An Eye Tracking Study: Are AI Overviews Changing Search Behavior?

Sara Allawati
RMIT University
Naarm/Melbourne, Australia
sara.allawati@student.rmit.edu.au

Dana McKay
RMIT University
Naarm/Melbourne, Australia
dana.mckay@rmit.edu.au

Mark Sanderson
RMIT University
Naarm/Melbourne, Australia
mark.sanderson@rmit.edu.au

Paul Thomas
Microsoft
Australia
pathom@microsoft.com

Johanne R. Trippas
RMIT University
Naarm/Melbourne, Australia
j.trippas@rmit.edu.au

Abstract

We conduct a lab eye-tracking study to examine how users interact with search engines that place Generative Artificial Intelligence (GenAI) results above traditionally ranked search results, also known as "ten blue links", and we use an engagement scale to evaluate their experience. Our aim is to study how users interact with search engine interfaces that incorporate GenAI content, assess users' willingness to scroll past GenAI content to view the traditional search results, and explore how these interactions differ from existing literature on scanning search engine result pages. We show that GenAI content is changing how people search, but the "golden triangle" remains valid, where the top-left section of the search page attracts the most attention. Searchers are still engaging with the blue links in patterns consistent with the literature; however, they engage significantly more with GenAI content. Finally, we outline future directions to deepen our understanding of search behavior in the era of GenAI.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Eye Tracking, AI Results Summaries, User Study

ACM Reference Format:
Sara Allawati, Dana McKay, Mark Sanderson, Paul Thomas, and Johanne R. Trippas. 2026. An Eye Tracking Study: Are AI Overviews Changing Search Behavior?. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809530

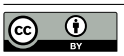
1 Introduction

The intersection of Generative Artificial Intelligence (GenAI) and information access has been defined as Generative Information Retrieval (GenIR) [7, 29, 39]. We still lack a clear understanding of how users interact with GenIR systems. Building clear, evidence-based models of user behaviors can improve how we evaluate these systems and their usability. This is why a number of eye-tracking and cognitive user studies were conducted on traditional search engines with the Traditionally Ranked Results also known as "ten blue links" [2, 3, 12, 17, 42].

However, previous findings do not directly apply to modern search engines that incorporate GenIR systems, which often summarize multiple results in the text they generate. Interactive Information Retrieval (IIR) researchers have therefore asked whether GenAI's diverse interaction possibilities will see widespread adoption or stay a research interest [7]. If so, these new developments in search engine functionality may signal a generational shift in how people search for information. We could then ask, is relying on the traditional "ten blue links" becoming less common, or does it depend on the nature of the tasks at hand?

While some studies have explored user interactions with GenIR tools [29, 37, 44], we did not find any lab studies focused on eye-tracking or any other physiological devices in this context. By researching human interactions with GenIR systems, we can contribute to developing these systems into true assistants [38, 39]. Search interfaces are constantly evolving, with the design space for search engine interfaces being vast, but interaction techniques and user interfaces with GenIR remain under-researched and poorly understood. Between 2020 and 2024, 750 preprints related to Large Language Models (LLMs) were published on arXiv in the field of Information Retrieval (IR), with only 22 mentioning "user interface" in their abstracts [7].

Information-seeking behavior has changed as search engines have introduced AI-generated summaries and conversational search features in place of simple keyword matching. These features enable more natural interaction and may attract users by saving time, reducing cognitive effort, and improving search efficiency. This shift has contributed to the emergence of GenIR as a distinct research area within IR. Unlike traditional IR systems, which primarily return ranked lists, GenIR systems generate written answers that synthesize information from multiple sources. Many papers in top IR journals and conferences note that user studies in this area remain scarce [7, 16]. To our knowledge, no eye-tracking studies have examined how users interact with these interfaces. This gap is important because eye-tracking is an established method in IR and IIR for studying search behavior [17].



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809530

Eye-tracking studies of Search Engine Results Page (SERP) have provided important insights into how users identify relevant or irrelevant documents [20], how their search behavior can be modeled [2, 21, 33], and how attention is shaped by contextual factors [12]. One of the well-known findings in the area is the Golden Triangle Pattern or F-pattern [31], showing that users tend to focus their attention on the top-left portion of the results, even if those results were not relevant. Insights from these user studies informed the development of algorithms and user models [33].

This study examines how users interact with SERP that integrate GenAI content, followed by the Traditionally Ranked Results. It addresses an important gap, as it remains unclear how current search behavior differs from patterns reported before GenAI was incorporated into SERPs. To investigate this, we conduct an exploratory eye-tracking study on SERPs that places GenAI content above the Traditionally Ranked Results.

RQ1: How do users scan and navigate search engine results pages that have Generative AI content placed before the Traditionally Ranked Results?

RQ2: To what extent do perceived user engagement levels correlate with users' gaze behavior on search engine results pages containing Generative AI content before the Traditionally Ranked Results?

RQ3: What are the perceived levels of user engagement when interacting with search engine results pages that have Generative AI content placed before the Traditionally Ranked Results?

By addressing our research questions, we contribute the following:

- Demonstrate that users continue to engage with Traditionally Ranked Results in patterns consistent with the literature, but show significantly higher engagement with GenAI content when both appear on the same page.
- We release our experimental setup for SERPs with GenAI content and Traditionally Ranked Results, to support reproducibility across other user populations.[1]

2 Related Work

In this section, we discuss research on Human–GenAI interactions, primarily in IR and non-IR, along with previous eye-tracking studies on how people interact with search engines.

In IR, several studies investigate how users interact with GenIR systems [29, 37, 40, 44]. A study [37] uses Bard log submissions from 95 crowd workers, considering factors such as gender, English skills, education level, search skills, and usage of search engines and LLMs. Another User Study [40] methodology consists of user groups from college students and crowd workers from various backgrounds. The study developed a ChatGPT-like interface with GPT-3.5 Turbo and integrated questionnaires and supportive functions, such as Perception Articulation, Prompt Suggestions, and Conversation Explanation, to determine which of those functions supported different user groups. We are also witnessing the publication of various resources and tools in IR to facilitate the study of user behavior in the era of GenAI [28, 46]; however, it's unclear whether they are suitable tools when running eye tracking studies.

Work from Yang et al. [44] and Liang et al. [29] looked at how users interact with the GenIR system in comparison with the traditional IR system, but the GenAI interaction component was mainly via a chat-based system. A recent study by Wardle et al. [41] researched how people meet their health-related information needs via search engines, including Google's "AI Overviews" (Powered by Gemini), ChatGPT, and Alexa. The main methodology was structured observation through the think-aloud protocol. It is important to note that the study mentioned that the AI Overviews were rolled back while data were being collected. Xu et al. [43], looked at GenAI content's influence on public opinion. None of the studies discussed above has utilized eye tracking devices to measure cognitive load.

Several eye-tracking studies have extensively analyzed gaze patterns across search engine interfaces [2, 3, 12, 15, 17, 30]. None of these studies explored user interfaces for information access with GenAI either; however, their findings can be compared with ours to examine the generational shift. Abualsaud and Smucker [2] used eye tracking to analyze user behavior while viewing SERP, on both desktop and mobile platforms. Their research tracked participants from the moment they entered a search query to their subsequent actions, which typically involved either clicking a result or refining their query. The study discussed the relationship between user behavior and queries to understand their impact on re-query decisions. Liu et al. [30] utilized eye tracking to analyze fixation patterns and examine user focus on various elements of search interfaces. They collected comprehensive interaction data, including mouse clicks, typed queries, eye movements tracked via eye tracking, and perceived task difficulty via self-report ratings.

3 Methodology

To the best of our knowledge, no eye-tracking studies have explored how users interact with search engine interfaces that have GenAI content placed before the Traditionally Ranked Results. In this section, we present the methodology for the user study setup, which will help us answer RQ1–RQ3. We conducted an in-lab user study, collecting gaze data using a Tobii Pro Fusion eye-tracking device and Tobii Pro Lab software. To ensure the validity of our results, we follow the instructions and guidelines provided by Tobii[2]. We use Qualtrics[3] for the surveys, which include the user study tasks [4].

3.1 Study Overview

We used a within-subjects design, in which the same participants completed all ten tasks. We collected *pre-study* participant demographics and characteristics, including age, gender, education, search experience, and prior LLM use. Each task (see below) included a pre-task questionnaire and controlled "backstory" as well as a single central SERP. After each task, participants completed a *post-task* questionnaire rating their search experience, including perceived relevance of the results and perceived results difficulty. In the *exit* stage, participants completed the User Engagement Scale (UES) [35], which measures aesthetic appeal, focused attention, perceived usability, and reward. Finally, they answered open-ended questions about their impressions of the SERP, trust in AI overviews

[1] https://github.com/Sara-Al-Lawati/Sara-Al-Lawati-SIGIR-26-AI-Overviews-Eye-Tracking

[2] https://www.tobii.com

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

[4] The setup was reviewed and approved by RMIT University's ethics board (ID: 28583).

```
Topic: average charitable donation

You regularly make charitable donations to a range of causes,
including medical research funds, local homeless support agencies,
and international aid funds. But you are a little concerned that
your support is less than others give, and decide to find out
how much the average USA individual gives each year in donations.
```

Figure 1: Backstory 1.

versus traditional links, and if the AI overview or traditional links met their information needs.

Tasks and Searches. Participants completed ten tasks. Each task consisted of three stages *(i)* completing the *pre-task* questionnaire, *(ii)* reading a backstory, and *(iii)* completing the search task by reviewing a simulated SERP while we logged the interaction behaviors. For each task we used a different search task (i.e., backstory) from the UQV100 test collection[5]. These backstories (i.e., short scenario descriptions to motivate and contextualize a search task, see Figure 1) represented real-world scenarios in which participants would need to search for information [9]. We selected backstories to provide variety in both topic and word count, and excluded topics related to health, medicine, and politics to minimize potential discomfort. Backstories were categorized by task complexity (see below). Participants were assigned 10 tasks, all participants completed the same set of tasks, and the task order was randomized per participant. For each task, participants first read the backstory and then clicked a link that issued a pre-typed query (see Figure 2 for an example SERP). They reviewed the results on a simulated SERP and could open results pages until they felt their information need was met. Participants were given unlimited time per task and could not switch between tasks once started.

The simulated SERP was based on Google's interface at the time[6] and presented GenAI overview or GenAI content, also called "AI Overview", above the Traditionally Ranked Results. The pre-written queries were selected from an earlier eye-tracking study conducted in early 2024 with different users, during a period when GenAI was becoming more widely adopted. In that study, users wrote queries corresponding to the same set of backstories used in this eye-tracking study. The selection was based on two criteria: their NDCG@10 scores as a measure of their effectiveness [45] and whether they produce AI Overviews from Google. We used user-written queries from the GenAI era to capture more realistic user experiences.

The Anserini toolkit was used for retrieval processes[7]. We used BM25 on the ClueWeb12-B corpus, with parameters set to $k1 = 0.9$ and $b = 0.4$ to retrieve 1,000 documents per query. BM25 matched the terms between queries and documents to determine relevance. For this study, we selected NDCG@10 to evaluate participants' performance and query effectiveness, similarly to [6, 30, 45]. The NDCG@10 values ranged from 0.1 to 0.6, with a range of 0.5 and a standard deviation of 0.1.

[5]Backstories used in this study are denoted as B1 to B10, corresponding to TREC Topic Numbers as follows: B1: 205, B2: 209, B3: 212, B4: 293, B5: 210, B6: 257, B7: 211, B8: 294, B9: 291, B10: 206 [9]

[6]https://blog.google/products/search/generative-ai-google-search-may-2024/

[7]https://github.com/castorini/anserini

Since the UQV100 test collection and its relevance judgments over the ClueWeb12-B dataset are based on queries collected in 2016, we also ask participants in this user study to self-rate the relevance of the results on a 5-point scale (1 = low relevance, 5 = high relevance). Overall, the rating average across all SERPs and participants is 4.2, standard deviation is 1, with the lowest average rating a SERP received is 3.4, and the highest is 4.6.

We submitted the queries to Google and saved the resulting SERPs as PDFs similar to how Wu et al. [42] collected SERPs from Google. These PDFs, collected in January 2025, were used to develop custom HTML code that replicated the Google interface at the time, including AI Overviews and Traditionally Ranked Results. While the original Google AI Overviews included reference links within the GenAI text, these were excluded due to additional complexities that were not compatible with the eye-tracking equipment used.

During each search, we recorded gaze data as participants read the backstory, interacted with SERPs, and reviewed the pages they selected. In parallel, we logged interaction behavior, including session duration, total mouse clicks, clicks on blue links, the number of unique websites visited, and the number of follow-on visits after each blue-link click.

Task Complexity. To control for variation in reading difficulty, we assessed backstory readability using Flesch-Kincaid scores [26]. We selected backstories spanning approximately 7th-grade to college level, ensuring the texts were suitable for participants, given our inclusion and exclusion criteria. We also controlled for cognitive task complexity. Following prior work on UQV100 backstories and the link between backstories and their associated queries [1], and consistent with research highlighting the importance of task complexity in IR and IIR [25, 32], Bloom's taxonomy was used to label each backstory [27]. Using the labeling approach as previous research [1], an independent categorized the backstories used in our lab study as *remember*, *understand*, or *analyze*. These levels correspond to retrieving relevant knowledge from long-term memory, constructing meaning by interpreting and summarizing information, and breaking information into parts to understand relationships and purpose [25].

SERP Length. The percentage of the page content above the fold across all 10 SERPs was computed by extracting the viewport directly from Tobii's raw data. On average across the 10 SERPs, 44% was above the fold (s.d. 4%, range 38–49%). The page fold location on a sample SERP is illustrated in Figure 2.

3.2 User Study

The researcher provided a verbal explanation of the study and presented written instructions. Participants reviewed the study protocol and then signed the consent form. To minimize distractions, the researcher left the lab before participants began the tasks. Participants could contact the researcher at any time via a microphone.

Study apparatus. The participants sat 65cm away from a computer with an on-screen-mounted eye-tracking device to complete the lab experiment. The eye-tracking device recorded gaze positions at 120Hz. We calibrated the eye tracker for each participant within Eye

Tracking Manager[8] and Tobii Pro Lab[9]. Participants were advised to minimize movements once calibration was completed to ensure the collection of high-quality eye-tracking data. Participants needed calibration values of 0.5 to 0.8 degrees for accuracy and 0.15 to 0.3 degrees for precision, per Tobii's guidelines.[10]

Recruitment. We recruited 43 participants via advertisements, including flyers, online postings, the researcher's social media, and word of mouth. Participants completed the study individually in a usability lab at RMIT University and received an AUD 25 gift voucher for their participation.

Prospective participants were screened, and only people *(i)* with working English proficiency, *(ii)* aged above 18, *(iii)* with normal vision (people wearing glasses are considered to have normal vision), *(iv)* with no existing health issues relating to the eye and cognition, and *(v)* with no eye surgery history was allowed to participate in the study. Our criteria ensured that participants had the necessary cognitive and physical abilities for consistent data collection.

Following the guidelines provided by Tobii, participants were informed of potential factors that could impact their performance, such as *(i)* sleeping less than seven hours, *(ii)* wearing heavy makeup or large eyelashes, *(iii)* taking medications affecting pupil dilation, or *(iv)* having certain medical conditions or drug use. Participants could opt out at any time.

Pilots. We conducted extensive piloting to ensure a robust experimental setup. We iteratively tested individual study components and then conducted full end-to-end runs of the procedure. For example, the pilots indicated that participants took longer to complete the tasks when the researcher remained in the room, informing our decision to have the researcher leave the lab during the session. Seven participants took part in pilots and the analyses in this paper are based on the remaining 36 participants[11]. Data from all 36 participants were collected between November and December 2025.

3.3 Participants

The sample (N = 36) was primarily young adults. Participants were aged 18–24 (30%), 25–28 (44%), or 29 and older (25%). Gender was reported as female (52%), male (41%), or non-binary (5%). Education level was high overall, with 10 participants holding a bachelor's degree (27%), 18 holding a master's degree (50%), and 8 holding a doctoral degree (22%). Participants also reported their search experience and LLM use. Most rated their search skills as moderate (22, 61%), with smaller proportions reporting low (9, 25%), high (3, 8%), or neutral (2, 5%) skill. Search frequency was predominantly high (24, 66%), followed by moderate (10, 28%) and low (2, 5%). LLM usage frequency was similarly high, with 20 participants (55%) reporting high use, 12 (33%) moderate use, 2 (5%) low use, and 2 (5%) reporting no prior LLM use.

3.4 Data Collection, Cleaning, and Analysis

We used Tobii Pro Lab for data collection and Python for data analysis. Since eye-tracking software typically provides x and y coordinates, we developed a Python script based on prior research [3, 11, 19, 22–24] to map participants' fixations, saccades, and pupil dilation as physiological measures. Physiological measures have been used in different disciplines, including IR, to model, evaluate, and predict user behavior, and they can be complementary to user subjective measures such as those in our questionnaires [18, 34]. We identified and extracted the text, and drew bounding boxes to define our Areas of Interest, similar to [22]. To map Areas of Interest on SERPs, we used Selenium[12] to extract element positions directly from the Document Object Model. The Areas of Interest mapping on a sample SERP can be shown in Figure 2.

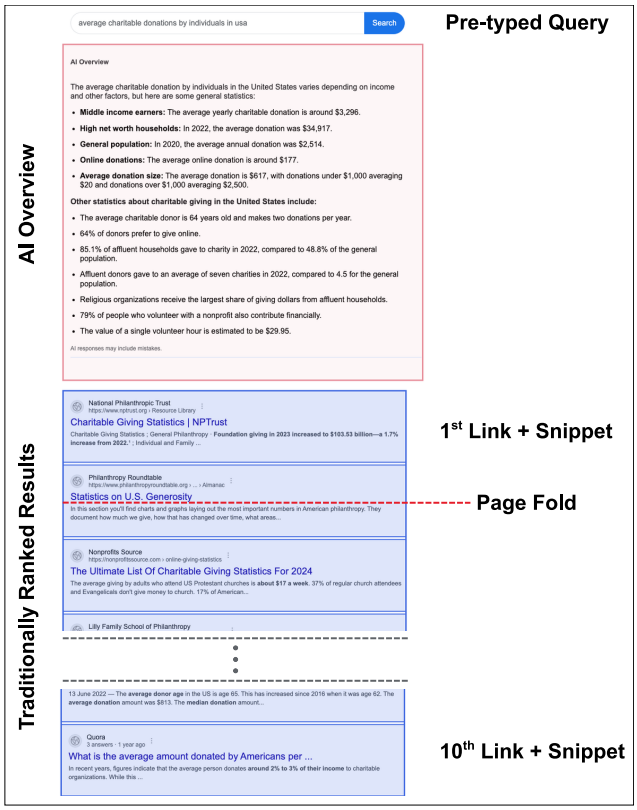


Figure 2: The SERP for backstory 1, showing AI Overview, 10 of the Traditionally Ranked Results, and the page fold (43.57% of the SERP is visible without scrolling) with Areas of Interest boundaries.

We cleaned and processed the pupil data using an established procedure and report Relative Pupil Dilation, defined as the change in pupil diameter over time relative to baseline [23]. We also measured saccade time, the duration of rapid eye movements between fixations, in milliseconds [20]. Saccade length was calculated in pixels as the distance between successive fixations [19].

Our code outputs were validated by directly comparing them with Tobii Lab Pro outputs. The raw gaze dataset for this study can be provided upon request. The repository includes code for cleaning raw eye-tracking data and Qualtrics data, synchronizing

[8] https://developer.tobiipro.com/eyetrackermanager.html
[9] https://www.tobii.com/products/software/behavior-research-software/tobii-pro-lab
[10] https://www.tobii.com/resource-center/data-quality#cta-section
[11] Note, six participants had minor missing data and are included in the analysis.
[12] https://www.selenium.dev/

eye-tracking and Qualtrics timestamps, extracting mouse clicks, and validating our pipeline.

It is important to note the multiple validation steps used in this study: (1) participants fixated on a cross displayed between all pages; this cross appeared on the screen for approximately 4 seconds, and our data showed that participants fixated on it for around 3 seconds, similar to earlier work [23]. This fixation period was used to align Tobii and Qualtrics timestamps with millisecond-level precision. (2) Heatmaps were generated following Tobii's documentation; the code and references are provided in the repository, and the resulting heatmaps were visually compared with Tobii's heatmaps. (3) When the study was run using Tobii Lab Pro, the entire screen was recorded; this screen recording was used as ground truth for selected analyses, such as determining how many websites were opened from the main SERP and how many were nested within other websites.

Qualitative Data. Open-ended questions were analyzed using the general inductive method [36], to extract key reasons for believing that either AI Overviews or Traditionally Ranked Results were more trustworthy and more useful.

4 Results

First, we explore differences in gaze metrics while browsing the SERPs. We analyze gaze metrics on the SERP across Areas of Interest. Second, we investigate gaze patterns and clicks on the SERP. Finally, we examine subjective self-reported ratings for pre- and post-task questionnaires, and UES. The Areas of Interest used to produce the findings of this paper can be visualized in Figure 2. The Areas of Interest consist of AI Overviews and the Traditionally Ranked Results, which consist of 10 Ranked Results. Each Ranked Result consists of a link and a snippet.

4.1 SERP Gaze Metrics

We compute statistics and then compare them by complexity and different Areas of Interest, including AI Overviews and the first 10 Traditionally Ranked Results.

SERPs Viewing Statistics. We report overall averages, see Table 1. People spent on average 15 seconds reading the backstories and about 46 seconds on the SERP. Since the SERPs have two main components, AI Overviews and Traditionally Ranked Results, we compare their significance using the Wilcoxon signed-rank test, given the nonparametric nature of the dataset, which is common in eye-tracking data. AI Overviews received significantly higher fixation and saccade times than Traditionally Ranked Results; however, this was not the case for saccade length. The saccade length being greater for the Traditionally Ranked Results can be explained by the area they occupy on the SERP. On average, across the ten SERPs, the Traditionally Ranked Results region is 2.7 times larger in pixel area and contains 2.8 times more words than the AI Overview region.

SERPs AI Overviews-Traditionally Ranked Results Comparison by Complexity. We compare the differences between gaze metrics on SERP when backstory tasks are categorized into different cognitive complexities (Remember, Understand, and Analyze), see Table 2. We also compared gaze metrics during reading backstories; however, no significant differences were found across the three levels of complexity. Since we are comparing three different groups, we applied a Bonferroni correction to account for multiple comparisons [14]. However, since this is an exploratory analysis, we report the results with and without Bonferroni since Bonferroni reduces false positives but may miss real effects due to increased false negatives. When the task was classified as Remember, AI Overviews received higher fixation and saccade times compared to AI Overviews for tasks classified as Analyze or Understand. In contrast, for Traditionally Ranked Results, Understand tasks received higher fixation and saccade times when compared with Remember and Analyze. Regarding saccade length, AI Overviews for Understand tasks showed higher values compared to Remember and Analyze.

Analyze seems to have lower fixation and saccade times when compared with remember and understand. This contrasts with the literature [8], which notes that "Remember" is the least cognitively demanding, followed by "Understand" then "Analyze", and our results partially align with this. We considered the question, are people looking elsewhere to other webpages navigated from the same SERP? What about the effects of self-reported rating pre- and post-task? What about the user engagement? Therefore, we investigate further.

Fixation times were consistently longer for AI Overviews compared to Traditionally Ranked Results across all complexity levels.

Table 1: Overall eye tracking statistics for SERP viewing, averaged at participant-level. Values are reported as mean (M) and standard deviation (SD). Statistically significant differences between gaze values mapped to AI Overviews and Traditionally Ranked Results (TRR) are shown in bold ($p < 0.05$).Wilcoxon signed-rank test was used.

Region	Fixation Time (s)	Relative Pupil Dilation	Saccade Time (s)	Saccade Length (px)
SERPs	45.73 (21.60)	-0.18 (0.07)	5.57 (6.01)	148.96 (19.06)
AI Overviews	**27.34 (17.47)**	-0.18 (0.07)	**3.15 (3.63)**	143.09 (24.53)
TRR	14.82 (9.68)	-0.19 (0.07)	1.85 (1.70)	**162.92 (33.84)**

Table 2: Pairwise comparisons on SERP by task complexity (Remember, Understand, Analyze) were conducted separately for AI Overviews and Traditionally Ranked Results (TRR). Friedman test with Wilcoxon post-hoc (uncorrected, exploratory). n.s. = not significant; bold with ↑ = significantly higher; † = not significant after Bonferroni correction.

		Task Complexity		
Region	**Metric**	R vs U	R vs A	U vs A
AI Overviews	Fixation Time	**R↑**	**R↑**	n.s.
	Relative Pupil Dilation	n.s.	n.s.	n.s.
	Saccade Time	**R↑**	**R↑**	n.s.
	Saccade Length	**U↑**†	n.s.	**U↑**
TRR	Fixation Time	**U↑**	n.s.	**U↑**
	Relative Pupil Dilation	n.s.	n.s.	n.s.
	Saccade Time	**U↑**	n.s.	**U↑**
	Saccade Length	n.s.	n.s.	n.s.

Figure 3: Aggregated heatmap of fixations across all 10 Search Engine Results Page (SERPs). Aggregated heatmap of fixations across 34 participants for Backstory 1 (B1) SERP.

For the Remember level, the mean fixation time was 32.9s for AI Overviews versus 13.5s for Traditionally Ranked Results ($p < 0.001$), and for Analyze, it was 26.0s versus 14.1s ($p < 0.001$), both showing significant differences. For Understand, the mean fixation time was 25.9s for AI Overviews and 18.5s for Traditionally Ranked Results, which was not statistically significant ($p = 0.127$). These comparisons were performed using the Wilcoxon signed-rank test, in which mean fixation times per participant were calculated for AI Overviews and Traditionally Ranked Results and then compared to assess whether the differences were significant. This explains why, when comparing the AI Overviews among three complexity groups, Remember received higher fixation and saccade times. This is an initial analysis of the effects of task complexity on SERPs for AI Overviews and Traditionally Ranked Results; further analysis is recommended to investigate these findings.

4.2 SERP Golden Triangle and Navigation

We aggregate fixations on heatmaps, take into consideration SERP word count, and report navigation results, such as clicks and arrival times to results, at the participant level on SERPs.

Aggregated Heatmaps and Golden Triangle. Figure 3 shows that the classic "golden triangle" or "F" pattern is found when aggregating all fixations of all the participant interactions across all ten SERPs. The pattern is well attested in earlier work [e.g., 3, 12], and the addition of AI Overviews does not seem to have changed this overall behavior. When AI Overviews appear in search results, they receive more visual attention than the 1st Ranked Result. This suggests that the traditional "golden triangle" or "F-pattern" of attention shifts upwards on the search results page. We investigate this attention pattern further in the subsequent analyses.

Fixation Duration Normalized by Word Count Analysis. To exclude the possibility that word count played a role in higher fixations, we compute the Fixation Time Normalized by Word Count (FN) for AI Overviews, Traditionally Ranked Results (TRR), and each Ranked Result (RR). This metric represents fixation time relative to each region's total word count, not fixation time mapped to individual words. The formula below was used to normalize fixation time by word count in each Area of Interest. First, we computed (1) Averages at the participant level (across tasks), (2) global averages across all participants, and finally, (3) computed drop analysis between AI Overviews, TRR, and between consecutive RR.

$$FN = \frac{\text{Fixation Duration (ms)}}{\text{Word Count}}$$

Absolute Drop (AI Overview to TRR):

$$\Delta_{\text{AI Overview}\to\text{TRR}} = FN_{\text{AI Overview}} - FN_{\text{TRR}}$$

Percentage Drop (AI Overview to TRR):

$$\%\Delta_{\text{AI Overview}\to\text{TRR}} = \frac{FN_{\text{AI Overview}} - FN_{\text{TRR}}}{FN_{\text{AI Overview}}} \times 100$$

Absolute Drop (RR $i \to i+1$):

$$\Delta_{RR_i \to RR_{i+1}} = FN_{RR_i} - FN_{RR_{i+1}}$$

Percentage Drop (RR $i \to i+1$):

$$\%\Delta_{RR_i \to RR_{i+1}} = \frac{FN_{RR_i} - FN_{RR_{i+1}}}{FN_{RR_i}} \times 100$$

Data are reported in Table 3 and shown visually as a bar plot in Figure 4. As can be seen, the attention received by the AI Overview is higher than that for each of the Ranked Results.

Fixation Duration Above and Below the Page Fold. Mean fixation duration (how long each individual fixation lasts) was similar on either side of the fold (231.66ms above and 232.25ms below). However, the mean *total* fixation time per participant (sum of all fixation durations) was very different (339.81s above, 82.15s below; Standard Deviation (SD) 164.47s and 64.53s). Participants fixate 4 times longer above the fold. In other words, when participants do fixate somewhere (above or below the fold), each fixation lasts about the same duration (232ms). But participants make many more fixations above the fold, so the total accumulated time is much higher.

Click Rank and Fixation Time. Figure 5 plots, for each rank in the traditional results, the mean number of clicks; the mean fixation time; and the mean time when participants first fixated on that rank (arrival time). Click rate and rank are strongly correlated (Spearman's $\rho = 0.94$, $p = 0.0001$). Again, however, behaviors past the AI overview are very similar to behaviors in earlier studies; at least in aggregate, searchers seem to read traditional links in the same way.

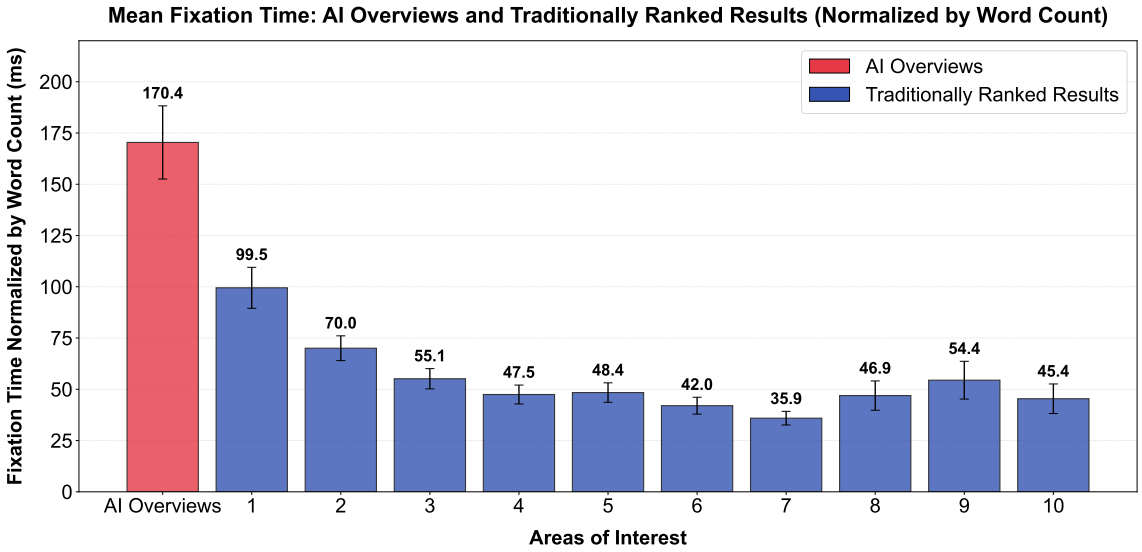


Figure 4: Visualization of AI Overviews and each Traditionally Ranked Results normalized by word count.

Table 3: Fixation Time Normalized by Word Count across Areas of Interest Regions: AI Overviews and Traditionally Ranked Results (TRR). The metric represents total fixation time (ms) divided by the total word count of each region, enabling fair comparison across regions with different text lengths. Drop (%) indicates percentage change from the previous Areas of Interest. N is number of participants.

Areas of Interest	Mean	SD	N	Drop (%)
AI Overviews	170.39	107.10	36	—
TRR	33.99	22.27	36	↓80.1%
Ranked Result 1	99.50	59.95	36	—
Ranked Result 2	70.04	35.72	35	↓29.6%
Ranked Result 3	55.14	29.22	35	↓21.3%
Ranked Result 4	47.46	26.86	34	↓13.9%
Ranked Result 5	48.40	25.99	30	↑2.0%
Ranked Result 6	42.00	22.55	30	↓13.2%
Ranked Result 7	35.91	18.54	31	↓14.5%
Ranked Result 8	46.93	38.49	29	↑30.7%
Ranked Result 9	54.44	48.75	28	↑16.0%
Ranked Result 10	45.39	36.08	25	↓16.6%

Note: Mean and SD are in units of ms per total word count. Drop (%) shows change from previous row. ↓ indicates a decrease (less fixation time per word count), ↑ indicates an increase.

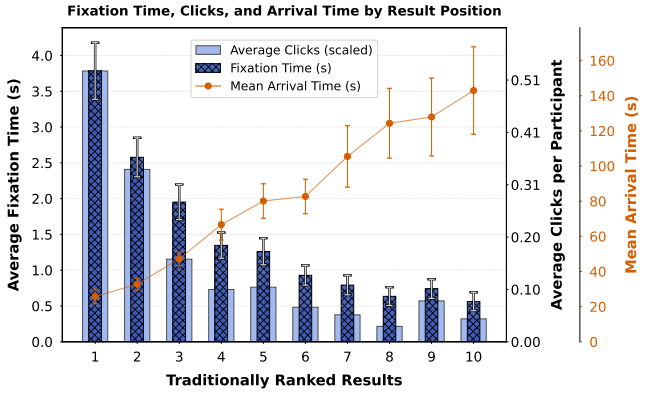


Figure 5: Average fixation times in seconds (s) and click count per Ranked Results on SERP, includes the arrival time for the participant's gaze to arrive at that result rank.

4.3 Subjective Dimensions and UES Analysis

We report on subjective dimensions, which are the self-reported pre- and post-task ratings and the UES questionnaires.

Pre-task and Post-task Questionnaire. As part of the study setup, each participant reported their interest, familiarity, difficulty, and curiosity regarding the backstory. We classified those self-rating questions as part of the pre-task questions for each task. Once participants finished exploring each SERP and felt their information need was met, they were directed to post-task questions, including rating the difficulty and relevance of the results on the SERP. Our findings are summarized in Table 4. A Repeated Measures Correlation (rmcorr) was used to compare each participant with themselves [10], as each participant completed the same set of tasks and answered the same self-reported ratings before and after each task. For AI Overview gaze data, both Relative Pupil Dilation and Saccade Length show a weak but significant positive correlation with interest, familiarity, and curiosity. Fixation time, however, is negatively correlated with both difficulty and relevance of the AI Overview results. For Traditionally Ranked Results, fixation and saccade times show a significant negative correlation with difficulty

and relevance of the results. Saccade length also shows a small but significant positive correlation with interest and curiosity.

UES Questionnaire. We also used the UES [35] to capture the dimensions of engagement: *aesthetic appeal*, *focused attention*, *perceived usability*, and *reward*. UES was collected from participants after they completed all search tasks. Participants rated each question on a scale from Strongly Disagree (1) to Strongly Agree (5). We report the averages and SDs of each dimension: *perceived usability* (4.19) and *reward* (3.90) are higher than *aesthetic appeal* (3.13) and *focused attention* (3.09). All three dimensions, *perceived usability* (SD = 0.70), *reward* (SD = 0.70), and *aesthetic appeal* (SD = 0.79), exhibited relatively similar variability, while *focused attention* demonstrated greater variability (SD = 0.95). Overall, the UES (3.58) had a SD of 0.54, the minimum score was 2.25, and the highest was 4.92. Regarding correlations with fixation time, relative pupil dilatation, and saccades, we did not observe any significant association between gaze patterns and UES responses, overall or on any individual dimension.

4.4 Trustworthiness and Utility

Of our 36 participants, 33 responded in writing to questions about the trustworthiness and utility of AI Overviews compared with Traditionally Ranked Results. These responses were analyzed using the general inductive method [36].

Trustworthiness. Of the 33 participants, 15 found search results more trustworthy, 12 found AI more trustworthy, and 6 gave mixed responses or said they felt the two were the same. Participants gave a range of rationales for finding AI untrustworthy. One common rationale was that AI is machine-generated: "Less trustworthy, as you know it's not written by a person" (P98), "Less trustworthy since it is a soulless summary that may contain what you are looking for but can be wrong" (P194). In contrast to this, participant 202 noted similar levels of trust because the AI summarizes human content "Similar levels of trust since overview id derived from human-made results and I am inherently more trusting of something written by a human". Another reason participants gave for not trusting AI summaries was to do with the underlying data, e.g., participant 117, who noted that they did not feel AI covered the same range of material "I think AI Overview is less trustworthy in this case; traditional search results usually vary from different years and areas etc" or P207 who said "what search results it summarizes I don't know". Another variation on this was being concerned that AI "did not cite its source" (P206), a concern shared by three other participants. Other participants were concerned about misrepresenting the underlying data, from the simple comment from P190 "less because some trends didn't match" to the more in-depth reflection from P98, who notes "I do not know where exactly they got their information from or how it might be biased or if it hallucinated. If it is a traditional result, at least I know where it is coming from and identify certain shortcomings or biases", or P112's comment "as it often misrepresents facts and can get data with missing or incomplete context".

In contrast with the variety of reasons for not trusting an AI overview, nearly every response saying AI was trustworthy noted its convenience "In most scenarios the AI Overview has given a reliable, easy to understand summarized version of the traditional results" (P100), "the good thing is that at least I don't have to click on the webpage to see what's further content there, and all information are listed in points". The other reason given was that the summaries are aggregates of underlying search results, e.g., "The content provided by the AI is trustworthy because they are information taken from legit websites and summarized" (P191)

A cross cutting reflection on trustworthiness was the need to check AIs summaries, mentioned by eight participants, e.g., P184 who was neutral on whether AI was more trustworthy and wrote "I did feel like I used the information from the traditional search results to help confirm or cross-check the information I was given by the AI overview", or P189 who trusted AI but said "However, I do double-check with the traditional search results to confirm the information". Overall, participants do not yet trust AI summaries completely, but they strongly recognize their convenience and readability.

Utility. Of our 33 participants, 15 found AI overviews more useful, 12 preferred search results, and 6 were neutral or gave mixed results. While this looks like a small move from trustworthiness to usefulness, 10 respondents felt differently about utility than trustworthiness. Of those 10, the majority moved from being neutral to having a preference or vice versa, e.g., P200 who was neutral on trustworthiness but commented "AI overview is better, as it gives my a concise and short summary of information I am looking for. Reading traditional results is time-consuming and may not get what I want precisely" or P206, who found traditional search more trustworthy, but said "For a quick search and introduction to the topic, AI Overview does help". Only two participants had a strict dichotomy between utility and trustworthiness. Of these, one provided no explanation of their utility score, even though they found summaries more useful and results more trustworthy. The other participant said they found AI more trustworthy, but "traditional search results as they are quite detailed with different information in different sites".

This analysis presents mixed results, with neither search nor AI emerging as more useful or more clearly trustworthy than the other tool.

5 Discussion

We discuss the implications of our findings, and compare them with previous eye-tracking literature on SERPs.

We found that AI Overviews received significantly longer eye fixation and saccade times than Traditionally Ranked Results; however, this was not the case for saccade length. We found that *Remember* tasks elicited higher fixation and saccade times in AI Overviews, while *Understand* tasks showed the highest fixation and saccade times in Traditionally Ranked Results and the highest saccade length in AI Overviews. When this was later investigated, it appears that some users may have spent a longer time on Traditionally Ranked Results for *Understand* tasks in comparison to *Remember* and *Analyze*. This may indicate that users prefer the Traditionally Ranked Results rather than AI Overviews for *Understand* tasks. However, some confounding variables need to be investigated in our future work, such as controlling for complexity by ensuring equal numbers of tasks in each complexity category.

Table 4: Summary of Repeated Measures Correlation (rmcorr) results for SERP data (Backstory Level): AI Overview and Traditionally Ranked Results (TRR). Values are correlations r_{rm}, bold values are statistically significant ($p < 0.05$).

		Topic				Search Results	
Region	**Measure**	Interest	Familiarity	Difficulty	Curiosity	Difficulty	Relevance
AI Overviews	Fixation Time (ms)	0.02	0.06	0.07	0.07	**-0.13**	**-0.11**
	Relative Pupil Dilation	**0.18**	**0.17**	0.05	**0.17**	0.03	0.04
	Saccade Time (ms)	0.05	0.05	0.08	0.08	-0.10	-0.10
	Saccade Length (px)	**0.20**	**0.13**	0.02	**0.17**	0.05	-0.01
TRR	Fixation Time (ms)	0.04	-0.01	0.00	0.06	**-0.16**	**-0.33**
	Relative Pupil Dilation	0.09	0.11	0.08	0.08	0.01	-0.01
	Saccade Time (ms)	-0.00	-0.03	0.00	-0.02	**-0.14**	**-0.19**
	Saccade Length (px)	**0.13**	0.06	-0.06	**0.17**	-0.02	0.03

Our aggregated heatmap from our experiments shows the classic "golden triangle" and F-shape of user behavior still holds even though the underlying content our users examined is new compared to past work. Buscher et al. [12] found that when sponsored ads appeared before ten ranked results, the first ranked result received the most attention, not the ads. When less interesting content is returned by a searching system, it would appear that users change their attention behavior. We speculate that our users examined AI Overview content due to interest, not just because the content appeared at the top.

Wu et al. [42] examined the effect of presenting a direct answer at the top of a SERP, followed by Traditionally Ranked Results. We computed the proportions of fixation durations on AI Overviews and Traditionally Ranked Results in our eye-tracking study and compared these with proportions reported in prior work, where participants were presented with either only Traditionally Ranked Results or a single direct answer followed by Traditionally Ranked Results (see Figure 6). It is important to note that our study setup, interfaces, and participant sample differed from those in prior work.

When one answer was provided, the proportion of fixation duration was 0.50, compared to 0.59 for the AI Overview condition. Participants in the condition with only Traditionally Ranked Results, compared to those in the one-answer condition, spent considerably more time on the first-ranked result (0.31 vs. 0.20). Furthermore, when comparing the one-answer condition with the AI Overview, fixation on the first-ranked result was lower (0.20 vs. 0.09). Overall, we observe a substantial decrease in fixation on the first-ranked result across the three SERP conditions. The proportion of the first Ranked Result (0.09) in the AI Overview Condition is comparable to the proportions of the 4th and 5th Ranked Result positions (0.11 and 0.07) in the condition with only Traditionally Ranked Results. This may suggest that the 1st Ranked Result is effectively shifted to the 4th or 5th position when AI Overviews are placed above the Traditionally Ranked Results. The F-pattern remains consistent, with the top-left position continuing to attract the most attention, as observed in the literature, but that position is occupied by an AI Overview instead of the first-ranked result, as further supported by the aggregated fixation heatmaps in our findings.

First-ranked results attracted the most attention and clicks, with the fastest arrival time, while attention and clicks decreased and arrival time increased down the ranked list. This is consistent with

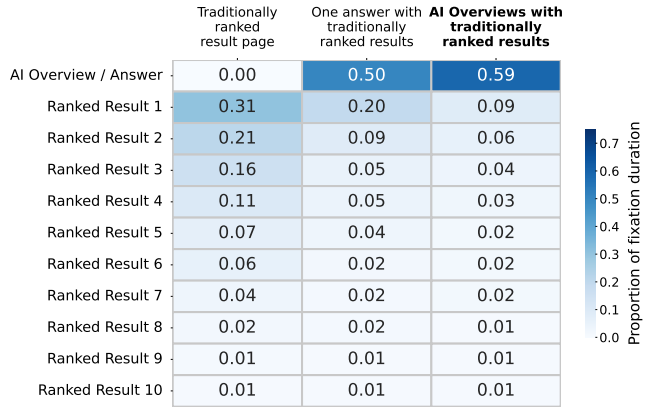


Figure 6: Proportions of fixation duration of AI Overviews with Traditionally Ranked Results (in bold) directly compared with Wu et al. [42] results.

other findings in the literature, where the mean arrival time usually increases as users scan down, as the mean fixation time is always higher at the first ranked result position, with declining fixations and clicks count the lower the results are ranked [13, 17]. However, differences in the slope of attention decline across the Traditionally Ranked Results, compared to those reported in prior literature, require further investigation.

Our results show that AI Overviews receive significantly higher attention than Traditionally Ranked Results. This not only has implications for new user models but also for other risks related to the content users consume when seeking information, especially the risk of AI hallucinations.

Regarding pre-task and post-task questionnaires, our results show weak associations with the different gaze metrics. For AI Overviews, higher interest, familiarity, and curiosity are associated with greater pupil dilation and longer saccade lengths, while longer fixation times are associated with lower difficulty and relevance of the search results. For Traditionally Ranked Results, longer fixation and saccade times correspond with lower difficulty and relevance, and longer saccade lengths are associated with higher interest and curiosity. A noteworthy distinction is that, for AI Overviews, higher interest, familiarity, and curiosity are associated with greater

pupil dilation, whereas this pattern is not observed for Traditionally Ranked Results. An increase in pupil dilation has been associated with higher cognitive load [23]. Our results suggest that AI Overviews may be associated with different cognitive load patterns compared to Traditionally Ranked Results. As for the UES, participants experienced moderate engagement with the search interface, with *perceived usability* and *reward* receiving higher ratings compared to *aesthetic appeal* and *focused attention*. The relatively lower scores for *aesthetic appeal* and *focused attention* may be attributed to the controlled experimental setup, which simulates a search engine rather than a commercial one. It might have been more effective to ask participants to rate the UES at the end of each task rather than once at the end, given the relatively small sample size (N = 33) for a survey question-based analysis.

Limitations. Our study is exploratory, so it is important to acknowledge several limitations. These include potential task confounding variables, the inability for users to perform subsequent search sessions or refine queries, and a lack of access to alternative search systems such as conversational search. Our sample sizes were unequal across levels of complexity and AI Overviews size. In addition, the AI Overviews content relevance was not evaluated against multiple measures other than NDCG@10 scores for prewritten queries and self-reported ratings of search results relevance.

Participants also had to conduct the study in a controlled lab environment, which may differ from performing searches in their own setting. Other limitations relate to ecological validity, which is a common challenge in controlled studies. For example, our design may not fully account for individual differences in domain expertise, device usage [4], anomalous states of knowledge, or cognitive biases [5]. Researchers have attempted to overcome the complexity of understanding human search behavior by creating personas and simulating user behavior [46]. However, this raises concerns that simulated users may not accurately reflect real human behaviors, and assessing the validity of simulated users still remains a challenge [4]. Conducting different types of studies—separately and in combination—is important, as each approach can provide valuable insights and collectively contribute to the body of knowledge to understand GenIR interactions.

Some of these limitations could be addressed through further analysis of the current data, while others require future studies to enhance user interactions with GenIR and identify gaps in this emerging field.

6 Conclusions

This eye tracking study examines how users interact with a SERP that presents GenAI content at the top of the page, immediately above the Traditionally Ranked Results. The study provides empirical evidence that AI Overviews are changing the way people search, but not to the extent that the golden triangle or the F-pattern observed in SERP are no longer valid. Users still engage with the Traditionally Ranked Results in patterns consistent with the literature [13, 17]; however, they engage more with AI Overviews than with Traditionally Ranked Results. Despite our study providing evidence that users still interact with the Traditionally Ranked Results in patterns similar to those reported in the literature, further research is required to understand the extent to which those patterns are similar. As for the qualitative analysis of trustworthiness and utility, we observed mixed results, with neither AI Overviews nor Traditionally Ranked Results perceived as more trustworthy or useful than the other.

Given the rapid evolution of interfaces integrating GenAI, future studies with our setup could explore AI Overviews that contain multiple links as sources and that appear in different SERP positions. Future work could investigate GenIR systems across specific tasks, diverse user groups, or different ecologically valid settings. It could also group users by survey responses, UES ratings, and qualitative feedback, such as reported trust in the system and perceived usefulness. This would clarify how search behavior and gaze patterns vary across groups and which factors most strongly shape these patterns.

This work would help the IR community in making recommendations for optimizing GenAI systems based on user preferences and behaviors, developing a framework of how people are influenced to search in the era of GenAI, and working toward making these systems true assistants.

Acknowledgments

The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands this research was conducted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community. This research was partially supported by the Australian Research Council through the ARC Centre of Excellence for Automated Decision-Making and Society CE200100005. We sincerely thank Nuha Abu Onq for categorizing the backstories' task complexity.

References

[1] Nuha Abu Onq, Mark Sanderson, and Falk Scholer. 2025. Classifying Term Variants in Query Formulation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2395–2406. https://doi.org/10.1145/3726302.3729924

[2] Mustafa Abualsaud and Mark D. Smucker. 2019. Patterns of Search Result Examination: Query to First Action. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, Beijing China, 1833–1842. https://doi.org/10.1145/3357384.3358041

[3] Mustafa Abualsaud, Mark D. Smucker, and Charles L. A. Clarke. 2021. Visualizing Searcher Gaze Patterns. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, Canberra ACT Australia, 295–299. https://doi.org/10.1145/3406522.3446041

[4] Marwah Alaofi, Negar Arabzadeh, Charles L. A. Clarke, and Mark Sanderson. 2025. *Generative Information Retrieval Evaluation*. Springer Nature Switzerland, Cham, 135–159. https://doi.org/10.1007/978-3-031-73147-1_6

[5] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Madrid Spain, 2850–2862. https://doi.org/10.1145/3477495.3531711

[6] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Taipei Taiwan, 1869–1873. https://doi.org/10.1145/3539618.3591960

[7] Mohammad Aliannejadi, Jacek Gwizdka, and Hamed Zamani. 2025. *Interactions with Generative Information Retrieval Systems*. Springer Nature Switzerland, Cham, 47–71. https://doi.org/10.1007/978-3-031-73147-1_3

[8] Jaime Arguello, Sandeep Avula, and Fernando Diaz. 2016. Using query performance predictors to improve spoken queries. In *European Conference on Information Retrieval*. Springer, 309–321. https://doi.org/10.1007/978-3-319-30671-1_23

[9] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, Pisa Italy, 725–728. https://doi.org/10.1145/2911451.2914671

[10] Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated Measures Correlation. *Frontiers in Psychology* 8 (April 2017). https://doi.org/10.3389/fpsyg.2017.00456

[11] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do People and Neural Nets Pay Attention to the Same Words: Studying Eye-tracking Data for Non-factoid QA Evaluation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, Virtual Event Ireland, 85–94. https://doi.org/10.1145/3340531.3412043

[12] Georg Buscher, Susan T. Dumais, and Edward Cutrell. 2010. The good, the bad, and the random: an eye-tracking study of ad quality in web search. ACM, Geneva Switzerland, 42–49. https://doi.org/10.1145/1835449.1835459

[13] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. ACM, San Jose California USA, 407–416. https://doi.org/10.1145/1240624.1240690

[14] Nicola Ferro and Mark Sanderson. 2024. Uncontextualized significance considered dangerous. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 261–270. https://doi.org/10.1145/3626772.3657827

[15] Lingyue Fu, Jianghao Lin, Weiwen Liu, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. An F-shape Click Model for Information Retrieval on Multi-block Mobile Pages. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1057–1065. https://doi.org/10.1145/3539597.3570365

[16] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 1916–1929. https://doi.org/10.1145/3626772.3657849

[17] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 478–479. https://doi.org/10.1145/1008992.1009079

[18] Jacek Gwizdka. 2010. Distribution of cognitive load in Web search. *Journal of the American Society for Information Science and Technology* 61, 11 (Nov. 2010), 2167–2187. https://doi.org/10.1002/asi.21385

[19] Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. *Journal of the Association for Information Science and Technology* 68, 10 (Oct. 2017), 2299–2312. https://doi.org/10.1002/asi.23904

[20] Jacek Gwizdka and Yinglong Zhang. 2015. Differences in Eye-Tracking Measures Between Visits and Revisits to Relevant and Irrelevant Web Pages. ACM, Santiago Chile, 811–814. https://doi.org/10.1145/2766462.2767795

[21] Jiaman He, Marta Micheli, Damiano Spina, Dana McKay, Johanne R. Trippas, and Noriko Kando. 2026. Characterizing Personality from Eye-Tracking: The Role of Gaze and Its Absence in Interactive Search Environments. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (CHIIR '26)*. Association for Computing Machinery, New York, NY, USA, 193–203. https://doi.org/10.1145/3786304.3788842

[22] Daniel Hienert, Dagmar Kern, Matthew Mitsui, Chirag Shah, and Nicholas J. Belkin. 2019. Reading Protocol: Understanding what has been Read in Interactive Information Retrieval Tasks. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, Glasgow Scotland UK, 73–81. https://doi.org/10.1145/3295750.3298921

[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*. ACM, Washington DC USA, 1006–1017. https://doi.org/10.1145/3626772.3657793

[24] Kaixin Ji, Danula Hettiachchi, Falk Scholer, Flora D. Salim, and Damiano Spina. 2025. SenseSeek Dataset: Multimodal Sensing to Study Information Seeking Behaviors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3 (2025), 92:1–92:29. https://doi.org/10.1145/3749501

[25] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ACM, Northampton Massachusetts USA, 101–110. https://doi.org/10.1145/2808194.2809465

[26] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report.

[27] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (Nov. 2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[28] Yidong Liang, Zhijing Wu, Yuchen He, Fengming Liang, Kexin Liu, and Jiaxin Mao. 2025. A Flexible User Study Platform for Generative Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 4066–4070. https://doi.org/10.1145/3726302.3730140

[29] Yidong Liang, Zhijing Wu, Fan Zhang, Dandan Song, and Heyan Huang. 2025. How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Padua Italy, 634–644. https://doi.org/10.1145/3726302.3729998

[30] Ying-Hsang Liu, Paul Thomas, Tom Gedeon, and Nicolay Rusnachenko. 2022. Search Interfaces for Biomedical Searching: How do Gaze, User Perception, Search Behaviour and Search Performance Relate?. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Regensburg Germany, 78–89. https://doi.org/10.1145/3498366.3505769

[31] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59, 7 (2008), 1041–1052. https://doi.org/10.1002/asi.20794

[32] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2014. Assessing the Cognitive Complexity of Information Needs. In *Proceedings of the 2014 Australasian Document Computing Symposium*. ACM, Melbourne VIC Australia, 97–100. https://doi.org/10.1145/2682862.2682874

[33] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: what observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, San Francisco California USA, 659–668. https://doi.org/10.1145/2505515.2507665

[34] Heather L. O'Brien, Jaime Arguello, and Rob Capra. 2020. An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing & Management* 57, 3 (May 2020), 102226. https://doi.org/10.1016/j.ipm.2020.102226

[35] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (April 2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[36] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748

[37] Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 2703–2707. https://doi.org/10.1145/3626772.3657914

[38] Johanne R. Trippas and J. Shane Culpepper. 2025. Report from the 4th Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *SIGIR Forum* 59, 1 (2025), 1–68. https://doi.org/10.1145/3769733.3769739

[39] Johanne R. Trippas, Damiano Spina, and Falk Scholer. 2025. *Adapting Generative Information Retrieval Systems to Users, Tasks, and Scenarios*. Springer Nature Switzerland, Cham, 73–109. https://doi.org/10.1007/978-3-031-73147-1_4

[40] 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 ACM SIGIR Conference on Human Information Interaction and Retrieval*, Vol. 55. ACM, Sheffield United Kingdom, 370–375. https://doi.org/10.1145/3627508.3638344

[41] Claire Wardle, Shaydanay Urbani, and Eric Wang. 2025. Evolving Health Information–Seeking Behavior in the Context of Google AI Overviews, ChatGPT, and Alexa: Interview Study Using the Think-Aloud Protocol. *Journal of Medical Internet Research* 27 (Oct. 2025), e79961. https://doi.org/10.2196/79961

[42] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1635–1644. https://doi.org/10.1145/3340531.3412017

[43] Yiwei Xu, Saloni Dash, Sungha Kang, Wang Liao, and Emma S. Spiro. 2025. AI summaries in online search influence users' attitudes. arXiv:2511.22809 (Dec. 2025). https://doi.org/10.48550/arXiv.2511.22809 arXiv:2511.22809 [cs].

[44] Yuyu Yang, Kelsey Urgo, Jaime Arguello, and Robert Capra. 2025. Search+Chat: Integrating Search and GenAI to Support Users with Learning-oriented Search Tasks. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Melbourne Australia, 57–70. https://doi.org/10.1145/3698204.3716446

[45] Oleg Zendel, Melika P. Ebrahim, J. Shane Culpepper, Alistair Moffat, and Falk Scholer. 2022. Can Users Predict Relative Query Effectiveness?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in*

Information Retrieval. ACM, Madrid Spain, 2545–2549. https://doi.org/10.1145/3477495.3531893

[46] Saber Zerhoudi, Adam Roegiest, and Johanne R. Trippas. 2026. Simulation of Interactive Information Retrieval: A Guided Tour. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval (CHIIR '26)*. Association for Computing Machinery, New York, NY, USA, 434–436. https://doi.org/10.1145/3786304.3787892