Search+Chat: Integrating Search and GenAI to Support Users with Learning-oriented Search Tasks

Yuyu Yang
University of North Carolina at Chapel Hill
North Carolina, USA
yuyu18@live.unc.edu

Kelsey Urgo
University of San Francisco
California, USA
kurgo@usfca.edu

Jaime Arguello
University of North Carolina at Chapel Hill
North Carolina, USA
jarguello@unc.edu

Robert Capra
University of North Carolina at Chapel Hill
North Carolina, USA
rcapra@unc.edu

Abstract

Generative AI (GenAI) technologies such as ChatGPT are changing the ways people interact with information. To illustrate, popular search engines (e.g., Google) have started integrating responses from GenAI tools with the traditional search results. In this paper, we explore the integration of GenAI technology with traditional search in the context of a learning-oriented task. We report on a between-subjects study ($N = 40$) in which participants completed a complex, learning-oriented search task. Participants were assigned to one of two conditions. In the SearchOnly condition, participants used a traditional web search system to gather information. In the Search+Chat condition, participants used an experimental system that combined a traditional web search component and an interactive GenAI-based chat component (Chat AI). The study investigated seven research questions. RQ1-RQ3 focused on differences between groups: (RQ1) post-task perceptions, (RQ2) search behaviors, and (RQ3) learning outcomes. To measure learning, participants completed a multiple-choice test before the search task, immediately after, and one week later (to measure retention). RQ4-RQ7 delved deeper into participants' behaviors and experiences in the Search+Chat condition: (RQ4) motivations for (and gains from) engaging with the Chat AI; (RQ5) the phases during which participants engaged with the Chat AI; (RQ6) the types of queries issued to each component; and (RQ7) perceptions about the information returned by each component.

CCS Concepts

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

Keywords

Generative AI, search-as-learning, search behavior, mixed-methods

CHIIR '25, March 24–28, 2025, Melbourne, VIC, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1290-6/25/03
https://doi.org/10.1145/3698204.3716446

ACM Reference Format:
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 *2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25), March 24–28, 2025, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3698204.3716446

1 Introduction

Our research in this paper lies at the intersection of Generative AI (GenAI) and search-as-learning (SAL). GenAI technologies such as ChatGPT have revolutionized the ways people interact with information. People can interact with GenAI technologies as standalone services. However, GenAI technologies are also being *integrated* into existing search systems. For example, Google has begun to show GenAI-based responses above the search results. Information retrieval (IR) researchers have also argued that GenAI tools should not be seen as a replacement for traditional search systems. Instead, we should consider how GenAI tools can be *integrated* into traditional search systems to support users with complex information-seeking tasks [36]. In this paper, we explore the integration of a GenAI-based chat with a standard web search interface. Specifically, we explore this integration in the context of a complex, learning-oriented search task.

SAL research explores how people use search systems to learn. SAL studies have focused on a wide range of topics, for example: (1) understanding the contexts in which people search to learn; (2) understanding factors that may impact learning during search; (3) understanding search behaviors that predict learning during search; and (4) understanding how experimental tools can encourage and support learning during search. Our research in this paper belongs to this final category—tools to support learning during search.

SAL studies have explored how different tools and features can support learning, including note-taking tools [11, 28, 29], visualizations [8, 20, 30], goal-setting tools [34, 35], and self-assessment tools [32]. To our knowledge, however, no SAL study has explored the integration of GenAI-based chat and traditional web search to support learning during search. Here, there are several open questions. How does this integration impact perceptions, search behaviors, and learning outcomes? How do searchers engage with an embedded GenAI-based chat? Specifically, when do they engage, for what reasons, and what do they gain? Are there specific needs that prompt users to engage with one component versus the other? How do users evaluate the information provided by each component in terms of its quality (e.g., accuracy and credibility)?

GenAI tools are already being used by people (e.g., students) to support their learning [7, 18, 26]. Outside of SAL, studies have explored the impact of standalone GenAI tools on learning and have found mixed results. Some studies have found positive effects on learning [1, 23, 41] and others have found the opposite [4, 19]. In light of these findings, it is important to understand how learning is impacted by the integration of GenAI-based chat and traditional web search.

In this paper, we report on a study in which participants ($N = 40$) were asked to complete a learning-oriented search task. Participants were asked to learn about the biological concepts of diffusion and osmosis. The study employed a between-subjects design, and participants were assigned to one of two *system conditions* (20 participants per condition). In the SearchOnly condition, participants used a standard web search system. In the Search+Chat condition, participants used an experimental system that integrated a standard web search component and a GenAI-based chat component referred to as Chat AI. In both conditions, participants were given 40 minutes to gather information using the assigned system and take notes. To measure learning, participants completed a multiple-choice test called the Osmosis and Diffusion Conceptual Assessment (ODCA) [10] before the search task, immediately after, and one week later (to measure retention). The study investigated seven research questions. RQ1-RQ3 focus on differences between participant groups (i.e., SearchOnly versus Search+Chat). Additionally, to gain insights about the effects of integrating search and Chat AI, RQ4-RQ7 focus on participants' motivations, gains, and behaviors in the Search+Chat condition. Our seven research questions are as follows:

- **RQ1:** What were the effects of the system condition on participants' post-task perceptions?
- **RQ2:** What were the effects of the system condition on participants' search behaviors?
- **RQ3:** What were the effects of the system condition on participants' learning outcomes?
- **RQ4:** In the Search+Chat condition, what were participants' motivations for engaging with the Chat AI and what did they gain from engaging with the Chat AI?
- **RQ5:** In the Search+Chat condition, during which phases did participants decide to engage with the Chat AI and why?
- **RQ6:** In the Search+Chat condition, what types of queries did participants issue to the Chat AI component versus the web search component of the system?
- **RQ7:** In the Search+Chat condition, what were participants' perceptions of the information returned by the Chat AI component versus the web search component of the system?

Our results found several interesting trends. First, in terms of post-task perceptions (RQ1), we did not find significant differences between groups. Second, in terms of search behaviors (RQ2), participants in the Search+Chat condition spent more time on the search interface (versus reading documents) and engaged less with the search results in the web search component of the system. Third, in terms of learning outcomes (RQ3), participants in the Search+Chat condition showed greater improvements in their ODCA scores immediately after the search task. However, these improvements were less pronounced one week later. Finally, our close examination of participants' behaviors in the Search+Chat condition found that participants: (RQ4) engaged with the Chat AI to ask specific questions, to avoid searching, to get easy-to-understand information, and to save time and energy; (RQ5) used the Chat AI during all the five phases that we asked about; (RQ6) issued different types of queries to the Chat AI versus web search component; and (RQ7) had less trust in the information returned by the Chat AI versus web search component.

2 Related Work

2.1 Tools to Support Learning during Search

SAL studies have explored how different search tools can support learning during search, including: (1) note-taking tools, (2) visualizations, (3) goal-setting tools, and (4) self-assessment tools.

Note-taking Tools: Freund et al. [11] investigated the effects of different reading environments on reading comprehension. Participants had better learning outcomes when they read documents in plain text versus HTML, which included distracting elements (e.g., ads). However, participants had *similar* learning outcomes when provided with tools to highlight text and make "sticky notes". Roy et al. [29] investigated the effects of two tools on learning. One tool enabled participants to highlight text and see a summary of their highlights. A second tool enabled participants to take notes. Access to either tool improved learning. However, access to *both* tools did not improve learning, possibly due to cognitive overload. Qiu et al. [28] assigned participants to one of four conditions. One manipulation involved having participants use a text-based conversational search interface versus a traditional search interface. A second manipulation involved having participants take notes versus not take notes. Participants had the greatest knowledge gains when using the standard search interface and instructed to take notes.

Visualizations: Kammerer et al. [20] investigated the effects of a search system that enabled participants to filter results using social tags. Participants had better learning outcomes with the experimental system versus a baseline system without tags. Câmara et al. [8] developed a visualization that presented participants with their coverage of topics explored during the session. With the visualization, participants explored more topics but did so *superficially* and did not have better learning outcomes. Salimzadeh et al. [30] found that displaying entity cards on the SERP did not improve learning.

Goal-setting Tools: Urgo and Arguello [34] experimented with a tool called the Subgoal Manager (SM). The tool was designed to help searchers break apart a learning objective into specific subgoals and take notes with respect to each subgoal. The study had three conditions. In one condition, participants used the SM with prepopulated subgoals. In a second condition, participants used the SM and set their own subgoals. In a third condition, participants were not told anything about subgoals and used a simple text editor to take notes. Participants had slightly better learning outcomes when they used the SM and set their own subgoals. In a follow-up study [35], participants were assigned to one of two conditions. In one condition, participants used the SM and set their own subgoals.

In a second condition, participants were not told anything about subgoals and used a simple text editor to take notes. Participants had better learning outcomes in the SM condition, particularly in terms of retention. A qualitative analysis of search sessions revealed that participants in the SM condition engaged in more self-regulated learning processes like activating their prior knowledge and monitoring their progress.

Self-Assessment Tools: Syed et al. [32] experimented with a reading environment that dynamically prompted participants to answer questions about paragraphs read during the session, predicted using eye-tracking. The study included conditions that prompted participants to answer manually curated questions and automatically generated questions. Prompting participants to answer questions (manually curated or automatically generated) improved knowledge retention. However, this was only true for participants with low prior knowledge. Additionally, participants had better learning outcomes with automatically generated questions because they were more specific.

Our study extends this prior work by experimenting with a search interface that combined traditional web search with a generative AI tool.

2.2 Generative AI & Learning

Generative AI (GenAI) tools are already being used by students in the learning process [7, 18, 26]. However, little is known about the effects of GenAI tools on learning. In the learning sciences and education, researchers are *beginning* to explore these effects with a handful of studies. The results of these studies are complex and open questions remain. For instance, some studies have found that GenAI tools have positive effects on learning outcomes [1, 23, 41] and other studies have found the opposite [4, 19].

As an example of a study with positive effects, Yilmaz and Karaoglan Yilmaz [41] explored the effects of a GenAI tool on students' motivations, self-efficacy (i.e., confidence in completing a task), and computational thinking skills. Students completed weekly programming practices with or without ChatGPT. Students who used ChatGPT had higher levels of motivation, self-efficacy, and computation thinking skills. However, the study did not *objectively* measure learning.

On the other hand, as an example of a study with negative effects, Bastani et al. [4] had nearly 1,000 student participants complete math practice problems in one of three conditions: (1) ChatGPT, (2) GPT Tutor, and (3) baseline (no GenAI support). In the GPT Tutor condition, for each practice problem, ChatGPT was prompted with: (1) the correct answer to the problem; (2) common misconceptions about the problem; and (3) instructions to only provide hints and not the answer. After the practice problems, participants completed a closed-notes exam. Performance on the practice problems was highest in the GPT Tutor condition, followed by ChatGPT and baseline conditions. However, performance on the exam was similar in the GPT Tutor and baseline conditions and *lowest* in the ChatGPT condition. These results suggest that guardrails may be necessary for GenAI tools to support learning.

2.3 Generative AI & Search

Several studies have examined why users switch between traditional search engines and GenAI tools. Zhou and Li [44] found that users abandon traditional search systems when the retrieved information does not align with the task goals and due to information overload. Additionally, they found that users turn to GenAI tools because they provide well-structured, concise information and allow for interactive exchanges. However, these findings were based on online questionnaires rather than observations of real-time behavior.

Yen et al. [40] explored how programmers decide between traditional search systems and GenAI tools during problem-solving tasks. They found that programmers prefer traditional search when domain knowledge is low and when the task is amorphous (vs. well-defined). Additionally, web search is often used to validate the correctness of GenAI results. Again, these findings emerged from interviews rather than observations of real-time behavior.

Other work has investigated how systems can integrate GenAI-based features into a search environment to support the information-seeking process. Liu et al. [22] investigated the Selenite system, which uses LLMs to generate overviews and suggestions for decision-making in unfamiliar domains. Participants completed tasks faster and identified more relevant criteria with Selenite compared to traditional web search. However, without an objective measure of learning, what users internalized and learned remains unclear.

Park et al. [25] developed ChoiceMates, which allows users to engage with multiple LLM-powered agents offering diverse perspectives. Results found that participants explored more diverse viewpoints and made more confident decisions than with traditional web search. Additionally, participants noted that interacting with different agents helped with avoiding multiple searches.

Zheng et al. [43] created DiscipLink to support interdisciplinary information seeking. DiscipLink helps users explore literature across diverse fields by generating exploratory questions and organizing retrieved papers. Users reported that the system facilitated deeper exploration of unfamiliar topics and was particularly valuable in early-stage research. Participants also noted that the system helped them to generate ideas for future studies.

In prior work [6], we conducted a small study ($N = 10$) with a system that also integrated traditional web search and GenAI-based chat. The interface in our Search+Chat condition was inspired by this system. After completing tasks, participants reported using chat for concise answers but expressed uncertainty about trusting its responses, particularly when they had low prior knowledge. Our current study in this paper builds upon Capra and Arguello [6] by comparing against a baseline condition (SearchOnly), including more participants (enabling statistical comparisons), investigating the effects on learning and retention, and focusing on additional aspects—analyzing the phases during which participants engaged with the chat component and the types of queries issued to each component (search vs. chat).

3 Methods

To investigate RQ1-RQ7, we conducted a between-subjects remote study with 40 participants. Participants were assigned to one of two conditions: SearchOnly or Search+Chat (Section 3.2). The study was conducted over the Zoom videoconferencing platform. Participants were recruited through an opt-in mailing list of undergraduates at our university. The study was approved by our university's Institutional Review Board (IRB).

Our 40 participants were aged 18-46 and the median age was 20. Twenty-seven identified as female, 9 as male, and 4 as non-binary. In terms of highest level of biology course completed, 1 had completed a biology course at the 8th grade level or lower, 23 had completed a high school course, and 16 had completed an undergraduate course. Finally, participants in both conditions were asked about their use of GenAI tools such as ChatGPT. Two participants reported using GenAI tools multiple times a day, 12 multiple times a week, 12 multiple times a month, and 14 only a few times ever.

3.1 Study Protocol

First, after signing a consent form, participants completed a demographics questionnaire. Then, participants watched a video describing the study. Next, to capture their prior knowledge about diffusion and osmosis, participants completed a validated, multiple choice test called the Osmosis and Diffusion Conceptual Assessment (ODCA) (Section 3.5). After this, participants were shown the task description and were asked to read it aloud (Section 3.3). Then, participants completed a pre-task questionnaire about their perceptions of the task (Section 3.4). Next, participants watched a video introducing the system associated with their assigned condition. During the SearchOnly condition, the video demonstrated features of the web search component. During the Search+Chat condition, the video demonstrated features of the web search component and the Chat AI component. It also explained how the web search and Chat AI components were interconnected (Section 3.2). Participants then completed the main learning-oriented search task (Section 3.3). During the task, participants were asked to gather information and take notes. Participants were provided with a custom-built note-taking tool resembling a standard text editor. Participants were given 40 minutes to work on the main task and were notified by the moderator when they had 5 minutes remaining. Next, participants completed a post-task questionnaire about their experience during the main search task. Finally, to measure learning, participants completed the ODCA again (post-task). Participants were given a US$30 Amazon gift card for completing this phase of the study. To measure retention, one week after the study session, participants were emailed a link to complete the ODCA a third time. Participants were instructed to complete the ODCA within 48 hours. Participants were given a US$10 Amazon gift card for completing the retention ODCA, and all participants completed it.

3.2 System Conditions

Participants were assigned to one of two conditions: SearchOnly and Search+Chat. We start by describing the Chat AI-integrated search interface in the Search+Chat condition.

Figure 1 illustrates the interface in the Search+Chat condition. During the Search+Chat condition, participants used an experimental system that included a search component on the left and a Chat AI component on the right. The task description was displayed at the top of the interface (A) for reference. The web search component was implemented using the Bing Web Search API and enabled participants to issue queries (B) and get back web results (C). The interface displayed 10 results per page and included pagination controls at the bottom (not shown in Figure 1). The Chat AI component was implemented using the OpenAI GPT-3.5-Turbo API.

To encourage participants to engage with the Chat AI component, we connected the web search and Chat AI components in several ways. First, queries issued to the web search component were also issued to the Chat AI component. We prompted ChatGPT to produce a paragraph of background information about the query and also to return a brief list of important related concepts. This feature is illustrated in Figure 1. In response to the query "diffusion" issued to the web search component (B), the Chat AI component automatically displays a definition of "diffusion" (D) and a list of related concepts (E) (e.g., "passive transport", "osmosis", "concentration gradient" , etc.). Second, the related concepts displayed in the Chat AI output were designed to be clickable. Clicking a related concept resulted in: (1) the concept being issued as query to the web search component—producing a new set of web results—and (2) the concept being issued to the Chat AI component—producing a new paragraph of background information about the concept and a new set of clickable related concepts.

Participants could also interact with the Chat AI component by asking direct questions. For example, as shown in Figure 1, one might ask a follow-up question (F) about previous information returned by the Chat AI component (D & E).

As described in Section 3.1, participants were provided with a custom-built tool (not shown in Figure 1) to take notes while searching and learning. A "show notes" button (H) on the interface opened the note-taking tool in a new browser window. The note-taking tool saved changes automatically.

The interface in the SearchOnly condition looked exactly the same, but did not include the Chat AI component. As per Figure 1, it only included elements (A), (B), (C), and (H) as described above.

3.3 Search Task

Participants completed the following learning-oriented search task, which included a scenario to contextualize the task and an explicit learning objective.

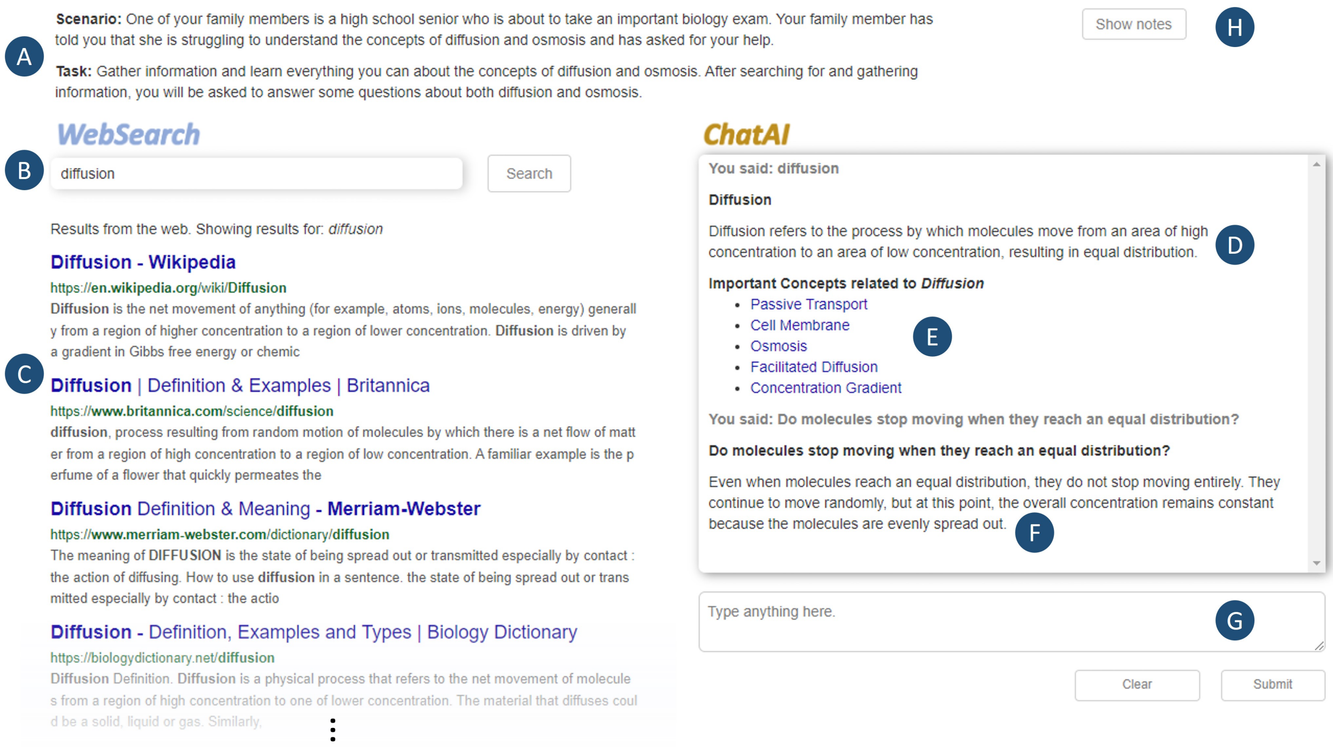
Scenario: One of your family members is a high school senior who is about to take an important biology exam. Your family member has told you that she is struggling to understand the concepts of diffusion and osmosis and has asked for your help.

Learning Objective: Your goal is to use this search system to learn everything you can about the concepts of diffusion and osmosis. After searching and gathering information, you will be asked to answer some questions about both diffusion and osmosis.

3.4 Pre- & Post-task Questionnaire

Participants completed a pre- and post-task questionnaire before and after the main search task. In both questionnaires, participants responded to agreement statements on a 7-point scale ranging from strongly disagree (1) to strongly agree (7). The full text of both questionnaires is available online.

Pre-task Questionnaire: Participants in both conditions completed the same pre-task questionnaire. Participants completed the pre-task questionnaire before knowing any details about the system they would use to complete the task. The pre-task questionnaire asked about: (1) interest in the task (1 item), (2) prior knowledge (3 items), (3) expected difficulty (4 items), and (4) *a priori* determinability (6 items). *A priori* determinability relates to the extent to which



Scenario: One of your family members is a high school senior who is about to take an important biology exam. Your family member has told you that she is struggling to understand the concepts of diffusion and osmosis and has asked for your help.

Task: Gather information and learn everything you can about the concepts of diffusion and osmosis. After searching for and gathering information, you will be asked to answer some questions about both diffusion and osmosis.

diffusion

Results from the web. Showing results for: *diffusion*

Diffusion - Wikipedia
https://en.wikipedia.org/wiki/Diffusion
Diffusion is the net movement of anything (for example, atoms, ions, molecules, energy) generally from a region of higher concentration to a region of lower concentration. **Diffusion** is driven by a gradient in Gibbs free energy or chemic

Diffusion | Definition & Examples | Britannica
https://www.britannica.com/science/diffusion
diffusion, process resulting from random motion of molecules by which there is a net flow of matter from a region of high concentration to a region of low concentration. A familiar example is the perfume of a flower that quickly permeates the

Diffusion Definition & Meaning - **Merriam-Webster**
https://www.merriam-webster.com/dictionary/diffusion
The meaning of **DIFFUSION** is the state of being spread out or transmitted especially by contact : the action of diffusing. How to use **diffusion** in a sentence. the state of being spread out or transmitted especially by contact : the actio

Diffusion - Definition, Examples and Types | Biology Dictionary
https://biologydictionary.net/diffusion
Diffusion Definition. **Diffusion** is a physical process that refers to the net movement of molecules from a region of high concentration to one of lower concentration. The material that diffuses could be a solid, liquid or gas. Similarly,

You said: diffusion

Diffusion

Diffusion refers to the process by which molecules move from an area of high concentration to an area of low concentration, resulting in equal distribution.

Important Concepts related to *Diffusion*

- Passive Transport
- Cell Membrane
- Osmosis
- Facilitated Diffusion
- Concentration Gradient

You said: Do molecules stop moving when they reach an equal distribution?

Do molecules stop moving when they reach an equal distribution?

Even when molecules reach an equal distribution, they do not stop moving entirely. They continue to move randomly, but at this point, the overall concentration remains constant because the molecules are evenly spread out.

Type anything here.

Figure 1: Search+Chat interface. (A) - learning-oriented search task; (B) - search field; (C) - search results; (D) - AI-generated response; (E) - query-related concepts; (F) - participant follow-up question; (G) - chat field; and (H) - button to open notes editor.

aspects of the task are known in advance (e.g., requirements, goals, strategies for completion, etc.) [5]. The groups of items for prior knowledge, expected difficulty, and *a priori* determinability had high internal consistency (Cronbach's $\alpha \geq .80$). Therefore, groups of responses were averaged to form three composite measures.

Post-task Questionnaire: The post-task questionnaire was organized into three parts. The first part asked about: (1) interest increase (1 item), (2) knowledge increase (3 items), and difficulty (4 items). The groups of items for knowledge increase and difficulty had high internal consistency (Cronbach's $\alpha \geq .79$). Therefore, groups of responses were averaged to form two composite measures.

The second part of the post-task questionnaire asked about the extent to which participants perceived to have engaged in different cognitive and metacognitive activities: (1) setting goals, (2) deciding how to begin the search, (3) connecting new information to existing knowledge, (4) relating topics, (5) comparing different explanations of similar ideas, (6) deciding when new information matched previously encountered information, (7) tracking progress toward their goals, (8) gauging their understanding of information, (9) judging whether information was useful, and (10) evaluating their approach to the task. These items were analyzed individually.

The third part of the post-task questionnaire was different depending on the system condition. In the SearchOnly condition, participants were asked whether the information returned by the web search system was perceived as: (1) factual, (2) trustworthy, (3) accurate, (4) up-to-date, (5) reliable, (6) credible, and (7) unbiased. In the Search+Chat condition, participants were asked these seven questions about the web search component and were also asked the same seven questions about the information returned by the Chat AI component of the system. The questionnaire included screenshots of the web search and Chat AI components to ensure that participants knew which component we were asking about.

Finally, participants in the Search+Chat condition were asked three open-ended questions about their engagement with the Chat AI component of the system. The first two questions asked about their motivations for engaging with the Chat AI and what they gained from the Chat AI. The third question asked participants if they engaged with the Chat AI during specific phases of the task and, if so, to provide examples. Participants were provided with the following phases/descriptions: (1) initiation—getting an initial understanding of the task; (2) planning—deciding how to approach the task; (3) pursuing—searching for specific information; (4) verifying—verifying the accuracy, completeness, or credibility of information already found; and (5) stalling—feeling stuck and unsure about what to do next.

3.5 Learning Assessment

To measure prior knowledge, learning, and retention, participants completed the multiple-choice Osmosis and Diffusion Conceptual Assessment (ODCA) [10] before the search task, immediately after the task, and one week later. The ODCA includes 18 questions about diffusion and osmosis. The questions are organized in pairs. Each pair contains a knowledge question and a reasoning question. The knowledge question is designed to assess the test taker's comprehension of specific concepts and processes. The reasoning question is designed to assess the test taker's justification for their answer to the knowledge question. In other words, the knowledge question focuses on "what?" and the reasoning question focuses on "why?".

The ODCA was used for two reasons. First, it was developed with the help of expert biology faculty and targets common misconceptions that students have about diffusion and osmosis [10]. Second, the ODCA is a valid and reliable instrument. ODCA items have been found to have high internal consistency across student cohorts [10]. The ODCA is also included in our online appendix.

To measuring learning, we used participants' pre- and post-task scores on the ODCA to compute *normalized gain:*

$$\text{Normalized Gain} = \frac{(\text{PostScore} - \text{PreScore})}{(1 - \text{PreScore})}, \qquad (1)$$

where PreScore and PostScore are the percentage of correct answers in the pre- and post-task ODCA, respectively. Similarly, to measure retention, we used the same normalization. That is, we used Equation 1 but replaced PostScore with RetScore—the percentage of correct answers in the retention ODCA. This type of normalization is common in education research [16] and search-as-learning studies [12, 39, 42]. Normalized gain accounts for participants' prior knowledge based on their pre-task scores. It answers the question: "Of the percentage a participant could have gained, what percentage did they actually gain?" On rare occasions, the normalized gain can be a negative value (i.e., PostScore < PreScore). This can happen, for example, if participants guess correctly on the pre-test and not the post-test.

3.6 Search Behaviors

RQ2 investigated the effects of the system condition on participants' search behaviors. To this end, we logged participants' interactions with both systems and computed the following measures:

(1) Number of web search queries.
(2) Number of abandoned web search queries.
(3) Number of web results clicked.
(4) Average rank of web results moused over.
(5) Time (seconds) to the first web result clicked in the session.
(6) Average time (seconds) between each web search query and the first web result clicked for the query (if any).
(7) Completion time.
(8) Time spent on web pages (minutes).
(9) Time spent on the search interface (minutes).
(10) Time spent taking notes (minutes).
(11) Length of notes taken (words).

The above measures were available in both system conditions. Several measures need additional clarification. In the Search+Chat condition, measure #1 excludes queries issued to the Chat AI component and related concepts clicked from the Chat AI output; measures #5 and #6 consider only web results clicked (not related concepts in the Chat AI output); and measure #9 considers the amount of time participants spent on the interface, which included both the web search and Chat AI components.

3.7 Analysis of Queries

To address RQ6, we analyzed all queries issued by participants in Search+Chat condition *directly* to either the web search component or Chat AI component of the system. These exclude queries associated with clicks on the related concepts displayed on the Chat AI output, which automatically triggered a web search and Chat AI query. In total, the 20 participants assigned to the Search+Chat condition issued 301 queries. Of these, 131 were issued to the web search component and 170 were issued to the Chat AI component. To understand the context of each query, each search session was represented as a sequence of timestamped events (i.e., queries and clicks).

Our analysis of queries in the Search+Chat condition proceeded as follows. First, all authors independently analyzed all queries from four participants. Each author developed their own coding scheme based on interesting phenomena observed. Next, all authors met and developed a coding guide with codes, definitions, and examples. Most of our codes (12 out of 14) were designed to be mutually exclusive and were associated with the participant's intent (e.g., the participant was looking for a *definition* or an *example* of a concept). Two of our codes were designed to be orthogonal to the participant's intent. In some cases, participants issued queries requesting information in a *specific format* (e.g., "bulleted list of notes about diffusion and osmosis"). Additionally, in some cases, participants issued queries that assumed the system's awareness of the *context* (e.g., "how does temperature impact *that*"). After developing our coding guide, two of the authors (referred to as A1 & A2) independently coded 100% of the data. Across our mutually exclusive codes (12/14), agreement was high (Cohen's $\kappa = 0.821$). Agreement varied for our two orthogonal codes. Agreement was high for coding queries that leveraged the search session history ($\kappa = 0.849$). Agreement was low for coding queries that requested information in a specific format ($\kappa = 0.215$). This low agreement was mostly due to this code being rare (it only occurred 8 times) and to A1 & A2 disagreeing about "test questions" being a specific format. Finally, all authors met to discuss and resolve all cases where A1 & A2 disagreed. Ultimately, we decided that "test questions" are a specific format. The following list describes our codes and provides examples.

- **Overview:** The query mentions a single concept, suggesting that the participant wants a general overview (e.g., "diffusion")
- **Definition:** The participant wants a definition of a concept by including terms like "definition" or "define" (e.g., "osmosis *definition*").
- **Explanation/Clarification:** The participant wants to understand the outcome of a process (e.g., "how does dye disperse in water") or wants to resolve a point of confusion (e.g., "is diffusion active or passive movement?").
- **Differentiate Concepts:** The participant wants to understand the similarities/differences between concepts (e.g., "difference between osmosis and diffusion", "solvent vs solute vs solution").
- **Cause and Effect:** The participant wants to understand a causal relation (e.g., "concentration gradient impact on diffusion").
- **Hypothesis Verification:** The participant wants to verify a specific hypothesis or proposition. The query could be answered with a simple yes or no (e.g., "does movement stop in diffusion").

- **Examples:** The participant requests examples of a concept or process (e.g., "what is an *example* of osmosis?").
- **Representation Retrieval:** The participant requests a specific representation, such as an graphic or video (e.g., "*diagram* of osmosis").
- **Simplification:** The participant wants a simplified version of specific information (e.g., "Can you give this to me in *layman's terms*?").
- **Generate Ideas:** The participant wants ideas about things to search for (e.g., "now that I know [...] *what else can I learn*?").
- **Verbatim Task:** The participant issues a query with text copied and pasted from the task description.
- **Test Knowledge:** The participant wants to test their knowledge (e.g., "sample *questions* on diffusion and osmosis").
- **Use of Context:** The participant issues a follow-up query that leverages the history of interaction with the system (e.g., "why does *this* happen?").
- **Specific Format:** The participant wants textual information in a specific format (e.g., "make a *list* of similarities and differences between diffusion and osmosis.")

3.8 Analysis of Open-ended Responses

In the Search+Chat condition, the post-task questionnaire asked participants three open-ended questions about their motivations for engaging with the Chat AI, gains obtained from the Chat AI, and the phases during which they engaged with the Chat AI.

To address RQ4, responses to the first two questions about motivations and gains were analyzed as follows. Participants responded similarly to both questions. That is, participants commented on engaging with the Chat AI because they wanted XYZ and commented on gaining XYZ. Therefore, we analyzed responses to both questions together. Three of the authors independently analyzed all responses and came up with individual sets of themes. Then, all authors met to develop a cohesive set of 15 total themes. However, several themes overlapped with: (1) our information-seeking phases for RQ5 (see Section 3.4) and (2) our query codes for RQ6 (see Section 3.7). To reduce redundancy, we limited our analysis to the five remaining themes that did not overlap with RQ5 and RQ6. These are described in Section 4.5. To generate counts (i.e., number of participants who commented on each theme), one of the authors assigned themes to responses.

Finally, to address RQ5, we manually segmented participants' responses based on the five phases that we *explicitly* asked about: (1) initiation, (2) planning, (3) pursuing, (4) verifying, and (5) stalling.

3.9 Statistical Analysis

RQ1-RQ3 focused on differences between conditions. Most of our dependent variables were not normally distributed. Therefore, we decided to use non-parametric Mann-Whitney U tests to check for statistically significant differences between groups. In addition to reporting p-values, we report U statistic values. Our U statistic values can be interpreted as the "number of pairwise wins" by participants in the Search+Chat condition versus participants in the SearchOnly condition. Given that we had 20 participants in each condition, our U statistic values are in the range 0 to 400 (20 × 20). Values closer to 200 (the midpoint) imply no significant differences between groups, and values closer to 0 or 400 imply the opposite. In RQ6, we investigate whether participants in the Search+Chat condition issued different types of queries to the web search versus Chat AI component of the system. Here, we used chi-squared tests to check for significant differences in counts. Finally, in RQ7, we investigate whether participants in the Search+Chat condition perceived differences in the quality of information returned by the web search versus Chat AI component of the system. Since these are paired comparisons, we used Wilcoxon signed rank tests to check for statistically significant differences. Given the exploratory nature of our study, all tests were two-tailed tests. RQ4 and RQ5 did not involve significance testing.

4 Results

In the following sections, we present our results for RQ1-RQ7. To conserve space, with one exception (learning retention), we only include figures for outcome measures with significant or marginally significant differences. All figures are box plots. In the text, we use M_{sc} and M_{so} to denote the median for the Search+Chat and SearchOnly condition, respectively.

4.1 Prior Knowledge & Pre-task Perceptions

Before presenting results for RQ1-RQ7, we report on differences in prior knowledge and pre-task perceptions between groups. Given the between-subjects nature of our study, we were curious about differences between participants assigned to each condition.

To measure their prior knowledge, participants completed the multiple-choice ODCA at the beginning of the study session. Prior knowledge was measured based on the percentage of ODCA questions answered correctly (out of 18). Differences in prior knowledge were not statistically significant ($M_{\text{sc}} = 0.50$, $M_{\text{so}} = 0.61$, $U = 152.5$, $p = .201$).

After reading the task description, participants completed a pre-task questionnaire about their perceptions of the task (i.e., interest, prior knowledge, expected difficulty, and *a priori* determinability). Here, we found two significant differences. Participants in the Search+Chat condition reported having less prior knowledge about diffusion and osmosis ($M_{\text{sc}} = 2.00$, $M_{\text{so}} = 2.50$, $U = 122$, $p = .034$) and expected the task to be more difficult ($M_{\text{sc}} = 3.75$, $M_{\text{so}} = 3.13$, $U = 281.5$, $p = .028$). Participants completed the pre-task questionnaire before receiving any information about the system they would use to complete the task. Thus, these differences are due to the random assignment of participants to conditions.

To summarize, *objectively* speaking, participants in both conditions had similar levels of prior knowledge based on their performance on the pre-task ODCA. Due to random chance, participants in the Search+Chat condition *perceived* to know less about the subject of the task and expected the task to be more difficult. However, we do not believe that these differences impacted our RQ1-RQ3 results, which focused on differences between groups. First, while significant, these differences in perceptions are small (i.e., differences in median values are about half a point on a 7-point scale). Second, as discussed next in Section 4.2, participants reported similar levels of knowledge increase and experienced difficulty after completing the task.

4.2 RQ1: Post-task Perceptions

After the task and before the post-task ODCA, participants completed a post-task questionnaire. Here, we focus on questions that were common to *both* conditions. These included questions about participants' perceptions of: (1) interest increase, (2) knowledge increase, (3) difficulty, (4) engagement in specific cognitive and metacognitive activities, and (5) the quality of information returned by the web search component. We did not find significant differences for any of these perceptions.

4.3 RQ2: Search Behaviors

To address RQ2, we compared the search behaviors described in Section 3.6 between system conditions. Six differences between groups were statistically significant and are shown in Figures 2a-2f. Participants in the SEARCH+CHAT condition: (a) clicked on fewer web search results ($M_{sc} = 5$, $M_{so} = 9.5$, $U = 98$, $p = .006$); (b) took longer to click on the first web search result in the session ($M_{sc} = 18.72$, $M_{so} = 7.26$, $U = 331$, $p = .003$); (c) took longer to click on the first web search result after each web search query ($M_{sc} = 20.23$, $M_{so} = 7.93$, $U = 379$, $p < .001$); (d) had more "shallow" mouseovers ($M_{sc} = 1.92$, $M_{so} = 2.87$, $U = 137$, $p = .042$); (e) spent more time on the SERP ($M_{sc} = 10.1$, $M_{so} = 3.96$, $U = 345$, $p < .001$); and (f) spent less time viewing pages ($M_{sc} = 13.18$, $M_{so} = 21.35$, $U = 117$, $p = .026$).

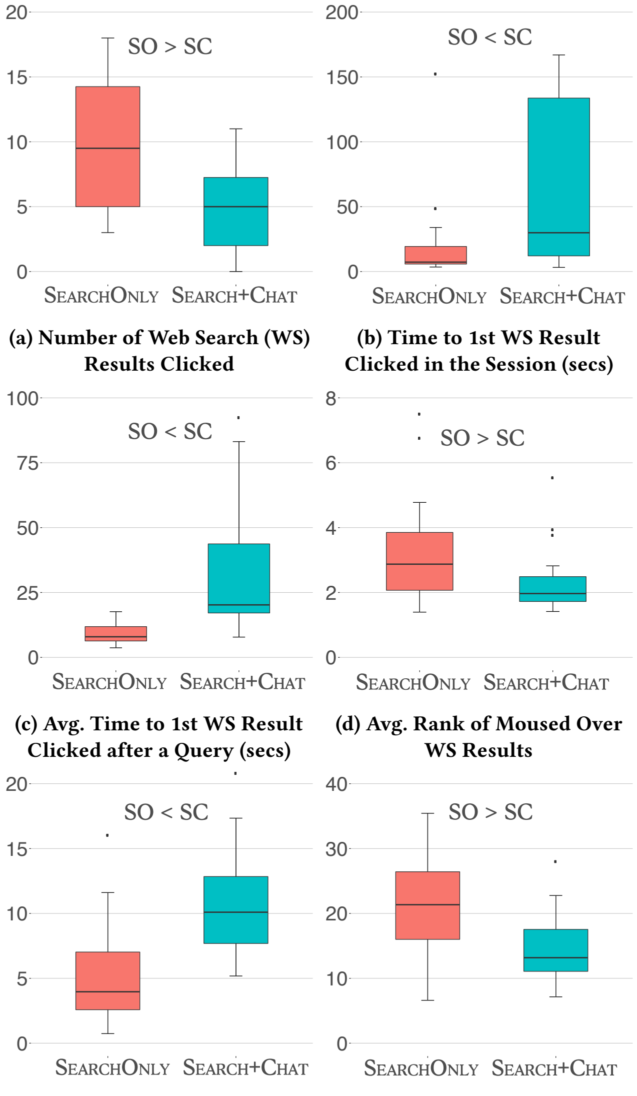
The results above suggest that participants in the SEARCH+CHAT condition engaged less with the results returned by the web search component, but spent more time on the SERP. As might be expected, this is because they engaged with the Chat AI component of the system. Participants issued an average of 8.5 ($S.D. = 6.8$) queries directly to the Chat AI component and clicked on an average of 2.15 ($S.D. = 2.23$) related concepts from the Chat AI output. Out of the 20 SEARCH+CHAT participants, 19/20 issued at least one chat query, 13/20 clicked at least one related concept, and 20/20 interacted with the chat in some way. Two participants did not click on a single web search result and engaged exclusively with the Chat AI component.

4.4 RQ3: Learning Outcomes

To address RQ3, we analyzed participants' performance on the ODCA taken immediately after the search task and one week later (to measure retention). Learning and retention was measured using *normalized gain* (Equation 1), which accounts for prior knowledge based on participants' scores on the pre-task ODCA.

Figure 3 show differences in normalized gains immediately after the search task and one week later. In both conditions, normalized gains were largely positive (i.e., normalized gain > 0). Participants were able to improve upon their pre-task ODCA scores both immediately after the search task and one week later. Thus, both participant groups were able to learn and retain some of their new knowledge.

Comparisons between conditions show an interesting effect. Immediately after the search task, participants in the SEARCH+CHAT condition had higher normalized gains than participants in the SEARCHONLY condition (Figure 3a). This difference is visually salient in Figure 3a and was found to be marginally significant ($M_{sc} = 0.47$, $M_{so} = 0.26$, $U = 268$, $p = .067$). This suggests that participants learned more in the SEARCH+CHAT condition. However, as shown in Figure 3b, the differences in normalized gains one week after the search task were less pronounced and not statistically significant ($M_{sc} = 0.38$, $M_{so} = 0.17$, $U = 247$, $p = .208$).



(a) Number of Web Search (WS) Results Clicked

(b) Time to 1st WS Result Clicked in the Session (secs)

(c) Avg. Time to 1st WS Result Clicked after a Query (secs)

(d) Avg. Rank of Moused Over WS Results

(e) Time on Interface (mins)

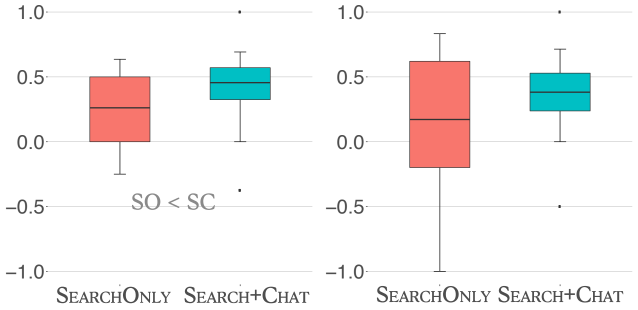
(f) Time on Pages (mins)

Figure 2: Effects of the SEARCH+CHAT condition on search behaviors

The above results suggest that participants in the SEARCH+CHAT condition were able to learn more (marginally significant). However, both groups retained similar levels of new knowledge a week after the session.

4.5 RQ4: Chat AI Motivations and Gains

As described in Section 3.4, the first two open-ended questions in the post-task questionnaire asked participants in the SEARCH+CHAT condition about their motivations for engaging with the Chat



(a) Post-task Normalized Gain **(b) Retention Normalized Gain**

Figure 3: Effects of the Search+Chat condition on Learning Outcomes

AI component and what they gained from their engagement. Responses to these two questions were very similar. Therefore, we decided to analyze them together (i.e., motivations and gains). As described in Section 3.8, we identified five themes that were unique to RQ4 and did not overlap with RQ5 and RQ6. Below, we describe these themes. The numbers in parentheses indicate the number of participants (out of 20) who mentioned each theme.

Answers to Specific Questions (4): Participants mentioned being able to get answers to specific questions. For example, P26 noted: “I was able to get answers that exactly catered to what I was looking for.” Similarly, P28 noted: “I was trying to ask more specific questions.”

Avoid Searching (3): Related to the previous theme, participants mentioned engaging with the Chat AI to avoid scanning through search results and/or searching for specific information on a webpage. For example, P34 noted: “It automatically presented me with the information I was searching for.” Similarly, P40 noted: “[It helped me] avoid reading long articles that AI could easily summarize.”

Connecting to Prior Knowledge (3): Participants mentioned being able to learn about new topics related to those they already knew about or had already learned about. For example, P32 mentioned: “[I was able to] enhance my knowledge further such as connecting known topics with new topics.” Similarly, P16 noted: “I gained knowledge of topics related to those I had already asked about.”

Easy to Understand (8): Participants mentioned being able to get information in simple terms. For example, P30 mentioned: “[The] information [...] was easier to digest.” Similarly, P18 noted: “[It was able] to explain the concepts in more layman’s terms.”

Save Time & Energy (11): Finally, participants mentioned that the Chat AI returned concise answers that saved them time and energy. For example, P2 noted: “I mainly used this function to get concisely presented information.” Similarly, P20 mentioned: “[It provided] short responses.” Finally, one participant (P28) noted that the Chat AI enabled them to get information without having to switch between websites: “[...] instead of having to change to a different link/website).”

4.6 RQ5: Phases of Chat AI Use

As described in Section 3.4, the last open-ended question in the post-task questionnaire asked participants in the Search+Chat condition about the phases during which they decided to engage with the Chat AI component. Participants were provided with the following options: (1) initiation, (2) planning, (3) pursuing, (4) verifying, and (5) stalling. All participants reported engaging with the Chat AI during at least one phase. The numbers in parentheses indicate the number of participants (out of 20) who reported engaging with the Chat AI during each phase.

Initiation (13): Participants engaged with the Chat AI to get a basic understanding of the task topic. For example, P2 stated: “I used it first during [the] initiation phase to get a brief overview and definition of the terms and their relationship to each other.” Similarly, P34 noted: “I used the Chat AI to ask opening questions and to get a basic understanding of the topics.”

Planning (9): Participants engaged with the Chat AI to decide how they should learn about the topic. In some cases, participants asked the Chat AI explicitly for advice. For example, P8 stated: “I used [it] to ask examples for how I should learn and it told me to look at practice questions and pictures.” Similarly, P30 stated: “I planned by asking what further topics I needed to know about osmosis and diffusion.” In other cases, the related concepts returned by the Chat AI (see Figure 1E) helped participants plan their future searches. For example, P18 noted: “The system returned additional topics (semipermeable membrane) that helped me understand the overall topic and gave me a path to pursue.” Similarly, P38 noted: “[It] told me related topics, so I knew what to research next.”

Pursuing (15): Participants engaged with the Chat AI to ask specific questions that arose during the learning process. Participants also asked clarifying questions. For example, P34 noted: “I asked what molecules move through diffusion and then learned it was many different types.” Participants also asked for examples to deepen their understanding. To illustrate, P22 noted: “[I asked it for] examples of diffusion and osmosis.” Finally, participants asked about new concepts encountered. For example, P14 noted: ‘It also helped me with pursuing certain definitions like concentration gradient.”

Verifying (9): Participants engaged with the Chat AI to corroborate information found online. For example, P18 noted: “[I used it for] fact-checking what osmosis and diffusion were from Britannica.” Similarly, P30 noted: “[I] asked the AI to verify my findings.” Participants also engaged with the Chat AI to assess their own understanding of material. For example, P34 noted: “ [I used it to] confirm my thoughts on the topics.”

Stalling (8): Finally, participants engaged with the Chat AI when they were unsure about what to do next. As with planning, some participants explicitly asked for ideas. For example, P8 noted: “I asked it what I should learn about next.” Other participants used the related concepts to become unstuck. For example, P16 noted: “[I used it] when I didn’t know what else to search for and would scroll through the terms it suggested I follow up on.” Similarly, P26 noted: “I did not know what to search for and the suggested hyperlinks helped me keep going.”

4.7 RQ6: Query Characteristics

In RQ6, we investigate whether participants in the Search+Chat condition issued different types of queries to the Chat AI component versus the web search component of the system. As described in Section 3.7, our qualitative analysis of queries resulted in 14 different codes. Of these, 12 are associated with the participants' intent and are mutually exclusive. The remaining two codes (i.e., use of context and specific format) are orthogonal. Table 1 shows the number of queries associated with each code that were issued to the Chat AI versus web search component. We used chi-squared tests to test whether queries associated with each code were more frequently issued to one component versus the other. The last row shows the total number of queries issued to each component. As shown, participants were significantly more likely to issue queries to the Chat AI versus web search component.

Several codes were significantly more frequent for queries issued to the web search component. Participants were significantly more likely to query the web search component for *overviews* (e.g., "diffusion") and *definitions* (e.g., "define thermal energy"). The Chat AI component was only designed to produce text. Therefore, participants were significantly more likely to query the web search component to retrieve specific *representations* such as images and videos (e.g., "videos about how osmosis works"). Finally, participants were significantly more likely to query the web search component for material to *test their knowledge* (e.g., "biology diffusion osmosis practice questions"). However, it is interesting that two queries issued to the Chat AI by different participants had this intent (e.g., "give me osmosis examples for me to solve").

Other codes were significantly more frequent for queries issued to the Chat AI component. First, participants were significantly more likely to query the Chat AI for *explanations* about specific processes and concepts. Many of these questions were clearly motivated by information encountered during the search session (e.g., "how does water move with osmosis"). Second, participants were significantly more likely to query the Chat AI when they had a specific *hypothesis* to test. These were queries that could be hypothetically answered with a yes or no response (e.g., "does osmosis involve only the movement of water molecules?"). Third, participants were significantly more likely to query the Chat AI for *examples* of specific concepts (e.g., "can you give an example of diffusion?"). Fourth, participants were significantly more likely to query the Chat AI when they wanted information in *simple terms* (e.g., "explain osmosis in elementary") and when they wanted to *generate ideas* about things to search for next (e.g., "Now that I know what diffusion and osmosis are, what else can I learn?"). Finally, participants were significantly more likely to query the Chat AI in ways that assumed the system's awareness of previous interactions (*use of context*). The conversational nature of the Chat AI component was leveraged by participants to ask follow-up questions (e.g., "In diffusion, it's different. Right?"). Interestingly, however, it was also used by participants to request new information (e.g., "can you teach me *more* about osmosis and diffusion") and to request modifications to previously returned information (e.g., "now explain everything you just told [me] *in detail*").

Table 1: Analysis of Queries issued to Search & Chat AI

	Search	Chat AI	χ^2	p-value
overviews	**39**	**22**	**4.74**	**.030***
definitions	**13**	**2**	**8.07**	**.005***
explanation/clarification	**17**	**54**	**19.28**	**.000***
differentiate concepts	19	27	1.39	.238
cause and effect	5	12	2.88	.090
hypothesis verification	**10**	**21**	**3.90**	**.048***
examples	**7**	**17**	**4.17**	**.041***
representation retrieval	**12**	**2**	**7.14**	**.008***
simplification	**0**	**4**	**4.00**	**.046***
generate ideas	**0**	**6**	**6.00**	**.014***
verbatim task	0	1	1.00	.317
test knowledge	**9**	**2**	**4.46**	**.035***
use of context	**0**	**20**	**20.00**	**.000***
specific format	5	3	0.50	.480
total	**131**	**170**	**5.05**	**.025***

4.8 RQ7: Quality of Information

Participants in the Search+Chat condition were asked about their perceptions of the information returned by the web search component and the Chat AI component. Participants were asked about the information being: (1) factual, (2) trustworthy, (3) accurate, (4) up-to-date, (5) reliable, (6) credible, and (7) unbiased. The same set of questions were asked about each component. Because these are *paired* comparisons, we tested for statistically significant differences using Wilcoxon signed rank tests.

As shown in Figure 4a, participants perceived the information returned by the Chat AI component as being significantly *less* trustworthy ($M_{\text{chat}} = 5$, $M_{\text{search}} = 6.5$, $V = 116$, $p = .010$). A similar trend was found for three other perceptions. However, those differences were only marginally significant. Participants perceived the information returned by the Chat AI component as being *less* accurate ($M_{\text{chat}} = 6$, $M_{\text{search}} = 6$, $V = 45.5$, $p = .068$); *less* credible ($M_{\text{chat}} = 5.5$, $M_{\text{search}} = 6$, $V = 81$, $p = .072$); and *less* reliable ($M_{\text{chat}} = 5$, $M_{\text{search}} = 6$, $V = 93$, $p = .059$). Interestingly, we did not find any significant differences in participants' perceptions about the information being factual, up-to-date, or unbiased. This may be due to the topic of the task. Diffusion and osmosis involve factual knowledge that is objectively true and has not changed in recent years. We might have observed a different trend with a heavily debated topic, involving differences of opinions and perspectives.

5 Discussion

In this section, we summarize our results, compare them to results from prior work, discuss their implications, and suggest opportunities for future research.

RQ1: Post-task Perceptions: We did not find significant differences in post-task perceptions. After the task, participants in both conditions reported similar levels of interest increase, knowledge increase, difficulty, and engagement in specific cognitive and metacognitive activities. There are two possible reasons for why we did not observe significant differences in post-task perceptions.

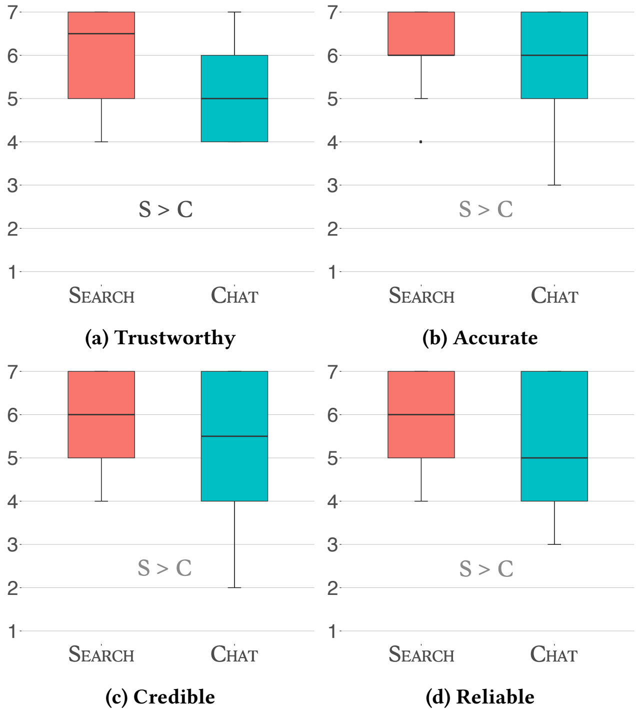


Figure 4: Perceptions about the information returned by the search and Chat AI components in the Search+Chat condition.

First, our study employed a between-subjects design (i.e., participants were assigned to one of two conditions). Therefore, when responding to our post-task questionnaire, participants did have a chance to compare their experiences between conditions. We chose a between-subjects design for practical reasons. Having participants complete two 40-minute, learning-oriented search tasks (along with questionnaires and knowledge assessments) would have taken too much time.

Second, specific to perceptions of knowledge increase, participants in Search+Chat condition had higher normalized gains immediately after the search task (marginally significant) but did not report higher levels of knowledge increase. Results from prior studies suggest that perceptions of learning do not always align with *actual* learning [27, 33].

RQ2: Search Behaviors: In terms of search behaviors, participants in the Search+Chat condition had less interaction with the web search results, spent less time reading pages, and spent more time on the search interface. This is because they had high levels of engagement with the Chat AI component. To illustrate, they issued an average of 8.5 queries directly to the Chat AI. These results show that when the Chat AI was included, participants changed their traditional search behaviors, reducing their interaction with web search results and engaging more with the Chat AI.

RQ3: Learning Outcomes: In terms of learning outcomes, we found mixed results. Participants in the Search+Chat condition demonstrated higher levels of learning immediately after the task (marginally significant). The same trend was found one week later. However, the differences were less pronounced and not significant. Our RQ3 results have several important implications.

First, our results underscore the importance of administering a retention assessment. Knowledge gains measured immediately after the search task may not align with knowledge gains measured one week later. Several SAL studies have found that measuring retention can provide key insights [28, 31, 35]. In some cases, experimental systems can have a *stronger* positive effect on knowledge gains measured one week after the search task versus immediately after [35]. In our study, we found the opposite—integrating GenAI and web search had a *stronger* positive effect on knowledge gains measured immediately after the search task versus one week after. Additional work is needed to understand factors that have positive and negative effects on retention after a learning-oriented search task.

Second, our results suggest that integrating GenAI and traditional search tools has the *potential* to improve learning. However, additional work is needed. As mentioned in Section 2.2, studies of standalone GenAI tools have also found mixed effects on learning [1, 4, 19, 23, 41]. There are several directions to explore. First, GenAI tools could be prompted to interact more like an educator in the domain of the task. As an example, Bastani et al. [4] had students complete math practice problems with the help of ChatGPT. In one condition, ChatGPT was prompted with common misconceptions about each problem and instructions to provide hints and not the answer. Second, additional research is needed to understand behaviors that should be encouraged and discouraged when learners interact with GenAI tools. Finally, prior work has found that effectively engaging in self-regulated learning (SRL) processes has positive effects on learning [2, 3, 14]. SRL is an active, reflective process in which a learner monitors and controls their own learning [13, 38]. SRL processes include goal-setting, monitoring progress, and adjusting learning strategies as needed. Later in this section, we describe how GenAI tools may encourage and support effective SRL.

RQ4: Motivations & Gains: In RQ4, we sought to understand what motivated participants to engage with the Chat AI and what they gained from it. Some themes overlapped with the phases during which participants engaged with the Chat AI (RQ5) and the types of queries they issued to the Chat AI (RQ6). However, we also identified motivations/gains unique to RQ4. Participants were motivated to engage with the Chat AI: to ask specific questions, to avoid searching, to get easy-to-understand information, and to save time and energy.

Our RQ4 results have similarities and differences to prior work. A prior study also found that participants interacted with a GenAI tool to get answers to specific questions, to get information in simple terms, and to get synthesized information [6]. Zhou and Li [44] found that people favor GenAI tools (vs. traditional search tools) when they have a highly specific objective and to avoid information overload.

Capra and Arguello [6] observed two trends that we did not observe in our study. First, some participants commented on favoring a GenAI tool when they were running out of time. We did not observe this motivation, possibly because participants had 40 minutes to complete the task. Second, Capra and Arguello [6] observed that several participants (out of 10) did not engage with the GenAI tool at all. In our case, *all* participants engaged with the Chat AI in some way or another. Capra and Arguello [6] conducted their study shortly after the first release of mainstream GenAI tools. It may be that people are "warming up" to GenAI tools, possibly because they are improving.

RQ5: Phases of Use: In RQ5, we sought to understand the phases during which participants decided to engage with the Chat AI. Participants engaged with the Chat AI during *all* phases that we asked about: (1) initiation (e.g., to get a basic overview); (2) planning (e.g., to plan what topics to investigate); (3) pursuing (e.g., to clarify concepts and get examples); (4) verifying (e.g. to verify information found through web search); and (5) when they were stalled (e.g., to get ideas about what to look for next).

To contrast our RQ5 results with prior work, Huurdeman et al. [17] investigated the effects of the task phase on participants' use of different search assistance tools. Results found that different tools were used at different phases. Conversely, in our study, participants commented on using the Chat AI component during all phases. Our RQ5 results suggest that the Chat AI was a versatile tool that participants could use to achieve different objectives. Our RQ5 results on phases of use are also corroborated by our RQ6 results on query types issued to the Chat AI. For example, participants issued the following types of queries to Chat AI that align with phases of use: (1) queries to get overviews and definitions (aligns with initiation & planning); (2) to get examples and clarifications (e.g., aligns with pursuing); (3) to verify a hypothesis (aligns with verifying); and (4) to generate ideas (aligns with stalled).

RQ6: Query Types: In RQ6, we sought to understand the types of queries issued by participants to the Chat AI versus web search component. Participants were significantly more likely to query the Chat AI for explanations, examples, simplifications, hypothesis verification, and to generate ideas. Conversely, Participants were significantly more likely to query the web search component for overviews, definitions, specific representations (e.g., videos), and to find questions to test their knowledge. To contrast with prior work, Yen et al. [40] found that computer programmers prefer to use web search (versus GenAI tools) when they have low domain knowledge and when the task is not well-defined. While our task was different from programming tasks, our participants also preferred to query the web search component for overviews and definitions (i.e., when their domain knowledge was low) and preferred to query the Chat AI for answers to specific, well-defined questions (e.g., "does osmosis involve only the movement of water molecules").

Our RQ6 results suggest several interesting trends. First, participants considered the strengths and weaknesses of each component when deciding which to use. GenAI tools are known for their ability to provide ideas, explanations, clarifications, simplifications, and examples. Thus, participants queried the Chat AI for those purposes. Similarly, people know that there are webpages that provide overviews and definitions of educational concepts like diffusion and osmosis. Therefore, participants queried the web search for overviews and definitions. Second, our RQ6 results suggest that *both* components played a role in participants' information-seeking behaviors. Participants moved fluidly between using the web search and Chat AI. They also took advantage of their integration. When they issued a web search query, they also engaged with the Chat AI output, which included a background summary and clickable related concepts. When they clicked on a related concept in the Chat AI output, they sometimes clicked on a new web result returned for that topic. Finally, many participants took advantage of the Chat AI's ability to use the previous context of the chat to streamline their interactions (e.g., "tell me more about *that*").

RQ7: Quality of Information: Finally, in RQ7, we sought to understand participants' perceptions of the quality of information returned by each component. Participants viewed the information returned by the Chat AI as significantly less trustworthy, and (marginally significantly) less accurate, less credible, and less reliable. GenAI tools are known for their tendency to hallucinate. Our results are consistent with prior work, which found that people mistrust the output of GenAI tools, especially when they do not link to sources [6]. Interestingly, we did not find differences in perceptions about the information from each component being factual, up-to-date, and unbiased. This is likely because of the task topic. Knowledge about diffusion and osmosis is largely factual, static, and agreed-upon.

Opportunities for Future Work: Our analyses revealed several trends that suggest opportunities for future work. First, participants preferred web search to find overviews, definitions, and specific representations of information. On the other hand, the Chat AI was preferred for generating explanations, simplifying information, verifying hypotheses, and generating ideas. These findings suggest that: (1) each tool was used for complementary roles, and that (2) additional *integration* of GenAI and web search is an area to explore. Second, our results did not find significant differences between conditions in terms of learning *retention*. This may be because the Search+Chat system was not explicitly designed to support critical SRL processes known to promote deep learning, such as planning, goal-setting, monitoring, and judging understanding [2, 3, 14]. In future work, the Chat AI component could be enhanced to promote these types of SRL processes. Below, we present several options to explore.

First, one potential integration is to offer tools that allow users to query the Chat AI directly from a page in strategic ways. For example, users could highlight text on a page and choose options like simplify, clarify, or generate ideas—functions our participants found valuable with the Chat AI. This could allow users to leverage the Chat AI's strengths without interrupting their learning process, making it easy to seek clarifications or further insights as they engage with web content. For clarifications or simplifications, this functionality could be paired with SRL *Judgment of Learning* [21, 37] prompts that could encourage users to reflect on whether their understanding improved after the clarification or simplification. Such reflections are critical metacognitive monitoring processes [2, 3, 14].

Second, the Chat AI could dynamically highlight and elaborate on passages within search results that connect to what users have already learned. The Chat AI could present context-sensitive prompts like, "Remember we discussed concentration gradients in osmosis? This concept is applied differently here in diffusion." This could stimulate SRL *Prior Knowledge Activation* [15, 35]. Such prompts could guide learners to draw connections between new material and what they have previously learned.

Third, the system could suggest logical next steps in learning, guiding users from one topic to the next. For example, a user exploring tonicity might be guided to investigate how osmosis supports cellular function as a next step in their learning process. Suggesting logical next steps could be enhanced further to support SRL *Planning & Goal-Setting* [9, 24, 34]. Prompts could encourage users not only to follow suggested next steps but also to take an active role in shaping their learning pathway by setting specific subgoals with measurable success criteria.

6 Conclusion

We conducted a between-subjects study with a SearchOnly and a Search+Chat system to: (1) compare participants' perceptions, search behaviors, and learning outcomes between the two conditions; and (2) gain insights about participants' use of the Chat AI component in the Search+Chat condition. Our results found that participants in the Search+Chat engaged less with the web search results because they engaged with the Chat AI. Participants in the Search+Chat condition had higher levels of learning immediately after the task, but this effect dissipated in a retention test one week later. Participants in the Search+Chat condition used the Chat AI tool to ask specific questions, to avoid searching, to get easy-to-understand information, and to save time. They also used it across different phases of the search process. Compared to web search component, the Chat AI component was used more for explanations, examples, simplifications, verification, and to generate ideas. Finally, participants viewed the Chat AI information as less trustworthy, accurate, credible, and reliable than the web search results. Our results have implications for the integration of GenAI tools into search systems and for the design of GenAI-based tools to support learning during search.

Acknowledgments

This work was supported by NSF grant IIS-2106334. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsor.

References

[1] Yazid Albadarin, Mohammed Saqr, Nicolas Pope, and Markku Tukiainen. 2024. A systematic literature review of empirical research on ChatGPT in education. *Discover Education* 3, 1 (May 2024), 60. https://doi.org/10.1007/s44217-024-00138-2

[2] Roger Azevedo, Daniel C. Moos, Jeffrey A. Greene, Fielding I. Winters, and Jennifer G. Cromley. 2008. Why Is Externally-Facilitated Regulated Learning More Effective than Self-Regulated Learning with Hypermedia? *Educational Technology Research and Development* 56, 1 (2008), 45–72. http://www.jstor.org/stable/25619907 Publisher: Springer.

[3] Maria Bannert and Christoph Mengelkamp. 2013. Scaffolding Hypermedia Learning Through Metacognitive Prompts. In *International Handbook of Metacognition and Learning Technologies*, Roger Azevedo and Vincent Aleven (Eds.). Springer, New York, NY, 171–186. https://doi.org/10.1007/978-1-4419-5546-3_12

[4] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning. https://doi.org/10.2139/ssrn.4895486

[5] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (March 1995), 191–213. https://doi.org/10.1016/0306-4573(95)80035-R

[6] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? arXiv:2307.03826 [cs.HC] https://arxiv.org/abs/2307.03826

[7] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (July 2023), 43. https://doi.org/10.1186/s41239-023-00411-8

[8] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2021. Searching to Learn with Instructional Scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 209–218. https://doi.org/10.1145/3406522.3446012

[9] Billie Eilam and Irit Aharon. 2003. Students' planning in the process of self-regulated learning. *Contemporary Educational Psychology* 28, 3 (July 2003), 304–334. https://doi.org/10.1016/S0361-476X(02)00042-5

[10] Kathleen M. Fisher, Kathy S. Williams, and Jennifer Evarts Lineback. 2011. Osmosis and Diffusion Conceptual Assessment. *CBE—Life Sciences Education* 10, 4 (Dec. 2011), 418–429. https://doi.org/10.1187/cbe.11-04-0038 Publisher: American Society for Cell Biology (lse).

[11] Luanne Freund, Rick Kopak, and Heather O'Brien. 2016. The effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science* 42, 1 (Feb. 2016), 79–93. https://doi.org/10.1177/0165551515614472 Publisher: SAGE Publications Ltd.

[12] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 2–11. https://doi.org/10.1145/3176349.3176381

[13] Jeffrey Alan Greene and Roger Azevedo. 2007. A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions. *Review of Educational Research* 77, 3 (Sept. 2007), 334–372. https://doi.org/10.3102/003465430303953 Publisher: American Educational Research Association.

[14] Jeffrey Alan Greene and Roger Azevedo. 2009. A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology* 34, 1 (Jan. 2009), 18–29. https://doi.org/10.1016/j.cedpsych.2008.05.006

[15] Jeffrey Alan Greene, Cheryl Mason Bolick, William P. Jackson, Alfred M. Caprino, Christopher Oswald, and Megan McVea. 2015. Domain-specificity of self-regulated learning processing in science and history. *Contemporary Educational Psychology* 42 (July 2015), 111–128. https://doi.org/10.1016/j.cedpsych.2015.06.001

[16] R. Hake. 2002. Relationship of Individual Student Normalized Learning Gains in Mechanics with Gender , High-School Physics , and Pretest Scores on Mathematics and Spatial Visualization, Vol. 8. 1–14. https://www.semanticscholar.org/paper/Relationship-of-Individual-Student-Normalized-Gains-Hake/ab557de0fdafe5def057a795c25264e74ac0e332

[17] Hugo C. Huurdeman, Max L. Wilson, and Jaap Kamps. 2016. Active and Passive Utility of Search Interface Features in Different Information Seeking Task Stages. Association for Computing Machinery, 3–12.

[18] Heather Johnston, Rebecca F. Wells, Elizabeth M. Shanks, Timothy Boey, and Bryony N. Parsons. 2024. Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity* 20, 1 (Dec. 2024), 1–21. https://doi.org/10.1007/s40979-024-00149-4 Number: 1 Publisher: BioMed Central.

[19] Qirui Ju. 2023. Experimental Evidence on Negative Impact of Generative AI on Scientific Learning Outcomes. https://doi.org/10.48550/arXiv.2311.05629 arXiv:2311.05629 [cs].

[20] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H. Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, Boston, MA, USA, 625–634. https://doi.org/10.1145/1518701.1518797

[21] Asher Koriat. 1997. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General* 126, 4 (1997), 349–370. https://doi.org/10.1037/0096-3445.126.4.349 Place: US Publisher: American Psychological Association.

[22] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–26. https://doi.org/10.1145/3613904.3642149

[23] Duong Thi Thuy Mai, Can Van Da, and Nguyen Van Hanh. 2024. The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach. *Frontiers in Education* 9 (Feb. 2024). https://doi.org/10.3389/feduc.2024.1328769 Publisher: Frontiers.

[24] Daniel C. Moos and Roger Azevedo. 2008. Monitoring, planning, and self-efficacy during learning with hypermedia: The impact of conceptual scaffolds. *Computers in Human Behavior* 24, 4 (July 2008), 1686–1706. https://doi.org/10.1016/j.chb.2007.07.001

[25] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. ChoiceMates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions. http://arxiv.org/abs/2310.01331 arXiv:2310.01331 [cs].

[26] Iris Cristina Peláez-Sánchez, Davis Velarde-Camaqui, and Leonardo David Glasserman-Morales. 2024. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. *Frontiers in Education* 9 (June 2024). https://doi.org/10.3389/feduc.2024.1392091 Publisher:

Frontiers.

[27] Adam M. Persky, Edward Lee, and Lauren S. Schlesselman. 2020. Perception of Learning Versus Performance as Outcome Measures of Educational Research. *American Journal of Pharmaceutical Education* 84, 7 (July 2020), ajpe7782. https://doi.org/10.5688/ajpe7782

[28] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Towards Memorable Information Retrieval. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 69–76. https://doi.org/10.1145/3409256.3409830

[29] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 229–238. https://doi.org/10.1145/3406522.3446025

[30] Sara Salimzadeh, David Maxwell, and Claudia Hauff. 2021. On the Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR on International Conference on Theory of Information Retrieval*. ACM, 10.

[31] Rohail Syed and Kevyn Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New Brunswick, NJ, USA, 191–200. https://doi.org/10.1145/3176349.3176397

[32] Rohail Syed, Kevyn Collins-Thompson, Paul N. Bennett, Mengqiu Teng, Shane Williams, Dr. Wendy W. Tay, and Shamsi Iqbal. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1693–1703. https://doi.org/10.1145/3366423.3380240

[33] Kelsey Urgo. 2023. *Investigating the Influence of Subgoals on Learning During Search*. Ph.D. The University of North Carolina at Chapel Hill, United States – North Carolina. https://www.proquest.com/docview/2854268320/abstract/9393604F5EFD4DB5PQ/1 ISBN: 9798380133043.

[34] Kelsey Urgo and Jaime Arguello. 2023. Goal-setting in support of learning during search: An exploration of learning outcomes and searcher perceptions. *Information Processing & Management* 60, 2 (March 2023), 103158. https://doi.org/10.1016/j.ipm.2022.103158

[35] Kelsey Urgo and Jaime Arguello. 2024. The Effects of Goal-setting on Learning Outcomes and Self-Regulated Learning Processes. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Sheffield United Kingdom, 278–290. https://doi.org/10.1145/3627508.3638348

[36] Ryen W White. 2024. Advancing the Search Frontier with AI Agents. *Commun. ACM* (2024).

[37] Philip H. Winne. 2004. Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research* 41, 6 (Jan. 2004), 466–488. https://doi.org/10.1016/j.ijer.2005.08.012

[38] Philip H. Winne and Allyson F. Hadwin. 1998. Studying as self-regulated engagement in learning. In *Metacognition in educational theory and practice*.

[39] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2020. How Does Team Composition Affect Knowledge Gain of Users in Collaborative Web Search?. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20)*. Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/3372923.3404784

[40] Ryan Yen, Nicole Sultanum, and Jian Zhao. 2024. To Search or To Gen? Exploring the Synergy between Generative AI and Web Search in Programming. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. https://doi.org/10.1145/3613905.3650867

[41] Ramazan Yilmaz and Fatma Gizem Karaoglan Yilmaz. 2023. The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence* 4 (Jan. 2023), 100147. https://doi.org/10.1016/j.caeai.2023.100147

[42] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 75–84. https://doi.org/10.1145/3209978.3210064

[43] Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 2024. DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration. http://arxiv.org/abs/2408.00447 arXiv:2408.00447 [cs].

[44] Tao Zhou and Songtao Li. 2024. Understanding user switch of information seeking: From search engines to generative AI. *Journal of Librarianship and Information Science* (April 2024), 09610006241244800. https://doi.org/10.1177/09610006241244800 Publisher: SAGE Publications Ltd.