Query Better

In this section we perform a pre-registered user study[8] to investigate how boost interventions can empower users to act more effectively based on the search tactics discussed earlier. Based on the results from our previous analyses, we draw four key conclusions, with the mapping to user study conditions provided in italics:

(1) Initial queries were more effective when the algorithm had access to the headline and first paragraph, but not the rest of the article *(Read 1st)*.
(2) Strong queries tend to have less overlap with the text of the source article *(Own Words)*.
(3) Weak initial queries improve by reviewing the entire results page, which is more likely to yield reliable documents than focusing solely on the top results *(Read All)*.
(4) Submitting more queries within a session improves result quality by the fifth query, regardless of the reformulation strategy used *(Multiple Queries)*.

Inspired by the literature, we developed boost messages (one for each condition) that aim to encourage the successful behaviours observed in our previous analyses (see Table 3 for an overview and Figure 5 for the boost presentation in the SERP).

5.1 Materials and Setup

Participants performed the same task as in Aslett et al., searching to assess the trustworthiness of news articles. They were assigned to only one condition (between groups) and evaluated only a single article. All articles used in this study were identified as misleading, aligning with the prior analyses. Articles were selected from outlets included in the Aslett et al. dataset that provided articles their fact-checkers classified as misleading. Both conservative and liberal sources were used, including the following sources: *The Federalist Papers*, *ZeroHedge*, *Natural News*, *WND*, *Stillness in the Storm*, *Palmer Report*, *GNews*, *Occupy Democrats*, *Townhall*, and *Newsmax*.

To identify suitable articles, these websites were scraped, and URLs of current articles were collected. A predefined prompt was used with GPT-4 as a tool to assist in the evaluation process. A team of three researchers manually examined fresh articles (no more than two days old) and identified candidates they believed were fake news. The prompt provided guidance by offering an automated likelihood score based on source credibility, content analysis (bias, sensationalism, unsupported claims), and cross-verification with credible sources. Articles flagged by the researchers as potentially fake news were then sent to professional fact-checkers for

[8] https://osf.io/x9g74/?view_only=c1cef259191c4a8dabd0602b1a2c1470

Table 3: Study Conditions with either a vanilla SERP or a boost containing a search tip.

Boost	Wording
No boost	Vanilla SERP w/o boost
Own Words	Users in our pre-study, who formulated their queries in their own words —rather than simply extracting keywords from the source document— were far more successful at detecting fake news.
Read 1st	Users in our pre-study who *fully* read the first paragraph of an article before formulating their queries were far more successful at detecting fake news.
Read All	Users in our pre-study who *read all search results* before crafting their query were far more successful at detecting fake news
Multiple Queries	Users in our pre-study who issued the most queries were the most successful at detecting fake news.

verification, ensuring that only articles verified to be misleading were included in the study[9].

By using fresh articles, we assume that lower-quality results remain in the search index, addressing an issue observed in the simulated study. We attained four articles from four different media outlets. Searches were conducted through an experimental system (see Figure 5) powered by the Bing API, enabling us to record detailed interaction data including queries submitted, result clicks, and timestamps. Participants were provided with the news article in html form on the right-hand pane and could search using the interface on the left side. The boost was provided in a prominent position in the top-left of the screen to make it more likely to be read. As in Aslett et al [6], after searching participants evaluated the article and provided demographic information and self-estimated digital literacy and veracity scores.[10]

5.2 Hypothesis

We define the following hypothesis H1: *Users in the boost conditions will submit queries resulting in higher average NewsGuard scores.*

We focused on NewsGuard scores because we assumed they are less influenced by user factors, such as political views and education, or article characteristics, such as presentation quality. Additionally, Aslett's data show that mean NewsGuard scores strongly predict search outcomes when these other variables are controlled.

5.3 Participants

The number of participants was established by means of a power analysis. Given that the number of queries submitted will vary per participant, we initially considered a mixed-effects model. However,

[9]The prompt and fact-checker reports on the articles are included in our repository: https://anonymous.4open.science/r/sigir-aslett-misinfo-BC03

[10]See Appendix J in Aslett et al's paper [6] for the questions

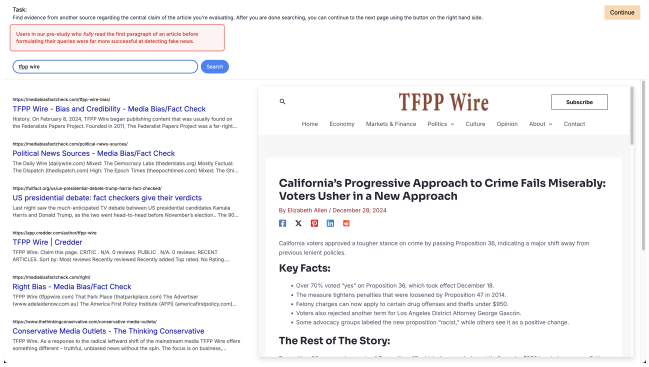


Figure 5: The search interface used in our study, boost shown in red in the upper left-hand corner

due to the low number of repeated measures for many participants (40% submitted only 1 query), attempts to fit a random-effects model based on the data simulated for the power analysis resulted in singular fits. Consequently, we opted for a fixed-effects model (one-way ANOVA) to compare the five experimental conditions. A power analysis determined that 200 participants are required to achieve 80% power at an alpha level of 0.05, assuming a medium effect size ($f = 0.25$). Full details can be found in the pre-registration.

A total of 260 participants were recruited via Prolific to obtain 200 participants (120 male, 77 female, and 3 diverse/other) who passed the attention check. These were all US-based native or fluent English speakers, 106 of whom hold a bachelor's degree, 31 a master's, 49 a high school diploma, and 9 a doctorate. Ages ranged from 19 to 72 years, with a median of 36 and a mean of 37.17 (IQR: 26.75–45). Participants reported a wide range of occupations, including roles in IT (e.g., software engineers, data analysts), healthcare (e.g., nurses, physicians), education (e.g., teachers, researchers), business (e.g., project managers, administrators), with 11 participants identifying as students and 7 as unemployed.

On average, participants took 7.09 minutes to complete the study ($SD = 4.73, Mdn = 5.68$). This varied slightly based on the condition with those in the 'Multiple Queries' condition ($M = 8.72, SD = 6.81, Mdn = 6.00$) taking the longest to complete and 'Read 1st' ($M = 6.21, SD = 3.93, Mdn = 5.53$) being the fastest (No Boost: $M = 6.24, SD = 3.46, Mdn = 5.61$, Own Words: $M = 6.66, SD = 3.67, Mdn = 5.39$, Read All: $M = 7.77, SD = 4.87, Mdn = 6.67$).

Looking at educational background, the number of participants who correctly identified the article as fake news was relatively similar (high school: 65.3 %, bachelor's degree: 60 %, master's degree: 51.5 %, doctorate: 55.6 %, other forms of education: 60 %).

Using Aslett's scales, the sample shows a slightly liberal bias ($M = -0.41, SD = 2.18$) with 23 extreme conservatives, 36 extreme liberals, and 30 participants who did not report their ideology. Digital literacy scores of participants ranged from 8 to 44 ($M = 26.7, SD = 5.99$).

5.4 Results

The majority of the participants (59. 2 %) were able to correctly identify the presented articles as fake news, while 31.3 % believed

the article to be true and the remaining 9.5 % could not determine its veracity. This is a slightly higher percentage than was observed in Aslett's study 5, which may be partially explained by the boost interventions. However, we believe it is more likely to relate to the way in which news articles were sampled (see discussion below).

Table 4 shows the mean NewsGuard scores for the search results by condition. There appears to be a slight increase in the score for the 'Multiple Queries' and 'Read All' conditions. However, the ANOVA results indicate that these differences are not statistically significant ($F(4, 1680) = 0.219, p = 0.931$).

Similar trends are observed when examining the mean number of queries submitted by users. The 'No Boost' condition had the fewest queries overall, while the boost conditions saw slightly higher query counts. Again, the 'Read All' condition was the joint highest in terms of number of queries submitted. The percentage of participants who correctly identified the article also varied across conditions. The 'Read All' condition had the highest percentage, but the other three boost conditions were actually lower than the 'No Boost' condition. This suggests that the boosts did not have the effect we had expected, but advising users to read all of the search listings before crafting their queries holds the most promise. We did not conduct statistical tests on number of queries or task success because they were not included in our pre-registered analysis plan. This decision was made to avoid drawing misleading conclusions from multiple comparisons that were not planned in advance.

Table 4: Summary of Mean and SD of NewsGuard Scores, Num of Queries per User and Percentage of Participants who Identifed their Article as Fake News by Condition. Highest values are bolded

Cond.	Mean NG	SD	Mean Queries	SD	% Ident.
No Boost	90.4	13.7	1.27	0.686	60.0%
Read 1st	90.3	18.2	**1.47**	0.910	57.1%
Own Words	90.8	14.6	1.37	0.888	55.1%
Mult. Queries	91.3	13.3	1.42	0.683	51.3%
Read All	**91.9**	15.8	**1.47**	0.878	**77.8%**

Our setup also enabled us to explore behaviours not captured in the Aslett data, such as click-through data. Table 5 shows how click-based metrics differ across the experimental conditions. The first metric is the mean NewsGuard score for the viewed results, and the second is the percentage of clicked results that have an associated NewsGuard score. The second metric assumes that leading media outlets have available NewsGuard scores, and that unknown results are likely of lower quality. The findings align with those above: the 'No Boost' condition scores the lowest in both metrics, while 'Read All' performs well on both counts, but again the differences are not significant $F(4, 163) = 0.685, p = .603$.

Finally, a correlation analysis reveals a negative relationship between the overall NewsGuard score (`avg_score`) and the Jaccard overlap for headlines (-0.26, $p < 0.0001$), while the correlations between `avg_score` and the first paragraph (-0.09, $p = 0.16$) and full text (-0.12, $p = 0.06$) are weaker, suggesting that NewsGuard scores are more closely aligned with headline content than with other document parts.

Table 5: Summary of Mean and SD of NewsGuard Scores of the pages participants clicked on and percentage of clicked results with associated NewsGuard score. Highest values are bolded

Cond.	Mean NG	SD	% with Score
No Boost	86.2	15.9	67.7
Read 1st	88.4	21.3	83.7
Own Words	**93.0**	9.0	69.8
Mult. Queries	87.7	18.1	79.2
Read All	90.7	16.6	**84.4**

6 Discussion

Taking the findings from the various investigations together reveals that validating misleading news articles is a challenging task. In our study fewer than 60% of participants were able to say with certainty that the article they were assigned was fake news and over 30% believed it to be truthful. Aslett et al. discovered even higher percentages of participants believing misleading news.

The evidence shows that searching to validate news is not per se problematic. The three studies (i.e., Sections 3, 4 and 5) consistently show that the way people create search queries impacts their ability to evaluate news. Using vocabulary from the source article, particularly the headline, often leads to lower-quality search results and increased belief in misinformation. Analysis of Aslett et al.'s data revealed that when participants believed a fake news article, their queries closely matched the headline. Although the headline+first paragraph conditions in the simulation performed best, further analysis revealed that higher quality results were linked to queries with less overlap with both the headline and initial paragraph (see Table 2). Similar correlations were found in the boost study.

In general, the boost strategies used in this study showed limited success in modifying user behaviour and outcomes, an unexpected result given the effectiveness of similar interventions in influencing other search behaviours and outcomes [16, 63], including query generation [41]. We consider reasons for this lack of impact. One potential explanation concerns how the messaging was perceived. While some studies suggest that boosts with tips with non obvious knowledge (e.g., informing users that content from trusted sources, such as .gov domains, is more reliable than .com [63, 91]) are acted on, it is possible that our boost messages were seen as unnecessary by certain participants. This may be attributed to overconfidence in their ability to identify misleading news, which is a well-documented phenomenon [51]. Participants who felt assured in their own information literacy skills might have dismissed the tips as irrelevant or patronising, thinking, "I don't need this".

In light of this, interventions that promote intellectual humility—encouraging participants to acknowledge the limits of their knowledge—may be more effective. These could include boost interventions, similar to those proposed by Rieger et al.'s study [72], or social nudges, where users compare their performance to that of experts, as seen in Bateman et al's work [12].

Despite the lack of significant results, consistent trends in the data suggest that encouraging users to 'Read All' search results before forming a query is beneficial. Participants in this condition

not only identified misleading articles most frequently but also submitted the most queries on average. Additionally, the articles they viewed were more likely to have an associated NewsGuard score, and the NewsGuard scores for the returned results were the highest in this condition.

It is important to note that 'Read All' was an intervention targeting query reformulation, and many participants submitted only a single query. Therefore, we suspect that for those influenced by the boost, it likely not only affected their querying behaviour but also implicitly encouraged them to interact with a broader set of results, promoting a more thorough approach to information evaluation. Combining this boost with one aimed at improving the initial query could potentially enhance its effectiveness.

7 Limitations

Although we present three complementary studies of different types, there are a number of limitations to our work, which we will discuss here along with their potential impact.

One limitation is that we focused exclusively on fake news. Our decision to centre the study on fake news was driven by the desire to build upon the main message from Aslett et al's work -that searching made people more likely to place their faith in misleading news. However, the querying strategies we explored may not be universally applicable to all types of articles. For instance, using vocabulary directly from the source article in queries may be particularly problematic for fake news but may have the opposite effect if the article is trustworthy.

In the simulated study, we used the same articles as those in [6]. Although these articles were outdated, we chose them to ensure comparability with previous studies. In hindsight, it might have been better to use the approach we applied in our final user study, which involved more current articles. Even with this method, the articles were not entirely "fresh", as the fact-checking process meant the articles were already 1-3 days old by the time the study took place. This is comparable to the original Aslett study.

Despite these limitations, the patterns we observed were consistent across all three analyses. Encouraging users to use vocabulary not directly contained in the source article appears to be key to improving the quality of the search results they receive.

Another limitation is that our user study focused on only four articles. While we aimed to capture differences in user behaviour by collecting multiple data points per article, the power analysis indicated that 200 data points were needed, with 50 per article being plausible. However, these four articles may not be representative of all fake news articles, and future research could expand the study to include a broader sample. In fact, we believe our article sampling process may have been biased towards selecting more obviously misleading articles, as we worked to achieve high agreement between three researchers and professional fact-checkers. This process proved challenging, as many article authors crafted misleading narratives without making explicit claims. Instead, they carefully curated quotes to advance a particular narrative. We selected articles where the claims were more clearly defined, which makes it all the more concerning that a significant percentage of our participants still rated these articles as trustworthy.

Additionally, we only simulated query generation based on specific parts of the document. While this was a sensible first step, there are many other strategies that could be explored in future studies, such as crafting queries to specifically reflect claims or involving negation.

We tested only four boost strategies. Other approaches, not explored in this study but mentioned in our discussion, might offer more promising results and could be valuable for future research.

Lastly, we used NewsGuard as a proxy for search quality, which we believe was a well-justified choice. However, NewsGuard does not capture the semantic aspect of search quality, such as whether queries drift off topic. We are currently developing other metrics that focus on the claims made in search results and how they confirm or contradict those in the source document. We believe these new metrics will provide deeper insights into user behaviour.

While our studies provide valuable insights, these limitations highlight areas for improvement and future exploration.

8 Ethical Considerations

Boost interventions help users make better decisions without restricting freedom, unlike result filtering. While boosts promote transparency and knowledge, they can still have negative effects, as shown in Table 4, where participants in three of four boost conditions were less likely to identify fake news.

9 Future work and Conclusions

Beyond testing the ideas presented in the discussion, an obvious direction for future research is to explore how Generative AI systems, which are increasingly central to information-seeking, fit into this ecosystem. AI agents, such as co-pilots, could potentially help users create better queries, especially if they understand the task at hand. Furthermore, generative AI interfaces like ChatGPT may be used instead of traditional search engines to validate news articles. This raises many questions, such as how users would interact with these systems for this purpose and how successful they would be.

In summary, our work highlights the role of querying behavior in validating news articles. The evidence suggests that query crafting matters, and systems could be designed to improve this process. While the best approach is unclear, it could complement measures like media literacy training, fact-checking tools, source credibility indicators, scepticism nudges, and collaboration with expert networks to enhance information literacy.

10 Open Science

We have made all resources including the data, code, articles and processes available in two anonymised github repositories:

See https://github.com/markusbink/sigir-aslett-misinfo/ for Sections 3 and 5.
See https://github.com/MarcosFP97/sim-sigir for Section 4.

Acknowledgments

MFP & DL thank the funding by MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF).

References

[1] Alexander Acht. 2024. *Who is benefitting from fact-checking on social media – user or platform? Examining the impact of different fact-checking approaches on social media platforms on user's perception of trust.* Master's thesis. Harvard University Division of Continuing Education.

[2] 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. https://doi.org/10.1145/3331184.3331246

[3] J. Allan, E. Choi, D. Lopresti, and H. Zamani. 2024. *Future of Information Retrieval Research in the Age of Generative AI CCC Workshop Report.* Technical Report. Computing Research Association (CRA). https://cra.org/wp-content/uploads/2024/12/Future-of-Information-Retrieval-Research-in-the-Age-of-Generative-AI.pdf

[4] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[5] 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.

[6] 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.

[7] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 1259–1276. https://doi.org/10.1145/3299869.3300079

[8] Anne Aula. 2003. Query Formulation in Web Information Search.. In *ICWI*. 403–410.

[9] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. 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, 455–462.

[10] Sumitra Badrinathan. 2021. Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review* 115, 4 (2021), 1325–1341.

[11] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[12] Scott Bateman, Jaime Teevan, and Ryen W White. 2012. The search dashboard: how reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1785–1794.

[13] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic chatnoir: Search engine for the clueweb and the common crawl. In *European Conference on Information Retrieval*. Springer, 820–824.

[14] Janek Bevendorff, Michael Völske, Benno Stein, Alexander Bondarenko, Maik Fröbe, Sebastian Günther, and Matthias Hagen. 2020. Webis at TREC 2020: Health Misinformation Track. In *Proceedings of the 29th Text REtrieval Conference (TREC)*.

[15] 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. https://doi.org/10.1145/3209978.3210063

[16] 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.

[17] 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* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/3576840.3578323

[18] 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* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/3498366.3505766

[19] Joel Breakstone, Mark Smith, Priscilla Connors, Teresa Ortega, Darby Kerr, and Sam Wineburg. 2021. Lateral reading: College students learn to critically evaluate internet sources in an online course. *The Harvard Kennedy School Misinformation Review* (2021).

[20] E. Broda and J. Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association* 48, 2 (2024), 139–166. https://doi.org/10.1080/23808985.2024.2323736

[21] Ben Carterette, Ashraf Bah, and Mustafa Zengin. 2015. Dynamic Test Collections for Retrieval Evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton, Massachusetts, USA) *(ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/2808194.2809470

[22] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[23] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[24] 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. https://doi.org/10.1145/1277741.1277767

[25] Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. 2021. Overview of the TREC 2021 Health Misinformation Track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[26] Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. 2020. Overview of the TREC 2020 Health Misinformation Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[27] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior* 42 (2020), 1073–1095.

[28] Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (2024), eadq1814.

[29] Gregor Donabauer and Udo Kruschwitz. 2023. Exploring Fake News Detection with Heterogeneous Social Media Context Graphs. In *European Conference on Information Retrieval*. Springer, 396–405.

[30] Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User Preference-aware Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada). Association for Computing Machinery, New York, NY, USA, 2051–2055. https://doi.org/10.1145/3404835.3462990

[31] Tim Draws, Nava Tintarev, Ujwal Gadiraju, 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.

[32] E.-A. Dumitru, L. Ivan, and E. Loos. 2022. A generational approach to fight fake news: In search of effective media literacy training and interventions. In *Human aspects of IT for the aged population. Design, interaction and technology acceptance*, Q. Gao and J. Zhou (Eds.). Springer International Publishing, 291–310. https://doi.org/10.1007/978-3-031-05581-2_22

[33] David Elsweiler, David E. Losada, José C. Toucedo, and Ronald T. Fernandez. 2011. Seeding simulated queries with user-study data for personal search evaluation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 25–34.

[34] Lisa Fazio. 2020. Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review* 1, 2 (2020).

[35] Brian J Fogg. 2003. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*. 722–723.

[36] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[37] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. 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. https://doi.org/10.1073/pnas.1920498117

[38] Sebastian Günther, Paul Göttert, and Matthias Hagen. 2022. Exploring LSTMs for Simulating Search Sessions in Digital Libraries. In *Linking Theory and Practice of Digital Libraries: 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022* (Padua, Italy). Springer-Verlag, Berlin, Heidelberg, 469–473.

[39] Sebastian Günther and Matthias Hagen. 2021. Assessing Query Suggestions for Search Session Simulation. In *Joint Proceedings of the Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) Workshops 2021. Co-located with the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM-SIGIR 2021) (Sim4IR)*.

38–45.
[40] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From search session detection to search mission detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (Lisbon, Portugal) *(OAIR '13)*. Paris, FRA, 85–92.
[41] Morgan Harvey, Claudia Hauff, and David Elsweiler. 2015. Learning by example: training users with high-quality query suggestions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 133–142.
[42] A. Helfers and M. Ebersbach. 2022. The differential effects of a governmental debunking campaign concerning COVID-19 vaccination misinformation. *Journal of Communication in Healthcare* 16, 1 (2022), 113–121. https://doi.org/10.1080/17538068.2022.2047497
[43] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science* 12, 6 (2017), 973–986.
[44] 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.
[45] Jyun-Yu Jiang and Wei Wang. 2018. RIN: Reformulation Inference Network for Context-Aware Query Suggestion. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/3269206.3271808
[46] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, 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. https://doi.org/10.1145/1229179.1229181
[47] Joel Kaplan. 2025. More Speech and Fewer Mistakes. Meta Newsroom. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ Accessed: 2025-01-09.
[48] Markus Kattenbeck and David Elsweiler. 2019. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management* 71, 3 (2019), 368–391.
[49] Joshua Klayman and Young-Won Ha. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review* 94, 2 (1987), 211.
[50] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M Herzog, Ullrich KH Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, et al. 2024. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* (2024), 1–9.
[51] Don Latham and Melissa Gross. 2008. Broken Links: Undergraduates Look Back on Their Experiences with Information Literacy in K-12 Education. *School Library Media Research* 11 (2008).
[52] 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.
[53] S. Lewandowsky, J. Cook, U. Ecker, D. Albarracin, P. Kendeou, E. J. Newman, and M. S. Zaragoza. 2020. The Debunking Handbook 2020. https://digitalcommons.unl.edu/scholcom/245
[54] Stephan Lewandowsky, Laura Smillie, David Garcia, Ralph Hertwig, Jim Weatherall, Stefanie Egidy, Ronald Robertson, Cailin O'Connor, Anastasia Kozyreva, Philipp Lorenz-Spreen, Yannik Blaschke, and Mark Leiser. 2020. *Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making*. European Commission. https://doi.org/10.2760/709177
[55] David Maxwell. 2019. *Modelling search and stopping in interactive information retrieval*. Ph. D. Dissertation. University of Glasgow, UK.
[56] Valeria Mazzeo, Andrea Rapisarda, and Giovanni Giuffrida. 2021. Detection of fake news on COVID-19 on web search engines. *Frontiers in physics* 9 (2021), 685730.
[57] Sarah McGrew. 2024. Teaching lateral reading: interventions to help people read like fact checkers. *Current Opinion in Psychology* 55 (2024), 101737.
[58] 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.
[59] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 639–649.
[60] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.
[61] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Information Processing & Management* 50, 1 (2014), 218–234.
[62] Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery. https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-latest.pdf
[63] Anna-Marie Ortloff, Steven Zimmerman, David Elsweiler, and Niels Henze. 2021. The effect of nudges and boosts on browsing privacy in a naturalistic environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 63–73.
[64] João Palotti, Allan Hanbury, Henning Müller, and Charles E Kahn. 2016. How users search and what they search for in the medical domain: understanding laypeople and experts through query logs. *Information Retrieval Journal* 19 (2016), 189–224.
[65] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
[66] Marinella Petrocchi and Marco Viviani. 2023. ROMCIR 2023: Overview of the 3rd workshop on reducing online misinformation through credible information retrieval. In *European Conference on Information Retrieval*. Springer, 405–411.
[67] Marinella Petrocchi and Marco Viviani. 2024. Report on the 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2024) at ECIR 2024. *SIGIR Forum* 58, 1 (Aug. 2024), 1–9. https://doi.org/10.1145/3687273.3687285
[68] M Petrocchi, M Viviani, et al. 2022. Overview of ROMCIR 2022: the 2nd workshop on reducing online misinformation through credible information retrieval. In *ROMCIR 2022 CEUR Workshop Proceedings*, Vol. 3138.
[69] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA 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*. 209–216.
[70] Ronak Pradeep and Jimmy Lin. 2024. Towards Automated End-to-End Health Misinformation Free Search with a Large Language Model. In *European Conference on Information Retrieval*. Springer, 78–86.
[71] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2066–2070.
[72] 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. https://doi.org/10.1145/3465336.3475101
[73] J. Roozenbeek, S. Van Der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science advances* 8, 34 (2022), eabo6254. https://doi.org/10.1126/sciadv.abo6254
[74] F Saracco, M Viviani, et al. 2021. Overview of ROMCIR 2021: workshop on reducing online misinformation through credible information retrieval. In *ROMCIR 2021 CEUR Workshop Proceedings*, Vol. 2838.
[75] Sebastian W Schuetz, Tracy Ann Sykes, and Viswanath Venkatesh. 2021. Combating COVID-19 fake news on social media through fact checking: antecedents and consequences. *European Journal of Information Systems* 30, 4 (2021), 376–388.
[76] 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. https://doi.org/10.1145/1978942.1979127
[77] Siddhant Bikram Shah, Surendrabikram Thapa, Ashish Acharya, Kritesh Rauniyar, Sweta Poudel, Sandesh Jain, Anum Masood, and Usman Naseem. 2024. Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords. *IEEE Access* (2024).
[78] Jae-Seung Shim, Yunju Lee, and Hyunchul Ahn. 2021. A link2vec-based fake news detection model using web search results. *Expert Systems with Applications* 184 (2021), 115491.
[79] Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. 2012. Reliability prediction of webpages in the medical domain. In *European conference on information retrieval*. Springer, 219–231.
[80] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. 2018. Overview of the CLEF eHealth evaluation lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*. Springer, 286–301.
[81] Richard H Thaler and Cass R Sunstein. 2021. *Nudge: The final edition*. Yale University Press.
[82] C. S. Traberg, J. Roozenbeek, and S. van der Linden. 2022. Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science* 700, 1 (2022), 136–151. https://doi.org/10.1177/00027162221087936
[83] Marco Viviani and Gabriella Pasi. 2017. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data*

mining and knowledge discovery 7, 5 (2017), e1209.

[84] R. Weeks, P. Sangha, L. Cooper, J. Sedoc, S. White, S. Gretz, and N. Bar-Zeev. 2023. Usability and credibility of a COVID-19 vaccine chatbot for young adults and health workers in the United States: formative mixed methods study. *JMIR human factors* 10, 1 (2023), e40533. https://doi.org/10.2196/40533

[85] 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* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/2484028.2484053

[86] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*. 132–141.

[87] Ryen W White and Dan Morris. 2007. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 255–262.

[88] Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. *Teachers College Record* 121 (2017).

[89] 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. https://doi.org/10.1145/1978942.1979126

[90] Jingwen Zhang, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine* 145 (2021), 106408.

[91] 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 (CHIIR '19)*. 283–287.