Can We Delegate Learning to Automation?: A Comparative Study of LLM Chatbots, Search Engines, and Books

Yeonsun Yang
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
diddustjs98@dgist.ac.kr

Ahyeon Shin
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
ahyeon@dgist.ac.kr

Minchel Kang
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
presidentmc@dgist.ac.kr

Jiheon Kang
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
kangjiheon@dgist.ac.kr

Jean Y. Song
Humanities, Arts, and Social Sciences, Yonsei University
Incheon, Republic of Korea
jeansong@yonsei.ac.kr

Abstract

Learning is a key motivator behind information search behavior [8]. With the emergence of LLM-based chatbots, students are increasingly turning to these tools as their primary resource for acquiring knowledge. However, the transition from traditional resources like textbooks and web searches raises concerns among educators. They worry that these fully-automated LLMs might lead students to delegate critical steps of search as learning. In this paper, we systematically uncover three main concerns from educators' perspectives. In response to these concerns, we conducted a mixed-methods study with 92 university students to compare three learning sources with different automation levels. Our results show that LLMs support comprehensive understanding of key concepts without promoting passive learning, though their effectiveness in knowledge retention was limited. Additionally, we found that academic performance impacted both learning outcomes and search patterns. Notably, higher-competence learners engaged more deeply with content through reading-intensive behaviors rather than relying on search activities.

CCS Concepts

• **Human-centered computing → Interaction paradigms**; *Interactive systems and tools.*

Keywords

Large Language Models, Search-as-Learning

ACM Reference Format:
Yeonsun Yang, Ahyeon Shin, Minchel Kang, Jiheon Kang, and Jean Y. Song. 2024. Can We Delegate Learning to Automation?: A Comparative Study of LLM Chatbots, Search Engines, and Books. In *arXiv September, 2024.* ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Real learning often begins beyond the classroom walls [59]. To deepen their understanding or bridge knowledge gaps left in classes, students embark on a journey of information-seeking behavior, relying on tools such as books and the web to guide their search. Commonly, students visit libraries or search the internet, engaging in mainly four interacting stages: identifying gaps in their knowledge, seeking relevant materials, evaluating and selecting pertinent information, and synthesizing this knowledge to enhance their understanding [51, 67, 72]. This self-directed search and learning is crucial because they allow students to explore topics in depth, personalize their learning experience, and develop the independence and problem-solving skills necessary for both academic success and lifelong learning [52].

However, as students progress into higher education, they need to invest significantly more time and effort into the search-as-learning process due to the increasing complexity and difficulty of course material. This demand may overwhelm students who struggle to dedicate extra time to learning. In response, with the advancement of large language models (LLMs), students are increasingly shifting their primary resources from traditional search tools (such as textbooks and web search engines) to LLM-based chatbots, attracted by their speed, convenience, and ease of use. These automated tools reduce the cognitive burden on learners by scaffolding or replacing specific steps in the search-as-learning process, such as information retrieval, evaluation, and synthesis for knowledge acquisition.

Despite this rising popularity and growth of LLM-based chatbots, educators have sincere concerns about incorporating LLMs into learning. Many believe that for effective learning, students must actively engage with the material by going through trial and error and spending adequate time reflecting on it. They worry that relying too much on the efficiency and convenience of LLMs could lead to over-reliance, ultimately negatively affecting learning outcomes [4, 6, 42, 62]. Consequently, some institutions have banned the use of GPT or blocked access on campus, opting for alternatives such as requiring handwritten assignments or solving problems during class [54, 66]. Despite ongoing research demonstrating the educational potential of LLMs and their gradual introduction into educational settings, these concerns persist.

Our study seeks to examine whether, as some educators caution, the delegation of search-as-learning processes to automated tools is detrimental to the learning outcomes. First, we conducted a survey analysis to understand educators' perceptions of LLM-based chatbots more accurately. From this analysis, we developed a two-dimensional model that explains how specific factors of search tools interact with educators' expectations for successful search-as-learning. The results revealed that educators have three major concerns: (1) lack of reliability, (2) insufficient systematic organization, and (3) weak cognitive engagement.

To investigate whether relying on LLM-based chatbots impacts learning outcomes as educators concern, we conducted a mixed-design empirical study with 90 students and compared three learning tools with varying degrees of automation: books, the web, and ChatGPT. The results show that while the LLM-based chatbot did not affect understanding key concepts, it was less effective than books in supporting long-term retention of the concepts. Despite a significant part of search-as-learning being automated by using LLM-based chatbot, the results show that the students did not adopt a passive learning approach. In addition to comparing the impact of different degrees of automation on learning tools, we further examined the relationship between students' competence level and their patterns in the search-as-learning process. We found that students with higher academic performance achieved better learning outcomes regardless of the tool used. We report distinct patterns in the search and learning strategies between high-competence students and those with lower competence.

Based on the findings, we discuss how LLM-based chatbots can be effectively utilized during the search-as-learning process and propose design implications to consider when developing future educational tools. To summarize our contributions:

- We analyze educators' concerns on automating the learning process through a survey with over 75 educators and identified a two-dimensional in-depth model that shapes effective search and learning.
- We systematically and structurally investigated the validity of educators' concerns of automating the process of search-as-learning through a 3-by-3 within-subject experiment with 92 participants.
- We found that passive learning is influenced more by individual student competence than by the learning tool itself, and through behavioral analysis, we identified specific differences in strategy patterns.
- Based on these findings, we offered suggestions and design implications for how automated tools can be leveraged in helping concept learning.

2 BACKGROUND AND RELATED WORK

2.1 Search as Learning

Human information behavior [76] plays a crucial role in both our work and everyday lives. According to information science literature and theory, information-seeking occurs when there is a need for more information or knowledge reconstruction to resolve a problematic situation [23], such as mitigating the uncertainty (or gap), or addressing an anomalous state of knowledge in the context of problem-solving and sense-making, which are integral to human learning [7, 20, 34, 76]. Research in fields of information science and education has conceptualized information-seeking as a learning process [37, 52], and learning as a key outcome of information search and use [1, 33, 77].

To effectively engage in information-seeking and knowledge construction, learners leverage search systems to navigate the following steps: *(i) query formulation, (ii) material collection, (iii) selection, and (iv) organization* [51, 67, 72]. For example, learners first recognize their information needs and formulate their own queries. Next, they search for and browse relevant information by examining the returned search results. Afterwards, learners evaluate and select content based on relevance and reliability, ultimately transforming this information into meaningful knowledge. While textbooks were traditionally used for this process, the advent of the internet significantly broadened access to information, making web-based learning prevalent [31, 32, 79]. More recently, LLM-based search systems have emerged as a new, highly automated learning tool that not only assists with information retrieval but also offers personalized, adaptive responses that cater to individual academic pursuits. This growing automation has the potential to optimize learning by reducing the cognitive load involved in search, but the implications of such systems on learning outcomes remain underexplored.

Previous research has characterized traditional search systems as tools for learning and investigated their influence on learning outcomes [8, 27, 64, 69, 71]. However, there is limited research on the impact and potential of LLM-based search systems, especially in comparison to traditional methods. The shift from manually driven search processes to increasingly automated systems raises important questions about the role of automation in search-as-learning. This study aims to fill this gap by examining how LLM-based search systems affect both learning gain and search behavior in comparison to other levels of automation.

2.2 Large Language Models in Education

The advances of LLMs that leverage artificial intelligence and natural language processing technologies are rapidly transforming educational environments. Due to their fluency, naturalness in producing language, and versatility, students are increasingly turning to LLMs for support in various academic tasks, including completing homework, writing essays or academic reports, and even searching for information and concepts covered in coursework [38, 50]. As a result, there has been a growing body of research focused on developing LLM-based educational systems and demonstrating their effectiveness [5, 35, 40, 70], with LLM chatbots beginning to be adopted in real-world learning environments [30].

However, there remains considerable debate about the promises and perils of using LLMs as educational tools, and these discussions have sparked conflicts among educators [4, 6, 42, 54, 62]. While prior work has investigated educators' concerns, these are often domain-specific or overly general in nature. Furthermore, there is a lack of empirical research examining how these concerns translate into actual impacts on learning outcomes. Thus, we aim to investigate educators' concerns and explore how these translate into learning outcomes through a structured study.

Additionally, most research has concentrated on evaluating the effectiveness of LLM-driven tools at higher levels of learning, such as applying knowledge to tasks like problem-solving or creative activities (e.g., essay writing, programming support, and solving math problems) [5, 36, 39, 70]. Despite the importance of the foundational phase of learning [48] — where learners integrate new information into existing knowledge structures — this critical phase has been underexplored in terms of how LLMs can effectively support the development of a strong knowledge base. Therefore, our study seeks to assess the role of LLMs in concept learning and knowledge building through interaction with educational content.

2.3 Effective Information Searching Strategies

Educational and cognitive scientists have found that learners retain knowledge better and can apply it across various contexts when meaningful learning takes place. Ambrose et al. [2] introduced the principles of meaningful learning, emphasizing that motivation, metacognition, and self-regulation are critical factors in promoting deep understanding. In the context of search as learning, to foster meaningful learning beyond passive information consumption, comprehensive search practices are proposed, characterized by iterative, reflective, and integrative search sessions [65]. Further research has shown that successful learners—those who achieve higher levels of meaningful learning—exhibit distinct search patterns during the learning process. A commonly observed trend is that time spent reading pages (as opposed to searching) is associated with deeper learning outcomes. For example, [17, 28] shows that learners with high competence levels focus more on content pages, while [78] found that factors such as document retention, query length, and the average rank of selected results could be predictive of domain expertise. Additionally, [16] found that eye-gaze patterns could predict an individual's level of domain expertise, based on the cognitive effort associated with reading.

There has been increasing interest in studying how higher education students' information search and use behaviors affect and support their learning [73, 74, 80]. For instance, [80] reports that common patterns affecting learning outcomes include the reliance on a rudimentary search heuristic, consistently using the same simple search strategy regardless of the context, and habitual topic switching after superficial skimming without evaluating all search results. These growing interests at the intersection of searching and learning highlight the importance of understanding how learning occurs during the search process. However, while this correlation between individual search patterns and learning outcomes is well-studied in traditional search tools such as books or web search engines, the distinct search patterns associated with emerging tools like LLM-based chatbots have not yet been thoroughly explored. Therefore, in this study, we conducted additional analyses (see Section 6) to investigate whether differences in search patterns emerge when using LLM-based chatbots in a search-as-learning context, particularly across varying levels of learner competence, and to identify any notable patterns that may arise.

3 EXPLORATORY STUDY

In the early phase of our research, we conducted a survey study to explore educators' perspectives of different learning tools with varying levels of automation, including textbooks, search engines, and LLM-based chatbots. For a systematic and structured investigation, we particularly focused on understanding the concerns educators have about automation across four steps in the information behavior process: *(i) query formulation, (ii) material collection, (iii) selection, and (iv) organization* [51, 67, 72]. This was primarily to help us develop a set of specific research questions, as listed in Section 3.3.

Table 1: Demographic information of educators we interviewed for the exploratory study

Variable		n	%
Position	University Professor	16	21.33
	High School Teacher	34	46.67
	University Teaching Assistant	20	26.67
	Other	5	6.67
Age (M=39.88, SD=10.97)	20-29	19	25.33
	30-39	18	24
	40-49	22	29.33
	50+	16	21.33
Gender	Male	48	64
	Female	25	33.33
	Not disclosed	2	2.67
GPT Usage	Have used	47	62.67
	Haven't used	28	37.33
Subject Area	STEM	43	57.33
	HSS	25	33.33
	Other	7	9.33

We recruited 75 educators with varying educational backgrounds and experiences through snowball sampling and online advertising (Table 1 shows details of the educators who participated in our survey). In the survey, educators were given a scenario designed to reflect a situation where a student wants to study independently to achieve desired learning outcomes within a limited time frame. *"Dia, a senior university student, is taking 20 credits this semester. Among them, your course is unfamiliar to her, but she is committed to getting an A grade. Despite paying close attention in class, she still doesn't fully understand the materials, so she plans to allocate additional time to study on her own."* Afterward, they were asked to respond to the following two questions: (1) *Among the options of books, Google, and ChatGPT, what do you think are the most effective learning sources (ranked in order), and why is each source more effective compared to others?* (2) *Which steps in the information behavior process should not be delegated to automation (ranked in order), and why must the student perform these steps on their own?* Additionally, two open-ended questions were asked to educators to describe their thoughts on pros and cons of using LLM-based chatbots in a learning environment.

For quantitative analysis, we statistically examined the overall rankings and explored the data from various demographic variables, including position, gender, age, subject area, and GPT usage

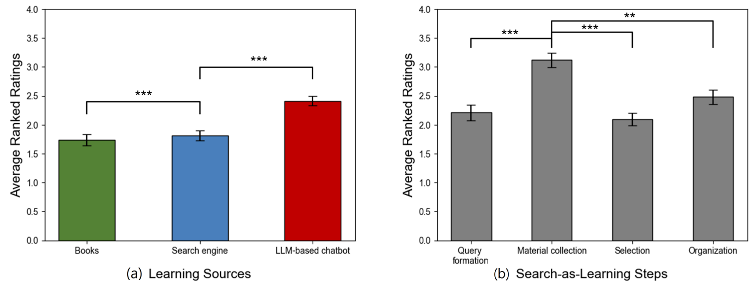


Figure 1: Bar chart showing the average ranked ratings from educators' survey responses. (1 indicates the highest rank). The left plot (a) represents the ranking of three different learning sources. The right plot (b) shows the ranking of the steps in the search-as-learning process where educators believe automation should not be applied. The symbol * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < .001$.**

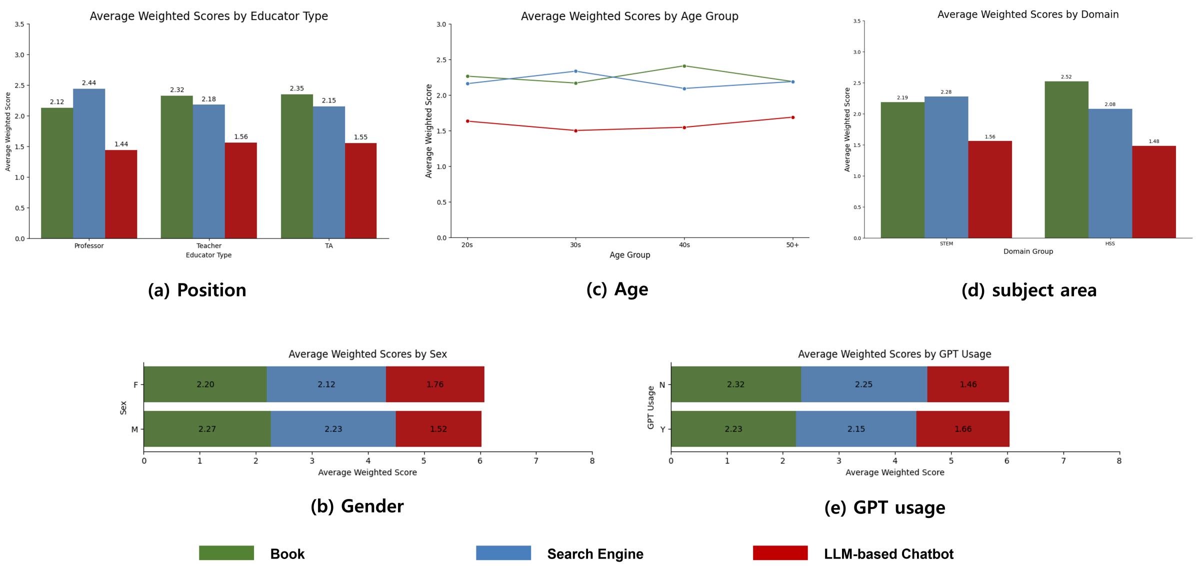


Figure 2: Summary of educators' perspectives on three different learning sources, showing weighted average ranked ratings (1st place = 3 points, 2nd place = 2 points, 3rd place = 1 point). The graphs are divided by various demographic variables: (a) position, (b) gender, (c) age, (d) subject area, and (e) GPT usage. Regardless of the variable, the preference for LLM-based chatbots is consistently low across all demographics..

(Figure 2). For qualitative analysis, three of the authors reviewed the responses and conducted theoretical coding [56] to uncover educators' expectations for effective learning and the key factors of learning sources that support these expectations (as shown in Figure 3). Conflicts were resolved through iterative discussions, and the three authors achieved inter-rater reliability of Krippendorff's alpha 0.89. The remainder of this section presents the findings identified from the survey analysis, and outlines our research questions based on them.

3.1 Statistical Analysis

Through quantitative analysis, we found that the preference for LLM-based chatbots is significantly lower than that for other learning sources, and none of the steps, except for *(ii) collection*, are considered suitable for automation.

Disapproval of ChatGPT is widespread and consistent. We present the comparative results in Figure 1(a) and Figure 2. Given that our data consists of ordinal rankings, we performed Friedman's test [26] instead of ANOVA that is not appropriate due to its assumptions about continuous data and homoscedasticity. Through a Friedman test, we observed that a statistically significant difference exists in ranked ratings across three learning sources ($X^2(2) = 21.62$,

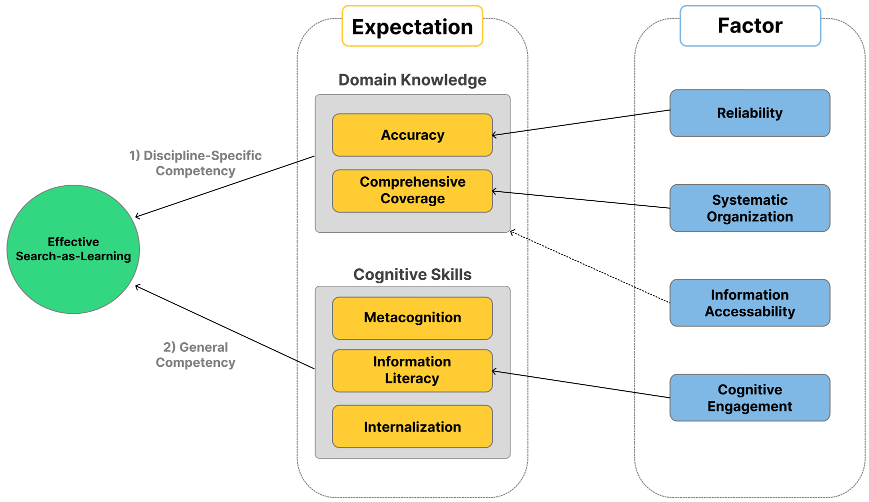


Figure 3: Two-dimensional model illustrating educators' expectations and related key factors for effective search-as-learning. The expectation dimension is based on (1) discipline-specific competency, focusing on accuracy and comprehensive understanding, and (2) general competency, which includes metacognition, information literacy, and internalization, applicable across disciplines. The model highlights learning source factors tied to these expectations, with the exception of information accessibility, which is perceived as covering domain knowledge but ranked lower in priority compared to other factors.

$p < 0.001$). Post-hoc analysis using Conover tests [18] with Bonferroni correction suggests that these differences exist between books and LLM-based chatbots ($p < 0.001$) and search engines and LLM-based chatbots ($p < 0.001$). These results show that compared to two traditional learning sources with lower automation levels, LLM-based chatbots are less preferred by educators. To further explore whether these preferences varied across educators' backgrounds and experiences, we also conducted a multifaceted analysis of the data by position, gender, age, subject area, and their GPT usage experience. As shown in Figure 2, we found consistent patterns in the perceptions of ChatGPT, regardless of the grouping method. While a majority of educators selected books (52%; 39 of 75) and Google (36%; 27 of 75) as the best learning sources, only 12% (9 of 75) chose LLM-based chatbots. Thus, educators are hesitant to adopt LLM-based chatbots in learning contexts and also perceive it as less effective compared to existing learning tools. This highlights a gap between the positive advancements and demonstrated potential of LLM-powered educational tools and the cautious perceptions held by educators.

Query formulation, selection, and organization are perceived as steps that must be carried out by the learners themselves without automation. Most educators (85%) are opposed to delegating any part of the learning process to LLM-driven automation. To discern their perspectives, we analyzed the ranked ratings with a Friedman test and Bonferroni-corrected Conover post-hoc comparisons (Figure 1(b)). There were significant differences between *(ii) collection* and the other steps — query formulation, selection, and organization ($p < 0.001$, $p < 0.001$, $p = 0.017$, respectively). This finding suggests that educators believe these three steps should remain learner-driven, implying sufficient time and effort is required rather than being fully automated. However, there was no consensus on the relative importance among those steps, as perspectives varied, reflecting personal differences.

3.2 Two-Dimensional Model Shaping Effective Learning

To uncover the underlying reasons for the concerns of educators with automated tools, we further conducted a qualitative analysis. Based on survey responses from educators, we derived a two-dimensional model through thematic coding, which comprises educators' expectations for learning and the key factors of sources that help achieve these expectations (Figure 3). This structured model explains how these dimensions interact to form an effective learning process.

3.2.1 Educators' Expectations: Developing Domain Knowledge and Cognitive Skills.

Domain Knowledge: Regarding the disciplines, 23 out of 75 educators expect students to build accurate domain knowledge first and foremost. E27 said, *"knowing the precise keywords is essential for expanding knowledge,"* and E8 added, *"Even if it takes longer, focusing on accurately grasping concepts is more important than quick acquisition."* Besides, E37 and E43 emphasized that acquiring correct information is crucial, especially during the early stages of learning: *"Unless students have developed enough foundational knowledge to identify errors or biases, they should not be allowed to use automated educational tools."* Educators (E7, E9, E15, E20, E41, E46, and E48) also prioritize that students construct their knowledge through comprehensive coverage, when all the fundamental content is thoroughly addressed. Notably, none of them placed importance on the quantity of knowledge. Overall, we found that

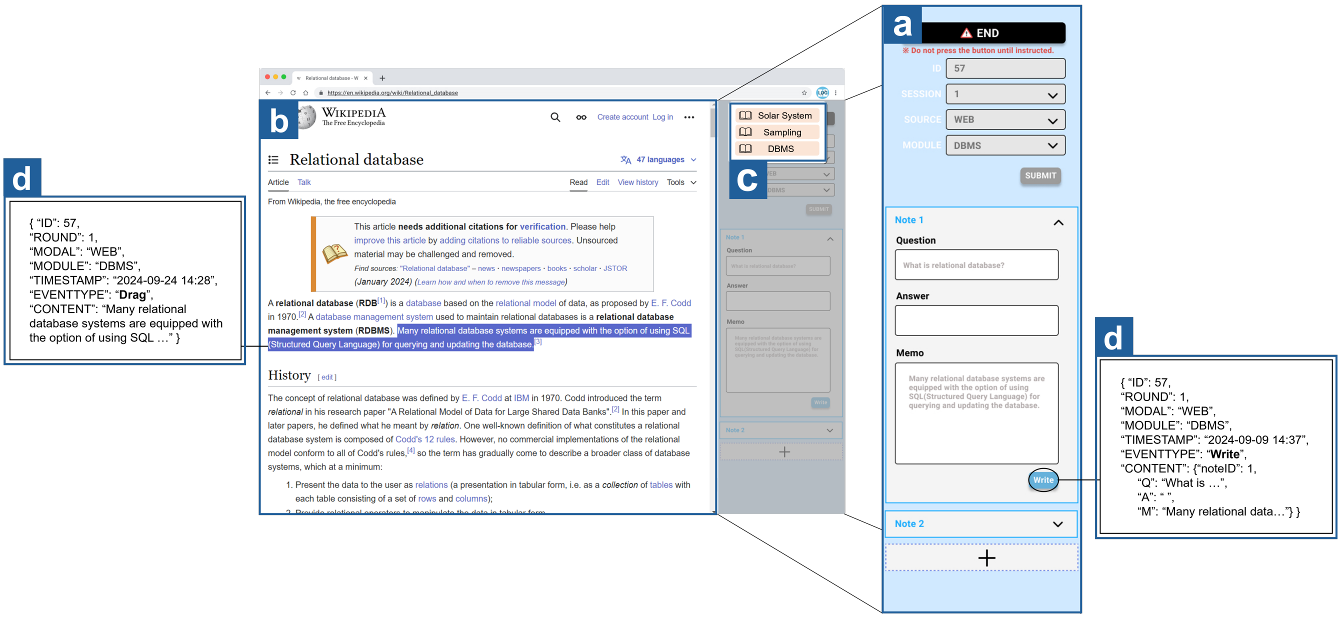


Figure 4: Overview of the SAL (Search-as-Learning) logger, consisting of two interface components: (a) the note panel and (b) the browser window. Additionally, the Chrome extension plug-in (c) was equipped with pre-selected textbooks (e.g., [14, 41, 61] for each module) for each module in PDF format. In the note panel, participants initially input their experimental information (ID, session, source, and module) and then begin a session with the pre-assigned learning source and module pair for the day. First, participants add a new note and write their query in the Question section. As they engage in self-guided search and learning on the browser, they can drag relevant information or organize it into their own words in the memo section. Once they find an answer to their query, participants write the answer in the note and then submit it. They can also add new notes by clicking the '+' button. (d) illustrates the JSON format in which these logs are sent to the server.

the primary competence expected of learners in the course is the accurate and comprehensive construction of knowledge.

Cognitive Skills: On the other hand, we discovered that a majority of them (52%) not only focus on discipline mastery but also expect students to enhance their cognitive skills. This is because, as noted by educators (E7, E13, E35, E68, and E72), these skills are transferable beyond the course and applicable across different disciplines and learning contexts. Specifically, educators emphasize three key abilities that are directly involved in the information-seeking and knowledge-building process: recognizing and addressing gaps in their understanding (metacognition), efficiently searching for and critically assessing large volumes of information (information literacy), and systematically integrating and embedding new knowledge (internalization).

3.2.2 Factors to Build Strong Domain Knowledge: Reliability, Systematic Organization, and Information Accessibility.

Reliability: Educators stated that the accuracy of knowledge depends on the reliability of learning sources. Fourteen of them affirmed that textbooks are a highly reliable source. For example, E13 and E68 explained that books compile the accumulated knowledge of experts who are thoroughly trained in their fields and undergo multiple iterations of refinement through rigorous review processes conducted by top professionals in the domain. Additionally, E37 said, *"GPT learns from processed web content, and web content is derived from books. Therefore, books, as the original source of data, are the most reliable."* Meanwhile, educators (E2, E63, and E70) also regard Google as a reliable tool because of its high-quality materials, such as publications, well-written posts by experts, and Wikipedia. In contrast, **educators (14 of 75) raise concerns about the reliability of ChatGPT's responses**, citing issues such as hallucinations and misinformation [10, 46, 53, 75].

Systematic Organization: Educators described that systematically organized content from educational resources can provide learners with comprehensive coverage of domain knowledge. Books are considered the most effective tool for guiding learners toward structured and inclusive knowledge-building. Specifically, educators (E7, E15, E20, and E21) said that, compared to other tools, books provide content in a systematically organized way: (1) sequential structuring from basic to advanced concepts, and (2) logical progression within topics, such as definitions, explanations, examples, and exercises. Whereas, educators expressed contradictory thoughts about ChatGPT. For example, while E41 and E48 said that *"ChatGPT readily categorizes information and provides responses in a structured format,"* E7 commented that *"ChatGPT is proficient at answering the questions users ask, but its responses often lack continuity and coherence, coming across as disjointed and isolated rather than part of an organic flow and sequential line of reasoning."* This reveals that **educators have conflicting views on whether ChatGPT serves as a systematic tool for supporting comprehensive coverage**.

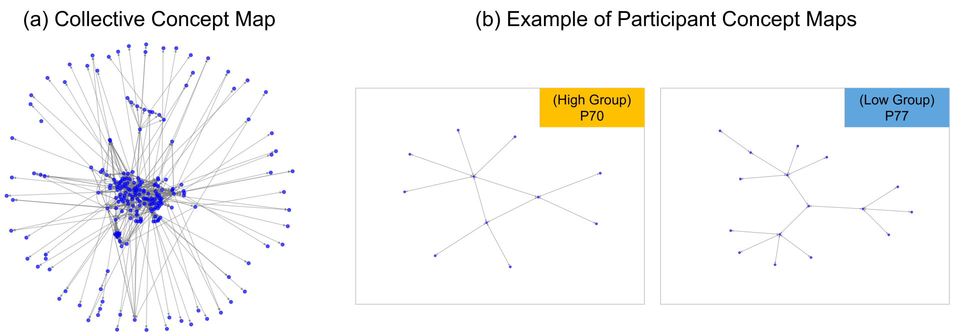


Figure 5: Concept maps on the Sampling module. (a) Collective concept map created by merging the concept maps of all 92 participants for concept map evaluation. (b) Example participant concept maps from both the high-competence group (left, participant P70) and the low-competence group (right, participant P77), as analyzed in Section 6. Each node represents concepts learned during the study phase, and edges represent connections made between these concepts through explanations. Despite differences in group competence, the concept maps of P70 and P77 show similar structural patterns in their concept representations.

Information Accessibility: Nineteen of the educators valued Google as an effective learning tool because of its efficiency in accessing information. For example, E50 and E59 said, *"With the ubiquity of the internet, accessing and searching for information has become much easier and faster, which is likely to be especially useful for students."* Moreover, educators appreciate the ease of access to a wide range of information, from the latest research to posts grounded in human understanding (E3, E36, and E63), along with various modalities beyond text, such as videos and illustrative images (E75). We found that while educators acknowledge ChatGPT's efficient accessibility, they do not encourage its use. This is because educators prioritize the reliability of the source over mere accessibility.

3.2.3 Factors believed to Enhance Cognitive Skills: Cognitive Engagement.

Out of 75 educators, 64 emphasized that certain phases of the search and learning process must be carried out by students themselves, as only through active and repeated participation can cognitive skills be developed and honed. For example, E46 said, *"Without dedicating sufficient time and effort to self-reflection and independent thinking, learning cannot occur."* E26 and E35 also said that *"By engaging in trial and error while evaluating materials and selecting the most relevant information, students acquire valuable know-how in information-seeking strategies."* Additionally, E68 remarked that *"While ChatGPT is a useful tool, it may cause students to passively accept information. The ability to judge and filter valuable information from the irrelevant remains crucial, regardless of AI advancements. Therefore, students should be responsible for identifying key points, selecting relevant information, and organizing it."* Unlike books or Google, which necessitate active participation in the learning process, **ChatGPT has the potential to offload cognitive effort, leading to passive information consumption**. This concern helps explain why most educators view books as the most effective tool, despite the significant time and energy they require, and why educators remain cautious about the implications of LLM-based chatbots.

3.3 Goals and Research Questions

Collectively, we identified three key concerns raised by educators regarding the integration of LLM-based chatbots into educational settings: (1) lack of reliability, (2) insufficient systematic organization, which educators worry may undermine the development of well-structured domain knowledge, and (3) week cognitive engagement, which they believe could impede the cultivation of essential cognitive skills. Given these concerns, the overarching goal of our study is to explore the potential impact of these highly automated LLM-based chatbots on learning outcomes. Specifically, we aim to determine whether their use might negatively influence the quality and effectiveness of learning. To address these goals, we formulated the following research questions:

- RQ1. Does learning with LLM-based chatbots result in **less accurate** and **less comprehensive** understanding of domain knowledge compared to traditional learning sources?
- RQ2. Does learning with LLM-based chatbots encourage learners to spend **less cognitive effort**, leading to passive information consumption?

4 METHODOLOGY

To investigate the gaps between educators' concerns and the actual effects of automation in search-as-learning contexts, we conducted an empirical study with 92 university students. The study aimed to observe the impact of different levels of automated tools on information-seeking behavior. An overview of the study procedure is illustrated in Figure 6.

Figure 6: Procedure of the mixed-method within-subject study with 92 students.

4.1 Study Design

In this work, we designed a counterbalanced mixed-method study that incorporates elements of both within-subjects and between-subjects designs. The study consisted of three sessions, each conducted on a separate day. In the between-subjects component, we compared learning outcomes across three types of sources: books (representing non-automated searching), web search engines (offering partial automation), and LLM-based chatbots (capable of fully automated searching). In the within-subjects component, each participant used a different learning sources in each session, thus experiencing all three conditions over the course of the study. The study conditions were defined as follows:

- **Condition 1: Book.** To ensure a uniform comparison with the other experimental conditions, we simulated a digital reading environment by providing textbooks in PDF format. In order to replicate the experience of reading physical textbooks, digital features such as keyword search were disabled, requiring participants to manually navigate and read through the material.
- **Condition 2: Web.** Participants use the web search engine to seek information by entering relevant keywords and accessing content on various websites. They are restricted to using Google, the most widely used and familiar search engine among participants, to ensure consistency in the data and insights gathered from our analysis.
- **Condition 3: ChatGPT.** Participants are restricted to use ChatGPT as LLM-based chatbots to search for information through a question-answering form. We selected ChatGPT-4o [60] because of its enhanced speed and multimodal capabilities, allowing efficient delivery of both text and visual materials.

To eliminate potential learning effects across sessions, we selected three distinct university-level modules from STEM subjects: (a) The Solar System (Astronomy), (b) Sampling (Statistical Mathematics), and (c) Database Management Systems (Computer Science). These modules were randomly assigned to the three learning tools, ensuring that each tool was paired with a different module across the three sessions. In addition, we defined three sequential learning objectives (LOs) for each session, which participants were required to meet. These objectives were aligned with the three levels of Bloom's taxonomy (Understand, Apply, Analyze) [9, 48]. A detailed summary of the learning objectives (LOs) and the rationale for selecting these levels of Bloom's taxonomy is provided in the Appendix A.

4.2 Task Design and Setup

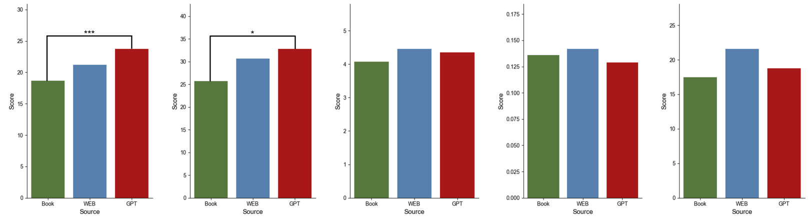
To address the research questions outlined in Section 3.3, the data collection and evaluation was conducted in two phases. The first phase focused on assessing participants' knowledge comprehension gained from each learning source (RQ1). The second phase examined their learning activities and experiences throughout the sessions (RQ2).

4.2.1 Measuring knowledge Accuracy and Comprehensiveness. According to the model of comprehension [15, 43–45], learners construct and integrate knowledge through multiple levels of mental representation. Specifically, learners first extract key concepts from the materials into working memory, then formulate propositions from those concepts. Finally, they integrate these propositions into a coherent mental model. Based on this, we designed two learning tasks to evaluate participants' knowledge comprehension along three hierarchical levels — concepts, connections, and the development of a coherent mental model.

- **Concept Map Drawing:** Concept maps are one of the most widely used tools for approximating learners' understanding within a particular domain or course material [13]. In this study, we employed Novakian concept mapping [12], a method often used to capture a learner's mental model by representing a network of connections between related concepts [19, 55, 57, 58, 63]. For example, participants map out their understanding of the Database Management Systems (DBMS) by including MySQL, Oracle, and MongoDB, using the relationship "example of" to logically connect them.
- **Post and Retention Test:** To assess the accurate and persistent construction of a mental model for essential knowledge aligned with our LOs, participants perform an immediate post-test after each session and a retention test two weeks later. The post-tests consist of a set of nine multiple-choice questions (MCQs), one of the most commonly used forms of assessment [11, 68]. The retention test consists of the same set of nine MCQs, with both the question order and the options randomized from the post-test. We set AI-generated (GPT) MCQs for the tests, as prior research has shown that LLMs can effectively generate high-quality MCQs that are well-aligned with specific LOs and comparable in quality to those crafted by experts [21, 22]. The process of automatic MCQ generation is detailed in Appendix B, including our prompt engineering, iterative quality evaluations, and examples.

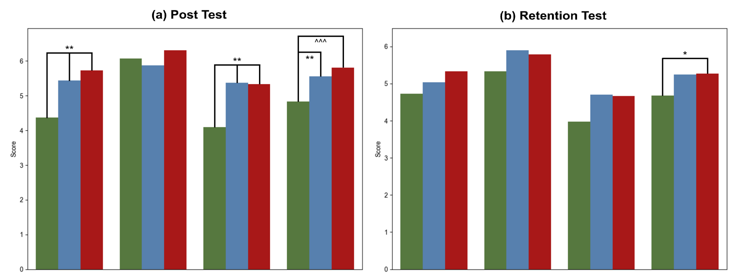
4.2.2 Interface for Learning Activity Logging. The SAL logger was instrumented as the apparatus, shown in Figure 5. This system records timestamped user interactions within the browser (e.g., keyboard input, mouse clicks, and dragging) to analyze learner activeness and behavior. We developed the SAL logger as a Chrome extension using HTML and JavaScript. It has a client-server architecture that enables authentication, stores search histories, collects learning logs, and provides study materials in PDF format. The server is implemented using Node.js and stores data in JSON format.

The log data collected by the system consists of the following key attributes: participant information (ID, session, assigned source, and module), timestamp, event type, and content. The event type field categorizes interactions, including *Info* (used to track when experimental information is submitted), *Drag* (to track which content



		Number of Nodes	Number of Edges	Level of Deepest Hierarchy	Node Degree	Edge Consensus
Book	*M*	18.67	25.69	4.07	0.136	17.49
	SD	7.71	15.42	1.06	0.044	11.00
Web	*M*	21.21	30.64	4.46	0.142	21.60
	SD	9.34	18.54	1.65	0.048	14.19
GPT	*M*	23.76	32.83	4.35	0.129	18.75
	SD	10.00	17.45	1.29	0.042	9.70

Figure 7: Results for the average concept map evaluation across five metrics. Significant differences between the book and GPT groups are observed only in the number of nodes and number of edges metrics. The symbol * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < .001$.**



		(a) Solar System	(a) Sampling	(a) DBMS	(a) Average		(b) Solar System	(b) Sampling	(b) DBMS	(b) Average
Book		4.37	6.07	4.09	4.83	*M*	4.72	5.33	3.97	4.67
		1.61	1.33	1.35	1.67	*SD*	1.81	1.77	1.56	1.78
Web		5.44	5.87	5.37	5.55	*M*	5.03	5.90	4.70	5.24
		1.61	1.33	1.54	1.27	*SD*	1.36	1.63	1.39	1.57
GPT		5.73	6.31	5.33	5.80	*M*	5.33	5.78	4.66	5.27
		1.67	1.50	1.27	1.55	*SD*	1.77	1.43	1.50	1.61

Figure 8: Overall results for the post-tests and retention-tests, presented both by individual modules and as an overall average. (a) The post-test results show significant differences between the book and GPT groups, as well as between the book and web groups, in the solar system and DBMS modules, as well as in the overall average, with the exception of the sampling module. (b) In the retention test, significant differences between the book and GPT groups are observed only in the overall average. This suggests that books have an advantage over automated tools in terms of retention. The symbol * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < .001$.**

was dragged), *Write* (to log entries made in the note panel), and *Web_url* (to capture website access when the source is web-based). The content field contains data specific to each event type, such as dragged text for a *Drag* event or a written note for a *Write* event.

4.3 Data Collection and Analysis

To analyze participants' search and learning patterns, the SAL logger automatically collected keyboard and mouse events, along with the content they interacted with (e.g., dragged text or clicked links) throughout the session, as described in Section 4.2.2. Additionally, the entire experimental process was video recorded to capture participants' real-time interactions with the system. Following this, participants' concept maps, created using a pen-and-paper approach, were collected and digitized using the NetworkX Python library [29] for quantitative analysis. At the end of the study, a post-test and survey were administered, which included open-ended questionnaires and Likert scale questions regarding participants' perceptions and experiences with the three different learning sources, learning gains, and insights for future design implications of LLM-powered tools for learning. Two weeks later, a retention test was conducted.

To quantitatively measure participants' learning gain, we analyze both the digitized concept maps and the post- and retention test results (scaled from 0 to 9). For assessing the concept maps based on content and structure, we employed five network analysis metrics: number of nodes, number of edges, deepest hierarchy level, node degree, and edge consensus. The first three metrics are simple count-based metrics present in the maps, indicating the number of acquired concepts, the connections made between them, and the depth of knowledge structuring. While these traditional metrics are commonly used in the literature to predict an individual learner's understanding, they have limitations in assessing correctness and comprehensiveness. To supplement this, we incorporated two additional metrics — node degree and edge consensus — based on a group consensus approach, comparing an individual learner's conformity to the larger group's collective understanding [25]. Node degree measures the number of edges connected to a node, while edge consensus evaluates the number of overlapping connections made by other learners. To apply these metrics, we merged all participants' maps (see Figure 6(a)), and measured how accurately individuals identified key nodes and edges, as determined by their peers. Previous research supports the use of these metrics, showing that a group's collective mental model can approximate that of an expert [3, 24, 47, 49]. Furthermore, to assess the degree of engagement throughout the sessions, we measured the number of submitted notes, as well as the average time and number of cognitive activities (i.e.,searching, reading, dragging, navigating, and note-taking) per cycle in the search-as-learning process, based on the collected logs with timestamps. We employed the Bonferroni correction for all statistical tests to avoid potential multiple comparison problems.

4.4 Particpants and Procedure

4.4.1 Participants. We recruited 92 participants from our university mailing lists and through online advertisements on social media (age=21±3, 46 males and 46 females). To control for prior knowledge, we recruited participants exclusively from our institute, where the modules used in our study are not part of the standard curriculum. We also asked participants to complete a pre-test consisting of three MCQs at the 'remember' level, which aligns well with assessing prior knowledge by requiring participants to retrieve relevant information from long-term memory [48]. We excluded applicants who scored three on the pre-test. The scores of participants were M=0.55, SD=0.68. Participants were randomly assigned to experimental conditions, and no significant differences were observed between the conditions in the pre-test (Kruskal-Wallis H=0.675, p=0.7).

4.4.2 Procedure. The study was conducted in a controlled setting, either in person or online. The study lasted for three days with 90-minute sessions each day, followed by a 30-minute session two weeks later. Participants received 66,000 KRW (i.e., approximately 49.2 USD) as compensation. Additionally, those whose test results ranked in the top 10% were offered a 20% incentive to further motivate their performance. The study protocol was approved by our institution's IRB, and all study materials used in the study were translated into Korean to prevent any language barriers and reduce unnecessary cognitive load.

The study procedure was organized into five phases (Figure 6). During the study phase, participants first installed the SAL logger on the Chrome browser. They were then asked to study the assigned learning objectives using a designated learning source. Participants followed a structured study process designed to track their search-as-learning behaviors. First, participants were instructed to formulate their own queries if they identified any knowledge gaps or internal questions that arose during the study to achieve the LOs. These questions were recorded in the question section of the note panel. To answer the query, participants conducted information searches using the learning source through the browser window. Participants were guided to move the highlighted cursor along with their gaze while browsing, and were asked to drag sections of text they focused on. Any key insights or pertinent information discovered during the search could be transferred or restructured in the memo section of the note panel for further clarification. Once participants felt that they had resolved their knowledge gap, they were asked to submit their answers in the answer section of the panel. After submitting the completed note, participants added a new note to the panel and repeated the process throughout the study phase. Previously submitted notes could be modified and resubmitted later.

Next, participants were given 5 minutes to review and reflect on their notes. Afterward, they moved on to the concept mapping phase, which lasted 20 minutes. During this phase, participants hand-drew labeled nodes to represent concepts and linked them to illustrate the relationships between the concepts they had acquired during the session. Following this, participants entered the test phase, where they had 15 minutes to complete a post-test consisting of 9 MCQs related to the module. After completing the test, participants used an online form to answer a post-survey, which included several questions about their experience and the effectiveness of each search tool for learning. This process was repeated over three consecutive days, and two weeks later, participants were asked to complete a 30-min retention test.

5 RESULTS

Overall, a comparative analysis of the outcomes from the three experimental groups (Book vs. Web vs. GPT) reveals that LLM-based chatbots were competitive at lower levels of understanding (i.e., formation of concepts and connections) and outperformed books in helping learners build coherent mental models. However, LLMs led to reduced retention of information compared to books (RQ1). Additionally, our results show there were no significant differences in the average cognitive effort expended per search-as-learning cycle among different learning sources (RQ2). The following outlines our findings for each research question, along with participants' perspectives and insights directly comparing their experiences with each learning source.

5.1 RQ1: Accuracy and Comprehensive Coverage in Knowledge Building

To investigate whether LLM-based chatbots hinder learners' understanding in terms of accuracy and comprehensive coverage, we first analyzed 276 concept maps (92 participants x 3 conditions) to assess fundamental level of understanding, which indicates how well learners identified and connected key concepts aligned with the given learning objectives (section 5.1.1). Next, we analyzed post-test and two-week-after retention test scores to assess a more advanced level of understanding (section 5.1.2), focusing on how well the knowledge was retained and applied within their mental models. Detailed module-level results are provided in Appendix C.

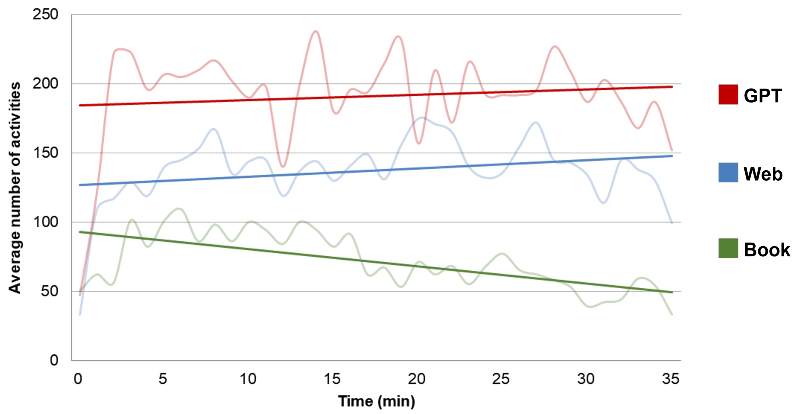
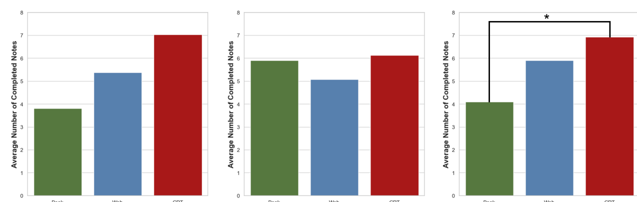


Figure 9: Average number of activity logs per condition, collected using the SAL logger during the 40-minute study phase. The time axis is normalized to 40 minutes for all participants. This graph shows the overall activeness across the three conditions. The GPT group generated the highest number of event logs, followed by the Web group, while the Book group recorded the fewest activity logs on average. Notably, the GPT and Web groups exhibited a slight upward trend in activeness over time, whereas the Book group showed a decreasing trend in activity. (For accurate activeness analysis, only the first 35 minutes of data are displayed, as participants were informed to conclude their study five minutes before the end of the session.



		Solar System	Sampling	DBMS
Book	*M*	3.80	5.90	4.09
	SD	2.25	3.00	3.03
Web	*M*	5.38	5.07	5.90
	SD	4.42	2.73	3.20
GPT	*M*	7.03	6.13	6.93
	SD	7.29	4.01	4.77

Figure 10: Average number of completed notes submitted by participants during the 40-minute study phase for each module, indicating the number of search-as-learning cycles completed. Post-hoc test results reveal a significant difference between the Book and GPT groups in the DBMS module. The symbol * indicates $p < .05$, ** indicates $p < .01$, and * indicates $p < .001$.**

5.1.1 Concept Map Evaluation. The results (Figure 7) show that the average number of nodes was significantly greater in the GPT group compared to the book group (GPT: M=23.76, SD=10.00; Book: M=18.67, SD=7.71; p<0.001). Similarly, the average number of edges was also significantly greater in the GPT group (GPT: M=32.83, SD=17.45; Book: M=25.69, SD=15.42; p<0.05). These results indicate that participants in the GPT group acquired more concepts and connections than those using books, and this trend was consistent across all three modules. To further understand how participants structured their acquired knowledge, we analyzed the structures of their concept maps. On average, there were no significant differences in the depth of maps across the three groups (p>0.1). However, at the module level, the GPT group built significantly deeper maps in the DBMS module compared to the book group (GPT: M=4.8, Book: M=4; p<0.05). Additionally, no meaningful differences were found in node degree or edge consensus (p>0.1). Although none of the results reached statistical significance when broken down by module, we observed that in the DBMS module, participants in the GPT and web groups identified more correct edges than those in the book group. Taken together, these results suggest that LLM-based chatbots help learners identify fundamental concepts and accurately connect them, comparable to traditional learning sources. Despite acquiring more concepts and edges in a given time, learners using GPT structured and understood these relationships just as effectively.

5.1.2 Post-Test and Retention-Test Redsults. As shown in Figure 8, immediate post-test scores were significantly lower in the book

Table 2: Demographic information of participants in high group and low group based on their GPA.

High Group				Low Group			
PID	Gender	Age	GPA	PID	Gender	Age	GPA
P18	F	25	4.30	P23	M	26	2.90
P28	M	23	3.90	P47	M	18	2.70
P50	F	22	3.90	P48	F	22	3.00
P53	M	28	3.99	P59	M	19	2.91
P65	F	22	4.03	P74	F	21	2.93
P66	M	28	4.22	P77	M	25	2.70
P70	M	20	4.14	P80	M	15	2.88
P85	M	20	4.00				

group compared to both the web group and the GPT group. Notably, these significant differences were observed in the astronomy and DBMS modules. This indicates that participants using books fell short in achieving equivalent test scores compared to the two higher levels of automated learning tools. We hypothesize that the non-automated nature of books may have hindered participants from building strong, coherent mental models within the limited session time, as the manual process of searching and learning was less efficient compared to the automated, streamlined approaches offered by the web and GPT tools.

In contrast, the two-week retention tests show no significant difference between the book and web groups, and the average difference between the book and GPT groups was reduced compared to the post-test results. In fact, when looking at the retention test scores broken down by module, we found no remarkable differences among the three groups. This suggests that while the book group initially struggled to achieve high scores, they may have retained the learned knowledge better over time, indicating an advantage in retention compared to the more automated learning sources. Therefore, while LLM-based chatbots may be less effective than books in terms of retention, they do not hinder learners' ability to accurately and comprehensively construct understanding in self-directed search environments. This suggests that, despite concerns about reliability and systematic organization, these highly automated tools did not appear to negatively impact knowledge comprehension, addressing our RQ1.

5.2 RQ2: Assessing Degree of Cognitive Engagement

To investigate whether LLM-based chatbots promote passive information consumption rather than active engagement into the search-and-learning process, we first refined 12,287 activity logs collected throughout the study. From this data, we analyzed participants' activeness levels during information-seeking behaviors and compared the number of completed notes. By understanding these two factors together, we aimed to assess the degree of cognitive engagement across different learning tools.

5.2.1 Assessing Activeness Levels. To assess activeness during the sessions, we quantitatively analyzed the average number of cognitive activities such as searching, navigating, dragging, clicking and note-taking for each participant over time (Figure 9). Surprisingly, contrary to educators' expectations, the results showed that the GPT group exhibited the highest level of activeness throughout the sessions, followed by the web group, and lastly the book group. Additionally, the trendline for participants using books showed a noticeable decline in activeness as the session progressed. In contrast, the activeness levels for participants using both GPT and web search engines remained steady, indicating sustained engagement throughout the session. From this observation, we suggest that the effort required to manually complete all steps with books may have caused participants to fatigue more quickly, resulting in decreased activeness in the latter part of the session. This contrasts with the GPT and web conditions, where automated features likely reduced the cognitive burden and helped participants maintain their engagement over time. However, these results alone cannot fully address our RQ2, as they may not necessarily reflect active cognitive engagement. Therefore, we further analyzed the number of completed notes within each session to assess the time and effort invested in each search-as-learning cycle. By examining the quantity of notes relative to the activeness level, we aimed to determine whether participants were deeply engaged in the learning process or merely exhibiting frequent interactions without substantial cognitive effort, as discussed in the next subsection.

5.2.2 Comparison of completed notes. To investigate whether the number of completed notes differed between learning sources, we performed one-way ANOVA tests (Figure 10). The results showed no significant main effect of learning sources for the solar system module ($F(2,89) = 3.041$, $p = 0.053$) or the sampling module ($F(2,89) = 0.867$, $p = 0.424$), but a significant difference was found for the DBMS module ($F(2,89)=4.611$, $p=0.012$, $\eta^2=0.094$). However, despite GPT's time-efficiency, no remarkable differences in the number of notes completed were observed compared to other sources. Similarly, the search engine, a partially automated tool, did not show a meaningful difference from the book group. This suggests that participants engaged in each search-as-learning cycle with similar time and effort, regardless of the level of automation in the learning source. Notably, although the activeness level for LLM-based chatbots was higher than for other sources, the number of completed notes did not differ significantly. This challenges educators' concerns that fully automated tools like GPT may encourage passive learning approaches. Therefore, our findings support RQ2, suggesting that LLMs do not reduce the cognitive effort invested compared to traditional sources.

5.3 Summary

Through the study, we found that LLM-based chatbots did not hinder understanding of domain knowledge; in fact, participants absorbed more content overall. Moreover, automation levels did not deactivate participants' search and learning behaviors. On the contrary, using non-automated tools throughout the entire process may lead to quicker exhaustion. Nonetheless, this self-navigated tool like books proved effective in promoting long-term memory. Given the lack of significant differences in learning outcomes across the tools, we conducted additional analysis in the next section on LLM-based search-as-learning patterns, focusing on participants' GPA. We explore whether academic performance influenced the

Table 3: Summary of LLM-based search-as-learning strategies. Based on [73, 74, 80], we defined search-as-learning patterns into read-related and write-related strategies. After analyzing 17 video recordings, the strategies were categorized through three iterations of behavior coding performed by four authors. These strategies capture the key behaviors observed in participants during the search and learning process, highlighting how learners interact with content through reading (TR, RV, SN) and writing (PH, TP, CP).

Mode	Strategy	Description	Observation
Read	Thorough Reading (TR)	Engaging deeply with the material	Dragging the mouse to highlight sections of text while reading from start to finish
	Revisiting (RV)	Returning to the content	Reviewing a previous response received
	Scanning (SN)	The simple act of reading	Reading content without mouse interaction
Write	Paraphrasing (PH)	Process of restating text	Organizing and summarizing self-learned content
	Typing (TP)	Typing the content without modifications	Typing the content as it was found, following along while reading without making any modifications.
	Copy-Pasting (CP)	Simply copying and pasting	Selecting and copying the content by dragging without making any modifications

effectiveness of each learning tool. This allowed us to identify potential differences between low-GPA and high-GPA learners.

6 EXPLORING THE IMPACT OF STUDENTS' COMPETENCE ON LLM-BASED SEARCH-AS-LEARNING

A rich body of research has shown the correlations between students' search strategies and their learning outcomes [16, 17, 28, 73, 74, 78, 80]. Drawing from the literature, we hypothesize that students with higher academic performance are likely to use more effective strategies that foster meaningful learning [2] regardless of the automation level of tools in search-as-learning contexts. To further investigate this, we conducted an additional in-depth analysis by quantitatively analyzing participants' learning outcomes and video recordings. This investigation revealed two key findings: (1) a correlation between students' competence and their knowledge acquisition, and (2) distinct search-as-learning strategies across different competence levels. These results provide valuable insights into how LLM-based search tools can be designed to support effective and meaningful learning, particularly for students with varying levels of competence.

6.1 Higher-Competence Students Exhibit Deeper Cognitive Investment and Retention

To examine the correlation between students' competence levels (as reflected by GPA) and learning outcomes, we performed a Pearson correlation test. While no significant correlations were found between GPA and concept map evaluations across all five metrics for knowledge comprehension, we observed a significant correlation between academic performance and post- and retention-test results, which represent higher-order understanding.This suggests that while basic understanding did not vary significantly by GPA, students with higher academic performance were more likely to develop and retain a cohesive mental model.

To further explore students' cognitive engagement, we analyzed correlation of activeness level with GPA. The results indicated no significant association between them. However, we found a notable correlation between GPA and the number of completed notes, suggesting that higher-performing students tended to invest more time and effort into each search-as-learning cycle. These findings imply that competence level influences search strategies, specifically in how actively students engage in learning tasks, potentially leading to more comprehensive understanding of domain knowledge. To better capture these patterns, we divided participants into high- and low-performance groups and analyzed their behaviors through video recordings, which will be detailed in the next section.

6.2 High-Competence Students Prioritize Deep Reading, While Low-Competence Focus on Task Completion

To explore the differences in information searching behavior patterns based on students' competence, we divided the participants into two groups: the high group (GPA $\geq$ 4.0) and the low group (GPA $\leq$ 3.1). The demographics of each group are presented in Table 2. Drawing from strategies reported in prior research on shaping meaningful learning, four of the authors conducted two rounds of iterations analyzing 17 videos of our participants engaging in search and learning using large language models (LLMs). Through this process, we identified six common strategies: three related to reading and three related to writing (see Table 3). We

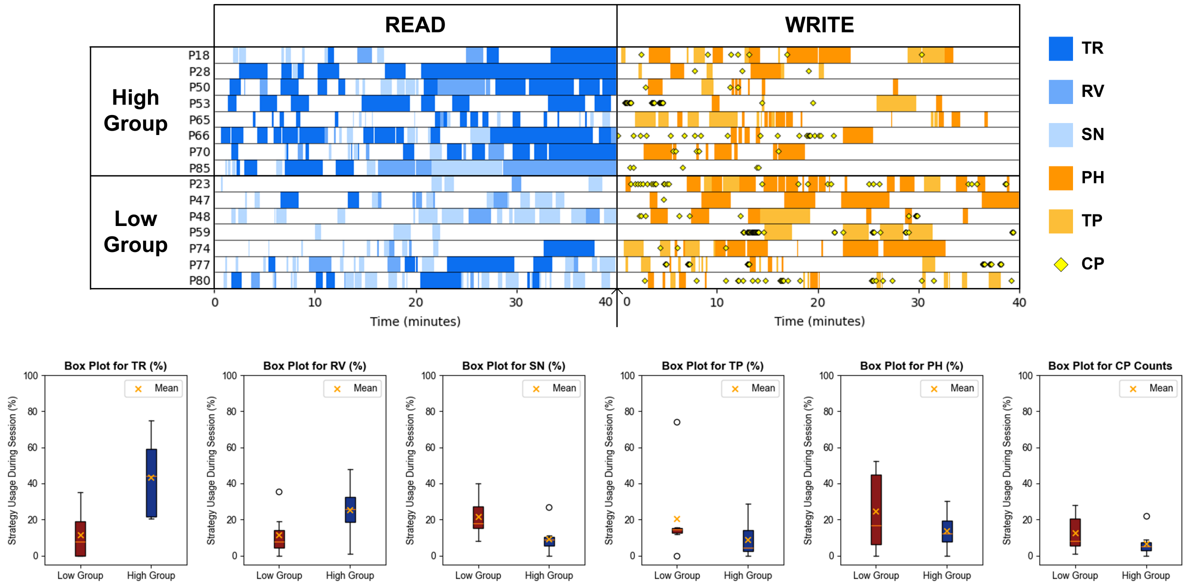


Figure 11: Search-as-learning patterns of participants in both high-competence (N=8) and low-competence (N=7) groups, derived from video recording analysis. The top portion illustrates participant strategies over the 40-minute experimental session, annotated with three read-related strategies (TR, RV, SN) and three write-related strategies (PH, TP, CP), as defined by four authors. Each block represents the occurrence of these strategies during the study phase.The bottom box plots depict how frequently each group employed these strategies over the course of the 40-minute session, expressed as a percentage of the total time, except for CP, which is shown as the number of occurrences since it is measured in seconds. In the Read mode (left side), the high-competence group frequently utilized TR and RV strategies, while the low-competence group predominantly used SN. In the Write mode (right side), the low-competence group spent more time on writing tasks (e.g., formulating queries, writing answers, and taking memos) compared to the high-competence group. Overall, the low-competence group maintained a consistent read-write process throughout the session, while the high-competence group increasingly focused on reading strategies as the session progressed.

used these strategies to code the videos, resolving any conflicts through discussion.

As shown in Figure 11, we observed distinct patterns between the groups in terms of both reading and writing behaviors. First, with regard to reading, the high group dedicated significantly more time to thorough reading (TR) and revisiting (RV), while the low group primarily focused on scanning. Additionally, as the sessions progressed, the high group displayed an increasingly reading-intensive pattern. Next, in terms of writing, the low group spent slightly more time on paraphrasing (PH) and typing (TP). However, we also found that copy-pasting (CP) occurred more frequently in the low group. Overall, the high group demonstrated a strategic approach to time allocation. Early in the session, they spent considerable time on information searching, synthesizing, and note-taking, but later shifted their focus to thoroughly reading and revisiting previous content to deepen their understanding. In contrast, the low group consistently alternated between reading and writing throughout the session, primarily concentrating on task completion through note-taking.

These findings suggest that the high group prioritizes internal processes during search as learning, whereas the low group seems to perceive the externalization of knowledge as the primary driver of learning gains. Synthetically, we propose design guidelines aimed at leveraging the strategies observed in the high group as benchmarks, while implementing guardrails to support the low group in effectively using LLM-based systems as tools for learning in Section 7.2.

7 DISCUSSION

7.1 The Promise of LLMs as a Starting Point for Search as Learning

Our study showed that LLM-based learning enabled learners to acquire a greater number of key concepts and their connections compared to other learning sources, without compromising the accuracy of their knowledge. Furthermore, we found that LLM-based chatbots effectively scaffolded comprehensive understanding of domain knowledge within a limited time frame. This suggests that LLMs are particularly effective for quickly grasping new concepts in a short period of time. However, in terms of retention, traditional books outperformed LLMs and search engines. This may be because the cognitive processes involved in reading books are more self-guided, and the structure is designed in a way that experts in the field believe is optimal for learning. While this high cognitive load can enhance retention, it can also lead to fatigue and reduced engagement, especially when compared to more automated tools. Based on these findings, we propose a learning approach where learners, particularly those lacking background knowledge, begin with LLMs to efficiently grasp key concepts and build an initial overview. Once this foundation is established, transitioning to book-based learning can further support retention and deepen understanding. This method makes engaging with books easier than starting from a point of unfamiliarity, ultimately maximizing learning outcomes. Several participants supported this approach with their feedback. Participant (P11) noted, "There were too many sections to look through, making it difficult," while another (P23) commented, "I felt tired and frustrated by the process." Additionally, one participant (P65) emphasized the benefits of starting with GPT, stating, "Using GPT in the early stages of learning was helpful because it provided answers even to abstract questions, which made it easier to navigate the material."

7.2 Design Implications Derived from Learner Competence Levels

Based on our analysis of search strategies, we found distinct interaction patterns between high-performing and lower-performing learners. The high-competence group increasingly focused on reading behaviors as the learning session progressed, while the low-competence group maintained a consistent pattern of alternating between writing and reading throughout each search-as-learning cycle. While it may be challenging for users to consciously adopt these patterns on their own due to the cognitive burden, we propose design implications for future LLM-based search systems that intuitively guide users to emulate the search behaviors of higher-competence learners. These systems would help users follow effective strategies naturally, while discouraging less productive behaviors observed in the lower-competence group. The design implications are as follows.

7.2.1 Support active reading. To encourage learners to invest time and effort into thoroughly reading materials, knowledge externalization strategies are a promising approach. Learning from text is inherently constructive, but as texts become more complex, building a cohesive mental model becomes increasingly challenging. Existing tools often promote passive reading, offering little support for engaging deeply with content. However, strategies such as diagramming, note-taking, and providing effective markdown formatting can facilitate this challenging process, helping learners actively organize and retain information. Therefore, we propose designing tools that foster more interactive engagement with content. These tools should allow users to modify the material directly, such as by annotating and creating diagrams from the searched content, which could enhance retention and comprehension.

7.2.2 Navigate learning histories. A commonly observed pattern in the lower-competence group is their tendency to initiate new interactions with LLMs rather than referring back to previously searched results. However, recalling previous interactions is a valuable strategy for enhancing retention and comprehension. The challenge lies in the cognitive effort required to remember past interactions, scroll back through long sessions, and revisit earlier content, which can be burdensome for learners. To address this, we propose implementing systems that allow users to easily access and navigate their learning histories in a more efficient and visually intuitive way, rather than relying on tedious scrolling. A clear, summarized overview of previous interactions, such as visual timelines or knowledge maps (e.g., tree-based structure), could help learners reference past material and seamlessly integrate it into their ongoing search and learning processes. Such tools would not only support better recall but also enhance continuity between sessions, enabling users to build upon previous knowledge more effectively.

7.2.3 Visualization of progress. Higher-competence learners tend to be driven by intrinsic motivation, which fosters self-directed, deeper learning. In contrast, lower-competence learners often rely more on external motivation, which explains their more task-oriented behavior. For these learners, seeing clear, tangible signs of progress is crucial for feeling a sense of accomplishment and learning gain. This is partly because reflective and integrative search strategies pose significant challenges for them. Therefore, a tool that visualizes their learning progress could be highly beneficial. By providing visual indicators of progress—such as milestones achieved, time spent on tasks, or key concepts covered—such a system would offer immediate feedback, helping learners gauge their performance. Research has shown that progress visualization not only boosts engagement but also fosters a sense of control over the learning process, making it easier for learners to overcome the difficulty of self-directed search and learning approaches. This would help guide them toward more productive and reflective learning behaviors, while easing the cognitive burden of tracking their own development.

7.3 Towards the Complementarity of Search Engines and LLMs

With the emergence of tools like Perplexity AI and Search GPT, integrating web search engines with LLMs offers a way to enhance both the reliability and efficiency of information retrieval. While LLMs can generate nuanced responses and assist with query formulation, they are prone to inaccuracies. In contrast, web search engines provide verifiable and traceable results. By combining the two, we can leverage the LLMs for generating and refining search prompts, while using search engines to ensure the information retrieved is accurate and reliable.

To maximize the benefits of this hybrid approach, users must develop strong prompt formulation skills, as crafting effective queries can deepen cognitive engagement and learning. Future systems that blend LLM-generated suggestions with reliable search engine results could help guide users toward more accurate and reflective search behaviors, enhancing both learning and information accuracy.

7.4 Limitations and Future Work

First, our study was not conducted longitudinally, which may have limited our ability to fully capture the modality's characteristics and more natural student usage within a 40-minute session. In particular, we may not have fully observed the advantages of using resources that require longer study times, such as textbooks. Future work could involve observing long-term usage patterns to address this limitation.

Second, our study was limited to a university-level STEM course, which focused on structured, fact-based content. This may have reduced the likelihood of encountering issues like hallucination or misinformation, which could be more prevalent in subjects that involve complex interpretations (e.g., humanities) or requires state-of-the-art knowledge not shared in the web (e.g., state-of-the-art semiconductor processing technology). Future work should explore LLM performance in a broader range of subjects, including the humanities, to better assess reliability and knowledge accuracy across diverse fields.

Third, while we conducted a large-scale study, our participants were limited to university students. This may limit the generalizability of our findings to learners with different purposes or at more advanced levels. Future work should expand the demographic diversity of the participant pool.

Lastly, while our study focused on GPT-based LLM chatbots, this was aligned with our objective to examine how LLMs facilitate autonomous learning and information retrieval in a structured, interactive format. However, as advanced hybrid search tools that combine LLMs with traditional search engines continue to emerge, it will be important to evaluate whether these technologies address the limitations of existing systems. Future work should explore the effectiveness of these advanced tools and whether educator concerns, particularly regarding reliability and educational impact, persist in the evolving landscape of LLM-based learning systems.

8 CONCLUSION

This study examines whether the changes in learning processes brought by LLMs, such as GPT, are negative, as some educators fear, or have a positive impact. To explore this, educators were surveyed on LLM-assisted learning, which revealed a two-dimensional model that explains how specific factors of search tools interact with educators' expectations for successful search-as-learning. Through a three-day experiment with university students using multiple learning sources, we investigate the impact of learning tools with different level of automation on students' learning outcomes. Participants created concept maps, completed MCQs, and provided feedback on each tool. The findings suggest LLM-based chatbots effectively support the acquisition and connection of key concepts without sacrificing accuracy or engagement. Despite educators' concerns, LLMs promote active learning comparable to traditional sources, although books performed better in long-term retention. Based on these results, we propose design implications for search-learning tools that combine LLMs' strengths in quick learning with strategies for deeper comprehension and retention.

References

[1] Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari. 2014. Evaluation methodologies in information retrieval dagstuhl seminar 13441. In *Acm sigir forum*, Vol. 48. ACM New York, NY, USA, 36–41.

[2] SA Ambrose. 2010. *How learning works: Seven research-based principles for smart teaching.* John Wiley & Sons.

[3] Payam Aminpour, Steven A Gray, Antonie Jetter, and Philippe J Giabbanelli. 2018. Is the crowd wise enough to capture systems complexities? An exploration of wisdom of crowds using Fuzzy Cognitive Maps. *Environmental Modelling & Software* 109 (2018), 240–255.

[4] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.

[5] Sarah W Beck and Sarah R Levine. 2023. Backtalk: ChatGPT: A powerful technology tool for writing instruction. *Phi Delta Kappan* 105, 1 (2023), 66–67.

[6] Brett A Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. 2023. Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1.* 500–506.

[7] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for information retrieval: Part I. Background and theory. *Journal of documentation* 38, 2 (1982), 61–71.

[8] Nilavra Bhattacharya. 2023. LongSAL: A Longitudinal Search as Learning Study with University Students. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–8.

[9] Benjamin S Bloom. 1968. Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation comment* 1, 2 (1968), n2.

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[11] Andrew C Butler. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition* 7, 3 (2018), 323–331.

[12] AJ Cañas, JD Novak, and FM González. 2004. Varieties of concept mapping. *Proceedings of the First International Conference on Concept Mapping* (2004).

[13] Alberto J Cañas, John W Coffey, Mary-Jo Carnot, Paul Feltovich, Robert R Hoffman, Joan Feltovich, and Joseph D Novak. 2003. A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support. *Report to the Chief of Naval Education and Training* (2003), 1–108.

[14] Bradley W Carroll and Dale A Ostlie. 2017. *An introduction to modern astrophysics.* Cambridge University Press.

[15] Michelene TH Chi. 2009. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science* 1, 1 (2009), 73–105.

[16] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091.

[17] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as learning (dagstuhl seminar 17092). (2017).

[18] WJ Conover and RL Iman. 1979. *Multiple-comparisons procedures.* Technical Report LA-7677-MS. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). https://doi.org/10.2172/6057803

[19] Harry S Delugach, Letha H Etzkorn, Sandra Carpenter, and Dawn Utley. 2016. A knowledge capture approach for directly acquiring team mental models. *International Journal of Human-Computer Studies* 96 (2016), 12–21.

[20] Brenda Dervin. 1983. An Overview of Sense-Making Research: Concepts, Methods, and Results to Date. In *International Communication Association annual meeting.* ICA, Dallas, TX. Presented at the International Communication Association annual meeting, Dallas, May 1983.

[21] Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, et al. 2024. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference.* 114–123.

[22] Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23084–23091.

[23] H Tristram Engelhardt. 1974. *The Structures of the Life-world.* Heinemann.

[24] Hayden Freedman, André van der Hoek, and Bill Tomlinson. 2022. Improving Wikidata with Student-Generated Concept Maps. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 205–215.

[25] Hayden Freedman, Neil Young, David Schaefer, Qingyu Song, André van der Hoek, and Bill Tomlinson. 2024. Construction and Analysis of Collaborative Educational Networks based on Student Concept Maps. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–22.

[26] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* 32, 200 (1937), 675–701.

[27] 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*. 2–11.

[28] 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*. 2–11.

[29] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

[30] Scharon Harding. 2023. *Harvard's New Computer Science Teacher Is a Chatbot*. https://www.pcmag.com/news/harvards-new-computer-science-teacher-is-a-chatbot Accessed: 2024-09-12.

[31] Alison Head and Michael Eisenberg. 2010. Truth be told: How college students evaluate and use information in the digital age. *Available at SSRN 2281485* (2010).

[32] Alison J Head. 2007. Beyond Google: How do students conduct academic research? *First Monday* (2007).

[33] William R Hersh, Diane L Elliot, David H Hickam, Stephanie L Wolf, and Anna Molnar. 1995. Towards new measures of information retrieval evaluation. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*. 164–170.

[34] Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of documentation* 52, 1 (1996), 3–50.

[35] Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education* 55, 1 (2023), 48–63.

[36] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–28.

[37] Cody Kalina and KC Powell. 2009. Cognitive and social constructivism: Developing tools for an effective classroom. *Education* 130, 2 (2009), 241–250.

[38] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[39] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[40] Rehan Ahmed Khan, Masood Jawaid, Aymen Rehan Khan, and Madiha Sajjad. 2023. ChatGPT-Reshaping medical education and clinical management. *Pakistan journal of medical sciences* 39, 2 (2023), 605.

[41] Hoil Kim. 2023. *Sampling Methodology*. Kyungmoon Publishing.

[42] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 385–404.

[43] W Kintsch. 1978. Toward a model of text comprehension and production. *Psychological Review* (1978).

[44] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review* 95, 2 (1988), 163.

[45] Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition* (reprinted 2000, 2003 ed.). Cambridge University Press, Cambridge.

[46] Will Knight. 2021. Ai can write disinformation now-and dupe human readers. *Wired, May* (2021).

[47] Ismo T Koponen and Maija Nousiainen. 2019. Pre-service teachers' knowledge of relational structure of physics concepts: finding key concepts of electricity and magnetism. *Education Sciences* 9, 1 (2019), 18.

[48] DR Krathwohl. 2002. A Revision Bloom's Taxonomy: An Overview. *Theory into Practice* (2002).

[49] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. ConceptScape: Collaborative concept mapping for video learning. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[50] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. 2022. Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*. 37–39.

[51] Gary Marchionini. 1995. *Information seeking in electronic environments*. Cambridge university press.

[52] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[53] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[54] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (2023), 333–334.

[55] Brian Moon, Charles Johnston, and Skyler Moon. 2018. A case for the superiority of concept mapping-based assessments for assessing mental models. In *Concept Mapping: Renewing Learning and Thinking. Proceedings of the 8th Int. Conference on Concept Mapping, Medellín, Colombia: Universidad EAFIT*.

[56] Michael J Muller and Sandra Kogan. 2012. Grounded theory method in human-computer interaction and computer-supported cooperative work. *The Human Computer Interaction Handbook (3 ed.), Julie A. Jacko (Ed.). CRC Press, Boca Raton, FL* (2012), 1003–1024.

[57] Joseph D Novak. 1990. Concept mapping: A useful tool for science education. *Journal of research in science teaching* 27, 10 (1990), 937–949.

[58] Debra L O'Connor, Tristan E Johnson, and Mohammed K Khalil. 2004. Measuring team cognition: Concept mapping elicitation as a means of constructing team shared mental models in an applied setting. *Proceedings of the First International Conference on Concept Mapping* (2004).

[59] American Association of School Librarians and Association for Educational Communications. 1998. *Information power: Building partnerships for learning*. American Library Association.

[60] OpenAI. 2024. Introducing GPT-4o and more tools to ChatGPT free users. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/. Accessed: 2024-05-22.

[61] Sungjin Park. 2023. *Introduction to SQL and NoSQL Databases*. Saengneung Publishing.

[62] Junaid Qadir. 2023. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1–9.

[63] Debbie Denise Reese. 2004. Assessment and concept map structure: Interaction between subscores and well-formed mental models. In *meeting of the American Educational Research Association, San Diego*.

[64] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[65] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[66] Kalhan Rosenblatt. 2023. ChatGPT banned from New York City public schools' devices and networks. *NBC News* (2023). Available at: https://www.nbcnews.com.

[67] Amanda Spink. 1997. Study of interactive feedback during mediated information retrieval. *Journal of the american society for information science* 48, 5 (1997), 382–394.

[68] Marcy H Towns. 2014. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education* 91, 9 (2014), 1426–1431.

[69] Meng-Jung Tsai and Chin-Chung Tsai. 2003. Information searching strategies in web-based science learning: The role of Internet self-efficacy. *Innovations in education and Teaching International* 40, 1 (2003), 43–50.

[70] Mark Bedoya Ulla, William F Perales, and Stephenie Ong Busbus. 2023. 'To generate or stop generating response': Exploring EFL teachers' perspectives on ChatGPT in English language teaching in Thailand. *Learning: Research and Practice* 9, 2 (2023), 168–182.

[71] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2020. The effects of learning objectives on searchers' perceptions and behaviors. In *Proceedings of the 2020 acm sigir on international conference on theory of information retrieval*. 77–84.

[72] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.

[73] Hannes Weber, Dominik Becker, and Steffen Hillmert. 2019. Information-seeking behaviour and academic success in higher education: Which search strategies matter for grade differences among university students and how does this relevance differ by field of study? *Higher Education* 77 (2019), 657–678.

[74] Hannes Weber, Steffen Hillmert, and Karin Julia Rott. 2018. Can digital information literacy among undergraduates be improved? Evidence from an experimental study. *Teaching in Higher Education* 23, 8 (2018), 909–926.

[75] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[76] Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation* 55, 3 (1999), 249–270.

[77] Thomas D Wilson. 2000. Human information behavior. *Informing science* 3 (2000), 49.

[78] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. 2011. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*. 1225–1226.

[79] Olga Zlatkin-Troitschanskaia, Johannes Hartig, Frank Goldhammer, and Jan Krstev. 2021. Students' online information use and learning progress in higher education–A critical literature review. *Studies in Higher Education* 46, 10 (2021), 1996–2021.

[80] Olga Zlatkin-Troitschanskaia, Johannes Hartig, Frank Goldhammer, and Jan Krstev. 2021. Students' online information use and learning progress in higher education–A critical literature review. *Studies in Higher Education* 46, 10 (2021),

1996–2021.

A LEARNING OBJECTIVES

For the evaluation, the Understand, Apply, and Analyze categories from the revised Bloom's taxonomy were selected, except for Remember taxonomy which has already been used. The categories were deemed suitable for assessment through MCQs, while higher-level taxonomies like Evaluate and Create were excluded due to the challenge they pose within the 40-minute learning period. Three learning objectives per module are detailed in Table 4.

	Solar System	Sampling	DBMS
LO 1	Define and classify **planets** and **dwarf planets**.	Define the concept of **sampling**.	Define the concepts of **databases** and tables.
LO 2	Explain the properties and characteristics of **planets in the Solar System**.	Explain and compare **probability sampling** and **non-probability sampling**.	Explain and compare **RDBMS** and **non-RDBMS**.
LO 3	Apply **Kepler's Laws** of planetary motion.	Classify various **probability sampling techniques** and apply their formulas.	Apply **CRUD** operations using MySQL with the basic syntax.

Table 4: When providing the learning objectives (LO), the main keywords were highlighted using bold formatting.

B MCQ GENERATION

Multiple-choice questions (MCQs) were developed to evaluate the learning outcomes of the experiment participants. A total of 27 questions were created, with 9 questions for each module, consisting of 3 questions from each of the selected taxonomy (Appendix A) categories. Most of the questions were designed to cover each subject's learning objectives (LO). According to previous studies, generating MCQs through large language models (LLMs) is preferable due to its time and cost efficiency, similar quality of results, and alignment with Bloom's taxonomy; therefore, the process was conducted using ChatGPT-4o [60].

B.1 Prompts

As shown in Figure 13, a text-based prompt outlining the question requirements was created and used, and to ensure format consistency across different GPT sessions, an image containing the format guidelines was included.

B.2 Quality Evaluation

The evaluation was conducted using two levels of metrics: question-level [21] and quiz-level [21, 22]. 81 questions were generated, with three times the required number for each subject. The MCQ set was finalized through a three-step refinement process. *Step 1)* Four authors cross-validated 81 initial questions using metrics as shown in Table 5, and two authors further reviewed the questions for alignment with learning objectives (LO) and taxonomy, removing unsuitable ones. *Step 2)* Missing questions were regenerated and re-evaluated using the same criteria, resulting in 9 questions per subject (27 total). *Step 3)* Professors from relevant fields conducted a final review based on criteria as shown in Table 6. In the evaluation of Structure, Redundancy, and Usefulness, the following scores were assigned: Astronomy (2, 3, 3), Sampling (3, 2, 3), and Database (3, 2, 3).

B.3 Examples of MCQs

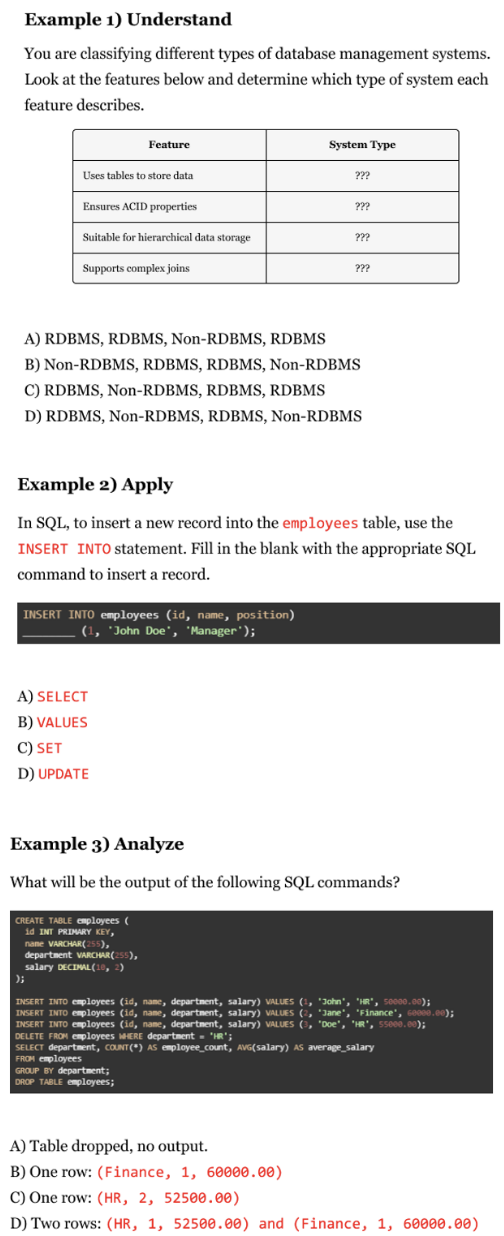


Figure 12: Three examples of post-test questions from the DBMS module, categorized by each taxonomy (Understand, Apply, and Analyze).

Prompt A) A text-based part

```
Create a multiple-choice question(MCQ) that reflects all the following
conditions. It should be well-aligned with the selected stage of the revised
Bloom's taxonomy.

- MCQ structure and output format
    - Refer to the uploaded images.
    - One shows a template for the section, and the other provides an example.
    - Omit the Visual data section if not needed.
- Number of MCQs: 3
```
Basic Settings

```
- Topic and learning objective
    - Topic:
        - SQL CRUD
        - Probability Sampling
        - The Solar System
    - Learning objective:
        - Define the concepts of **databases** and tables.
        - Explain and compare **RDBMS** and **non-RDBMS**.
        - Apply **CRUD** operations using SQLite 3 with the basic syntax.
        - Define the concept of **sampling**.
        - Explain and compare **probability sampling** and **non-probability
sampling**.
        - Classify various **probability sampling techniques** and apply their
formulas.
        - Define and classify **planets** and **dwarf planets.**
        - Explain the properties and characteristics of **planets in the Solar
System.**
        - Apply **Kepler's Laws** of planetary motion.
```
Modules | LOs

```
- Question type and revised Bloom's taxonomy
    - Revised Bloom's taxonomy:
        - Understand
        - Apply
        - Analyze
    - Question type:
        - Exemplify / Classify / Explain
        - Fill in the blank (The blank space is in the middle of code, figure,
or formula.)
        - Compare / Find the error / Choose a correct result
```
Taxonomy

```
- Conditions for stems
    - Set a scenario for the problem and then write down the situation in the
stem.
    - Actively use visual aids in the question; code, chart, mathematical
expression, or image.
- Conditions for options
    - 1 correct answer + 2 attractive incorrect answers + 1 simple incorrect
answer
```
QnA Settings

Prompt B) A structured format

Taxonomy / Qtype
The stem of the question
Visual data (ex: code, chart, mathematical expression, image)
Options (1 correct answer + 2 distracting incorrect answer + 1 simple incorrect answer)
Correct Answers / Distractors / Explanation

Prompt C) An example of the format

Taxonomy: Apply / **Qtype**: Fill in the blank

In Python, to implement binary search recursively, what condition should be checked first? Fill in the blank below.

```
def binary_search_recursive(arr, x, start, end):
    if ___:
        mid = (start + end) // 2
        if arr[mid] == x:
            return mid
        elif arr[mid] > x:
            return binary_search_recursive(arr, x, start,
                mid - 1)
        else:
            return binary_search_recursive(arr, x, mid + 1,
                end)
    return -1
```

A) start < end
B) start <= end
C) start === end
D) start > end

Correct Answer: B / **Distractors**: Two distracting incorrect answer
Explanation: Tell me why B is the correct answer.

Figure 13: Text (A) and images (B, C) were provided together to GPT-4. For the Modules | LOs and Taxonomy sections, only one color was retained for each to match their respective purposes.

	Rubric item	Question	Options
M1	Fluency	Is the language grammatically correct and clear?	(1) Yes, it is written in grammatically correct and clear language. (2) No, it is not written in grammatically correct and clear language. (3) I am unsure.
M2	Correct answer	Does the correct answer appear within the choices? If so, is the option marked as correct the right answer?	(1) Yes, the correct answer is present, and the option is marked as the 'correct' answer. (2) The correct answer is present, but it is not marked as the 'correct' option. (3) There are multiple correct answers. (4) No, the correct answer is not present among the options. (5) I am unsure.
M3	Unique choices	Are the answer choices distinct and unique from one another?	(1) Yes, the answer choices are completely distinct from one another. (2) Some choices are distinct, but others are too similar. (3) No, they all seem similar and appear to overlap. (4) I am unsure.
M4	No obviously wrong choice	Are there any answer choices that are incorrect or wrong?	(1) Yes, there are no incorrect answer choices. (2) Yes, but the correct answer is too easy to infer. (3) No, there are incorrect choices. (4) I am unsure.
M5	Correct material	If supplementary materials (e.g., code, formulas, images) are included in the question or choices, do they make sense grammatically and logically?	(1) Yes, the supplementary materials are grammatically and logically well-constructed. (2) There are minor issues. (3) No, the materials are incomprehensible. (4) I am unsure.
M6	LO alignment	Does this question contribute to achieving the learning objectives?	(1) Yes, it contributes to achieving the learning objectives. (2) It probably does, but there are significant gaps. (3) No, it does not help achieve the learning objectives. (4) I am unsure.
M7	Taxonomy alignment	Is the question appropriately aligned with the intended Bloom's taxonomy level?	(1) Yes, the question is aligned with the intended taxonomy. (2) No, the question is unrelated to the intended taxonomy. (3) I am unsure.

Table 5: The question-level metric was used to evaluate the appropriateness of the initial 81 questions and additional generated questions, resulting in the selection of 27 questions. Each question underwent cross-evaluation by at least three authors.

Metric	Definition	Evaluation
Structure	It measures whether the set of questions makes sense together.	Ordinal metric (1–3)
Redundancy	It measures if there is redundancy/repetition within the quiz	Ordinal metric (1–3)
Usefulness	It measures if a teacher would use the quiz in an assessment they create for their own class.	Ordinal metric (1–4)

Table 6: The quiz-level evaluation metric was used by subject matter experts to finalize the MCQ set. Each question was assessed based on three criteria: Structure, Redundancy, and Usefulness.

C Results of Concept Map Evaluation at Module

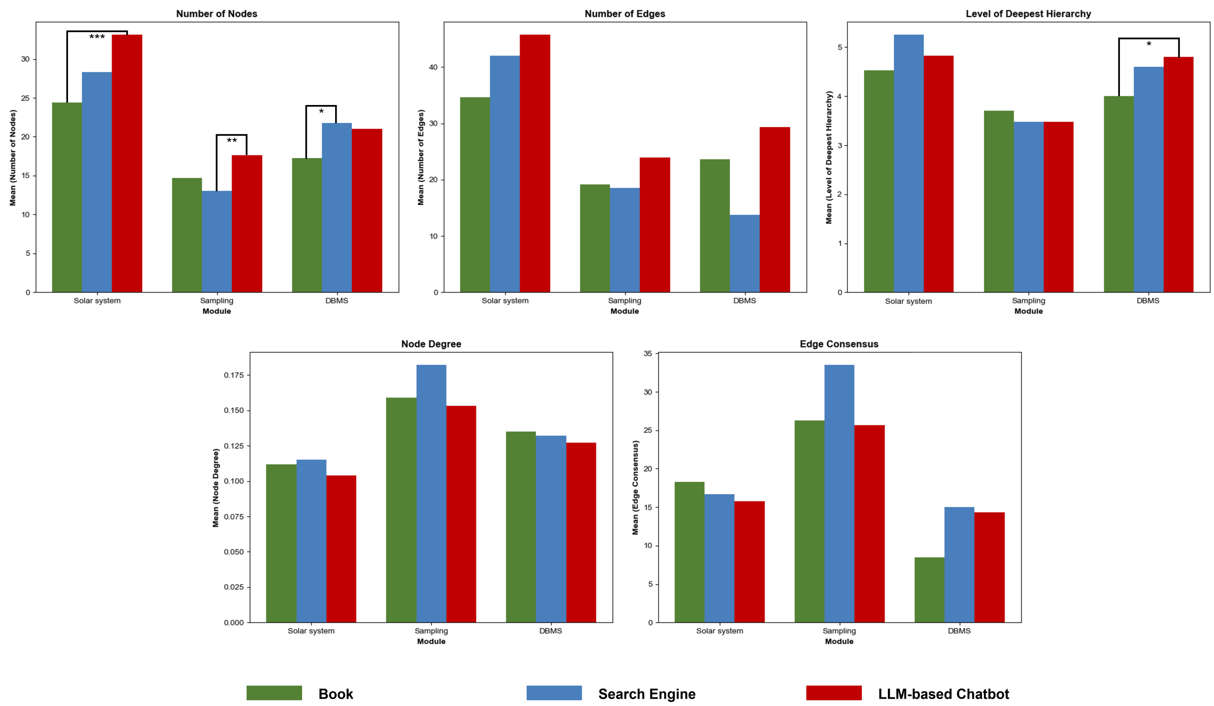


Figure 14: Results for the concept map evaluation at module level across five metrics.