http://dx.doi.org/10.14236/ewic/BCSHCI2025.13

From Queries to Prompts: Comparing User Experience in Generative AI Tools and Search Engines

Misbahu Zubair
Dept. of Computing and Mathematics
Manchester Metropolitan University
Manchester
UK
misbahu.zubair@mmu.ac.uk

Muhammad Abubakar Alhassan
School of Engineering and Computing
University of Central Lancashire
Preston
UK
maalhassan@uclan.ac.uk

Farid Bello
Dept. of Computer Science
University of York
York
UK
farid.bello@york.ac.uk

Recent advances in large language models (LLMs) and the rise of Generative Artificial Intelligence (GenAI) tools, such as ChatGPT and Copilot, are ushering in a significant shift in the way people interact with information seeking systems. This study presents a mixed-methods investigation aimed at comparing user experiences of GenAI tools and Conventional Search Engines (CSEs). Twenty-four participants completed fact-finding and browsing tasks using both types of tools. Quantitative data was gathered using Tobii Fusion eye tracking device and a paper-based NASA-TLX survey, while qualitative data was gathered through semi-structured interviews after task completion. Results revealed that GenAI prompts were significantly longer and more conversational, and GenAI tools imposed higher cognitive load during fact-finding, but less cognitive load during browsing tasks. Qualitative findings indicated that users value GenAI for abstract, creative and personalised tasks, but expressed concerns over accuracy, trust, and data privacy. This study expands the limited body of research on comparing user behaviour and experiences when seeking information using CSEs and GenAI tools. It offers a novel contribution by identifying differences in cognitive load associated with completing different task types across the different tool types, highlighting patterns in GenAI interaction behaviours, while also identifying the factors that influence user preferences, perceptions, and overall experience of GenAI tools. The paper concludes with a discussion of the implications of these findings and provides recommendations for designing GenAI tools to enhance user experience.

Generative AI, Search Engines, ChatGPT, Copilot, Google, User Experience, Usability, Cognitive Load, Information Seeking

1. INTRODUCTION

Generative AI (GenAI) tools like ChatGPT and Copilot are transforming the way users search for information. While Conventional Search Engines (CSEs) such as Google and Bing will generate a list of results for a search query from existing indexed documents, GenAI tools leverage natural language processing (NLP) and conversational interfaces to dynamically generate responses that are personalised and context-aware. These tools allow users to engage in interactive, multi-turn conversations, enabling the reformulation of prompts, asking clarifying questions, and evaluating responses in real-time (White 2024).

Notably, CSE information seeking behaviours and experiences have been extensively studied (Hsieh-Yee 2001; Thatcher 2006; Aula et al. 2010; Kim et al. 2015; Azzopardi 2021). However, the use of GenAI for this purpose, particularly from a HCI perspective, remains relatively unexplored. Therefore, the aim of this study is to compare the user experiences associated with the use of GenAI tools and CSEs for information seeking, as well as explore the factors that affect and influence it. Specifically, the study focuses on the following research questions:

RQ1 How do users prompt GenAI tools compared to their querying of CSEs?

RQ2 Does the cognitive load experienced by users differ when seeking information on GenAI tools compared to CSEs?

RQ3 What factors influence users' preferences, perceptions, and user experience of GenAI tools?

To answer the above research questions, we conducted a mixed-methods study with 24 participants. Participants took part in a within-subjects experiment where they completed fact-finding and browsing tasks using Copilot and Google Search. Following the tasks, semi-structured interviews were used to explore participants' use, experiences, preferences, and perceptions of GenAI tools.

This paper makes several key contributions. First, it expands the limited body of research on the similarities and differences in user behaviour when performing fact-finding and browsing tasks using CSEs and GenAI tools. Second, it offers a novel contribution by identifying differences in cognitive load associated with completing these tasks using Copilot and Google Search. Third, it adds to existing knowledge on the user experience of GenAI tools by uncovering the factors that influence user preferences, perceptions, and overall experience. Finally, this paper discusses the implications of these findings, proposes recommendations for the design of GenAI tools, and discusses opportunities for future research.

The rest of the paper is organised as follows: Section 2 follows this with a review of related work; Section 3 presents our research methodology; quantitative results from the experimental study are presented in Section 4, while qualitative findings from interviews are presented in Section 5; further discussion of results and findings is presented in Section 6; and a conclusion is presented in Section 7.

2. RELATED WORK

2.1. User Experience Evaluation

User experience has been a core part of research on information seeking systems for decades. This has been explored in the context of general-purpose conventional search engines (Kelly and Azzopardi 2015), context-specific systems (Jansson et al. 2022), intelligent smart devices (Pyae and Joelsson 2018) and more recently, GenAI systems (Skjuve et al. 2023).

Studies on information-seeking user experience can be conducted as user studies in controlled (Kelly and Azzopardi 2015) and uncontrolled settings (Papoutsaki et al. 2017), or as survey studies (Skjuve et al. 2023). User studies usually require participants to perform predefined information-seeking tasks designed to simulate real user tasks, for example, in the form of simple fact-finding tasks and broader browsing tasks (Thatcher 2006). Data is then measured objectively as tasks are performed, e.g., using eye tracking devices, collected through self-reports after tasks are completed, e.g., using the NASA Task Load Index (TLX), or using both methods. This allows the collection of measures that can be used to better understand and measure user experience, including behavioural metrics like the number of queries entered and performance metrics like time spent completing tasks (Kelly and Azzopardi 2015), and experience measures like cognitive load (Gwizdka 2010). Some studies also include the collection of rich qualitative data through think-aloud, focus groups or interviews to explore deeper insights around users' experiences and behaviours (Thatcher 2006).

The querying stage of information seeking has been found to be more mentally demanding than other stages (Gwizdka 2010). Unsurprisingly, several studies have explored how users query search engines or prompt generative AI tools. CSE querying has been well studied through large-scale query log analysis and lab-based evaluations. Early work in this area involved the analysis of search logs, such as the analysis of AltaVista (Silverstein et al. 1998) and Excite (Jansen et al. 1998). These studies found search queries to be short at less than 3 words on average, and found queries to be reformulated or modified less than 25% of the time. GenAI prompting, on the other hand, is more expressive and conversational, and thus could lead to even more cognitive load; however, we found no peer-reviewed research in the literature that compares the length of queries and prompts across multiple information-seeking task types.

Studies have been conducted to measure and investigate how cognitive load affects user experience across several information retrieval task types and task stages (Gwizdka 2010). (Kelly and Azzopardi 2015) used the NASA Task Load Index (TLX) to gather reported experienced workload when comparing different search engine results pages. Pupil diameter measured using eye-tracking devices has also been used as an indicator of cognitive load in information-seeking tasks (Ji et al. 2024; Al-Samarraie and Al-Hatem 2018). Similarly, task completion time has also been used to measure cognitive load in information-seeking tasks (Mendel and Pak 2009).

2.2. Comparing User Experience of CSEs and GenAI Tools

Several research studies on the applications and potential of GenAI tools have been conducted, with positive findings in diverse areas including information seeking, productivity, planning, research,

writing and learning (van den Berg and du Plessis 2023; Deng et al. 2024). The accuracy of GenAI tools as information seeking tools, is a popular area of research, especially in health information seeking.

Other studies have focused on understanding the perceptions and preferences of users around the use of GenAI tools (Zhang et al. 2025), and user experiences when using GenAI tools for information seeking (Wagwu et al. 2023).

Comparisons between GenAI tools and CSEs are also focused on the accuracy of results, especially in health information seeking. Only a few studies have compared these two groups of tools with the aim of evaluating differences in user behaviours and preferences associated with information-seeking tasks. Liu et al. (2024) evaluated the experiences of users completing academic information retrieval tasks in a between-subjects study; the group supported by a GenAI tool completed their tasks with fewer clicks and page visits, in less time, and with increased query length compared to the group that only used a CSE. Luo et al. (2025) compared user preferences for travel information seeking tasks using 4 between-subjects task-based studies. They found that participants' preference for GenAI reduced when completing tasks that are decision-based. More recently, Kaiser et al. (2025) conducted a study comparing ChatGPT and Google Search behaviours during browsing-type tasks; ChatGPT users were found to be faster and more likely to find correct answers, but participants still reported a preference for Google.

Our review of the literature showed no studies have been conducted to compare user querying and prompting behaviours, as well as the cognitive load associated with completing different types of information-seeking tasks in CSEs and GenAI tools.

However, a recent study which explored factors influencing users' intention to switch from using CSEs to GenAI tools found dissatisfaction as a result of low information fit and information overload, social factors, and the perceived value of GenAI tools due to perceived interactivity, perceived anthropomorphism, and information quality as factors leading to switching (Zhou and Li 2024). A major limitation of this work is that it does not consider security, privacy and ethical factors, which are significant issues from the perspective of users (Huang et al. 2023).

3. METHODOLOGY & PROCEDURE

We employed a two-part mixed-methods research design with 24 participants to answer our research questions; each part is described in detail below, and a visual summary is presented in Figure 1.

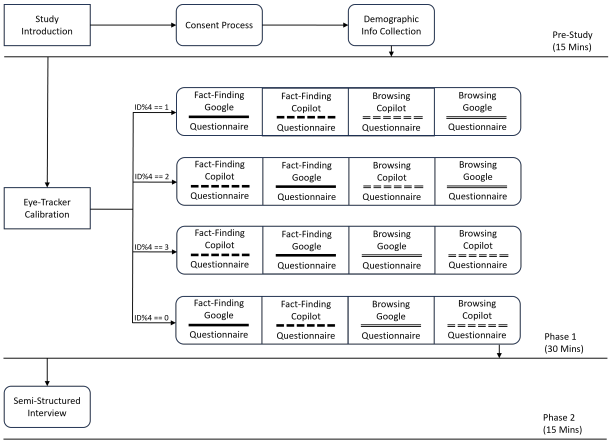

Figure 1: *Overall research procedure.*

The first part of the study employed a within-subjects experiment with two independent variables with two levels each: the information-seeking tool (Microsoft Copilot as the GenAI tool or Google Search as the CSE) and the task type (fact-finding or browsing); and four dependent variables: input length, NASA-TLX score, time to complete task and pupil diameter. Each participant completed four tasks (2 x fact-finding and 2 x browsing) using both tools (1 x fact-finding and 1 x browsing on each tool). Input length was examined to understand the effect of tool type on user interaction across fact-finding and browsing tasks, while the other three variables were analysed to evaluate the effect of tool type on cognitive load across fact-finding and browsing tasks.

A semi-structured interview was then conducted in the second part of the study to explore participants' use, experiences, preferences, and perceptions of GenAI tools.

Google Search was chosen as the CSE in this study due to its dominance in the field and for being synonymous with web search (Lewandowski 2023). While much of existing literature focuses on ChatGPT when exploring user experience of GenAI tools, e.g. (Kaiser et al. 2025; van den Berg and du Plessis 2023; Deng et al. 2024; Liu et al. 2024; Huang et al. 2023), this study chose to explore a popular yet underexplored example of GenAI in the form of Copilot. This was done to allow for the comparison of findings with ChatGPT-focused studies and to give participants who are primarily familiar with ChatGPT an opportunity to experience a different GenAI tool before participating in the interviews.

Ethical approval for this study was granted by the Science and Engineering Research Ethics and

Governance Committee at Manchester Metropolitan University.

3.1. Participants

Recruitment was done through multiple mediums including posting adverts on social media groups, online communities and forums for students and staff of Universities in Northwest England; distributing physical recruitment posters; and taking a snowball approach to reach additional participants through those recruited. A total of 24 participants (10 University students and 14 University staff) took part in the study. The average age of participants was 34.17, 3 were female and 21 male, 23 reported using CSEs daily, and 1 several times a week, 4 reported using GenAI tools daily, while 2 reported rarely or never, and the remaining 19 reported using GenAI tools either weekly or monthly.

3.2. Tasks

Two fact-finding and two browsing tasks, similar to those used by Thatcher (Thatcher 2006) were developed for this study as shown in Table 1. Fact-finding tasks were designed to simulate common information-seeking behaviours that involve retrieving specific and factual responses; they required the user to retrieve a single information unit. Browsing tasks on the other hand were designed to be exploratory in nature, and simulated information-seeking behaviours that require complex and multiple iterations of retrieving, analysing, and evaluating information.

3.3. Data Collection

A short Microsoft Forms survey was used to gather demographic information and information-seeking behaviours from participants.

In the first part of the study, participants completed tasks on a Windows 10 Desktop PC with a 1920 x 1080 display screen. Tobii Fusion eye-tracking device and the Tobii Pro Lab v1.241.54542.0 software were used to capture eye-gaze data and participants' screen recordings. Self-reported cognitive load for each task completed was collected using a paper-based NASA-TLX questionnaire.

In the second part of the study, each participant took part in a semi-structured interview for 15-20 minutes. The interviews explored participants' perceptions and preferences regarding the use of GenAI tools and CSEs, and strategies for querying, prompting and evaluating responses and results. All interviews were audio recorded.

3.4. Procedure

The study took place in a controlled UX Lab environment at Manchester Metropolitan University. Each recruited participant chose a 60-minute block of time using a dedicated Doodle poll.

Each session began with the researcher explaining the nature of the research and the session's activities and procedures. After this, and if the participant consents to participate, they were assigned an ID between N1 and Nn (where n is the total number of participants), and asked to complete a short demographic information form before going through the calibration process for the eye-tracking equipment. They are then presented with a pre-configured screen layout consisting of two Google Chrome browser windows positioned side by side as shown in Figure 2.

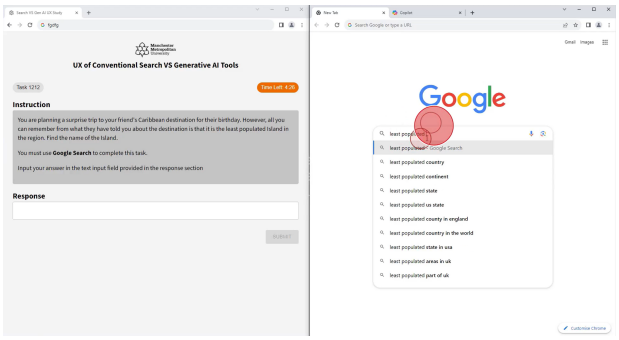

Figure 2: *Screenshot showing the two Google Chrome browser windows positioned side.*

The leftmost window contained the task app, a single-page application developed using ReactJS[1] and deployed on Vercel[2], used in this study to present tasks. The app was designed to utilise an incomplete counterbalancing measure using 4x4 Latin Square to determine the orderings of task-tool pairs (see Table 2). The order in which task types are performed was kept constant; all participants completed fact-finding tasks before browsing tasks to ensure no learning effect was caused by first completing the more complex and time-consuming browsing tasks.

The main menu of the app consisted of four task cards, with at most one (the next task) enabled at any time. Completed tasks remained disabled with their details visible, while upcoming tasks were disabled and had their details hidden, as shown in Figure 3. When a participant selected the enabled card, they were presented with instructions for the task, including whether to use Google Search or Microsoft Copilot, along with a response box to complete once the task was finished, as shown on the left

[1] https://react.dev/
[2] https://vercel.com/

Table 1: *Fact-finding and Browsing tasks completed by participants.*

Type	Task	Task ID
Fact-Finding	You are visiting Lisbon next week. One of your friends has suggested that you visit the city's oldest bookstore. Find the name of the bookstore.	1111
Fact-Finding	You are planning a surprise trip to your friend's favourite Caribbean destination for their birthday. However, all you can remember from what they have told you about the destination is that it is the least populated Island in the region. Find the name of the Island.	1212
Browsing	You have found out that your favourite charity has organised a marathon to raise funds and you are thinking about participating in it. However, this will be your first time running a marathon so before you make up your mind you want to find out as much as you can about training for it.	2424
Browsing	You have just moved to a new house with a large garden and you are considering growing fruits and vegetables. However, you have no previous gardening experience, so you want to gather as much information as you can on how to get started.	2525

Table 2: *Incomplete counterbalancing of Task-Tool pairings using a 4x4 Latin Square design (C = Copilot, G = Google)*

Task 1	Task 2	Task 3	Task 4
1212 - G	1111 - C	2525 - C	2424 - G
1111 - C	1212 - G	2424 - C	2525 - G
1212 - C	1111 - G	2525 - G	2424 - C
1111 - G	1212 - C	2424 - G	2525 - C

window in Figure 2. Participants were then expected to use the appropriate tool, opened in the rightmost browser window, to complete the task and enter their response in the response box. Once a response was submitted, the app prompted the participant to complete the NASA-TLX questionnaire for the completed task and then returned the participant to the updated main menu.

The window arrangement was kept constant throughout, and participants were asked not to move or resize windows. This was to ensure that Areas of Interest (AOI) could be easily and consistently defined when analysing the eye-tracking data.

The first part of the study concluded once participants had completed all tasks and the associated NASA-TLX questionnaires, which lasted 20 minutes on average. Finally, each participant took part in a short semi-structured interview; this lasted 18 minutes on average.

3.5. Data Analysis

3.5.1. Eye-tracking Data

On reviewing screen recordings for all participants' tasks before the analysis, N12's browsing task data was discarded for using a single tool for both tasks and all of N6's task data was discarded for using task descriptions as input rather than formulating their own prompts and queries.

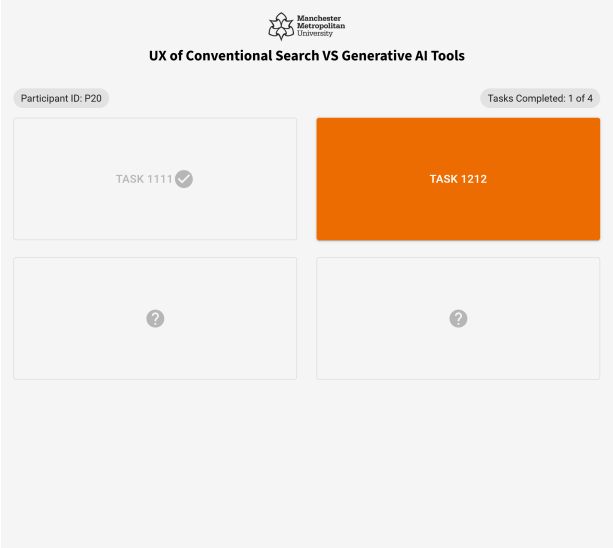

Figure 3: *Main menu showing completed, next, and upcoming tasks.*

Tobii Pro Lab was used to extract relevant logged data from the remaining participants' recorded data. An AOI was first established over the browser window in which all tasks were completed. Then, for each recording, Times of Interest (TOI) were created to denote the start point (i.e., when the participant keyed their first prompt/query character) and end point (i.e., when the participant submitted their response) for all tasks. This made it possible to export eye gaze data for individual tasks by exporting recorded data between all start and end points as separate CSV files. Each CSV file contained rows of data points each with several columns of data such as recording timestamp (in milliseconds), pupil diameter left (pupil diameter for left eye in millimetres), pupil diameter right (pupil diameter for right eye in millimetres), left validity (specifies whether the left eye was found using either Valid or Invalid), right validity (specifies whether the right eye

was found using either Valid or Invalid), and AOI hit (uses 0 or 1 to specify whether gaze was within the AOI when the current row data was recorded).

Analysis was conducted using Python and Google Colab. The time to complete each task was calculated by creating a new column in each dataset that stored the difference between the timestamp associated with each row of data and the timestamp of the previous row. Then the sum of time differences was taken for rows where both Validity left and Validity right were valid, and AOI hit was 1. The result was divided by 1000 to get the time participants spent looking at the window in which they completed the task in milliseconds. Similarly, the average pupil diameter for each task was calculated by taking the average of the left and right pupil diameters in each row where both "Validity left" and "Validity right" were valid, and AOI hit was 1, and then calculating the average of these values. Wilcoxon Sign Rank Tests were then used to compare (within subjects) time to complete tasks and average pupil diameters with tools and task types as independent variables.

3.5.2. Survey Data
Data gathered through the NASA-TLX questionnaires was transferred to an Excel CSV file. Each dimension rating was converted to a score out of 100 by subtracting 1 from the rating and multiplying the result by 5. The CSV file was then uploaded to the Google Colab project for further analysis which involved calculating an overall score for each response by averaging all 6 dimension scores, calculating an average score for each dimension for each task type and tool pairing, and making within-subject comparisons using the Wilcoxon Signed Rank Test with tools and task types as independent variables and average dimension scores and overall scores as dependent variables.

3.5.3. Interview Data
Since interviews were recorded using Microsoft Teams meetings, the first transcription step was done using Microsoft Teams' automatic transcription feature. Transcripts were then manually reviewed to correct errors and ensure accurate speaker attributions. Once all transcripts were reviewed and verified, they were uploaded to NVivo for analysis. A thematic analysis approach, as outlined by Braun and Clarke (2006), was used to identify patterns and themes within the data. Specifically, a theoretical approach to thematic analysis was used to ask questions about the data on how participants perceive Gen AI tools and how they used them compared to CSEs.

4. QUANTITATIVE RESULTS

The Wilcoxon Signed Rank Test, a non-parametric test, was used to make within-subject comparisons (n=22) of measured metrics. Comparisons were made across two dependent variables, tool (Copilot or Google) and task type (fact-finding or browsing).

4.1. Input Length

Only participants' first inputs were analysed due to the differences in the number of inputs used by participants per task. First Copilot prompts (mean = 7.50) were found to be significantly longer than first Google queries (mean = 5.31) for fact-finding tasks (w=35.50, p=0.008). Similarly, for browsing tasks, the first Copilot prompts (mean = 14.95) were found to be significantly longer than the first Google queries (mean = 7.77) (w=43.50, p=0.023).

4.2. NASA-TLX Scores

The average scores for all 6 NASA-TLX dimensions reported for Copilot fact-finding tasks were higher than those reported for Google fact-finding tasks, but the opposite was found to be the case for browsing tasks as shown in Figure 4. However, only the reported average frustration score for Copilot fact-finding tasks was found to be significantly higher than for Google (w = 23.0, p = 0.03).

Similar to the average dimension scores, the overall scores for Copilot fact-finding tasks (Mean = 19.47) were found to be higher than those reported for Google fact-finding tasks (Mean = 12.91), and those reported for Google browsing tasks (Mean = 43.79) were higher than those for Copilot browsing tasks (Mean = 35.98). However, both the fact-finding (w = 63.0, p = 0.12) and browsing (w = 67.50, p = 0.05) differences were not statistically significant.

4.3. Time to Complete Tasks

The average time taken to complete fact-finding tasks using Copilot (Mean = 63.60s) was found to be higher than when using Google (Mean = 53.53s); although this difference was not statistically significant (w = 98, p = 0.37).

On the other hand, the average time taken to complete browsing tasks using Copilot (Mean = 203.92s) was found to be less than when using Google (Mean = 226.79s), but this difference was also not statistically significant (w = 98, p = 0.37).

4.4. Pupil Diameter

The mean of the average recorded pupil diameters for fact-finding tasks completed in Copilot (Mean = 2.62mm) was found to be higher than for those

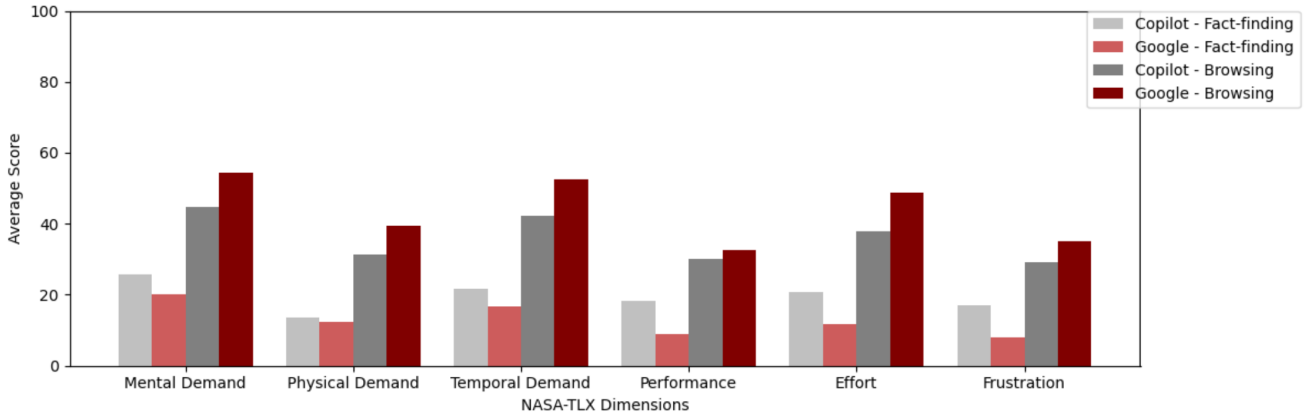

Figure 4: *Average scores for NASA-TLX Dimensions for all task types and tools.*

completed in Google (Mean = 2.61mm), although not statistically significant (w = 114.0, p = 0.70). There was also no statistically significant difference found in the average recorded pupil diameters for browsing tasks (w = 82.0, p = 0.16), although a higher mean was found for tasks completed using Google (Mean = 2.60mm) compared to Copilot (Mean = 2.58mm).

5. QUALITATIVE FINDINGS

Thematic analysis findings, in the form of themes made of patterns in participants' (n=24) reports, related to RQ1 are presented in 5.1 while those related to RQ3 are presented in 5.2.

5.1. User Interaction

Themes representing GenAI interaction strategies reported by participants are presented below:

5.1.1. Converse

Participants reported interacting with GenAI tools conversationally as if they were "talking to someone" (N1). With conventional search engines on the other hand, participants reported using "keywords" or "key terms" and avoiding "fluff" (N2) as they found that to "obscure the results" (N3). Pleasantries were also reported to be used in prompts, including "'please' and 'thank you' at the end" (N15). Conversational prompting was reported to be "easier because it is like you're chatting with a real human being, unlike with Google where you have to think about what's the best way to query it so that it gives you the right result" (N16).

5.1.2. Provide Context

Participants noted that providing contextual information is important for getting good responses from GenAI; "the more context you give Gen AI, the more useful the results it brings back, whereas it is the opposite with Google" (N12). Providing "maximum possible information" (N17) and "more detail about what you are looking for" (N18) are some of the approaches to providing contextual information reported by participants.

5.1.3. Review

Participants reported several approaches for reviewing content generated by GenAI tools. One approach is to look at "the source of the result and the quality of the source" (N24) similar to how participants reported reviewing results provided by conventional search engines; however, this is only possible when GenAI platforms provide links to their sources. Another approach reported was to use CSEs to "cross-verify" (N20) claims made by GenAI. Participants also reported using self-verification as a way of reviewing GenAI outputs, e.g., by testing generated programming code and "if it runs, the code is fine" (N12). Lastly, participants may use "human judgement to decide how much you are relying on generated output" (N13).

5.1.4. Iterate

Participants reported going through the process of conversational prompting, providing context, and reviewing responses iteratively, up to a certain point. N10 described the process as a "feedback cycle" where they continuously review and provide feedback with specific instructions such as "Can you phrase it like this instead? and Can you expand on that?" until the GenAI tool "reaches the memory limits of the model and it starts giving the same answers again and again and again".

5.2. Factors Affecting Perception, Preference and User Experience

Factors affecting participants' preferences, perceptions and user experience of GenAI tools were categorised into 3 major themes based on their source, these themes are presented below:

5.2.1. Tool Factors

Several participants reported conventional search engines as their default tools, over GenAI tools, due to their ease of access, i.e. users do not have to log in to access their full features; and the fact that they can easily access search engines on their smartphones.

Participants shared concerns about the accuracy of information provided by GenAI tools for several reasons, including a reported lack of up-to-date knowledge. This resulted in participants opting to use conventional search engines in situations where "very recent" (N2) information is required to provide accurate responses. Participants did not completely trust conventional search engines either or think every search engine result was accurate, but preferred taking advantage of the search engine features that let users "go to the sources and then decide" (N20). Participants found this to be a more trustworthy process than the way in which GenAI tools provide "responses rather than choices" (N24), in addition to the possibility that GenAI tools may "recreate information in a new way, and in the process of doing that, lead to misinformation" (N12). However, some participants perceived higher accuracy in newer GenAI models, for example, N2 mentioned that they "don't always feel like ChatGPT is accurate", but they "trust the new 4.o more than the old one".

Lastly, participants found GenAI tools' ability to understand prompts in natural language made them easier to interact with.

5.2.2. Task Factors

Participants reported they were more likely to utilise GenAI tools when they had problems or tasks that were abstract or creative. Providing advice (N1, N10), generating and validating ideas (N3, N12, N14, N13, N16), making plans (N2, N10, N19), and improving user-generated content (N12, N13, N14, N16, N19, N24) were some of the tasks described by participants as suitable for GenAI platforms. On the other hand, most participants reported a preference for conventional search engines when tasks involve facts, or as described by N3, when "more tangible or specific" responses are needed.

Participants also reported positive experiences associated with GenAI's ability to handle tasks that require responses tailored to individual needs. It was highlighted that conventional searches struggle to "find specific examples" (N15) of niche problems or issues, and N10 pointed out that in some situations (e.g. debugging code) identifying examples of "what's been done and how other people have done things" may not be enough because the user may "hit a mistake that [others] hadn't hit, and then you wouldn't be able to get any more feedback on it". GenAI tools mitigate this by generating results appropriate for users' unique situations that they "wouldn't be able to get with just searching on Google" (N1).

The use of GenAI for creating highly personalised playful content was also reported. This included both visual content e.g., "generating highly personalised joke images to send to friends, you can respond to someone with an image that is basically like a meme but it's literally got their name in it" (N1), as well as textual content e.g. "to write a little short novel on cheese in a semi sort of romantic way which was quite fun" (N3).

Participants reported a preference for conventional search engines when a "quick result" is needed, especially for fact-based questions. They found generative AI tools to be slow, "painstakingly so" (N2), and likely to provide unnecessary contextual information. However, they preferred GenAI tools in situations where critical evaluation is more important than quick task completion time, as they can avoid "going through the different search results and reading everything" (N14).

Lastly, participants reported turning to GenAI for tasks and questions that they had very little understanding or knowledge of. Combining GenAI and conventional search was also reported as a useful approach when not enough is known about a fact-based task or when one does not know what they are even searching for, therefore they "describe it, ChatGPT or Copilot can tell you what you're actually talking about and then you Google that"(N1).

5.2.3. User Factors

Most participants mentioned CSEs, specifically Google, as their default information seeking tool for habitual reasons, e.g. because they "are used to using it" (N13), or just because they "always used it" (N16). Similarly, N17 mentioned ChatGPT as their default GenAI tool because it "was the first option" they used.

Participants' understanding of GenAI, its capabilities and potential affected their initial perception and use of GenAI tools. Several participants reported initially using GenAI tools just because they found the technology "fun"(N1) or "interesting" (N3), but then "started to use it more as a tool than just something to mess around with" (N15). Others started using it "with the expectation that it's AI and it knows everything, but then realised that's not necessarily true" (N16).

While there was a lot of praise for GenAI for the increase in productivity it has allowed participants to gain, some participants still feared becoming too

reliant on GenAI platforms for everyday tasks that they used to perform independently. For example, N19 admitted to being worried because the first thing they think about when they have a task is “how can I use ChatGPT to do that?”. There were also concerns that the effectiveness of GenAI platforms in supporting users could have long-term consequences e.g. by making users “lazy” (N17, N23).

Participants also shared concerns about data privacy, security, and anonymity, including GenAI platforms’ handling of data from sensitive interactions with users around topics like health or relationships, and whether requests “to permanently delete this data” (N19) can be made and honoured. Similarly, registering an account or linking existing accounts to use GenAI platforms concerned N2 because they were “not too trustworthy of the fact that they fully get rid of that information or anonymise it”.

Reluctance to share personal information with GenAI tools was reported due to participants’ fear that their personal information could be used to train models. They questioned what made up the public data used in training GenAI models, whether “the information you put becomes public” (N12), and whether inputting your personal information as part of a prompt or input means ”someone else could search who is [name] and they could get your personal information” (N20).

6. DISCUSSIONS

We found significant differences in the length of inputs used to interact with Copilot and Google across both task types. This is in line with previous findings showing ChatGPT prompts to be significantly longer than Google queries in a study comparing the user experience and search performance of the two tools (Xu et al. 2023).

We also found that on average, fact-finding tasks were completed faster using Google, likely due to Google’s efficient indexing, information retrieval, and the manner in which it presents results, making it suitable for simple, well-defined queries (Lewandowski and Kammerer 2021). Although the difference in time taken to complete this task using both tools is not significant, this is still an interesting finding when considered together with other findings of this study. Several participants reported frustrations with GenAI tools’ method of communicating results when all they need is a “quick fact”, but have to read through the context-filled natural language response. This is likely a significant factor that influenced how long participants took to complete these tasks. Additionally, the reported frustration when using Copilot to complete fact-finding tasks is significantly higher than when Google is used further revealing participants’ feelings about GenAI tools and fact-finding tasks. In fact, while not statistically significant, the scores for all other NASA TLX dimensions measured after completing fact-finding tasks were higher for Copilot than Google, and the mean of the average pupil diameters recorded when participants used Copilot to complete fact-finding tasks was higher than when they used Google. All these point towards a higher cognitive load by users when using Copilot for fact-finding tasks compared to Google.

Browsing tasks, on the other hand, are similar to the abstract, creative and planning-type tasks that participants reported as suitable for completing with GenAI tools. They are also suited to the GenAI interaction style identified by this study, which allows users to iteratively prompt and review until their information needs have been met. Browsing tasks were, on average, completed faster on Copilot than on Google; however, this difference was not statistically significant. Although Kaiser et al. (2025) also found that ChatGPT users were faster in completing browsing tasks than Google users. All other quantitative measures were also consistent, although without statistical significance, in pointing towards Copilot as the tool that requires less cognitive load to complete browsing tasks: all NASA TLX dimensions were lower for Copilot than for Google, and the mean of average pupil diameter was lower for Copilot than for Google. This aligns with the existing studies suggesting that GenAI can reduce cognitive load by directly providing synthesised responses rather than requiring users to scan through multiple documents (White 2024). In general, we found that most of the features of GenAI tools that participants associated with positive experiences were aligned with the nature of browsing-type tasks.

Factors reported by participants as influencing their choice of using CSEs, Google in particular, over a GenAI tool include ease of access, habit and familiarity. This can be easily understood once the ubiquity of Google is considered, and common features on personal devices, such as the search functionality of browser URL bars and the search widgets on mobile devices.

There were several concerns around accuracy, trustworthiness, and data privacy that were raised and have the potential to be key barriers to the broader adoption of GenAI tools. Potential inaccuracies in generated responses, particularly for factual questions and questions requiring up-to-date and verifiable information, were a cause of

concern to participants. Although recent updates in mainstream GenAI tools such as ChatGPT mean they no longer solely depend on their training data and can search the internet for up-to-date information. However, the risk of hallucination may still exist and could still lead to misinformation.

Participants feared that personal data entered into GenAI platforms could be misused or used for training purposes. These concerns align with broader societal concerns around the ethical use of data in AI systmes (Al-kfairy et al. 2024). This highlights the need for clearer and more robust privacy guarantees from GenAI tool providers. While Copilot does not allow temporary or incognito interactions, ChatGPT has a temporary chat feature, which does not store chat history or use data from interactions to train models, however, data may be stored for 30 days for security purposes. Further research to understand whether this changes the perception of trust for users would be useful.

Lastly, participants also raised fears of over-reliance on GenAI tools, expressing concerns about the potential for these tools to erode, not just their's but society's, critical thinking skills and independent problem-solving capabilities.

6.1. Recommendations

Based on the results and findings from this study, the following recommendations are proposed for designers, and developers of GenAI tools.

6.1.1. Design for Task Fit

GenAI tools should offer users the option to switch between detailed conversational responses and concise factual outputs; or automatically switch based on context. This could help reduce unnecessary cognitive load during fact-finding tasks where speed and precision are valued.

6.1.2. Support User Verification

The fact that participants valued the transparency of CSEs, which allow direct access and review of sources, could be used as a foundation for addressing the accuracy and trust issues with GenAI. Developers should prioritise features that provide clearer source citations and empower users to verify information.

6.1.3. User-centred Privacy Settings

In addition to features that allow temporary interactions and the ability to request the deletion of data, GenAI platforms should empower users with the option to specify, in granular detail, how and what aspects of their data are processed. For example, users could choose to exclude any names and physical characteristics of persons included in prompts from being stored and processed, while allowing all other data to be retained to improve future interactions. This control could be enforced at a global or session level through settings, or even at the prompt level using inline commands.

6.1.4. Design to Support, not Replace User Skills

To address concerns around over-reliance and potential impact on skills development and retention, GenAI tools should be designed to support and enhance users' critical thinking and independent problem-solving. Restrictive modes tailored to provide support towards solutions, while encouraging user engagement and critical reflection, should be explored. These could be used or even enforced in certain contexts through policies or organisational controls, for example in academic institutions or during examinations. A good example is ChatGPT's recent Study Mode (OpenAI 2025), designed to support learning by prompting users to think through problems instead of simply providing direct answers.

7. CONCLUSION

This study provides novel insights into how users interact with GenAI tools compared to CSEs when completing fact-finding and browsing tasks, and the differences in associated cognitive load. Quantitative eye-tracking, input and cognitive load measures were combined with rich qualitative data gathered through interviews with 24 participants. Findings highlight users' preference for GenAI tools for creative, abstract, and personalised tasks, despite reports of challenges related to accuracy, trust, and privacy. Findings also showed a preference for CSEs for fact-finding tasks, and a reported lower cognitive load compared to fact-finding tasks in GenAI. We recommend that, as GenAI systems continue to evolve, designers and developers ensure that these tools are transparent, suitable for various task types, trustworthy, and designed to complement and develop rather than replace human skills.

7.1. Limitations

As with any research study, this study has its limitations. Firstly, only two platforms were used by participants to complete tasks in this study, Copilot and Google, and these tools do not represent all existing GenAI tools and conventional search engines. Although this limitation was not applied when participants were interviewed, most of their responses were associated with a small subset of GenAI tools and search engines as well. Secondly, the two types of tasks completed by participants do not fully represent all possible task types and response formats, image search and generation, for instance, were not covered by this study. Additionally, participants in this study were predominantly male

and were either students or staff at universities in Northwest England, and therefore not fully representative of the diverse population of GenAI and conventional search users.

The rapid advancement of GenAI tools means that several changes have been made to mainstream tools and search engines, including those used in this study and mentioned by participants, since the research was conducted. As a result, the results and findings of this study may no longer reflect the current state of the art, or fully capture the performance, features, or user experiences of current GenAI tools and search engines. This highlights the need for ongoing research to ensure that findings remain relevant and reflective of the most up-to-date developments in this field.

7.2. Future Work

Future studies should explore a wider range of tasks involving different information formats completed on a wider range of platforms and include a larger group of participants that is more diverse in terms of levels of expertise, and professional and cultural background to enhance the generalisability of the findings. Similar studies with people with disabilities, including physical and cognitive, should contribute to knowledge on the accessibility and accessible design of GenAI tools. Lastly, future studies should also consider longitudinal designs to explore how user experience and preferences evolve.

REFERENCES

Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., and Alfandi, O. (2024). Ethical challenges and solutions of generative ai: An interdisciplinary perspective. In *Informatics*, volume 11, page 58. Multidisciplinary Digital Publishing Institute.

Al-Samarraie, H. and Al-Hatem, A. I. (2018). The effect of web search result display on users' perceptual experience and information seeking performance. *The Reference Librarian*, 59(1):10–18.

Aula, A., Khan, R. M., and Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 35–44.

Azzopardi, L. (2021). Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 27–37.

Deng, R., Jiang, M., Yu, X., Lu, Y., and Liu, S. (2024). Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies. *Computers & Education*, page 105224.

Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology*, 61(11):2167–2187.

Hsieh-Yee, I. (2001). Research on web search behavior. *Library & Information Science Research*, 23(2):167–185.

Huang, K., Zhang, F., Li, Y., Wright, S., Kidambi, V., and Manral, V. (2023). Security and privacy concerns in chatgpt. In *Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow*, pages 297–328. Springer.

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. In *Acm sigir forum*, volume 32, pages 5–17. ACM New York, NY, USA.

Jansson, M., Liisanantti, J., Ala-Kokko, T., and Reponen, J. (2022). The negative impact of interface design, customizability, inefficiency, malfunctions, and information retrieval on user experience: A national usability survey of icu clinical information systems in finland. *International Journal of Medical Informatics*, 159:104680.

Ji, K., Hettiachchi, D., Salim, F. D., Scholer, F., and Spina, D. (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*, pages 1006–1017.

Kaiser, C., Kaiser, J., Schallner, R., and Schneider, S. (2025). A new era of online search? a large-scale study of user behavior and personal preferences during practical search tasks with generative ai versus traditional search engines. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Kelly, D. and Azzopardi, L. (2015). How many results per page? a study of serp size, search behavior and user experience. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 183–192.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., and Yoon, H.-J. (2015). Eye-tracking analysis of user behavior and performance in web

search on large and small screens. *Journal of the Association for Information Science and Technology*, 66(3):526–544.

Lewandowski, D. (2023). Search engines that give users control over their results. *Information Today Europe*.

Lewandowski, D. and Kammerer, Y. (2021). Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. *Behaviour & Information Technology*, 40(14):1485–1515.

Liu, Z., Wang, X., and Li, L. (2024). Chatgpt-assisted information retrieval: A comparative study of user behavior in academic information retrieval. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–3.

Luo, X., Xu, D., Li, Y., and Wan, L. C. (2025). Advancing information search through genai: the roles of search type, travel motive and genai customization level. *International Journal of Contemporary Hospitality Management*.

Mendel, J. and Pak, R. (2009). The effect of interface consistency and cognitive load on user performance in an information search task. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 53, pages 1684–1688. SAGE Publications Sage CA: Los Angeles, CA.

OpenAI (2025). Introducing study mode — openai.com. `https://openai.com/index/chatgpt-study-mode/`. [Accessed 31-07-2025].

Papoutsaki, A., Laskey, J., and Huang, J. (2017). Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 17–26.

Pyae, A. and Joelsson, T. N. (2018). Investigating the usability and user experiences of voice user interface: a case of google home smart speaker. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services adjunct*, pages 127–131.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. (1998). Analysis of a very large altavista query log. Technical report, Citeseer.

Skjuve, M., Følstad, A., and Brandtzaeg, P. B. (2023). The user experience of chatgpt: Findings from a questionnaire study of early users. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–10.

Thatcher, A. (2006). Information-seeking behaviours and cognitive search strategies in different search tasks on the www. *International journal of industrial ergonomics*, 36(12):1055–1068.

van den Berg, G. and du Plessis, E. (2023). Chatgpt and generative ai: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences*, 13(10):998.

Wagwu, V., Okpala, A. E., Oladokun, B. D., and Ajani, Y. A. (2023). User experience with chatgpt in a nigerian university library: Exploring users' satisfaction and feedback. *University of Ibadan Journal of Library and Information Science*, 6(2).

White, R. W. (2024). Tasks, copilots, and the future of search: A keynote at sigir 2023. In *ACM SIGIR Forum*, volume 57, pages 1–8. ACM New York, NY, USA.

Xu, R., Feng, Y., and Chen, H. (2023). Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*.

Zhang, Y., Yang, X., and Tong, W. (2025). University students' attitudes toward chatgpt profiles and their relation to chatgpt intentions. *International Journal of Human–Computer Interaction*, 41(5):3199–3212.

Zhou, T. and Li, S. (2024). Understanding user switch of information seeking: From search engines to generative ai. *Journal of Librarianship and Information Science*, page 09610006241244800.