The Effects of Goal-setting on Learning during Information Seeking with Generative AI

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

Yuan Li
University of Alabama
Alabama, USA
yuan.li@ua.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

Our research in this paper lies at the intersection of Generative AI (GenAI) and search-as-learning (SAL). GenAI technologies (e.g., ChatGPT) have revolutionized how people search for and interact with information. However, we do not yet fully understand how people use GenAI systems to *learn* about complex topics. SAL research has studied how different tools can support learning with traditional document retrieval systems. Our research closely relates to SAL work that has investigated the effects of goal-setting on learning during search. We explore the influence of goal-setting on learning during information-seeking sessions with a GenAI system. We report on a between-subjects crowdsourced study ($N = 120$) in which participants were asked to learn about a complex topic using a GenAI system. The study had four conditions that varied along two factors (a 2×2 design). The first factor involved displaying related web results in addition to the GenAI output. The second factor involved giving participants access to the Subgoal Manager (SM), a tool designed to help people develop subgoals and take notes. We investigated the effects of both factors on: (RQ1) perceptions; (RQ2) behaviors; (RQ3) learning and retention; (RQ4) the types of requests issued to the system; and (RQ5) participants' motivations for engaging (or not engaging) with the related web results. Results found that participants with access to the SM had higher post-task learning outcomes, did less copy/pasting into their notes, perceived the task as more difficult, and requested more examples and support for differentiating concepts from the GenAI system.

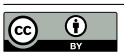
CCS Concepts

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

Keywords

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

ACM Reference Format:
Kelsey Urgo, Yuan Li, Jaime Arguello, and Robert Capra. 2026. The Effects of Goal-setting on Learning during Information Seeking with Generative AI. In *2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '26), March 22–26, 2026, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3786304.3788847



CHIIR '26, Seattle, WA, USA

ACM ISBN 979-8-4007-2414-5/2026/03
https://doi.org/10.1145/3786304.3788847

1 Introduction

Generative AI (GenAI) is reshaping the way people search for information and acquire knowledge. These conversational systems represent a fundamental shift in information seeking. Rather than navigating ranked lists of documents, users engage in multi-turn natural language exchanges to find information, answer questions, and construct understanding. Students increasingly turn to GenAI for help in their learning process [7]. Although GenAI systems feel helpful to students, emerging evidence suggests these systems have the potential to be detrimental to learning outcomes [4, 18]. Understanding how to design GenAI tools that genuinely enhance learning, rather than merely feeling helpful, has become a critical research priority.

The search-as-learning (SAL) community has long studied how people learn during information-seeking sessions with traditional search engines [35]. Researchers have developed interventions to support learning, including note-taking tools [11, 27, 28], visualizations [8, 17, 29], self-assessment tools [32] and goal-setting interfaces [33, 34]. However, nearly all SAL research has focused on document retrieval systems where learners formulate queries, evaluate results, and synthesize information across multiple documents. GenAI chat systems present a qualitatively different information seeking paradigm where synthesis happens within the system, documents may be hidden or absent, and the interaction is conversational rather than query-based. This raises a critical question: how do we support learning during information-seeking sessions with GenAI systems? Additionally, are successful interventions for traditional search *also* successful for this new form of interaction?

This paper investigates the effects of goal-setting, which offers particular promise for learning with GenAI. First, research from the learning sciences demonstrates that goal-setting plays a critical role in learning processes [30]. The Winne & Hadwin model of self-regulated learning identifies goal-setting as a key phase that enables learners to monitor progress, evaluate strategies, and adapt their approach [37]. Students who engage effectively in setting and monitoring their goals achieve better learning outcomes [9, 13]. Second, psychology research shows that goals enhance learning by honing task understanding, activating prior knowledge, directing attention to important information, and sustaining effort toward goal

achievement [21–23]. Third, prior SAL research with traditional search systems has demonstrated that goal-setting interventions improve learning outcomes, particularly knowledge retention, by encouraging greater engagement with self-regulated learning processes [34]. Despite this converging evidence, few studies have examined whether goal-setting interventions transfer effectively to information seeking with GenAI.

In this paper, we address this gap with a crowdsourced study in which participants ($N = 120$) were asked to complete a learning-oriented search task using a GenAI chat system. Participants were asked to learn about the biological concepts of diffusion and osmosis. The study used a 2×2 between-subjects design in which participants were assigned to one of four conditions: (1) Base, (2) +Web, (3) +SM, and (4) +Web+SM. Thirty participants were assigned to each condition. In the Base condition, participants used a GenAI chat system like ChatGPT and a text editor to take notes. In the +Web condition, participants had access to the same system plus an extra tab that displayed "Related Web Results". In the +SM condition, participants had access to the GenAI system and an experimental system called the Subgoal Manager (SM) that encouraged them to engage in effective goal-setting behaviors (detailed in § 3.2). In the +Web+SM condition, participants had access the GenAI chat system, the "Related Web Results" tab, and the Subgoal Manager.

Our 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' behaviors?
- **RQ3:** What were the effects of the system condition on participants' learning outcomes?
- **RQ4:** What were the effects of the system condition on the types of requests issued to the system?
- **RQ5:** In conditions that included related web results, what were participants' motivations for engaging with the web results?

2 Related Work

2.1 Tools to Support Learning during Search

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

Note-taking Tools: Freund et al. [11] conducted a study in which participants read articles in plain text versus HTML with distracting elements. Participants had higher reading comprehension scores in the plain text condition. However, this effect was attenuated when participants could highlight text and make "sticky notes". Roy et al. [28] investigated the effects of two tools on learning—one tool to highlight text and one to take notes. Access to either tool improved learning outcomes. However, access to both tools did not, possibly because of cognitive overload. Qiu et al. [27] investigated the effects of a note-taking tool within two search environments—a traditional search system and a conversational search system. Knowledge gains were greatest with the traditional search system and the note-taking tool.

Visualizations: Kammerer et al. [17] experimented with a search system that enabled participants to filter results using social tags. Learning gains were higher with the experimental system versus a system without social tags. Câmara et al. [8] investigated a visualization that displayed participants' coverage of subtopics during the search session. The visualization did not improve learning outcomes because participants explored more subtopics *superficially*.

Self-Assessment Tools: Syed et al. [32] experimented with a system that prompted participants to answer questions about passages read during the search session. Prompting participants to answer questions improved knowledge retention for participants with low prior knowledge.

Goal-setting Tools: Our work is related to prior research on the role of goal-setting on learning during search. Urgo and Arguello [33] developed the Subgoal Manager (SM) to help searchers break apart a complex learning objective into subgoals. In one study, they found that participants had the best learning outcomes when they had access to the SM and set their own subgoals [33]. In a separate study [34], they found that participants in the SM condition had greater knowledge retention scores and engaged in more self-regulation (e.g., prior knowledge activation and progress monitoring). Our work builds on this research by examining the role of goal-setting on learning with a generative AI system.

2.2 Generative AI & Learning

People are increasingly turning to GenAI tools to learn about complex topics [7, 14, 26]. Despite widespread adoption, researchers are only starting to understand how GenAI tools affect learning outcomes. Recent studies have shown mixed results. Some research suggests GenAI tools can enhance learning outcomes [1, 24, 42], while other work has found detrimental effects [4, 15, 18]. A significant limitation across this emerging work is the reliance on subjective measures. Rather than using objective assessments of knowledge or skill acquisition, most studies have measured self-reported experiences and perceptions of learning. Recent systematic reviews show psychological constructs (e.g., self-efficacy, motivation) are the predominant outcome measures [2, 38]. Objective measures of learning by closed- or open-ended assessments are notably absent [2, 38].

To underscore the distinction between subjective and objective learning outcomes, consider two contrasting studies. Yilmaz and Karaoglan Yilmaz [42] found that students using ChatGPT for programming exercises reported higher motivation, self-efficacy, and computational thinking skills. However, the study did not objectively measure learning. By contrast, Kosmyna et al. [18] used a rigorous multi-method set of measurements including EEG brain connectivity analysis, NLP-based essay scoring, and memory recall testing. Use of LLMs facilitated task completion but decreased memory consolidation and neural connectivity.

2.3 Generative AI & Search Behaviors

People engage with LLMs in ways that differ from traditional web search. During a product comparison task, Kaiser et al. [16] found that participants using ChatGPT wrote longer, more conversational prompts and visited fewer pages than participants using Google. Wazzan et al. [36] compared interactions with a traditional versus LLM-based search system during image geolocation tasks. Using the LLM-based search system, participants issued longer, more natural language queries and had shorter sessions. Additonally, during query reformulation, they tended to rephrase their queries instead of simply adding more terms. In a study of students solving physics

problems, Krupp et al. [19] found that ChatGPT users tended to copy/paste the full text of the question and ask for explanations, translations, summaries, and corrections. In contrast, Google users tended to divide the question into multiple queries.

Studies have also investigated why users turn to AI versus traditional search. Users often turn to AI when domain knowledge is low and tasks are vague [41], and when search results are irrelevant or overwhelming [44]. In contrast, users often turn to web results to verify an AI response [41].

In terms of outcomes, results are mixed. In their study on product comparison, Kaiser et al. [16] found that ChatGPT users finished faster and were more accurate. In their study on image geolocation, Wazzan et al. [36] found that participants were more accurate using traditional web search. In their study on physics problem-solving, Krupp et al. [19] found that ChatGPT users scored lower and over trusted its responses. In a study involving socio-scientific reasoning, Stadler et al. [31] found that ChatGPT users felt the task was easier, but web search users yielded stronger, better-supported conclusions. Finally, Yang et al. [40] found slightly better learning outcomes during a conceptual learning task for participants who used an LLM-enhanced search system.

3 Methods

To investigate RQ1-RQ5, we conducted a between-subjects crowdsourced study ($N = 120$) on the Prolific platform. Thirty participants were assigned to one of four conditions (§ 3.2): (1) Base, (2) +Web, (3) +SM, and (4) +Web+SM. We limited the study to Prolific workers in the U.S. who had completed at least 100 tasks with an acceptance rate of 95% or greater. Participants were aged 18-71 and the median age was 36. Sixty-one identified as female, 59 as male, and 0 as non-binary. Participants were asked about their familiarity with GenAI tools. Twenty-seven reported using GenAI tools multiple times a day, 56 multiple times a week, 27 multiple times a month, and 10 only a few times ever. The study was approved by the Institutional Review Board (IRB) of each author's institution.

3.1 Study Protocol

During the study, participants interacted with a "study workflow" page. The "study workflow" page did not allow participants to skip steps and included several instructional videos. The study protocol proceeded as follows. First, participants watched a video describing the study protocol. Second, participants completed a demographics questionnaire. Third, participants completed the multiple-choice Osmosis and Diffusion Conceptual Assessment (ODCA) (§ 3.5) to measure their prior knowledge of diffusion and osmosis. Fourth, participants were asked to read the learning task description (§ 3.3) and completed a pre-task questionnaire about their perceptions of the learning task (§ 3.4). Fifth, participants watched a video describing the system associated with their assigned experimental condition (§ 3.2). In the conditions with the Subgoal Manager, the video included a description of ideal subgoal characteristics (with examples) that make subgoals more achievable. Sixth, participants completed the main learning task. We did not enforce a time limit but told participants that the task should take about 40 minutes. Seventh, participants completed a post-task questionnaire (§ 3.4) about their experiences during the learning task. Then, to measure learning, participants completed the ODCA a second time. Finally, in the +Web and +Web+SM conditions, participants completed an exit questionnaire that asked about use of the "Related Web Results" tab (§ 3.4). Participants were paid US$30 for participating in the study. To measure their knowledge retention, participants were emailed an invitation to complete the ODCA a third time. A total of 107 participants completed the retention assessment: 28 in the Base condition, 26 in the +Web condition, 27 in the +SM condition, and 26 in the +Web+SM condition. Participants were given a US$10 bonus for completing the retention assessment.

3.2 Experimental Conditions

The study had four experimental conditions that varied along two factors. One factor manipulated whether participants were given access to related web results in addition to the AI agent. The second factor manipulated whether participants had access to the Subgoal Manager (SM) or a simple text editor to take notes. Thirty participants were assigned to each condition.

In the Base condition, participants only had access to the AI agent (no related web results) and a text editor to take notes. In the +Web condition, participants had access to the AI agent, related web results, and a text editor to take notes. In the +SM condition, participants only had access to the AI agent (no related web results) and the SM. Finally, in the +Web+SM condition, participants had access to the AI agent, related web results, and the SM.

Figure 1 illustrates the system in the +Web+SM condition. In all conditions, the interface included buttons for participants to revisit the task description (A) and finish the task (B). All conditions provided access to the AI Agent (C). The AI Agent was implemented using the OpenAI API with the `gpt-4o-mini` model. Participants could send requests to the AI Agent using the "Ask away" textbox at the bottom. The AI Agent worked similar to ChatGPT's web interface. Not shown in Figure 1, the AI Agent could also respond with images and explanations if explicitly requested (e.g., "give me a figure explaining osmosis"). In response to an explicit image request, the system used the OpenAI API to transform the request into a search query, which was then submitted to the Google Image Search API. The retrieved image URL was subsequently sent to the OpenAI API to present both the image and an explanation. The system saved the conversation history and participants could scroll up to see previous responses. Additionally, participants could create "new chats" and return to previous chats (E).

In the +Web and +Web+SM conditions, participants also had access to a "Related Web Results" tab (D). This tab displayed 10 web results related with the most recent AI agent request. The "Related Web Results" tab was implemented as follows. First, we used the OpenAI API to generate a search query associated with the most recent AI agent request. We prompted the OpenAI API with the conversation history in order to handle requests that relied on the conversational context (e.g., "give me a simpler explanation"). Then, we queried the Brave Search API to return 10 related web results. Participants could not issue queries directly to the Brave Search API. However, the related web results were updated after every AI agent request.

In the +SM and +Web+SM conditions, participants were given access to the Subgoal Manager (SM) on the right (F). The SM was designed to help learners break apart a complex learning task into

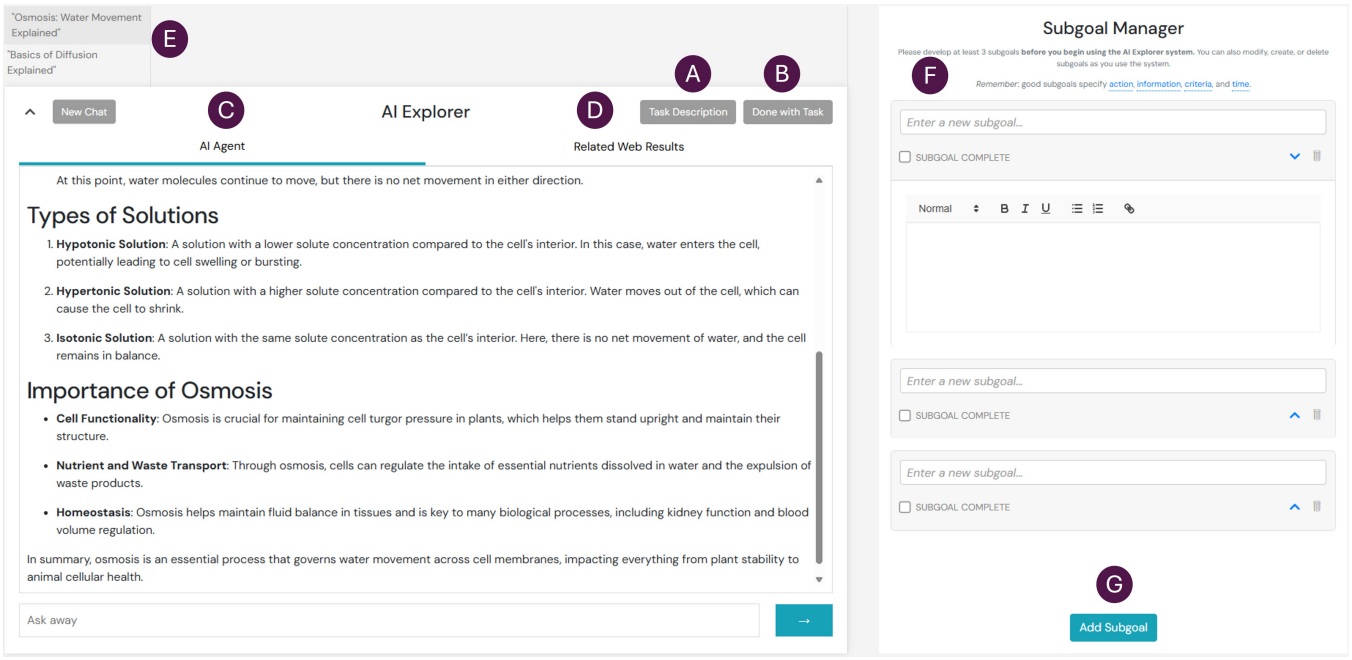


Figure 1: System Interface in the +Web+SM Condition

smaller, more concrete subgoals. The tool allowed participants to create subgoals, take notes with respect to each subgoal, and mark subgoals as completed. Checking the "subgoal complete" box collapsed the subgoal and turned it into a darker shade of gray. Participants could create new subgoals by clicking the "Add Subgoal" button (G) and could delete subgoals by clicking the trash button for the subgoal. Participants could also expand and collapse the notes associated with each subgoal. Finally, tooltips on the interface reminded participants to set subgoals with four ideal characteristics: (1) action, (2) information, (3) success criteria, and (4) approximate timeframe. Participants were instructed to set at least three subgoals before the task on their own. Subgoals could be added, modified, and deleted during the session. In conditions that excluded the SM (i.e., Base and +Web), participants were provided with a simple text editor that was positioned where the SM appears in Figure 1.

As mentioned in § 3.1, before the learning task, participants watched videos describing the system associated with their assigned condition. In conditions that included the SM (i.e., +SM and +Web+SM), the video also instructed participants to set subgoals with four ideal characteristics. Good goals are associated with a specific action, information, success criteria, and approximate timeframe. The video described these characteristics and provided examples. Finally, the video instructed participants to set at least three subgoals before beginning the task and reminded them that subgoals could be modified, added, and deleted during the task. In conditions that excluded the SM (i.e., Base and +Web), the system video did not mention anything about subgoals. Participants were simply instructed to use the text editor to take notes.

3.3 Search Task

During the study, participants completed a single learning-oriented task, which included the following contextualizing scenario and 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 Questionnaires

Participants completed questionnaires before and after the learning task. In both, participants responded to agreement statements on a 7-point scale ranging from "1-strongly disagree" to "7-strongly agree". Both questionnaires are available online.

Pre-task Questionnaire: After reading the task description, participants completed a pre-task questionnaire 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)—whether aspects of the task (e.g., requirements, goals, and strategies for completion) are known in advance [6]. Responses to the items about prior knowledge, expected difficulty, and *a priori* determinability had high internal consistency (Cronbach's $\alpha \geq .89$) and were therefore averaged to form three composite measures.

Post-task Questionnaire: The post-task questionnaire was organized in three sections. The first section asked about: (1) interest increase (1 item), (2) knowledge increase (3 items), and difficulty (4 items). Responses to the items about knowledge increase had high internal consistency ($\alpha = .94$) and were therefore averaged to form

a composite measure. Responses to the items about difficulty had lower internal consistency ($\alpha = .77$). However, dropping one item (i.e., difficulty in deciding when to end the task) resulted in higher internal consistency for the other three items ($\alpha = .82$). Therefore, responses to the other three items were averaged and the dropped item was analyzed separately.

The second section asked about participants' engagement in different cognitive and metacognitive activities: (1) setting goals, (2) deciding how to begin the task, (3) connecting information to existing knowledge, (4) relating topics, (5) comparing different explanations of similar ideas, (6) deciding whether new information matched previously encountered information, (7) tracking progress, (8) evaluating their understanding of information, (9) deciding whether information was useful, and (10) revising their approach to the task. These items were analyzed individually.

The third section asked whether the information returned by the system was: (1) credible, (2) trustworthy, (3) unbiased, (4) accurate, (5) factual, (6) reliable, and (7) up-to-date. Responses to these items had higher internal consistency ($\alpha = .93$) and were therefore averaged to form a composite measure.

Exit Questionnaire: In conditions that included the related web results in addition to the AI agent (i.e., +Web and +Web+SM), participants completed an exit questionnaire that asked two open-ended questions about their use of the "Related Web Results" tab: (1) what were they trying to accomplish by using the tab and (2) what types of information did they gain from the tab.

3.5 Learning Assessment

To measure learning and retention, participants completed the Osmosis and Diffusion Conceptual Assessment (ODCA) [10] before the learning task, immediately after, and one week later. The ODCA has 18 multiple-choice questions that are organized in pairs. The knowledge question asks "what?" and the reasoning question asks "why?". The ODCA was used for two reasons. First, it targets common misconceptions that biology students have about diffusion and osmosis [10]. Second, 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 combined pre- and post-task ODCA scores to compute:

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

where PreScore and PostScore are the percentage of correct answers in the pre- and post-task ODCA. To measure retention, we used the same normalization but replaced PostScore with RetScore—the percentage of correct answers in the retention ODCA. Normalized gain has been commonly used in SAL studies [12, 39, 40, 43].

3.6 Behavioral Measures

To address RQ2, we collected a broad set of behavioral measures. Our measures are grouped in three categories. First, the following measures apply to all four experimental conditions (i.e., Base, +Web, +SM, +Web+SM):

- `duration`: total time (in seconds) from the first request issued to the AI agent to the last event.
- `n_exchanges`: number of AI requests issued.
- `new_chats`: number of distinct chats created.
- `notes_all`: total character count in notes. In conditions with the SM, we combined the notes taken across subgoals.
- `paste`: number of paste events into notes.
- `avg_prompt_length`: average character count across AI requests.
- `unique_terms`: total number of unique terms across AI requests.
- `avg_term_specificity`: average IDF of words across AI requests. Higher values mean that the participant issued requests with more uncommon vocabulary.

Second, the following measures capture interactions with the related web results. Therefore, they only apply to conditions that included related web results (i.e., +Web, +Web+SM):

- `n_result_clicks`: number of web results clicked.
- `avg_click_rank`: average click rank.
- `n_web_tab_visits`: number of visits to the web results tab
- `avg_time_away_sec`: average time away from the system tab (i.e., mean of blur to focus intervals).
- `avg_chat_view_sec`: average dwell time on the AI agent tab before clicking on a different element

Third, the following measures capture interactions with the Subgoal Manager (SM). Therefore, they only apply to conditions that included the SM (i.e., +SM, +Web+SM):

- `n_subgoals`: number of subgoals created.
- `subgoal_note_chars`: total characters in SM note editors.
- `n_completed`: number of subgoals completed.
- `n_title_changes`: number of subgoal title edit events.
- `n_notes_changes`: number of subgoal note edit events.

Measures in the second and third groups examine whether access to the SM impacted interactions with the related web results and vice-versa. Measures in these groups were analyzed using Mann-Whitney tests to compare the respective pair of conditions.

3.7 Analysis of AI Requests

To address RQ4, we conducted a qualitative analysis of requests issued by participants to the AI agent. Participants issued a total of 1,139 requests. Our analysis of requests involved two authors (A1 & A2) and proceeded as follows. First, A1 analyzed requests from 10 participants and developed a codebook with codes, definitions, and examples. Then, A2 coded requests from the same 10 participants and A1 & A2 met to discuss disagreements. Several codes were refined and two were merged. After this, to test the reliability of our codes, A1 & A2 independently coded all requests from the same 24 participants. Our codes were designed to not be mutually exclusive. Requests could be assigned multiple codes. Therefore, we measured agreement per code. Six had agreement levels of *almost perfect* (Cohen's $\kappa > .80$) and eight had agreement levels of *substantial* ($.80 \geq \kappa > .60$) [20]. Given the high levels of agreement, the remaining requests were coded disjointly by A1 & A2.

Our codes are described below. Values in parentheses indicate the percentage of requests associated with each code. Some of our codes relate to the participant's intent (e.g., get a definition) and others relate to phenomena we observed (e.g., the participant asked for the response to be formatted a certain way).

(1) **Definition (9.92%):** the participant asked for the definition of one or more concepts (e.g., "can you define diffusion?").

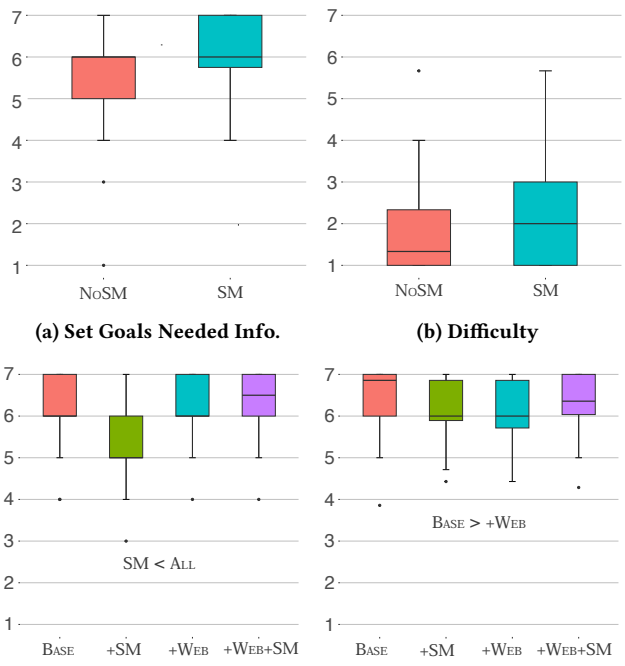
(2) **Examples (10.62%):** the participant asked for examples of a concept or process (e.g., "everyday examples of diffusion").
(3) **Explanation (23.79%):** the participant asked for a more in-depth explanation than a definition (e.g., "what are the components of the circulatory system and what roles do they play?").
(4) **Clarification (16.15%):** the participant asked a specific question about a point of confusion (e.g., "is water at the bottom of the ocean saltier than at the top?").
(5) **Hypothesis Verification (6.58%):** the participant wanted to verify whether their understanding was correct (e.g., "so diffusion happens faster in water than in syrup?").
(6) **Differentiate Concepts (8.60%):** the participant wanted to understand the similarities, differences, or relations between concepts (e.g., "differences between diffusion and osmosis.").
(7) **Cause and Effect (3.34%):** the participant asked about a causal relationship (e.g., "how does temperature affect diffusion?").
(8) **Ideas (6.41%):** the participant wanted ideas about things to learn (e.g., "is there anything else I need to know [...]").
(9) **Test Knowledge (7.55%):** the participant wanted the AI agent to test their knowledge (e.g., "give me a high school level multiple choice quiz [...] do each question separately and wait for me to answer."). This code was also applied when participants answered questions from the AI agent (e.g., "the answer is A").
(10) **Social (4.21%):** the request was a social nicety (e.g., "okay, thanks for all the info!").
(11) **Qualification (20.90%):** the participant qualified the type of information they wanted (e.g., "quick definition", "explain in a basic way", "this is way too complex!", "get more into specifics?").
(12) **Specific Text Format (6.32%):** the participant asked for the response to be formatted a certain way (e.g., "too long, one paragraph", "sum it up in three sentences", "write 3-4 paragraphs explaining everything [...]").
(13) **Visuals (8.60%):** The participant asked for visuals (e.g., "show me a diagram of diffusion vs. osmosis.").
(14) **Use of Context (23.88%):** the request leveraged the AI agent's ability to maintain conversational context (e.g., "can you sum this up?", "same question, but for diffusion", "make those 3 examples into exam questions.").

3.8 Analysis of Web Results Use Motivation

To address RQ5, we conducted an inductive thematic analysis [5] of participants' open-ended responses about their motivations for using the "Related Web Results" tab in conditions +Web and +Web+SM. The analysis involved two authors (A1 & A2). First, A1 familiarized themselves with the data and generated codes across all 60 responses (30 for +Web and 30 for +Web+SM). Then, A1 grouped codes into candidate themes and drafted theme descriptions. The full research team reviewed the initial codes and candidate themes. A2 checked each theme and its interpretations against the original responses. Finally, A1 & A2 discussed four discrepancies and resolved them through discussion. Motivations for engaging with the web results were not mutually exclusive.

3.9 Statistical Analysis

Our four experimental conditions varied along two factors: (1) access to the Subgoal Manager (SM) versus a simple text editor to take notes and (2) access to web results related to the latest AI agent request versus no access to web results. No dependent variable for



(a) Set Goals Needed Info. **(b) Difficulty**

(c) Decide Info. Useful **(d) Quality of Info.**

Figure 2: Post-task Perceptions

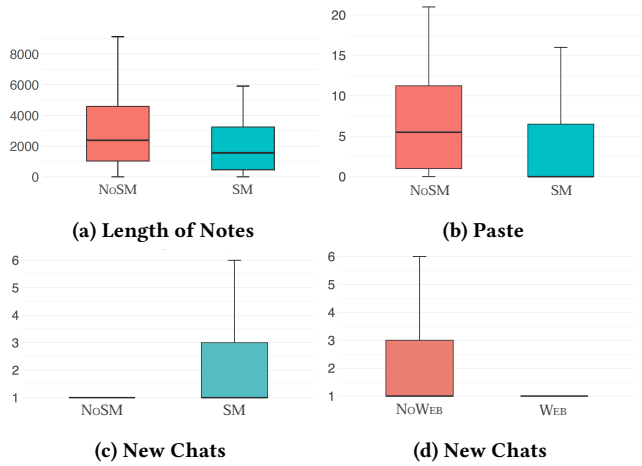
RQ1-RQ5 passed the Shapiro-Wilk test of normality. Therefore, for each dependent variable, we used Scheirer-Ray-Hare (SRH) tests to investigate the main effects of each of the factors above and their interaction. The SRH test is a non-parametric alternative to a multi-factorial ANOVA. For dependent variables with an interaction effect, we conducted Bonferroni-corrected Mann-Whitney tests to check for differences between all pairs of conditions.

4 Results

Before presenting our RQ1-RQ5 results, we report on differences in prior knowledge and pre-task perceptions between groups. Given that the study used a between-subjects design, we were curious about possible differences between groups. Pre-task ODCA scores were not significantly different between groups. Additionally, there were no significant differences in pre-task perceptions of interest in the task, prior knowledge, expected difficulty, and *a priori* determinability. Therefore, any significant differences between groups for RQ1-RQ5 cannot be attributed to participants in different groups having significantly different levels of prior knowledge and pre-task perceptions due to random chance.

4.1 RQ1: Post-task Perceptions

In terms of post-task perceptions, we found two significant main effects and two significant interaction effects. First, as shown in Figure 2a-2b, participants with access to the Subgoal Manager (SM) (i.e. in conditions +SM and +Web+SM) reported on engaging in more goal-setting ($H(1) = 9.54, p < .005$) and reported higher levels of difficulty ($H(1) = 3.90, p < .05$). One possible interpretation is that access to the SM made participants be more goal-oriented, which required effort.



(a) Length of Notes (b) Paste

(c) New Chats (d) New Chats

Figure 3: Behaviors

Second, there was a significant interaction effect on participants being able to decide whether information was useful ($H(1) = 4.58$, $p < .05$). As shown in Figure 2c, pairwise Mann-Whitney tests found that participants reported being less able to decide whether information was useful in condition +SM vs. the other three. One possible explanation is that SM made participants be more critical of the information returned by the AI agent but they were unable to verify information using the related web results.

Finally, there was a significant interaction effect on participants' perceptions of the quality of information returned by the system ($H(1) = 5.80$, $p < .05$). As shown in Figure 2d, pairwise Mann-Whitney tests found that participants rated the quality of the information significantly higher in condition Base vs. +Web. One possible explanation is that when participants were not goal-oriented (i.e., without the SM in conditions Base and +Web), they were more critical of the information returned by the AI agent when they could scrutinize it with the related web results (i.e., they rated the information quality lower in +Web vs. Base).

4.2 RQ2: Behaviors

We observed significant differences for four behavioral measures. Two measures (Figures 3a– 3b) had significant main effects from having access to the Subgoal Manager (SM). Participants in conditions +SM and +Web+SM (vs. Base and +Web) had significantly shorter notes ($H(1) = 6.10$, $p < .05$) and fewer paste events ($H(1) = 11.59$, $p < .001$).

Interestingly, the number of `new_chats` created (Figures 3c & 3d) had both a significant main effect from having access to the Subgoal Manager (SM) and a significant main effect from having access to the related web results (Web). The number of `new_chats` created were significantly higher ($H(1) = 6.45$, $p < .05$) in conditions with the SM and significantly lower ($H(1) = 8.04$, $p < .01$) in conditions with related web results.

4.3 RQ3: Learning Outcomes

As shown in Figure 4, participants with access to the SM (i.e., in conditions +SM and +Web+SM) had significantly larger normalized learning gains immediately after the task ($H(1) = 5.80$, $p < .05$). However, there were no significant effects on normalized learning gains one week after the task. These results suggest that goal-setting improved learning during the task. However, additional scaffolding may be needed for those gains to be long term.

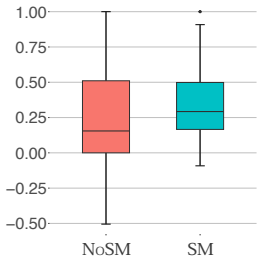
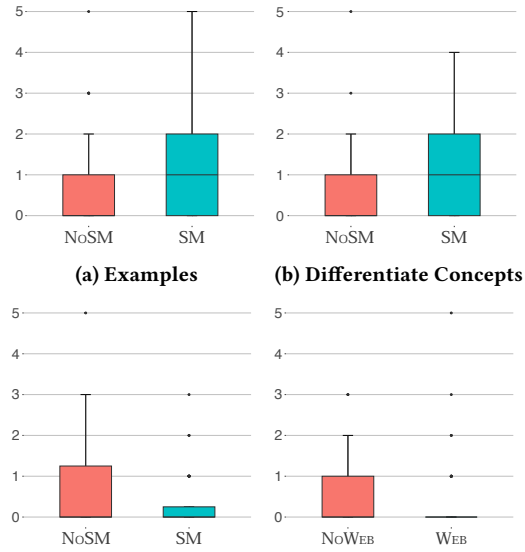


Figure 4: Post-task Normalized Gain



(a) Examples (b) Differentiate Concepts

(c) Ideas (d) Specific Text Format

Figure 5: Types of AI agent requests

4.4 RQ4: Types of Requests

We observed four significant differences in the types of requests issued by participants to the AI agent. As shown in Figures 5a-5b, participants with access to the SM (i.e., in conditions +SM and +Web+SM) issued more AI requests asking for examples ($H(1) = 5.24$, $p < .05$) and to differentiate concepts ($H(1) = 11.79$, $p < .001$). Additionally, as shown in Figure 5c, participants with access to the SM issued fewer AI requests asking for ideas about things to learn ($H(1) = 5.55$, $p < .05$). Finally, as shown in Figure 5d, participants with access to related web results (i.e., in conditions +Web and +Web+SM) issued fewer requests that asked for the response to be formatted in a specific way ($H(1) = 6.23$, $p < .05$).

4.5 RQ5: Motivations for Web Results (Non-)Use

RQ5 investigated why participants engaged or did not engage with the related web results. The numbers in parentheses indicate how many participants (out of 60) mentioned each theme.

Motivations for Using Web Results: Participants mentioned five reasons for engaging with the web results.

Seek additional information beyond AI response (27): Participants described using the web results to obtain information not covered in the AI response, such as details, examples, alternative formats (e.g., images), or explanations. They also used the web results to look for alternative viewpoints and sources. Participants made comments such as: "[to] find additional information and perspectives to complement what I learned from AI" and "[to] look for additional search terms that would bring up more information."

Validate AI's information (26): Participants used the web results to verify the accuracy of information returned by the AI agent. For instance, one participant wrote, "to check reputable pages to confirm that the information was correct and not a hallucination." Another wrote, "to verify where the AI was pulling its information."

Reduce overload from AI results (1): One participant described using the web results when they felt overwhelmed: "I felt overwhelmed by the information from the AI agent and needed to read the information in a smoother/more realistic way."

Out of curiosity (1): One participant described browsing the web results out of curiosity: "Out of curiosity I went and checked it out."

Motivations for NOT Using Web Results: Participants mentioned five reasons for not engaging with the web results.

Preference for AI retrieval (5): Five participants did not use the web results because they preferred having the AI collect and synthesize information instead of searching manually. One participant wrote, "I preferred letting the AI agent gather information for me." Another wrote, "I trusted my AI agent to do all the research."

AI responses were sufficient (3): Three participants felt that the AI's responses were sufficient. One participant wrote, "AI provided me with all the info I needed." Another wrote, "the AI agent was clear in its explanations and teachings."

Avoid being overwhelmed (3): Three participants mentioned not using the web results because they would be difficult to read and not worth the effort. One participant wrote, "with the internet, you need to read more and it is not in your face." Another wrote, "I felt the AI gave me a lot of useful information at once, whereas the other one [web] seemed to give me repetitive information."

Confident in prior knowledge (1): One participant did not use the web results because they felt confident in their own knowledge: "I can guess if the AI info is incorrect."

Focused on completing task (1): One participant did not interact with the web results because they were concentrated on task activities rather than branching into web browsing: "I was so focused on writing the notes from the AI tab."

5 Discussion

RQ1 Post-Task Perceptions: For RQ1, we found two significant main effects. First, participants with access to the SM reported on being able to set goals for what information they needed to find. This is not surprising given that the SM was designed to explicitly support goal-setting. Urgo and Arguello [33] also found that participants perceived the SM to help them with *planning*. Second, participants with access to the SM perceived the task to be more difficult. This can be explained by the fact that setting goals and making progress toward specific goals is an effortful activity.

Our RQ1 results also found two significant interaction effects. First, compared to the other three conditions, participants in condition +SM reported being *less* able to decide whether information would be useful to them. This might be explained by participants in condition +SM: (1) seeking specific information based on their subgoals *and* (2) not having related web results to verify goal-specific information returned by the AI agent. Based on the open-ended responses for RQ5, many participants in conditions with related web results (i.e., +Web and +Web+SM) *incorrectly* assumed that the AI agent was pulling information from the related web results. Participants often commented on visiting the "Related Web Results" tab to see which sources the AI agent was using.

Finally, participants perceived the quality of information from the AI agent to be highest in condition Base and lowest in +Web. This suggests that having access to related web results may lead people to be more critical of responses from the AI. This effect was not significant when the Subgoal Manager was used (e.g., +SM and +Web+SM), suggesting that goal-setting may mediate this effect.

RQ2 Behaviors: For RQ2, participants without the SM took significantly more notes and engaged in more copying and pasting. However, they had lower learning outcomes. This finding suggests that the quality of engagement matters more than the quantity of activity (e.g., more copying and pasting did not lead to better learning outcomes). This suggests that the SM promoted deeper, more deliberate cognitive processing. Rather than engaging in surface-level information gathering through extensive copying and pasting, participants with the SM appeared to have engaged in more selective, thoughtful note-taking.

Scaffolding tools such as the SM have the potential to increase the time needed for learners to meet their objectives. Indeed, participants in conditions +SM and +Web+SM had longer sessions if we account for the time they spent developing their initial subgoals (not discussed in § 4.2). However, they did *not* have longer sessions from the first AI request (i.e., after developing initial subgoals) to the end of the task. This finding suggests that an initial investment of ~5 minutes of goal-setting yielded better learning outcomes without requiring additional time during the learning process itself. This result resonates with Urgo and Arguello [34], who found that participants with the SM engaged in more time monitoring.

Finally, we found two main effects on the number of new chats created by participants. Participants created more chats with the SM. The SM may have encouraged participants to organize their learning into distinct, focused-inquiry sessions aligned with their subgoals. Conversely, participants created *fewer* new chats when they had related web results. One possibility is that the related web results drew participants' attention away from the "new chats" feature of the interface.

RQ3 Learning Outcomes: Our results demonstrate that encouraging and scaffolding goal-setting in a GenAI system may enhance learning outcomes compared to GenAI alone. Participants with access to the SM achieved significantly higher normalized learning gains immediately following the learning task. Such results indicate that structured goal-setting may have improved learning while interacting with the GenAI system. This significant difference, however, did not persist one week later, suggesting that while the SM effectively supports initial learning, additional scaffolding may be needed to promote long-term retention.

Our results for retention differ from prior work. Urgo and Arguello [34] found that access to the SM improved retention. Interestingly, that effect was *stronger* with an open-ended assessment (i.e.,

describe everything you learned). The ODCA used in our study may not capture everything participants learned and retained. Therefore, an open-ended assessment may have revealed differences in retention from having access to the SM.

RQ4 Types of Requests: Our qualitative analysis of AI requests (§ 3.7) revealed a wide range of ways in which people may be using GenAI to learn about complex topics. Some of our qualitative codes might be expected for a conceptual learning task. Participants asked the AI agent to define, exemplify, explain, clarify, and differentiate. However, other codes were more surprising. Participants asked the AI agent to provide ideas about things they should learn, verify whether a hypothesis was correct, test their knowledge with multiple choice questions, and respond to social exchanges. Additionally, participants often issued requests with extra-topical constraints [3]. For example, they asked for visuals; they qualified the type of information wanted (e.g., simple, in-depth, detailed); and they asked for responses to be formatted in a specific way (e.g., 3 paragraphs, bullet points). Finally, they issued requests that leveraged the AI agent's ability to maintain conversational context (e.g., "they sound alike to me."). Some of our more interesting codes were fairly common. For example, 20.90% of requests included a qualification and 23.88% leveraged the conversational context.

Our RQ4 results found four main effects: three from having access to the SM and one from having access to related web results. First, participants with access to the SM were more likely to request examples and support to differentiate concepts. These two results go hand in hand. Participants often asked for examples of one concept (e.g., diffusion) in one request and examples of another concept (e.g., osmosis) in a subsequent request (i.e., they were trying to differentiate through multiple requests). Given that osmosis is a special type of diffusion, exemplifying and differentiating are essential activities in learning about these concepts. This may explain why participants with the SM had better post-task learning outcomes (RQ3). Second, participants with access to the SM were less likely to ask for ideas about things to learn. This may be because they were more goal-oriented (i.e., top-down vs. bottom-up) and did not use the AI agent to brainstorm. Finally, participants with access to related web results were less likely to request the AI agent response to be formatted in a specific way. It may be that response format matters more when the AI agent is the *only* source of information (i.e., no related web results).

RQ5 Motivations for Web Results Use: The results from RQ5 revealed a nuanced picture of why participants did or did not interact with related web results. Participants used the related web results for additional details, examples, and perspectives not provided by the AI agent. This suggests that while GenAI is seen as a useful source, it is not always sufficient for learning. Participants also used the related web results to verify the accuracy and credibility of the AI agent's responses. Prior work observed a similar trend, noting that web results have more signals to assess credibility (e.g., up-votes in Q&A sites) [41]. This result suggests that people remain cautious about hallucinations from GenAI systems.

Among those participants who did not use web results, the majority indicated a preference for AI-led retrieval and synthesis, relying on the AI agent to collect and summarize information. This preference often had multiple layers: (1) participants wanted to delegate the search process to the AI agent; (2) they found the AI agent more user-friendly; and (3) they trusted the AI agent's capabilities. Some participants also expected the web results to be difficult to read. Prior work also found that people sometimes turn from search engines to GenAI to avoid information overload [44].

Opportunities for Future Work: Our findings suggest several directions for future work. First, additional analyses could examine how specific participant behaviors influenced learning outcomes. Future work could explore the relationship between the types of requests issued by participants to the system (e.g., testing knowledge) and learning outcomes. Additionally, prior work found that participants had better learning outcomes when they set high-quality subgoals in the SM (e.g., with measurable success criteria) [33]. In our study, participants with the SM improved their ODCA scores in the post-task assessment but not the retention assessment. Further analysis might examine whether retention was higher for participants who set high-quality subgoals.

Second, the SM could better support learners through enhanced goal-setting features. Effective goals specify an action, information, success criteria, and approximate timeframe. [25]. The system could provide feedback on goal quality, suggest additional goals, recommend the order in which goals should be pursued (e.g., diffusion before osmosis), and detect when goals are being neglected.

Third, AI responses could be enhanced to promote particular self-regulated learning strategies. The system could prompt learners to engage in productive request types observed in our study, such as requesting examples, differentiation, or knowledge testing. Additional features might include identifying prerequisite knowledge gaps, linking to corroborating sources, generating visual explanations when appropriate, and prompting learners to summarize their understanding (before or during the learning session) to surface misconceptions for the AI to correct.

6 Conclusion

This study investigated how goal-setting tools influence learning during information seeking with a GenAI system. We examined two factors. First, whether participants had access to the Subgoal Manager (SM) or only to note taking. Second, whether related web results were displayed alongside the GenAI output or not. Participants with access to the SM demonstrated greater knowledge gains immediately after the learning session. However, this trend was less pronounced (not significant) one week later, suggesting that additional scaffolding may be needed. Participants with access to the SM also exhibited distinct interaction patterns—they did less copy/pasting into their notes, they requested more examples and concept differentiation support, and they requested fewer ideas about things they should learn, suggesting more goal-oriented engagement. Additionally, participants with access to related web results were less likely to request formatted responses, suggesting that response formatting may matter more when the AI agent is the sole information source. When participants did engage with web results, they used them primarily to verify the accuracy of the GenAI output and to obtain additional perspectives. However, many participants preferred delegating search entirely to the AI agent for "finding" and summarizing information. These findings provide initial evidence that goal-setting scaffolds can influence learning behaviors and immediate outcomes in GenAI interfaces, while highlighting areas for further investigation.

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] Abdullah Alfarwan. 2025. Generative AI use in K-12 education: a systematic review. *Frontiers in Education* 10 (Sept. 2025). https://doi.org/10.3389/feduc.2025.1647573 Publisher: Frontiers.

[3] Jaime Arguello, Bogeum Choi, and Robert Capra. 2018. Factors Influencing Users' Information Requests: Medium, Target, and Extra-Topical Dimension. *ACM Trans. Inf. Syst.* 36, 4, Article 41 (July 2018), 37 pages. https://doi.org/10.1145/3209624

[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] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] 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

[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] Victor M. Deekens, Jeffrey A. Greene, and Nikki G. Lobczowski. 2018. Monitoring and depth of strategy use in computer-based learning environments for science and history. *British Journal of Educational Psychology* 88, 1 (2018), 63–79. https://doi.org/10.1111/bjep.12174 _eprint: https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjep.12174.

[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 A. Greene, Nikki G. Lobczowski, Rebekah Freed, Brian M. Cartiff, Cynthia Demetriou, and A. T. Panter. 2020. Effects of a Science of Learning Course on College Students' Learning With a Computer. *American Educational Research Journal* 57, 3 (June 2020), 947–978. https://doi.org/10.3102/0002831219865221 Publisher: American Educational Research Association.

[14] 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.

[15] 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].

[16] Carolin Kaiser, Jakob Kaiser, Rene Schallner, and Sabrina Schneider. 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*. 1–7.

[17] 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

[18] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. https://doi.org/10.48550/arXiv.2506.08872 arXiv:2506.08872 [cs].

[19] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Unreflected acceptance–investigating the negative consequences of ChatGPT-assisted problem solving in physics education. *arXiv preprint arXiv:2309.03087* (2023).

[20] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310 Publisher: [Wiley, International Biometric Society].

[21] Gary P. Latham, Terence R. Mitchell, and Dennis L. Dossett. 1978. Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology* 63, 2 (1978), 163–171. https://doi.org/10.1037/0021-9010.63.2.163 Place: US Publisher: American Psychological Association.

[22] Gary P. Latham and Lise M. Saari. 1979. Application of social-learning theory to training supervisors through behavioral modeling. *Journal of Applied Psychology* 64, 3 (1979), 239–246. https://doi.org/10.1037/0021-9010.64.3.239 Place: US Publisher: American Psychological Association.

[23] Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist* 57, 9 (Sept. 2002), 705–717. https://doi.org/10.1037/0003-066X.57.9.705 Publisher: American Psychological Association.

[24] 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.

[25] Lindsay McCardle, Elizabeth A. Webster, Adrianna Haffey, and Allyson F. Hadwin. 2017. Examining students' self-set goals for self-regulated learning: Goal properties and patterns. *Studies in Higher Education* 42, 11 (Nov. 2017), 2153–2169. https://doi.org/10.1080/03075079.2015.1135117 Publisher: Routledge _eprint: https://doi.org/10.1080/03075079.2015.1135117.

[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] 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

[28] 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

[29] 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.

[30] Traci Sitzmann and Katherine Ely. 2011. A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin* 137, 3 (2011), 421–442. https://doi.org/10.1037/a0022777 Place: US Publisher: American Psychological Association.

[31] Matthias Stadler, Maria Bannert, and Michael Sailer. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior* 160 (2024), 108386.

[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 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

[34] 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

[35] Kelsey Urgo and Jaime Arguello. 2025. Search as Learning. *Foundations and Trends® in Information Retrieval* 19, 4 (2025), 365–556. https://doi.org/10.1561/1500000084

[36] Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing traditional and LLM-based search for image geolocation. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 291–302.

[37] Philip H. Winne and Nancy E. Perry. 2000. Measuring self-regulated learning. In *Handbook of self-regulation*. Academic Press, San Diego, CA, US, 531–566. https://doi.org/10.1016/B978-012109890-2/50045-7

[38] Fan Wu, Yang Dang, and Manli Li. 2025. A Systematic Review of Responses, Attitudes, and Utilization Behaviors on Generative AI for Teaching and Learning in Higher Education. *Behavioral Sciences* 15, 4 (April 2025), 467. https://doi.org/10.3390/bs15040467

[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] Yuyu Yang, Kelsey Urgo, Jaime Arguello, and Robert Capra. 2025. Search+Chat: Integrating Search and GenAI to Support Users with Learning-oriented Search Tasks. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25)*. Association for Computing Machinery, New York, NY, USA, 57–70. https://doi.org/10.1145/3698204.3716446

[41] 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

[42] 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

[43] 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

[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.